



National Foundation for Educational Research

Investigating the relationship between A level results and prior attainment at GCSE

Tom Benton

Yin Lin

April 2011

Ofqual/11/5037

Ofqual
.....



This report has been commissioned by the qualifications regulators.

Contents

Contents	1
Executive summary and recommendations	1
Introduction	1
Aims	1
Key findings	1
Assessment of the current methodology behind prediction matrices	1
The accuracy of predicted grade distributions from prediction matrices	2
CCEA’s use of separate prediction matrices	3
Differences between awarding organisations	3
1. Introduction	5
1.1 Aims of analysis	5
1.2 Description of data	6
2 Is there a more appropriate measure of prior attainment than mean GCSE score?	7
2.1 Could mean GCSE score be used in conjunction with prior attainment in related subjects or English and Maths to provide better predictions?	10
2.2 Should Welsh as a second language be excluded from the measure of prior attainment for those students studying in Wales?	10
2.3 Should the measure of prior attainment give more weight to GCSE achievements with the same awarding organisation?	11
2.4 Summary	12
3. Could including additional information beyond prior attainment improve the accuracy of the methodology?	13
3.1 Alternative models	13
3.1.1 Prediction matrices (current methodology)	13
3.1.2 Proportional odds logistic regression	14

3.2	Comparing models	15
3.3	Exploring CCEA’s use of a separate prediction matrix.....	19
3.4	Which pieces of additional information are most valuable to include within prediction matrices?	20
3.5	Can the current prediction matrices methodology be improved by handling prior attainment differently?	23
3.6	Summary	25
4.	How accurate is the predicted grade distribution using the current methodology?	26
4.1	How do these estimates relate to currently used tolerances?.....	29
4.2	Summary	31
5.	Collating historical differences between awarding organisations and between years.....	32
5.1	Further investigation of the differences between CCEA and other awarding organisations.....	39
5.2	Summary	41

Executive summary and recommendations

Introduction

The use of statistical evidence is an important part of the way in which AS and A level grade boundaries are set. This approach requires an accurate view of how many students are expected to achieve each grade given their prior attainment at GCSE. It is therefore important that the relationship between GCSE attainment and A level grades is properly understood and any weaknesses in the ways in which this information is applied are identified.

Aims

Analysis in this report focuses on four broad questions:

- 1) What is the most appropriate measure of prior attainment to use within the process of predicting AS and A level grade distributions?
- 2) Can the process be made more accurate by using additional information about candidates and centres?
- 3) How accurate are the estimates of the expected grade distribution for each awarding organisation within each subject?
- 4) What historical differences are there between attainment levels for different awarding organisations once prior attainment and other candidate and centre characteristics are taken into account?

Within these broad aims were more specific lines of investigation where awarding organisations took an approach or were subject to a landscape that differed from the other awarding organisations. One of these related to whether candidates taking certain subjects were likely to generate different outcomes in the prediction matrices as a result of those subjects – specifically, whether in Wales candidates taking Welsh as a first or second language was likely to impact on the predicted grade outcomes for those candidates.

The other part of the analysis related to the current use by CCEA of a separate prediction matrix, and whether this is justified. This separate matrix is based solely on Northern Ireland candidates.

Key findings

Assessment of the current methodology behind prediction matrices

Analysis examined various alternatives to the current approach for producing predicted grade distributions. These included examining differing methods of calculating GCSE prior attainment (mean, total, mean excluding worst, total of best 4, etc). Analysis also explored incorporating additional variable within predictions

taking account of gender, centre type, region, whether or not the AS or A level subject had been studied at GCSE and whether or not any GCSE qualifications had been taken with the same awarding organisation as the AS or A level. Finally the analysis explored whether there was any value in changing the way in which students were split into groups based upon their prior attainment. The relative accuracy of the various alternative approaches was compared to identify whether any worthwhile improvements could be made.

These statistical analyses show that it is difficult to improve upon the method of prediction matrices as currently applied by awarding organisations. **We would recommend that this process is continued in its current form for the foreseeable future.**

One possible (though not necessary) change would be to switch from using mean GCSE score as the measure of prior attainment to using mean GCSE score once each candidate's worst grade was removed. This measure was found to have a very slightly higher correlation (0.003) with achievement at AS and A level and hence could be used as an alternative to mean GCSE score if there were non-statistical reasons for making this change.

The accuracy of predicted grade distributions from prediction matrices

Calculations were also conducted to estimate the standard errors associated with the predicted grade distributions that are generated by the current application of prediction matrices. These give an indication of how closely we would expect the actual final grade distribution of candidates to match with the predicted grade distribution within a given awarding organisation. It was shown that the accuracy of predictions may be influenced by the following factors (amongst others):

- 1) The number of candidates taking the qualification with the awarding organisation. As this number increases the precision of estimates will increase.
- 2) The number of candidates studying the subject with any awarding organisation in the previous year. Since prediction matrices are based on all candidates taking a given subject, an increased number of candidates will increase the precision of estimates.
- 3) The predicted percentage. Percentages that are predicted to be close to 50 per cent will tend to have lower precision than those that are close to 0 or 100.

On the basis of these calculations **we would recommend that the current tolerances for awarding organisations reporting outcomes from awards should be amended, and consideration should be given to different tolerances for grades A and E.**

CCEA's use of separate prediction matrices

The approach currently employed includes the use by CCEA of a separate prediction matrix that is based solely on Northern Ireland data. The use of this prediction by CCEA could be justified if the aim was simply to maintain year on year standards within CCEA. There is some evidence that the process of using a separate prediction matrix for CCEA may be justifiable in that, through maintaining the status quo, it leads to more accurate predictions of attainment. This is illustrated by the table below which shows that the rate at which grades are correctly classified for CCEA qualifications falls from 50.5 per cent to 45.8 per cent if the national matrix is used instead. There is also a 7.1 per cent increase in the deviance¹ of the predictive model.

Table: Comparing alternative prediction models for CCEA qualifications

	Prediction matrices based on Northern Ireland	National Prediction matrices	Percentage deterioration from using the national matrix
Deviance	88492	94785	7.1%
Correct classification rate	50.5	45.8	

Additional analysis brings out that there are historical differences between the attainment of Northern Ireland candidates as compared to candidates in England and Wales once prior attainment is accounted for. For example, analysis of 42 AS and A level subjects offered by CCEA in each of 2008, 2009 and 2010 showed that for 32 of these subjects attainment was significantly higher in Northern Ireland than in England and Wales in every year. The aim of the project was to ensure a consistent and fair approach to GCE awarding based on prior (GCSE) attainment. However, the differences in CCEA awarding in GCE cannot be justified based on an analysis of prior attainment and, therefore, a separate CCEA prediction matrix needs to be challenged.

Differences between awarding organisations

Further analysis also revealed statistically significant historical differences between the levels of achievement across different awarding organisations and years. This is true even after differences in prior attainment and other background characteristics

¹ Deviance is a statistical measure by which the accuracy of different statistical models can be compared. Higher deviances indicate a worse model fit. Further details are given within the main body of the report.

have been taken into account using multilevel modelling. A large proportion of these differences relate to the attainment of candidates taking qualifications with CCEA being higher than would be expected given their prior attainment and background characteristics. This apparent inconsistency between awarding organisations is undoubtedly related to CCEA's use of separate prediction matrices to the other awarding organisations. Whether or not this practice should be continued depends upon the reasons for the differences between CCEA and the other awarding organisations. There are several possible interpretations:

- 1) That CCEA attainment at AS and A level is inflated relative to other awarding organisations, given that the national predictions based on prior achievement at GCSE produce lower outcomes than the CCEA approach.
- 2) That attainment at GCSE of those candidates taking AS and A level with CCEA is deflated relative to other awarding organisations, given that the national predictions based on prior achievement at GCSE produce lower outcomes than the CCEA approach.
- 3) That the analysis has not fully accounted for the differences between candidates in Northern Ireland and candidates elsewhere.

Significant differences were also found for various subjects between the other four awarding organisations. If desirable it may be possible to reduce the number of such occurrences if awarding organisations were required to more strictly follow the percentage predicted to achieve each grade when deciding upon grade boundaries.

1. Introduction

The use of statistical evidence is an important part of the way in which AS and A level grade boundaries are set. This approach requires an accurate picture of how many students we expect to achieve each grade given their prior attainment at GCSE. As such it is important that the relationship between prior attainment and AS and A level grades is properly understood and any weaknesses in the ways in which this information is applied are identified.

In 2010 the three regulators for GCE qualifications (Ofqual, in England, DCELLS in Wales and CCEA in Northern Ireland) commissioned NFER to carry out an analysis of the current methodology for setting expectations based on prior achievement at GCSE. CCEA has a dual role in Northern Ireland, functioning both as a regulator and as an awarding organisation. These responsibilities are separately managed within the organisation.

The research described in this report has two overarching aims. Firstly, to determine the extent to which the current statistical approach to predicting the expected distribution of grades in AS and A levels could be usefully improved. Secondly to explore the level of consistency that is currently being achieved between different years and different awarding organisations in terms of the proportions of students awarded different grades once variations in the prior attainment of young people are taken into account.

1.1 Aims of analysis

Analysis in this report focuses on four broad questions:

- 1) What is the most appropriate measure of prior attainment to use within the process of predicting AS and A level grade distributions?
- 2) Can the process be made more accurate by using additional information about candidates and centres?
- 3) How accurate are the estimates of the expected grade distribution for each awarding organisation within each subject?
- 4) What historical differences are there between attainment levels for different awarding organisations once prior attainment and other candidate and centre characteristics are taken into account?

Within these broad aims were more specific lines of investigation where particular awarding organisations took an approach or were subject to a landscape that differed from the other awarding organisations. One of these related to whether candidates taking certain subjects were likely to generate different outcomes in the prediction matrices as a result of those subjects – specifically, whether in Wales candidates

taking Welsh as a first or second language was likely to impact on the predicted grade outcomes for those candidates.

The other awarding organisation-specific analysis related to CCEA's use of a separate Northern Ireland prediction matrix, to investigate whether there were any statistical differences in the outcomes following this approach.

1.2 Description of data

All analysis has been undertaken based on data provided by the awarding organisations. The data includes information on the achievement of all candidates taking AS and A levels in the summers of 2008, 2009 and 2010. Candidates with fewer than three matching GCSEs as well as those with partial absence² flags were removed from analysis. Furthermore, analysis did not consider any double awards or applied AS or A levels. AS and A level subjects with fewer than 500 entries in 2009 were also excluded from analysis with the exception of Welsh (first language). In order to include them within analysis, a number of languages with relatively low entry (including Chinese, Russian and Urdu) were combined and treated as a single subject labelled "Other Languages".

Once these exclusions were made the data set contained details of 4,926,175 AS and A level entries split roughly evenly between 2008, 2009 and 2010.

² Partial absence relates to subject results where candidates were absent for one or more individual units.

2 Is there a more appropriate measure of prior attainment than mean GCSE score?

Currently predictions of the expected distribution of AS and A level grades are based upon prior attainment as measured by a candidate's mean score³ across all of the GCSEs they took one year before (for AS grades) and two years before (for A level grades). The first stage of analysis was to explore whether any alternative ways of measuring prior attainment might be more productive. A number of potential alternative measures were considered within the following broad categories:

- 1) Alternative measures of the average. These might include median score or mean score once some of each candidate's best and worst grades are excluded.
- 2) Total GCSE scores. These reward candidates for taking a greater number of GCSEs. Total scores could either be across all GCSE subjects a candidate has taken or limited to a number of each candidate's best GCSE grades.
- 3) Measures with a specific focus on achievement in the core subjects of English and Maths. These include total scores from English and Maths alone as well as combining these with scores in other subjects. When using this measure it should be remembered that for a minority of candidates matched achievement in English and Maths was not available and so for these individuals it was necessary to impute English and Maths grades using their average GCSE score across other subjects⁴.
- 4) Achievement in subjects associated with the A or AS level under consideration. This was calculated by first deciding upon which GCSE subjects (if any) were associated with each of the AS and A level subjects under investigation. Subsequently it was possible to calculate, for each AS and A level subject a candidate was taking, the mean GCSE score they had attained in related subjects at GCSE. In some cases candidates did not have any matched data in associated subjects. In these cases it was necessary to impute average scores in related GCSEs using their average GCSE score across all subjects⁵.

³ By "score" we mean a numerical representation of grades where A* is worth 8 points, A is worth 7, and so on down to G being worth 1 point and U zero.

⁴ Five per cent of AS and A level entries were from candidates with no matching prior attainment in English and nine per cent were from candidates with no matching prior attainment in Maths. This is likely to be due to early entry within these subjects.

⁵ In total 38 per cent of all AS and A level entries were from candidates with no matching prior attainment within a related subject.

For each measure of prior attainment the correlation with AS and A level achievement⁶ across all subjects was calculated. This was done using combined data from each of the years 2008, 2009 and 2010. Results are shown in table 1. Across all of the various measures of prior attainment that have been explored it can be seen that the vast majority actually lead to a reduction in the correlation with achievement at AS and A level. Only two of the possible measures of prior attainment lead to any increase in correlation at all; the mean GCSE score excluding a candidate's worst grade at GCSE and excluding their worst two grades at GCSE. Even in these cases the improvement is extremely slight (only 0.003). On this basis we can conclude that there would be little advantage to replacing the use of mean GCSE score as the basis of prediction matrices with something else.

Having said this, using mean GCSE score excluding the worst grade may have some advantages in that it reduces the chances of pupils' prior attainment being depressed in circumstances relating to compulsory subjects that they may not have otherwise chosen. For this reason it may possibly be considered as a fairer measure of prior attainment and since our analysis shows it is empirically at least as good as mean GCSE grade this change could be made if it was considered to be a beneficial change and if it was agreed by all of the awarding organisations.

⁶ AS and A level grades were converted to numerical values to enable analysis. For consistency between years it was necessary to first combine grades A and A* as the A* grade was only introduced in 2010. Grades of A and A* were converted to 5 points, grade B was worth 4 points and so on down to grade E being worth 1 point and U zero.

Table 1: Correlations between various measures of prior attainment and achievement at AS and A level

Measure of prior attainment	Correlation with AS and A level grade
Mean GCSE grade	0.609
Median GCSE grade	0.582
Mean GCSE excluding worst	0.612
Mean GCSE excluding 2 worst	0.612
Mean GCSE excluding best and worst	0.608
Total GCSE points score	0.459
Total score from best 6 GCSEs	0.583
Total score from best 5 GCSEs	0.594
Total score from best 4 GCSEs	0.597
Total score from English, Maths and other best 4 grades	0.596
Total score from English, Maths and other best 3 grades	0.599
Total score from English, Maths and other best 2 grades	0.596
Total score from English, Maths and best other grade	0.582
Total score from English and Maths only	0.548
Mean GCSE score in related subjects	0.572

2.1 Could mean GCSE score be used in conjunction with prior attainment in related subjects or English and Maths to provide better predictions?

In order to explore this we split mean GCSE score into two component parts:

- 1) Mean GCSE score in related subjects (as described earlier)
- 2) Mean GCSE score in other subjects

If a candidate had either no matching prior attainment information in related subjects or had no matching prior attainment in unrelated subjects then both of the above measures were set to equal the mean GCSE score across all subjects. Once these two measures had been constructed, linear regression was used to obtain a weighted sum of the two measures with the weights chosen to optimise the correlation with achievement at AS and A level. The resulting measure (combining data from both elements of prior attainment) had a correlation of just 0.621 with AS and A level achievement. This is a tiny improvement (of only just over 0.01) over the correlation with mean GCSE score alone and so we can safely conclude that there is little to be gained by combining mean GCSE scores with achievement in related subjects to construct new measures of prior attainment.

Further analysis then explored whether there was any weighted sum of the above two measures along with prior attainment in English and Maths that might be more strongly correlated with achievement at AS and A level. Linear regression was again used to obtain four weights (one for each of the measures above and two more for English and Maths) chosen to optimise the correlation with achievement at AS and A level. The resulting measure had a correlation of 0.623 with AS and A level achievement implying once again that there is little to be gained by any replacement of mean GCSE score as the measure of prior attainment.

2.2 Should Welsh as a second language be excluded from the measure of prior attainment for those students studying in Wales?

As all candidates in maintained schools in Wales are required to study Welsh up to Key Stage 4 (either as a first or second language), it was noted that there was the potential for this to create an effect on the predictions that WJEC (the awarding organisation with the greatest proportion of candidates from Wales) would generate and use to guide its awards for all of its candidature. It was agreed that as this could

create an effect that would not apply to all awarding organisations, it would be worth exploring in the research.

In order to explore this issue mean GCSE score was recalculated excluding Welsh as a second language. The correlation between mean GCSE score both including and excluding Welsh as a second language and AS and A level grades was recalculated for all students studying in Wales⁷. The correlation using GCSE score including Welsh as a second language was found to be 0.591 whereas excluding Welsh as a second language the correlation was 0.590. This implies that whether or not Welsh as a second language is included in the measure of prior attainment makes very little difference in itself.

Of course this only answers one possible question on the use of Welsh as a second language within the calculation of prior attainment; whether or not it reduces the predictive power of the prior attainment measure. Another question might be whether or not the GCSE achievement of candidates in Wales was reduced by a consequential reduction in their GCSE choices (in that they are required to study Welsh as a first or second language). If this were the case, it would reduce the apparent prior attainment of the cohort and lead to overly severe standards being applied to AS and A level qualifications where a significant proportion of the candidature is from Wales (such as those offered by WJEC). Such a concern would be captured in there being a regional difference in AS and A level achievement after prior attainment has been controlled for⁸. Whether or not such differences exist and whether they should be accounted for within the application of prediction matrices is the subject of a later section (Section 3.2).

2.3 Should the measure of prior attainment give more weight to GCSE achievements with the same awarding organisation?

In order to explore this we split mean GCSE score into two component parts:

- 1) Mean GCSE score for subjects taken with the same awarding organisation as the A or AS level of interest
- 2) Mean GCSE score for subjects taken with a different awarding organisation

⁷ This includes all AS and A level subject entries within Wales not just those entered with WJEC.

⁸ This is true whether or not all students take Welsh as a second language. The point is that if *on average* the prior attainment of students in Wales is lower than it would be if Welsh was not compulsory, this should result in overachievement in later qualifications compared to other students with the same apparent level of prior attainment.

If a candidate had either no matching prior attainment information with the same awarding organisation as the A or AS level of interest or had no matching prior attainment from other awarding organisations then both of the above measures were set to equal the mean GCSE score across all subjects⁹. Once these two measures had been constructed, linear regression was used to obtain a weighted sum of the two measures with the weights chosen to optimise the correlation with achievement at AS and A level. The resulting measure (combining data from both elements of prior attainment) had a correlation of just 0.599 with AS and A level achievement. This is slightly lower than the correlation that was found with mean GCSE overall indicating that giving greater weight to prior attainment in subjects taken with the same awarding organisation does not lead to improved predictions.

2.4 Summary

Analysis in this section has established that none of the alternative prior attainment measures considered is likely to yield more accurate predictions than mean GCSE score. Having said this there are a few options, such as mean GCSE score excluding the worst grade, which are equally good and could be used as an alternative if there were reasons for preferring a different measure.

⁹ In total 13 per cent of all AS and A level entries were from candidates with no matching prior attainment with the same awarding organisation.

3. Could including additional information beyond prior attainment improve the accuracy of the methodology?

Having established that there is little to be gained from changing the measure of prior attainment used within the models, analysis next explored whether there is any value in making use of further information about young people and the centres they attend within the process of predicting the grade distribution within any A or AS level. To address this question analysis focussed on comparing the relative accuracy of a number of different methodologies.

3.1 Alternative models

Initially two different models were compared. For the purposes of this report we focussed on predicting 2010 AS and A level grade distributions based on models built using data on AS and A level candidates in 2009.

3.1.1 Prediction matrices (current methodology)

At present the following process is used:

- 1) All AS and A level entrants are split into ten groups according to their mean prior attainment. These are deciles of prior attainment amongst AS and A level entrants combined within each year¹⁰. The implicit assumption here is that the top ten per cent of students in 2010 are equivalent to the top ten per cent of students in 2009 in terms of their prior achievement at GCSE. As such, the methodology does not make direct use of the actual GCSE points score of any student but only their ranking amongst all students nationally. This approach is adopted to ensure that predictions are not linked to grade outcomes at GCSE, thereby allowing awarding organisations to set A level standards independently each year.
- 2) For each AS and A level subject the probability of entrants achieving each grade in 2009 within each decile is calculated. These tables of probabilities are known as prediction matrices.

¹⁰ The process actually used by awarding organisations is very slightly different to this in that it takes account of possible changes in the overall ability of AS and A level entrants each year. However, this is a minor difference in the context of this report and is ignored within the analysis reported here.

- 3) These probabilities are then applied to each entrant in 2010 depending upon their decile to give a predicted probability of them achieving each possible grade.
- 4) These percentages are then totalled across entrants for each grade to yield the expected number of entrants achieving that grade. For example, adding up every 2010 student's predicted probability of achieving an A provides an estimate of the total number of students we expect to achieve an A.

The above process is applied to the entire data set (from all awarding organisations) to yield separate predictions for each of AQA, Edexcel, OCR and WJEC. Predictions for CCEA are derived in essentially the same way with the exception that rather than constructing prediction matrices from the entire data set these are based only on entrants taking AS and A levels within centres located in Northern Ireland in 2009¹¹.

3.1.2 Proportional odds logistic regression

A revised model was also trialled taking account of the following additional student and centre characteristics:

- 1) Gender
- 2) Centre type. For the purposes of analysis centres were split into the following types: comprehensive schools, selective schools, independent schools, FE colleges, sixth form colleges, tertiary colleges and other.
- 3) Centre attainment. Using the matched data the average GCSE attainment level within each centre was calculated. Centres were then split into quintiles depending upon the average prior attainment of students.
- 4) Centre location. These were split into the following regions: North East, North West, Yorkshire, East Midlands, West Midlands, Eastern, London, South West, South East, Wales, Northern Ireland and Other.
- 5) Whether or not the student had studied a related subject¹² at GCSE.
- 6) Whether or not the student had studied a related subject at GCSE with the same awarding organisation.

¹¹ This data is not explicitly restricted to students taking CCEA AS and A levels. Having said this, the vast majority of AS and A levels within Northern Ireland are taken with CCEA.

¹² The definition of "related" was developed in consultation with awarding organisations and with the three regulators. It should be noted that exactly which GCSE subjects were related to which AS and A level subjects was not entirely agreed upon and should this element ever be used to produce actual predicted distributions further discussion and formal agreement would be required. However, the definitions that were reached are sufficient for exploratory analysis of the type described within this report.

Each of these variables was used along with decile of prior attainment¹³ within a proportional odds logistic regression model¹⁴. Logistic regression models allow us to explore the extent to which the odds of any student achieving at or above each grade in any subject are jointly related to all of the variables described above. As such it allows us to examine whether there is any potential benefit in including any of the above variables within predictions.

3.2 Comparing models

Once each of the above models were fitted to the data it was necessary to empirically compare their relative merits. In order to do this we must first define a criterion by which models may be compared.

There are two possible criteria that may appear appealing but are in fact inadequate for our purposes.

- 1) We may be tempted to simply compare the predicted distribution in 2010 to the actual distribution in 2010 and see which of the models yields a more accurate prediction. However, such an approach would be flawed for two reasons. Firstly it would ignore the fact that the 2010 distribution is at least partially guided by the results from prediction matrices in the first place, since the predictions would have formed part of the statistical evidence used to inform grading decisions. As such this approach would give an unfair advantage to the current methodology. Even if this were not the case the approach would still be inadequate as in situations where the characteristics of entrants do not vary between years all models would yield an identical prediction of the distribution. Indeed in such a situation we could achieve an accurate predicted distribution without building a model at all as the expected grade distribution would remain unchanged. What we require is a model that will provide accurate predictions regardless of the extent of changes in the characteristics of entrants.
- 2) It may also appear tempting to compare models in terms of the standard errors of the predictions that are provided. Standard errors represent the extent to which the same predictions would be provided by a given model across different samples. However this approach is also inadequate as it does not take account of the extent of bias in any model. For example, suppose it were found that girls tended to do better in a particular A level than boys with similar levels of prior attainment.

¹³ This was defined in exactly the same way as for the prediction matrices described earlier.

¹⁴ Multilevel models were deliberately not used within this strand of analysis as the aim is predict the distribution of grades nationally rather than within the average school. For the logistic models described within this report only main effects were included within analysis. Alternative models that also explored interactions between prior attainment and each background variable were also briefly explored but were found to have an inferior model fit and so are not described any further within this report.

Suppose further that in a particular year there was a marked increase in the percentage of girls taking that A level. Now if our model has not specifically taken account of the gender of entrants then it may not fully capture the change we would expect in the grade distribution. Hence, it could be said to be more biased than a model that did take account of this. However the model that did take account of gender may well display greater standard errors than the one that did not. Thus using standard errors alone tends to favour overly simplistic models.

As can be seen from the discussion above, attempting to directly assess the accuracy of any predicted grade distribution is difficult. For this reason we must instead assess the relative merits of different models more generally. This is done by assessing model fit for individual candidates. We have made use of two different ways of assessing model fit:

- 1) **Correct classification.** The basis of this criterion is that a good model will predict the A level grade of a candidate more accurately than a bad one. Thus by calculating the percentage of candidates in 2010 whose final grade is correctly predicted by a model fitted using 2009 data we can (to some extent) assess the relative merits of any one model against any other. Of course the aim of the model is not to predict the results of individual students but rather to predict the grade distribution. However, it is fairly reasonable to assume that a model that can accurately do the former should also provide a robust basis for the latter.
- 2) **Deviance.** Although intuitively easy to understand there are several drawbacks with using correct classification rate to assess the quality of different models. Foremost amongst these is that the situation where a pupil achieves an A but supposedly had only a 1 per cent chance of doing so is treated identically to a situation where a pupil achieves an A but had a 30 per cent chance of doing so¹⁵. Deviance addresses this since it is based on the likelihood of each grade a candidate achieves¹⁶. If many candidates are achieving grades that the model predicts to be highly unlikely this will lead to higher deviance (indicating a worse model fit) than if candidates are achieving grades that are predicted to be likely. One slight drawback of deviance is that if a candidate achieves a particular grade that has a predicted probability of zero this would theoretically result in infinite deviance (since a supposedly impossible event has occurred). To combat, this all probabilities were truncated to be in the range 0.001 to 0.999 before the calculation of deviance of was begun.

The models described in the previous section were compared on the basis of both deviance and correct classification rate. An additional variant of each model was also included to investigate whether or not the current practice of using separate prediction

¹⁵ Assuming that they had a greater probability of achieving another grade.

¹⁶ Technically deviance is calculated as minus 2 times the log of the probability of candidates achieving their given grades.

matrices for CCEA qualifications on the basis of location was empirically justifiable. Thus an alternative prediction matrix methodology where CCEA predictions were based on the same matrices as the other four awarding organisations was trialled. Similarly a variant of the logistic model which did not take account of centre region was trialled. The results of these analyses totalled across all subjects¹⁷ are shown in table 2. Results are provided overall as well as separately for qualifications provided by each awarding organisation.

There are a number of points to be made from the results here. First (and perhaps foremost) is the fact that the improvement in predictive power associated with incorporating all of the possible background information within the model is tiny. The overall deviance from using the logistic model is only 0.65 per cent lower than the deviance from the existing prediction matrix model; a tiny improvement. The extremely modest improvement in predictive power can also be seen from the fact that the overall correct classification rate increases from 40.2 per cent with the existing model to 40.6 once all variables are taken account of within the model. A similar pattern is seen within each of the separate awarding organisations. The largest improvement in deviance is seen for CCEA but even here we have an improvement of only 2.54 per cent; still a relatively modest improvement in accuracy¹⁸.

¹⁷ To avoid confusion it should be noted that all models were fitted separately for each A level and AS level subject. Results have been totalled across the different subjects but each subject was modelled independently.

¹⁸ The improvement in deviance for CCEA is not caused by including region within the logistic model as whether students are located in Northern Ireland or not is already accounted for within the current prediction matrices procedure.

Finally from the table below we can note that taking account of the region of students does not improve the accuracy of predictions for WJEC qualifications. In fact we see that the logistic model excluding region has slightly lower deviance than the logistic model including region. This implies that taking account of whether candidates studying with WJEC are located in Wales or not does not improve the accuracy of the model. This indicates that once other factors (particularly prior attainment) are taken account of there are not significant differences between the performance of candidates in Wales and the performance of candidates in other regions. This provides further evidence that the requirement on Welsh candidates to study Welsh up to Key Stage 4 (either as a first or second language) does not create an effect on the predictions that WJEC uses to guide its awards for all of its candidature.

Table 2: Comparing alternative prediction models

		Including Region			Excluding region			Number of 2010 entries included in analysis
		Prediction matrices	Logistic Model	Percentage improvement from logistic model	Prediction matrices	Logistic Model	Percentage improvement from logistic model	
Deviance	Overall	4628127	4598051	0.65%	4634419	4605474	0.63%	1682220
	AQA	2067184	2056381	0.53%	2067184	2056338	0.53%	742416
	CCEA	88492	86301	2.54%	94785	90831	4.35%	37798
	Edexcel	873203	865889	0.84%	873203	867774	0.63%	321545
	OCR	1198067	1189642	0.71%	1198067	1191049	0.59%	433476
	WJEC	401180	399839	0.34%	401180	399481	0.43%	146985
Correct classification rate								
Correct classification rate	Overall	40.2	40.6		40.1	40.5		
	AQA	39.1	39.4		39.1	39.4		
	CCEA	50.5	51.0		45.8	48.4		
	Edexcel	41.4	42.0		41.4	41.7		
	OCR	40.4	40.8		40.4	40.7		
	WJEC	40.4	40.5		40.4	40.5		

3.3 Exploring CCEA's use of a separate prediction matrix

We can next turn our attention to the question of whether CCEA should use a separate prediction matrix based on students within Northern Ireland. The increase in deviance associated with using national prediction matrices for CCEA, from 88,492 to 94,785 (roughly 7 per cent), indicates that using a separate prediction matrix clearly leads to improved model fit. This can be seen even more clearly by looking at correct classification rates which fall from 50.5 per cent (when a separate classification matrix is used) to 45.8 per cent when the national prediction matrix is used. This provides some justification for the current practice of using separate prediction matrices for CCEA qualifications¹⁹.

The question now arises as to whether this difference in predictive power is caused entirely by the fact that separate prediction matrices from the 2009 data were used to set grade boundaries in 2010. This might lead to the separate prediction matrices appearing more accurate. In other words, is the improved model fit a self-fulfilling prophecy since the separate prediction matrices model from 2009 data was what was used to define the 2010 grade distribution? To address this question multilevel modelling was used. The aim of this piece of analysis was to discover whether statistically significant differences between attainment in Northern Ireland, and that in England and Wales existed in the data from 2008 that was provided for the analysis. If such significant differences do exist this would show that they have not purely emerged in 2010 due to the way prediction matrices were applied in 2009 and would provide further evidence for consideration.

Analysis focussed on 42 AS and A level subjects offered by CCEA in each of 2008, 2009 and 2010. Multilevel models (using proportional odds logistic regression) then compared the level of achievement in Northern Ireland for these subjects to the level of achievement elsewhere. In each case comparisons were made using combined data from all awarding organisations. Prior attainment was accounted for within the multilevel models so that any differences that were found could be said to be statistically significant over and above the impact of prior attainment.

Results of the analysis are shown in table 3. As can be seen, in each year a clear majority of subjects display a significant difference between candidates in Northern Ireland and those elsewhere. This indicates that a significant difference between Northern Ireland and elsewhere existed throughout the years being studied and did not purely emerge in 2010. Additionally, analysis found that 32 of the 42 subjects

¹⁹ Why there are differences in attainment levels between Northern Ireland and elsewhere is irrelevant at this stage. Even if they have occurred because historically CCEA has used separate prediction matrices the fact remains that attainment levels in Northern Ireland *are* significantly different to elsewhere in the UK. It could be argued that the model used to predict the grade distribution should take account of this unless there are reasons to suspect that these differences are illegitimate.

displayed significantly higher attainment in Northern Ireland throughout all three of the years that were considered. This confirms that there is a historical difference and suggests that it is important to consider this regional difference in the construction of prediction matrices. Having said this, it is important to note that this analysis is not sufficient on its own to conclude that the use of separate prediction matrices is correct. All we can say is that there exist historical differences between the Northern Ireland data and the England and Wales data and that a change in approach from CCEA from using separate prediction matrices to using the same matrices as the other awarding organisations would certainly lead to a statistically significant change in the predictions for CCEA.

Table 3: Comparing alternative prediction models

Year	Number of subjects studied	Number displaying a significant difference between students in Northern Ireland and those elsewhere
2008	42	34
2009	42	37
2010	42	39

Returning to table 2 and looking at both the correct classification rate and deviance of the logistic models excluding region reveals that the reduction in performance associated with CCEA not using separate prediction matrices is partially (although not entirely) alleviated by including other variables within the predictive model. For example, we have already noted that for CCEA qualifications the correct classification rate falls from 50.5 per cent to 45.8 per cent when information on region is not used. The correct classification rate is increased a little when information other than region is used within the logistic model (up to 48.4 per cent). This implies that although some of the differences between CCEA and other awarding organisations can be explained in terms of the characteristics of schools and pupils there remains an element of this difference which is not accounted for by any of the school and pupil characteristics recorded within our data. This issue will be further explored in a later section (Section 5).

3.4 Which pieces of additional information are most valuable to include within prediction matrices?

Comparisons between prediction matrices and the logistic model (including region) are presented for each of 100 separate AS and A level subjects in appendix 1. These tables include a calculation of whether the improvement in model fit is statistically

significant²⁰. Overall in 51 subjects the improvement in predictive performance was statistically significant and only two were found where a statistically significant deterioration in model fit occurred. Having said this, the size of the improvement in model fit was generally very small. There were only 10 instances of a reduction in deviance of more than two per cent. Furthermore, many of the apparently larger reductions in deviance occurred in subjects with relatively small numbers of entrants. In such cases the estimate of model improvement is likely to be subject to some unreliability and hence the results may not be replicated if the process were repeated with a different sample of data.

Appendix 1 also presents the largest sheaf coefficient from the logistic model within each subject. Sheaf coefficients are a method by which the size of the relationship between an outcome and a group of variables within a regression model can be captured. They are useful in our context as our models produce numerous coefficients relating to various background characteristics and we wish to summarise these to understand which of these is the most important. For example, each of our models contains six coefficients relating to centre type²¹. However we wish to combine these six coefficients into a single number that tells us the relative importance of centre type against the other variables in the logistic model²².

Although potentially useful, a possible drawback of sheaf coefficients is that even if many of the coefficients associated with a particular variable are not statistically significant they can still contribute to the size of the sheaf coefficient. For example, the sheaf coefficient for region is based on 11 separate model coefficients for each subject whereas the sheaf coefficient for gender is based upon just one. In some cases this leads to region having the largest sheaf coefficient despite many of the coefficients being non-significant whilst gender may be less likely to appear as the largest sheaf coefficient despite regularly being statistically significant. This problem is particularly prevalent in subjects with relatively small numbers of entries.

Despite these drawbacks we can use sheaf coefficients to get an approximate summary of the relative importance of the different background variables on achievement at A level and AS level. To partially alleviate the problems noted in the

²⁰ Significance was calculated by examining the difference in deviance for each individual candidate. Analysis tested whether the average difference in deviance was significantly different from zero once the structure of the data (candidates grouped within centres) was taken account of.

²¹ The six coefficients capture the differences between comprehensive schools and each of selective schools, independent schools, FE colleges, sixth form colleges, tertiary colleges and other centre types.

²² Sheaf coefficients are calculated as follows: For each individual their category (such as comprehensive, selective, independent and so on) is replaced by their regression coefficient for the subject of interest. The standard deviation of the resulting variable is the sheaf coefficient. If there is wide variation between groups this will result in large coefficients and so the standard deviation (that is, the sheaf coefficient will be large). If the coefficients are universally small this will result in a small standard deviation and hence a low sheaf coefficient.

paragraph above this has been done using a weighted average of sheaf coefficients across the 100 subjects being studied with more weight given to subjects with a larger number of entries. The results are summarised in the table 4.

The first thing that can be seen from this analysis is that prior attainment is much more important as a predictor of AS and A level attainment than any of the other background characteristics; being over eight times as large as the next biggest average sheaf coefficient. This goes some way to explaining why including all of the additional information in the statistical models makes so little difference in terms of improving model fit.

Table 4: Summary of sheaf coefficients

Sheaf variable	Weighted Average Sheaf Coefficient
Decile of prior attainment at GCSE	1.55
Centre attainment at GCSE	0.15
Centre type	0.16
Region	0.19
Gender	0.14
Taken related subject at GCSE	0.07
Taken related subject at GCSE with the same awarding organisation	0.03

Secondly we can note that, beyond prior attainment, none of the sheaf coefficients particularly stand out as being much more important than the others. Each of centre attainment, centre type, region and gender has a roughly equal sheaf coefficient. This indicates that even the modest improvements in model fit displayed in table 2 can only be achieved by using data from all of these variables and could not be achieved by using any one of them alone.

Finally we note that that having (apparently) taken a related subject at GCSE is (generally speaking) of little importance in predicting future AS and A level results. One possible explanation for this is that within our data it is impossible to distinguish between students who have not taken a related subject at GCSE and those that have taken a related subject at GCSE but where this has not been successfully matched to the data we have recorded about them. For example, this might include students who have taken related subject as a non-GCSE examination. Lastly we note that whether a related GCSE has been taken with the same awarding organisation is of even less importance than whether it has been taken at all.

3.5 Can the current prediction matrices methodology be improved by handling prior attainment differently?

As described above, prediction matrices are currently constructed on the basis of using deciles from the population as a whole. Two possible alternatives have been explored:

- 1) **Constructing deciles separately for each AS and A level subject.** Under the current model certain subjects have a very uneven distribution of pupils across the ten prior attainment groupings. For example, a very large percentage of candidates taking Further Maths A level are drawn from the top decile of prior attainment. Previous research by AQA²³ has suggested that situations such as this can have a detrimental effect on the accuracy of the statistical models being employed. Defining deciles separately for each subject avoids this issue by ensuring that all prediction matrices are developed with approximately a tenth of entrants placed into each prior attainment group²⁴.
- 2) **Using a number of groups other than 10.** Although currently candidates are split into ten groups there is no reason why we couldn't instead split them into a different number of groups. In this analysis we trial all possible numbers of groups between 5 and 15.

Overall results comparing the various possibilities are shown in the table 5. For simplicity all of these estimates are based on a universal prediction matrices model applied to all awarding organisations including CCEA. The deviance of each model is shown along with the percentage improvement over the current prediction matrices model.

Once again it can be seen that making amendments to the current model leads to (at best) very minimal improvements in predictive power. Using separate deciles within each subject leads to a reduction of only 0.2 per cent and changing the number of prior attainment groups leads to a maximum reduction in overall deviance of 0.3 per cent.

²³ Pinot de Moira, A. (2008). Statistical Predictions in Award Meetings. How confident should we be? *AQA Internal Report*, RPA_08_APM_RP_013.

²⁴ To avoid needing to assume that deciles within each subject are exactly equivalent in each year (for example, that the top 10 per cent of Mathematics entrants in 2010 are equivalent to the top 10 per cent of Mathematics entrants in 2009) the following process was used. Candidates across all subjects were first split into fiftieths of prior attainment within each year (so we assume that the top 2 per cent of candidates in one year nationally are equivalent to the top 2 per cent in another). Within each subject these fiftieths were collapsed into deciles based on all entrants combined across both 2009 and 2010. These deciles were then used as the basis for revised prediction matrices.

Analysis was also conducted separately for each A level and AS level subject. Although 27 instances were found where creating separate deciles within each subject led to a statistically significant reduction in deviance, only 4 instances were found where the percentage reduction in deviance was greater than two per cent²⁵. On this basis we can conclude that generating separate prior attainment deciles for each A level and AS level subject would be of little value in improving the accuracy of the model.

Table 5: Alternative methods of handling prior attainment

	Deviance	Percentage improvement over current model
Current model	4634419	
Separate deciles by subject	4623449	0.2%
5 groups	4713777	-1.7%
6 groups	4686881	-1.1%
7 groups	4664423	-0.6%
8 groups	4649782	-0.3%
9 groups	4641315	-0.1%
10 groups ²⁶	4634419	0.0%
11 groups	4628794	0.1%
12 groups	4625857	0.2%
13 groups	4622528	0.3%
14 groups	4620564	0.3%
15 groups	4619182	0.3%

Analysis also explored whether it was the case that using a smaller number of prior attainment groups would be advantageous for subjects with lower numbers of entrants. However, little evidence was found to support this. For example, using five groups instead of ten resulted in only three instances²⁷ where the reduction in deviance was greater than two per cent and the maximum reduction in deviance found was less than 3.5 per cent.

²⁵ These were A level Art and Design (History), A level Latin, AS Level Welsh (First language) and AS level Latin.

²⁶ This model is, of course, equivalent to the current model. It is included in the table here purely to provide continuity in terms of the number of groups studied.

²⁷ These were A level Art and Design (History), A level Ancient History and AS level Welsh (1st language).

3.6 Summary

Analysis in this section has established that:

- 1) There is some empirical basis for the current practice of using separate prediction matrices based on pupils located in Northern Ireland for CCEA. Historically the achievement levels in Northern Ireland have been different from achievement levels elsewhere once prior attainment is accounted for. If the main aim above all others is to maintain consistent outcomes with previous years, then the analysis would suggest that this process should be continued. Having said this, further discussion of the differences between CCEA and other awarding organisations is an issue that will be returned to within a later section.
- 2) Making use of additional data regarding the characteristics of schools and pupils leads to only a very slight improvement in the predictive power of the models. Furthermore, even this slight improvement in predictive power appears to be dependent upon using all of the various additional pieces of information. None of the additional variables stands out as being particularly crucial over and above the others in improving model performance.
- 3) There is little to be gained in terms of predictive power from changing the way in which prior attainment is grouped within prediction matrices.

4. How accurate is the predicted grade distribution using the current methodology?

The aim of analysis described in this section is to calculate the standard errors around the predicted percentages of students to achieve each grade. The technical process used to calculate standard errors was balanced repeated replication. This method, which is also used to calculate standard errors around estimates from high profile international studies such as PISA²⁸, is useful in the current context as it allows us to take account of the structure of the data (that is, that candidates are clustered within centres) in situations where we have not used multilevel modelling. A more detailed description of this technique is given in appendix 5.

The standard errors associated with the predicted percentages achieving each grade in each subject for each awarding organisation are shown in appendix 2. Standard errors in this table are split into three parts:

- 1) **Model standard errors.** These represent the uncertainty in predictions arising from the fact that the sample of data analysed in 2009 only provides an estimate of the percentage we expect to achieve each grade in each decile. This source of error relates to uncertainty within the prediction matrices themselves.
- 2) **Innate standard errors.** Even if the expected percentage to achieve a given grade within each decile of attainment is known precisely there is still uncertainty surrounding the numbers that will actually achieve this grade in 2010²⁹.
- 3) **Approximate overall standard errors.** The two sources of uncertainty above can be combined to get an approximate estimate of the overall standard error associated with each predicted percentage.

Note that model standard errors largely depend upon the number of candidates from 2009 included within the construction of prediction matrices regardless of which awarding organisation they take their exam with. In contrast, innate standard errors

²⁸ The OECD Programme for International Student Assessment (PISA) is an internationally standardised assessment that was jointly developed by participating economies and administered to 15-year-olds in schools.

²⁹ To illustrate the difference between an expected percentage and an actual percentage consider the example of tossing a fair coin. We know that the expected percentage of times that we will get heads is 50 per cent. However, random chance also plays its part and the actual percentage of times that we get heads may be somewhat different to 50 per cent (particularly for a small number of coin tosses).

largely depend upon the number of candidates that take a qualification with the particular awarding organisation of interest in 2010.

Averages across all subjects of the approximate overall standard errors associated with each grade are shown for each awarding organisation in table 6. As can be seen the largest standard errors are associated with qualifications offered by CCEA. This is caused by the fact that not only are these qualifications taken by a relatively small number of candidates but also the prediction matrices used by CCEA are based on smaller numbers of candidates than the prediction matrices used by other awarding organisations³⁰. It is important to remember that the larger standard errors associated with CCEA qualifications should not be used (in isolation) to determine whether or not the decision to base prediction matrices only on students studying in Northern Ireland is a good idea. Analysis in the previous section has already shown that this approach leads to improved accuracy for CCEA, at least in terms of producing outcomes similar to those produced in previous years.

The next largest standard errors are associated with qualifications offered by WJEC. Note that these are somewhat lower than for CCEA since although the numbers of students entering qualifications with this awarding organisation are relatively small the prediction matrices used by WJEC are based on a large number of candidates from all regions and awarding organisations.

Table 6: Average standard errors across all qualifications within each awarding organisation

Number of entrants	Approximate standard error around predicted percentage at each grade						Number of qualifications analysed
	A	B	C	D	E	U	
AQA	1.13	.98	.93	.87	.70	.66	83
CCEA	3.43	3.00	2.54	1.79	1.07	.71	44
Edexcel	1.09	.92	.87	.78	.63	.51	52
OCR	1.49	1.32	1.21	1.08	.90	.75	78
WJEC	1.83	1.68	1.57	1.41	1.03	.87	69

The size of standard errors within each awarding organisation may be influenced by the following factors (amongst others):

³⁰ This is because the prediction matrices used by CCEA are based solely on candidates studying in Northern Ireland.

- 1) The number of candidates taking the qualification with the awarding organisation in 2010. As this number increases the innate standard errors will decrease and so the precision of estimates will increase.
- 2) The number of candidates studying the subject with any awarding organisation in 2009. Since prediction matrices (other than those for CCEA) are based on all candidates taking a given subject, an increased number of candidates will lead to reduced model standard errors and the precision of estimate will increase.
- 3) The predicted percentage. Percentages that are predicted to be close to 50 per cent will tend to have larger standard errors than those that are close to 0 or 100.

To illustrate the first of these points, table 7 summarises how the estimated standard errors around the predicted percentage at each grade for each awarding organisation are associated with the numbers of entrants in 2010. As can be seen the precision of estimates improves dramatically as the number of candidates taking a subject with a particular awarding organisation increases.

Table 7: Average standard errors across qualifications with different numbers of entrants within each awarding organisation

Number of entrants	Approximate standard error around predicted percentage at each grade						Occurrences in 2010 ³¹
	A	B	C	D	E	U	
200 or less	5.33	4.97	4.09	2.98	1.73	1.23	14
201-300	4.83	4.35	3.32	2.45	1.31	0.85	5
301-500	3.01	2.80	2.52	2.15	1.50	0.96	26
501-1000	2.31	2.03	1.87	1.65	1.27	1.01	44
1001-2000	1.72	1.52	1.39	1.19	0.94	0.79	64
2001-4000	1.19	0.98	0.95	0.85	0.67	0.52	58
4001 or more	0.74	0.61	0.59	0.54	0.45	0.47	114

³¹ For example there are 15 instances where an awarding organisation has fewer than 200 entrants for a qualification in 2010 (within the data used to generate prediction matrices).

4.1 How do these estimates relate to currently used tolerances?

The results in table 7 can be used to inform decisions regarding how closely the predicted percentage of students to achieve each grade should be followed for individual subjects within awarding organisations. Where grade outcomes fall outside of these recommended limits, awarding organisations are required to provide written explanations for this, which tend to relate to technical issues. At present such tolerances are defined for the percentage achieving grade A and above and also grade E and above. The current tolerances are listed in the two leftmost columns of table 8. This shows, for example, that if there are more than 1000 entrants for a particular subject within an awarding organisation, and the actual outcomes were more than one per cent from the predicted outcomes, an explanation would be provided to the regulators. If there are less than 200 entrants for a subject within an awarding organisation then the outcomes need not be reported.

Table 8: Current and recommended tolerances

Current tolerances		Recommendations		
Entry	Reporting tolerance for grades A and E	Entry	Reporting tolerance for grade A ³²	Reporting tolerance for grade E ³³
200 or less	No reporting	200 or less ³⁴	6.1%	1.4%
201 to 300	4%	201 to 300	5.6%	1.0%
301 to 500	3%	301 to 500	3.5%	1.1%
501 to 1000	2%	501 to 1000	2.7%	1.2%
1,001 or more	1%	1001 to 2000	2.0%	0.9%
		2001 to 4000	1.4%	0.6%
		4001 or more	0.9%	0.5%

Using the calculations within table 7 it is possible to provide recommendations for these tolerance levels and these are presented in table 8. These recommendations are

³² This is for the percentage achieving grade A or above (that is, A* or A). For the analysis in the current report these two grades have been combined meaning that tolerances can be based upon the standard errors for the percentage predicted to achieve grade A.

³³ This is for the percentage achieving grade E or above. Since this percentage is 100 minus the percentage being awarded grade U the standard errors around the grade U predictions provide a basis for the recommended tolerances presented here.

³⁴ Although standard errors have been calculated for such subjects it should be noted that almost all subjects with an entry of less than 500 candidates in total (across all awarding organisations) have been removed from analysis. In cases where there are fewer than 500 candidates entering a subject across all awarding organisations the estimated standard error shown here may not be reliable.

based upon 75 per cent confidence intervals around that predicted percentages meaning that they are calculated by multiplying the relevant standard errors by a factor of 1.15.

The following differences from the current tolerances should be noted:

- 1) Separate tolerance levels are recommended for the percentage achieving at grade A or above and the percentages achieving at grade E or above. This is due to the fact that the percentage predicted to achieve grade A or above is typically much closer to 50 per cent than the percentage predicted to achieve grade E or above, as the percentage achieving grade E tends to be above 95% and therefore is subject to a lesser degree of statistical standard error. For this reason the reporting tolerances for grade E should be much lower than the reporting tolerances for grade A.
- 2) The recommended reporting tolerances are larger for grade A than are currently used but lower for grade E.
- 3) Many of the recommended tolerances do not relate to whole numbers in terms of percentage points. For this reason tolerances are recommended to one decimal place.
- 4) It can be seen that the accuracy of predictions continues to improve as the number of entrants increases beyond 1000. For this reason two additional categories have been created relating to cases where the number of entrants is greater than 2000.

It should also be noted that the recommendations above do not take account of the number of candidates used to construct prediction matrices. Not taking account of this fact will have particular implications for CCEA as their predictions are generated from prediction matrices based upon a much smaller group of candidates than is the case for the other awarding organisations. Furthermore, the above recommendations do not take account of the actual percentage of candidates predicted to achieve grade A and above and grade E or above. These percentages may vary somewhat between different subjects meaning that ideally the tolerances should be adjusted accordingly. For example, the A levels offered by OCR in Performance Studies and in Classical Civilizations both have a similar number of entrants in 2010 (around 1,200) and so would each work towards the same reporting tolerances. However, the percentage predicted to achieve an A in Performance Studies (around 11 per cent) is much lower (and much further from 50 per cent) than the percentage predicted to achieve an A in Classical Civilizations (31 per cent). This results in the standard error associated with the predicted percentage to achieve grade A in Performance Studies being somewhat lower than the equivalent standard error for Classical Civilizations. Differences of this type between subjects are not accounted for in the recommendations contained in table 8.

As an alternative to the above table, recommended tolerances could be generated individually for each subject. A relatively simple formula to facilitate this process is suggested in Appendix 5. If using such a formula was considered practical it would be possible to generate specific tolerance levels for each subject dependent upon the

percentage of students predicted to achieve at each grade or above, the number of students entering the subject with each awarding organisation and the number of students used to construct the prediction matrix. This would ensure that the most appropriate tolerance levels were used in each specific set of circumstances for each subject.

4.2 Summary

Analysis described in this section has established that the level of precision associated with predicted grade distributions varies according to a number of factors, in particular the number of candidates taking the specified qualification with the awarding organisation. The number of candidates used to construct prediction matrices is also important and for this reason predictions for CCEA are generally less precise than for other awarding organisations. Finally, predicted percentages close to 50 per cent will tend to have higher standard errors.

The calculations described in this chapter can be used as a basis to recommend revisions to existing tolerances in terms of how closely predicted percentages should be followed when grade boundaries are determined. It is recommended that separate tolerances are used for the grade A and grade E boundaries. To maintain consistency with current practice, recommendations have been generated based upon the number of entrants for a subject within a given awarding organisation. However, such an approach neither takes account of the exact predicted percentage (and how close this percentage is to 50 per cent) nor takes account of the number of candidates used to build prediction matrices. More appropriate tolerance levels could be set using the relatively simple formula that is developed within Appendix 5.

5. Collating historical differences between awarding organisations and between years

Analysis in this section collates the historical differences between awarding organisations and between different years for each of the 100 AS and A levels already considered.

Descriptive analysis is shown in separate tables in appendix 3. Results from the first of these tables, relating to A level Biology, are displayed in table 9 as an example. This table details the following information for each awarding organisation and for each year³⁵:

- 1) The number of students (within our matched data) entering the subject.
- 2) The mean GCSE grade of entrants³⁶.
- 3) The mean GCSE grade of entrants achieving an A.
- 4) The percentage of entrants achieving a grade A.
- 5) The mean grade achieved³⁷.

Using this table we can begin to explore the differences between the various awarding organisations. The percentage of students achieving an A is seen to be noticeably higher for CCEA than for the other awarding organisations. Furthermore, whereas the mean GCSE grade of CCEA entrants is similar to the mean GCSE of entrants within other awarding organisations, the mean grade achieved at A level is noticeably higher. This provides some indication that achievement in CCEA is higher than expected compared to other awarding organisations.

Having said this, there are a number of weaknesses associated with drawing conclusions based on the descriptive table. Firstly we cannot tell whether the differences visible within the table are statistically significant. Secondly it is difficult to know how much difference in AS and A level grades to expect given differences in GCSE grades. Finally these raw descriptive comparisons do not take account of other potentially influential variables such as gender, centre type and the level of attainment within centres. Analysis in previous sections has already shown that these variables have a statistically significant (albeit quite small) relationship with achievement in AS and A levels and hence, for the purposes of this section, it is important that they are included within analysis.

³⁵ Within the tables in appendix 3, instances where an awarding organisation has less than 50 matched candidates entering a particular subject within our data are excluded from analysis.

³⁶ For the purposes of calculating mean GCSE grade A* was treated as being worth 8 points, A as 7, B as 6 and so on down to 1 point for G and zero for U.

³⁷ For the purposes of this analysis A or AS level grades of A (or A*) were treated as being worth 5 points, B as 4 points and so on down to 1 point for an E and zero for U.

Table 9: Descriptive analysis showing differences between awarding organisations for A level Biology

				Number of students entering subject	Mean GCSE grade	Mean GCSE grade of entrants achieving an A	Percentage of students achieving an A	Mean grade achieved
Biology	A level	2008	AQA	18455	6.6	7.3	28.2	3.31
			CCEA	1632	6.7	7.2	44.2	3.92
			Edexcel	8450	6.6	7.4	27.6	3.31
			OCR	13750	6.6	7.4	26.2	3.27
			WJEC	1614	6.7	7.4	29.1	3.43
		2009	AQA	18248	6.7	7.3	30.4	3.39
			CCEA	1685	6.7	7.2	45.2	3.94
			Edexcel	8495	6.7	7.4	27.8	3.36
			OCR	13408	6.6	7.4	27.3	3.29
			WJEC	1659	6.7	7.4	30.7	3.44
		2010	AQA	20152	6.7	7.4	29.6	3.43
			CCEA	1942	6.8	7.3	45.8	3.98
			Edexcel	6501	6.6	7.3	26.8	3.32
			OCR	16154	6.7	7.4	28.0	3.37
			WJEC	2553	6.7	7.4	30.9	3.49

In order to gain a more thorough understanding of the statistical significance of differences between awarding organisations we used multilevel modelling. Multilevel modelling is crucial within the context of this analysis as it provides an accurate way to calculate the statistical significance of differences between awarding organisations whilst taking account of the fact that candidates are grouped within centres. If this is not done there is the risk that (for example) extremely good results in a small proportion of centres may skew results and make acceptable differences between awarding organisations appear to be statistically significant.

For each of the 100 AS and A level subjects being analysed a multilevel proportional odds logistic regression model was fitted to the data exploring the relationship between the final grade achieved and the year/awarding organisation with which the subject was taken whilst taking account of deciles of prior attainment, gender, centre type and centre attainment. Region was deliberately excluded from the models as it is closely related with awarding organisation in that the vast majority of CCEA qualifications are taken within Northern Ireland and the vast majority of qualifications from other awarding organisations are taken elsewhere. The result of this is that the following analysis is subject to the caveat that it is not in general possible to

distinguish the impact of the centre location being in Northern Ireland from the impact of the awarding organisation being CCEA.

Whether or not a student has taken a related subject before at GCSE has not been taken account of within this analysis. Partly this is due to the idea of including this within models emerging too late within the timeframe for analysis for this to be included. Having said this there are two additional reasons why it may not be a good idea to include this piece of information within analysis. Firstly analysis in previous sections has shown that this variable has generally very little predictive power. Secondly (and perhaps more crucially) there is some doubt over the validity of this information as it is not possible to distinguish candidates who have not taken a related subject at GCSE from those candidates that have taken such a subject but where it has not been identified within the matched data.

In order to calculate the statistical significance of differences between years and awarding organisations we must first define what we are comparing them against. The typical approach to this problem is to define one category (in our case one year for one of the awarding organisations) as a reference group and to make all other comparisons against this. In our case this is not an acceptable approach as it would mean elevating the status of one of the awarding organisations over and above the others. As an alternative we have decided to compare the coefficient for each year and awarding organisation against the average of all coefficients for all years and awarding organisations included within analysis³⁸. For each subject we can then calculate which awarding organisations have significantly higher achievement than expected and which have significantly lower achievement than expected.

Results of analysis are detailed for each subject in appendix 4. For purposes of illustration the first set of results (for A level Biology) is displayed in table 10.

³⁸ In order to do this we must constrain the coefficients associated with years and awarding organisations to add up to zero.

Table 10: Results of multilevel modelling for A level Biology

Subject	Level	Years/Awarding organisations with significantly higher attainment than expected	Associated odds ratios	Years/Awarding organisations with significantly lower attainment than expected	Associated odds ratios
Biology	A Level	2008/CCEA	2.43	2008/AQA	.71
		2009/CCEA	2.60	2008/EDEXCEL	.72
		2010/CCEA	2.18	2008/OCR	.76
				2008/WJEC	.78
				2009/AQA	.83
				2009/EDEXCEL	.84
				2009/OCR	.74
				2009/WJEC	.83
				2010/AQA	.86
				2010/EDEXCEL	.90
				2010/OCR	.85

Table 10 displays those combinations of year and awarding organisation where achievement levels are significantly higher than expected given the characteristics of the candidates, as well as those where achievement levels are significantly worse than expected. As can be seen CCEA has significantly higher than expected attainment in each year. The extent of this difference is expressed in terms of odds³⁹ ratios. These imply that, with everything else⁴⁰ being equal, across all grades, the odds of achieving at or above any given grade are over twice as high for CCEA candidates as they are on average across all years and awarding organisations. One of the consequences of comparing all awarding organisations to the average is that if one awarding organisation has substantially higher achievement than all of the others this will inevitably lead to all other awarding organisations having significantly lower attainment than the average. Even if we accept the interpretation that these results relate to A level attainment being higher or lower than expected, it is impossible to

³⁹ Odds refer to the ratio of the number of times an event is expected to happen to the number of times it is expected not to happen. Although the odds of an event are directly related to the probability of it occurring they are not the same thing and should not be confused. For example, doubling the odds of achieving at or above any given grade is not equivalent to doubling the probability of achieving at or above that same grade.

⁴⁰ By “everything else” we mean all of the variables that have been accounted for in the model and also given the same effect of centre.

tell from these results alone whether the qualifications offered by CCEA or by all other awarding organisations are pitched at the right level.

The following caveats should also be noted:

- 1) It is impossible to distinguish between cases where AS and A level achievement is high and where GCSE attainment is low.
- 2) It is impossible to distinguish using these purely empirical methods between cases where AS and A levels are too easy and cases where the course offered by the awarding organisation is more engaging for students leading to genuinely improved levels of attainment.
- 3) Equally it is possible that other influences that have not been accounted for within the models (such as teaching quality) may provide an explanation for differences in levels of achievement. Further research would be required to explore alternative explanations for differences in achievement between awarding organisations.

Bearing these caveats in mind, results across all of the tables in appendix 4 are summarised in table 11. This table displays the number of times A levels and AS levels are found to have higher than expected and lower than expected level of attainment within each awarding organisation.

Qualifications awarded by CCEA are commonly found to have significantly higher attainment than expected including, for example, over 80 per cent of all AS levels offered by this awarding organisation. Qualifications from AQA, Edexcel and OCR are commonly found to have lower attainment than expected. A levels from WJEC are commonly found to have higher than expected attainment levels whereas for AS levels the distribution between those qualifications that have significantly higher attainment and those that have significantly lower attainment is more evenly spread.

Table 11 Summary of results from all multilevel models

		Number of qualifications analysed from 2008-2010⁴¹	Number with significantly higher attainment than expected	% with significantly higher attainment than expected	Number with significantly lower attainment than expected	% with significantly lower attainment than expected
A Level	AQA	119	14	11.8%	55	46.2%
	CCEA	66	51	77.3%	0	.0%
	EDEXCEL	78	4	5.1%	43	55.1%
	OCR	114	2	1.8%	75	65.8%
	WJEC	100	48	48.0%	6	6.0%
AS Level	AQA	133	9	6.8%	64	48.1%
	CCEA	66	55	83.3%	3	4.5%
	EDEXCEL	79	5	6.3%	40	50.6%
	OCR	116	4	3.4%	67	57.8%
	WJEC	104	23	22.1%	11	10.6%

⁴¹ Note that if the same subject is offered in each of 2008, 2009 and 2010 this counts as three courses as opposed to one. Thus although only 47 A level subjects and 53 AS level subjects are explored in analysis some of the numbers in this table are somewhat higher than this.

The same information based purely on 2010 qualifications is shown in table 12. This largely shows the same pattern as table 11. However, there is a clear reduction in the percentage of A levels from WJEC that have significantly higher attainment than expected. Furthermore, in 2010 the percentage of AS levels from WJEC that have significantly higher attainment than average matches the percentage of AS levels that have significantly lower attainment than average. This indicates that any differences between WJEC and the other awarding organisations have reduced over time.

Table 12 Summary of results from all multilevel models for 2010 entrants

		Number of qualifications analysed	Number with significantly higher attainment than expected	% with significantly higher attainment than expected	Number with significantly lower attainment than expected	% with significantly lower attainment than expected
A Level	AQA	39	4	10.3%	20	51.3%
	CCEA	22	18	81.8%	0	0.0%
	EDEXCEL	26	1	3.8%	12	46.2%
	OCR	38	0	0.0%	22	57.9%
	WJEC	34	10	29.4%	1	2.9%
AS Level	AQA	44	4	9.1%	23	52.3%
	CCEA	22	17	77.3%	1	4.5%
	EDEXCEL	26	0	0.0%	11	42.3%
	OCR	39	1	2.6%	25	64.1%
	WJEC	35	3	8.6%	3	8.6%

5.1 Further investigation of the differences between CCEA and other awarding organisations

The most striking result of the analysis so far has been the significant differences between CCEA and the other awarding organisations. One shortcoming of this analysis is that although we have attempted to control for any differences in the background characteristics of candidates taking AS and A levels with different awarding organisations it remains the case that the majority of CCEA qualifications are taken within Northern Ireland whereas the majority of qualifications from the other awarding organisations are taken within England and Wales. It is possible that there remains some unaccounted for differences between Northern Ireland and elsewhere that would explain the differences in achievement between CCEA and the other awarding organisations.

In order to further investigate this possibility analysis was undertaken based on comparing those candidates within Northern Ireland taking qualifications with CCEA to those candidates *within* Northern Ireland taking qualifications with any of the other awarding organisations. In order to undertake such analysis it is important that there are sufficient candidates within Northern Ireland both amongst those taking a given qualification with CCEA and amongst those taking a qualification with another awarding organisation. For this reason analysis was restricted to those subjects where at least 300 candidates could be found taking the qualification with CCEA in 2010 and at least another 300 could be found taking the qualification with other awarding organisations (combined). The following subjects were identified as being suitable for analysis:

- 1) Biology AS and A levels
- 2) Chemistry AS and A levels
- 3) Physics AS level
- 4) Mathematics AS and A levels
- 5) Business Studies AS and A levels
- 6) English Literature AS level

For each of these ten subjects multilevel models were run comparing the performance of CCEA candidates to the performance of all other candidates taking account of the same school and pupil characteristics as were included in the initial multilevel models. The results of these models showed that across the ten subjects studied:

- 1) Eight were found to have significantly higher attainment for CCEA candidates in 2010 than for similar candidates with other awarding organisations.
- 2) Four were found to have significantly higher attainment for CCEA candidates in 2009 than for similar candidates with other awarding organisations.

- 3) Three were found to have significantly higher attainment for CCEA candidates in 2008 than for similar candidates with other awarding organisations⁴².

These results show that even when analysis is restricted to Northern Ireland a large number of significant differences are found between CCEA and the other awarding organisations. This is particularly true for results from candidates in 2010. This implies that it is unlikely that the differences revealed earlier are due to differences between Northern Ireland in comparison to England and Wales, and that differences are therefore more likely to relate to the differing approaches taken by CCEA and the other four awarding organisations.

⁴² Just one subject (AS level Business studies) was found where the attainment of candidates studying with CCEA was significantly lower than for similar candidates with other awarding organisations. This was only true in 2008 and was the only such difference identified.

5.2 Summary

Analysis in this section has established that:

- 1) There exist numerous examples where there are statistically significant differences between awarding organisations after taking account of prior attainment alongside other candidate and centre characteristics.
- 2) Whilst the reasons behind this cannot be established for certain, differences between CCEA and the other awarding organisations were particularly evident. These differences are unlikely to be due to differences between Northern Ireland and other regions.

The last point is undoubtedly related to the fact that CCEA use separate prediction matrices to the other awarding organisations to set grade boundaries. However, analysis in previous sections has already shown that there is some historical justification to support this practice. In essence what these results tell us is that there is a tension between maintaining consistency with historical level thresholds that have been set by CCEA and trying to achieve greater consistency with the processes adopted by other awarding organisations. Sadly, pure statistical analysis cannot provide any means of resolving this tension.

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

First published by the Office of Qualifications and Examinations Regulation in 2011

© Crown copyright 2011

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	2nd Floor
Coventry Business Park	Glendinning House
Herald Avenue	6 Murray Street
Coventry CV5 6UB	Belfast BT1 6DN

Telephone 0300 303 3344

Textphone 0300 303 3345

Helpline 0300 303 3346