A COMPARABILITY STUDY IN GCSE FRENCH

A statistical analysis of results by Awarding Body

Dr Mudhaffar Al-Bayatti
Research and Policy Analysis
AQA 2005

A study based on the Summer 2004 examination and sponsored by QCA

Al-Bayatti, M. (2005). *A comparability study in GCSE French. A statistical analysis of results by awarding body*. A study based on the summer 2004 examinations. London: Qualifications and Curriculum Authority.

EXECUTIVE SUMMARY

The Qualification and Curriculum Authority (QCA) commissioned a comparability study, comprising three strands, to assess the differences in grading standards between Awarding Bodies in GCSE French. The GCSE syllabuses used in this study were AQA (3651), Edexcel (1226), and OCR (1925).

The first and second strands related to a qualitative analysis and a script review which are being carried out by Awarding Body nominees and independent consultants. The third strand involves a multi-level modelling analysis which is the subject of this research. Candidate level result data from the Summer 2004 GCSE French examinations in the above three specifications were analysed.

Three outcome variables were constructed: whether a candidate achieved a grade A or better, a grade C or better, and a grade F or better. A multi-level model was fitted to each outcome variable to assess the probability of exceeding a given grade threshold.

The results of the multi-level models after controlling for centre type, candidate's mean Key Stage 3 score and gender showed evidence that the chance of being awarded a grade A or better was higher for Edexcel than for both AQA and OCR. The chances of being awarded a grade C or better were higher for OCR than for Edexcel, the chances of both of which were higher than those for AQA. The chances of being awarded a grade F or better were higher for OCR and Edexcel than for AQA. These differences may be due to differences in standards or they may be due to legitimate, albeit unexplained, factors not able to be included in the models.

The explanatory variables, centre type, mean KS3 and sex were statistically significantly different from zero in all the fitted models. For every unit increase in mean Key Stage 3 result, the odds of exceeding the grades A, C and F thresholds increased by more than thirteen fold, nine fold and three fold respectively when the values on the Awarding Body variables were zero. Female candidates appeared to perform better than male candidates, and candidates from selective or independent centres had a higher probability of achieving grades A, C and F than those from non-selective centres. Interaction terms between mean Key Stage 3 result and Awarding Body variables were also sometimes significant. This indicates that the mean Key Stage 3 result has a differential effect for different Awarding Bodies on the predicted probability of achieving grades A, C and F.

The percentage of variation explained by fitted models was 95.6 per cent, 75.2 per cent and 52.1 per cent for grades A, C and F respectively. Although these percentages are reasonably high, some variation is left unexplained and other predictors might have a statistically significant effect on the outcome variable.

Introduction and data sources

The Qualification and Curriculum Authority (QCA) is conducting a comparability study in GCSE French to review the grading standards for GCSE French in syllabuses offered by the AQA (3651), Edexcel (1226) and OCR (1925). The study comprises three strands. The first and second strands related to a qualitative analysis and a script review which are being carried out by awarding body nominees and independent consultants. The third strand involves a multi-level modelling analysis which is the subject of this paper.

For all candidates who entered for these specifications in June 2004, the following data were extracted from the National Matched Database (Key Stage 3 (KS3) 2002 matched with GCSE 2004):

- DFES centre number;
- Candidates' number;
- Candidates' name;
- Awarding Body subject number;
- GCSE examination grade;
- Gender;
- Mean KS3 score;
- Number of KS3 tests sat;

The total number of entries in 2004 for the four specifications under consideration was 257,939 but, as a result of attrition in the matching process, only 198,664 candidates with complete records were included. The main reason for attrition was that independent centres are not obliged to enter their students for KS3 tests, rendering impossible the matching of the KS3 scores of such candidates to their GCSE grade. This phenomenon might especially have been expected to affect OCR since it attracts a relatively high proportion of candidates from independent centres to its specification. The matching rate for each Awarding Body, shown in Table 1, indicates that this was indeed the case.

Table 1: Percentage of GCSE candidates matched to KS3 scores by awarding body

Awarding Body	Match rate of GCSE to KS3 scores
AQA	78.3
Edexcel	79.1
OCR	70.5
Total	**77.0**

Statistical analyses

(a) Preliminary analysis

Table 2, which contains the raw cumulative grade distributions in 2004 for the four GCSE specifications under consideration, might suggest that significant differences in the grading standards between Awarding Bodies exist. For example, the proportion of candidates awarded each grade is higher in OCR than in all the other Awarding Bodies, although as noted, it attracts a relatively high proportion of candidates from independent centres who traditionally perform better than candidates from the maintained sector.

Table 2: Percentage of candidates exceeding a given grade by awarding body (all 2004 UK candidates)

Awarding Body	Syllabus	A*	A	B	C	D	E	F	G	Candidates
AQA	3651	6.9	16.9	30.2	49.8	69.0	83.4	93.2	98.4	162004
Edexcel	1226	9.0	21.7	39.2	60.1	76.7	88.1	95.5	98.9	49193
OCR	1925	11.3	24.7	42.4	70.2	85.5	93.8	98.0	99.5	46742
Total		**8.1**	**19.2**	**34.1**	**55.5**	**73.5**	**86.2**	**94.5**	**98.7**	**257939**

The grade distributions for only those candidates for whom matched data were available are shown in Table 3. The patterns of the distributions in Table 2 have recurred in Table 3 especially at grades C to G. The proportion of candidates awarded grades A*, A, and B is higher in Edexcel than in all the other Awarding Bodies in the matched dataset, reflecting probably the loss of independent centres' data from OCR.

Table 3: Percentage of candidates exceeding a given grade by awarding body (matched candidates)

Awarding Body	Syllabus	A*	A	B	C	D	E	F	G	Candidates
AQA	3651	4.2	12.5	25.2	45.1	65.9	81.9	92.8	98.4	126831
Edexcel	1226	5.6	16.6	33.8	56.0	74.3	87.2	95.3	98.9	38893
OCR	1925	4.9	14.5	31.6	63.1	82.2	92.5	97.6	99.4	32940
Total		**4.6**	**13.6**	**27.9**	**50.3**	**70.2**	**84.7**	**94.1**	**98.7**	**198664**

Table 4 shows a comparison of the average of mean KS3 test results between awarding bodies.

Table 4: Average of mean KS3 test results by awarding body

Awarding Body	Average of mean KS3 results	Std.	Candidates
AQA	5.523	1.062	126831
Edexcel	5.546	1.067	38893
OCR	5.580	1.080	32940
Total	**5.537**	**1.066**	**198664**

The correlation coefficient between GCSE grades and mean KS3 scores was 0.73 and statistically significant at 1% level. The mean KS3 results explain 53% of the variation in GCSE French outcomes.

Table 5, which presents the cumulative grade distributions for the male and female matched candidates, also suggests that differences between Awarding Bodies in their overall raw grade distributions might partially be accounted for by their differential rates of gender entry.

Table 5: Percentage of candidates exceeding a given grade by sex

Sex	A*	A	B	C	D	E	F	G	Candidates
Male	3.1	9.7	21.1	41.7	62.9	79.9	91.8	98.1	90842
Female	5.9	17.0	33.7	57.4	76.4	88.7	96.0	99.1	107822
Total	**4.6**	**13.6**	**27.9**	**50.3**	**70.2**	**84.7**	**94.1**	**98.7**	**198664**

Similarly, Table 6, which presents the cumulative grade distributions for independent/selective and non-selective centres, suggests that differences between Awarding Bodies in their overall raw grade distributions might partially be accounted for by their differential rates of entry by centre type.

Table 6: Percentage of candidates exceeding a given grade by centre type

Centre Type	A*	A	B	C	D	E	F	G	Candidates
Non selective	3.0	10.3	23.6	46.3	67.6	83.3	93.5	98.5	181220
Selective	20.7	48.3	73.0	91.5	97.8	99.4	99.8	100.0	17444
Total	**4.6**	**13.6**	**27.9**	**50.3**	**70.2**	**84.7**	**94.1**	**98.7**	**198664**

(b) Multi-level Modelling

There are many variables, besides difference in Awarding Bodies' awarding standards, which may contribute to the differences between their grade distributions. Cresswell (1996) and, more recently, Schagen & Hutchison (2003), for example, argue that many factors need to be controlled for before beginning fairly to compare the outcomes of any two examinations and infer anything about their relative standards.

In recent comparability studies such as this, relationships between these factors and candidates' performance at GCSE have been studied, and controlled for, through the use of multi-level models (Pinot de Moira (2000a, 2000b and 2000c), While and Fowles (2000), and Jones, Meadows and Al-Bayatti (2004)). Similarly in this study, multi-level models were fitted in order to assess the average probability of achieving each judgemental grade (A, C, F) in each Awarding Body after controlling for some of the above explanatory variables.

In order to undertake this analysis, a dichotomous response variable for each judgemental grade boundary was defined, comprising whether each candidate had achieved that grade or not. There were two levels of hierarchy for each model: candidate was level 1 nested within centre as level 2. The selection of the explanatory variables was based on the given data and reference to previous studies (see for example Pinot (2000c)) which also used various background measures to control for different entry patterns between Awarding Bodies.

In addition to Awarding Body, the three explanatory variables used in this study were as follows:

1. **Centre type:** a dummy variable with 1 = independent/selective centre and 0 = non-selective centre. This classification is based on the type of centre each candidate entered for GCSE examination.

2. **Mean KS3 score:** a continuous variable and calculated as the average scores obtained by a candidate in English, Mathematics and Science. This variable is centred around the mean of candidates' mean KS3 scores (i.e. 5.537).

3. **Sex:** a dummy variable with 1 = male and 0 = female.

For each of the dummy variables, the choice of which category is assigned the value of 0, and which the value of 1, is arbitrary although traditionally the largest category is assigned the value of 0. This was the case for the Awarding Body variable which was formulated as a set of binary contrasts, with AQA (the largest entry) assigned the 0 value and each of the other Awarding Bodies a value of 1 when contrasted with it. Thus the baseline model, referred to in the following tables and graphs, comprises a candidate with 0 values for all the dummy variables and the mean KS3 score. Other models are expressed as deviations from this baseline model. It should be emphasised that AQA was assigned the baseline, 0 value solely because of the size of its entry. It does not imply that its standard was deemed correct, and that of other Awarding Bodies severe or lenient in an absolute sense. The positions of the Awarding Bodies in relation to each other would have been the same had another Awarding Body been assigned the baseline, 0 value.

Multi-level logistic modelling is a regression analysis designed for dichotomous response variables which takes into account and models the hierarchical nature of the data in the

population (Hosmer and Lemeshow, 1989). In this study there were three models each with a response variable with two possible values: awarded/not awarded a grade A, C and F or better. (N.B. References in the report to the probability of being awarded a grade implies being awarded that grade or better.) A separate model was fitted for each response variable on data having a two-level hierarchical structure, candidates being nested within centres.

The model parameters were estimated using iterative generalised least squares (IGLS) (Goldstein (2003)) and the estimation procedure required a linearisation in which an approximation method should be used. The 1st order marginal quasi-likelihood (MQL) approximation was used first to obtain starting values for the 2nd order predictive quasi-likelihood (PQL) procedure. MQL was used first, then PQL, because the former method may lead to estimates which are biased, particularly if sample sizes within level 2 units are small. Although the latter method offers a better approximation, it is less stable and takes longer to converge. The 2nd order PQL procedure did not converge for modelling grade F due to large number of level 2 units with invariable response values, therefore 1st order MQL method was used which converge and produced estimates that gave a reasonable variance for the predicted probabilities.

Tables 7, 8, and 9 contain details of fitted models. Definitions of the terms used in these tables are given in Appendix 2.

Grade A

Table 7 contains details of the multi-level model and shows that the effects of centre type, mean KS3 score and sex on whether a grade A was awarded were all highly significant. Thus, after controlling for the effects of KS3 score and gender, the odds of being awarded a grade A were 234 per cent higher for candidates from independent/selective than from non-selective centres. The odds for centre type can also be expressed as the odds ratios of independent/selective to non-selective centres: the odds ratio indicates that 334 candidates from the former category were awarded a grade A for every 100 such candidates in the latter category. Similarly, for every unit increase in mean KS3 score, the odds of being awarded a grade A increased by more than thirteen fold, and the odds of being awarded a grade A for male candidates are 70 per cent lower than for females, the odds ratio indicating that only 30 males were awarded a grade A for every 100 female candidates.

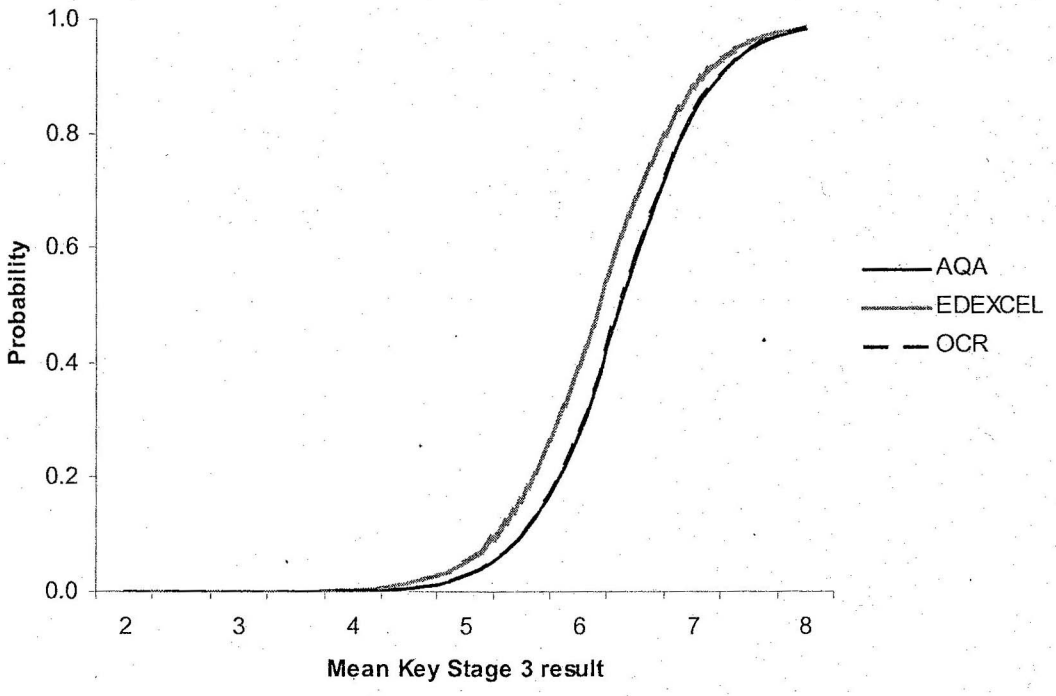
The joint chi-squared test showed evidence of statistically significant differences in the probability of being awarded a grade A dependent upon specification, although only the **Edexcel specification was significantly different from AQA's base line.** Terms of interaction between mean KS3 score and Awarding Body variables were, however, introduced to the model. The analysis showed that the interaction effect was statistically significant for Edexcel ($p = 0.001$), indicating that the relationship between candidates' mean KS3 scores and their chance of being awarded a grade A varies between Awarding Bodies. Edexcel candidates benefiting less from each unit increase in mean KS3 score.

Figure 1 shows the relative effect of Awarding Body on the probability of being awarded a grade A, with Edexcel candidates having the highest probability for all values of the mean KS3 score. Using the measure for the explained proportion of variance proposed by Snijders and Bosker (2000), the model explained 95.6 per cent of the variation in the outcome variable. The model fit statistics and residual plots are given in Appendix 2.

Table 7: Parameter estimates of GCSE French model for candidates attaining a grade A or higher

		β	se	p	Joint χ^2	p	Odds
Fixed Effects	Constant	-3.559	0.036	0.000			
	Centre type	1.205	0.063	0.000			3.337
	Mean KS3 result	2.593	0.022	0.000			13.370
	Sex	-1.200	0.022	0.000			0.301
	Edexcel	0.663	0.066	0.000	107.728	0.000	1.941
	OCR	0.003	0.077	0.969			1.003
	Edexcel*Mean KS3 result	-0.136	0.041	0.001	12.267	0.002	0.873
	OCR*Mean KS3 result	0.011	0.047	0.815			1.011
Random Effects	Centre level	0.937	0.036	0.000			

Figure 1: The probability of exceeding the grade A threshold in each awarding body dependent upon mean KS3 result (for a female candidate entered through a non-selective centre)



Grade C

Table 8 contains details of the multi-level model and shows that the effects of centre type, mean KS3 score and sex on whether a grade C was awarded were all highly significant. The odds of being awarded a grade C were 438 per cent higher for independent/selective candidates than for those from non-selective centres. Thus the odds ratio indicates that 538 candidates from the former category were awarded a grade C for every 100 candidates from the latter. Similarly, for every unit increase in mean KS3 score, the odds of being awarded a grade C increased by more than nine fold, and the odds for a male candidate were 71 per cent lower than for a female, the odds ratio indicating that only 29 male candidates were awarded a grade C for every 100 female candidates.

The joint chi-squared test showed evidence of significant differences in the probability of being awarded a grade C dependent upon specification, with **Edexcel and OCR both significantly different from AQA's baseline.** Allowing for a 95 per cent confidence interval (including a Bonferroni adjustment) around the multi-level parameter estimates associated with each Awarding Body also showed that the probability of being awarded a grade C was greater with OCR than with Edexcel. If the interaction terms in the model are assumed to be zero, the odds of obtaining grade C for a candidate from OCR were 4.03 and 1.62 times the odds for a candidate from AQA and Edexcel respectively.

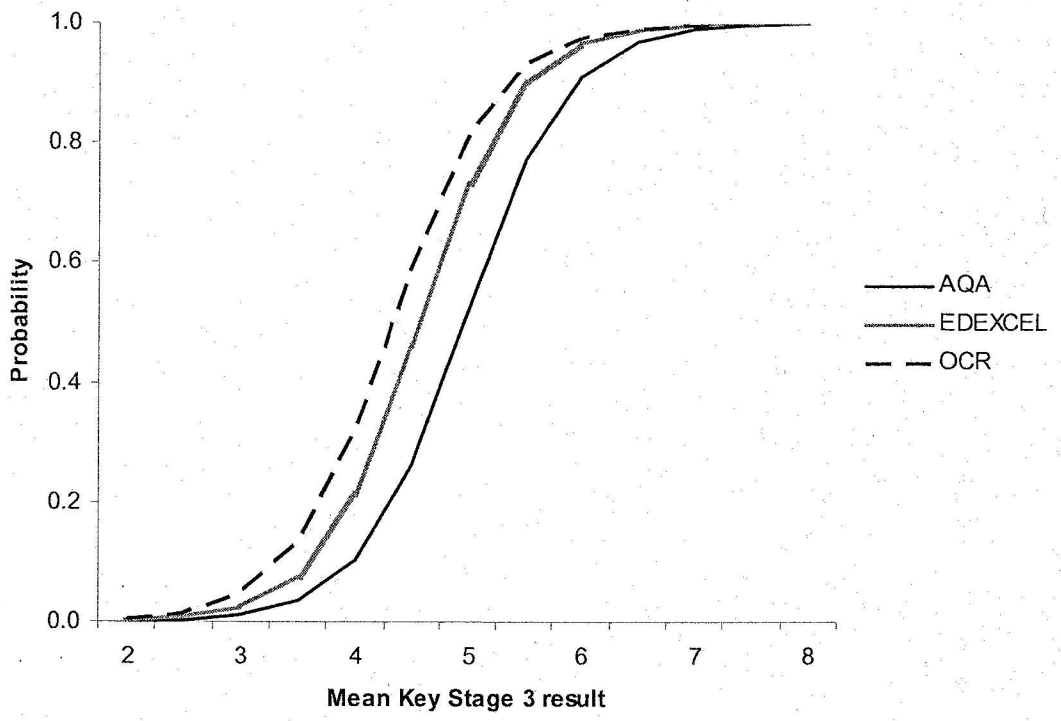
Terms of interaction between mean KS3 score and Awarding Body variables were, however, introduced to the model, and the analysis showed a statistically significant interaction effect between Edexcel and mean KS3 score ($p = 0.008$). The relationship between candidates' mean KS3 scores and their chance of being awarded a grade C thus varies between Awarding Bodies, with Edexcel candidates benefiting most from each unit increase in mean KS3 score. For those candidates who obtained grade C or better at GCSE, OCR's had the lowest mean KS3 score and AQA's had the highest.

Figure 2 shows the relative effect of Awarding Body on the probability of being awarded a grade C. OCR's and Edexcel's candidates had the highest probability for all values of the mean KS3 score. The model explained 75.2 per cent of the variation in the outcome variable. The model fit statistics and residual plots are given in Appendix 3.

Table 8: Parameter estimates of GCSE French model for candidates attaining a grade C or higher

		β	se	p	Joint χ^2	p	Odds
Fixed Effects	Constant	0.093	0.029	0.001			
	Centre type	1.683	0.078	0.000			5.382
	Mean Key Stage 3 result	2.251	0.014	0.000			9.497
	Sex	-1.248	0.014	0.000			0.287
	Edexcel	0.911	0.056	0.000	615.141	0.000	2.487
	OCR	1.393	0.064	0.000			4.027
	Edexcel*Mean KS3 result	0.077	0.029	0.008	8.324	0.016	1.080
	OCR*Mean KS3 result	-0.015	0.031	0.628			0.985
Random Effects	Centre level	1.153	0.038	0.000			

FIGURE 2: The probability of exceeding the grade C threshold in each Awarding Body dependent upon mean Key Stage 3 result (for a female candidate entered through a non-selective centre)



Grade F

Table 9 contains details of the multi-level model and shows that the effects of centre type, mean KS3 score and sex on whether a grade F was awarded were all highly significant. The odds of being awarded a grade F or better were 741 per cent higher for candidates from independent/selective than from non-selective centres, the odds ratio indicating that 841 candidates from the former category were awarded a grade F for every 100 from non-selective centres. Similarly, for every unit increase in mean KS3 score, the odds of being awarded a grade F increased by more than three fold, and the odds were 56 per cent lower for male than female candidates, the odds ratio indicating that 44 males were awarded a grade F for every 100 female candidates.

The joint chi-squared test showed evidence of statistically significant differences in probability of being awarded a grade F dependent upon specification, with **Edexcel and OCR both significantly different from AQA's baseline**.

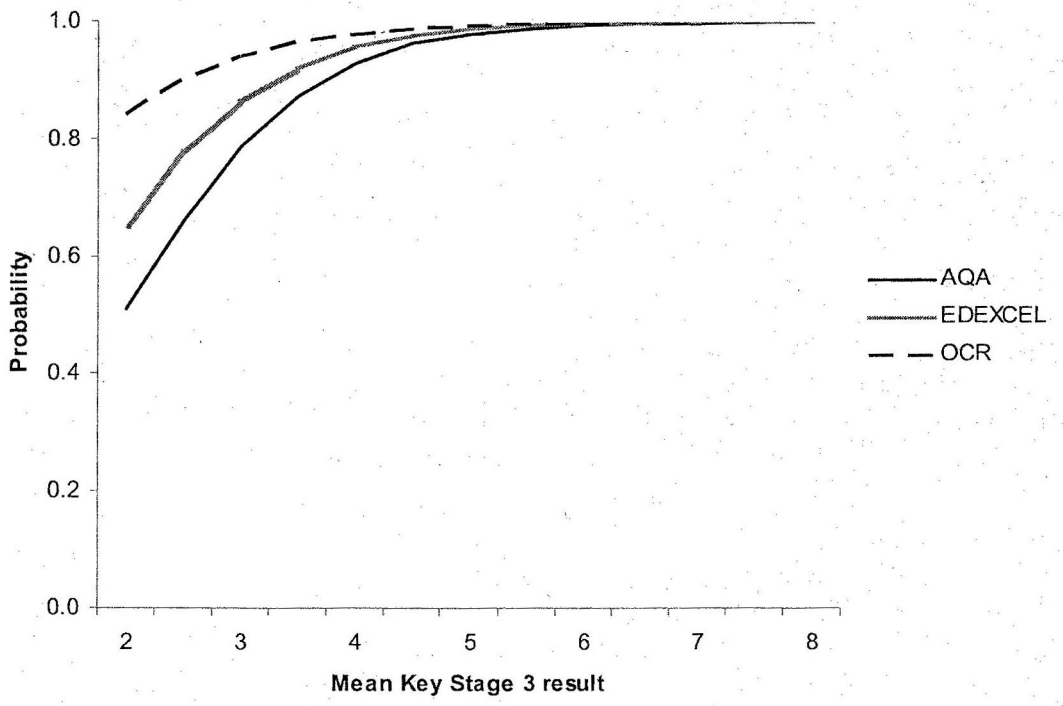
Terms of interaction between mean KS3 score and Awarding Body variables were introduced to the model, and the analysis showed a statistically significant interaction effect between OCR and mean KS3 score ($p = 0.000$). The relationship between candidates' mean KS3 scores and their chance of being awarded a grade F thus varies between Awarding Bodies, AQA candidates benefiting most from each unit increase in mean KS3 score. For those candidates who obtained grade F at GCSE, AQA's had the highest mean KS3 score.

Figure 3 shows the relative effect of Awarding Body on the probability of being awarded a grade F. OCR's and Edexcel's candidates had the highest probability for the lower values of the mean KS3 score. For values of five or more the probability of being awarded a grade F is almost the same for all Awarding Bodies. The model explained 52.1 per cent of the variation in the outcome variable. The model fit statistics and residual plots are given in Appendix 3.

Table 9: Parameter estimates of GCSE French model for candidates attaining a grade F or higher

		β	se	p	Joint χ^2	p	Odds
Fixed Effects	Constant	3.860	0.039	0.000			
	Centre type	2.129	0.222	0.000			8.406
	Mean Key Stage 3 result	1.274	0.014	0.000			3.575
	Sex	-0.827	0.024	0.000			0.437
	Edexcel	0.432	0.076	0.000	143.655	0.000	1.540
	OCR	1.084	0.095	0.000			2.956
	Edexcel*Mean KS3 result	-0.051	0.031	0.100	26.182	0.000	0.950
	OCR*Mean KS3 result	-0.190	0.038	0.000			0.827
Random Effects	Centre level	1.287	0.051	0.000			

FIGURE 3: The probability of exceeding the grade F threshold in each Awarding Body dependent upon mean Key Stage 3 result (for a female candidate entered through a non-selective centre)



CONCLUSION

Multi-level models were fitted to grades A, C and F, and Awarding Body appeared to be a statistically significant predictor of each outcome in the model. The award of the Edexcel specification appeared to be relatively lenient at grade A compared to that of the other specifications, and OCR and Edexcel appeared relatively lenient at grades C and F respectively.

At grade A, the AQA (3651) and OCR (1925) specifications appeared graded similarly to each other, and more severely than Edexcel (1226). The grading of the AQA specification was more severe than that of OCR and Edexcel at grades C and F, and Edexcel graded more severely than OCR at grades C and F.

The analyses showed that differences between Awarding Bodies appeared to exist after accounting for the hierarchical structure of the data and controlling for prior achievement at KS3, centre type and gender. These three explanatory variables were statistically significantly different from zero in all the models fitted for different grades.

Terms of interaction between Key Stage 3 result and Awarding Body variables were introduced to the model. The analysis showed that mean Key Stage 3 results have different effects for different Awarding Bodies on the predicted odds. OCR, Edexcel, and AQA candidates benefiting most from each unit increase in mean Key Stage 3 result at grades A, C and F respectively. It should be noted that some other variables might also have a significant effect on the probability of exceeding a given grade threshold which were not included in the analysis. The grade threshold was very small or zero in some level 2 units which gave evidence of under-dispersion at the candidate's level. This might violate the assumption of the models fitted. However the parameter estimates were not unbiased (see Goldstein, 2003).

References

Cresswell, M. J. (1996) *Defining, setting and maintaining standards in curriculum-examinations: judgemental and statistical approaches*. In H. Goldstein & T. Lewis (Eds), Assessment: Problems, Developments and Statistical Issues. Chichester: John Wiley & Sons Ltd.

Goldstein, H. (2003) *Multi-level Statistical Models*. 3rd edition. London: Arnold.

Hosmer, D. W. and Lemeshow, S. (1989) *Applied Logistic Regression.* New York: Wiley.

Jones, Meadows and Al-Bayatti (2004) *Report of the inter-Awarding Body Comparability Study of GCSE Religious Studies (Full Course) Summer 2003.* AQA, Manchester.

Pinot de Moira, A. (2000a) A *Comparability Study in GCSE Double Science: Statistical analysis of results by board.* A study based on the Summer 1998 examination and organised by AQA (SEG) on behalf of the Joint Forum for the GCSE and GCE.

Pinot de Moira, A. (2000b) *A Comparability Study in GCSE English: Statistical analysis of results by board.* A study based on the Summer 1998 examination and organised by AQA (SEG) on behalf of the Joint Forum for the GCSE and GCE.

Pinot de Moira, A. (2000c) *A Comparability Study in GCSE French: Statistical analysis of results by Awarding Body.* A study based on the Summer 1999 examination and organised by WJEC and AQA on behalf of the Joint Forum for the GCSE and GCE.

Schagen, I. and Hutchison, D. (2003) Adding Value in Educational Research - the marriage of data and analytical power. *British Educational Research Journal*. v29, n5.

Snijders, T. and Bosker, R. (2000). *Multilevel Analysis An introduction to basic and advanced multi-level modelling.* London: SAGE publications.

While, D., and Fowles, D. (2000) *A Comparability Study in GCSE Mathematics – Statistical analysis of results by board.* A study based on the Summer 1998 examination and organised by AQA (NEAB) on behalf of the Joint Forum for the GCSE and GCE.

APPENDIX 1: SOME STATISTICS FOR GCSE FRENCH DATA

Table A: Gender by Awarding Body

Awarding body	Male%	Female %	candidates
AQA	45.3	54.7	126831
Edexcel	46.5	53.5	38893
OCR	46.6	53.4	32940

Table B: Centre type by Awarding Body

Awarding body	Non selective %	Selective %	candidates
AQA	92.6	7.4	126831
Edexcel	90.5	9.5	38893
OCR	86.7	13.3	32940

APPENDIX 2: TECHNICAL TERMS USED IN MULTI-LEVEL MODELS

Using results in Table 7 for explanations:

β	Parameter estimate which is defined as the logged odds of success. A one unit increment in Mean KS3 score increases the logged odds of achieving grade A or higher by 2.593.
se	The standard error of the estimate which is defined as the standard deviation of the sampling distribution of the estimate.
p	The probability of the observed data when the null hypothesis is true. The null hypothesis here is: parameter = 0. The p-value corresponding to the Mean KS3 score is 0.000. This means it is very unlikely that the parameter for this variable would be zero and therefore the Mean KS3 score variable is significant in the model.
χ^2	The joint chi-squared test statistic used to assess the significance of a set of functions. The test statistic is 107.728 for testing the overall significance of Awarding Body variables in the model.
Odds	The ratio of the probabilities of the two states of a binary variable. In this context the odds of success is defined as $\text{Odds} = e^{\beta}$ In this instance, for every unit increase in mean KS3 score, the odds of exceeding the grade a threshold increased by more than thirteen fold, i.e. ((13.370 – 1) * 100) = 1237 per cent Odds ratio is the ratio of the odds for a binary variable in two groups of subjects, for example, males and females. In this instance, the odds ratio for sex indicates that 30 male candidates obtained grade A or higher per 100 female.

APPENDIX 3: MULTI-LEVEL MODEL FIT STATISTICS AND LEVEL 2 RESIDUAL PLOTS

The models used for the data carry an assumed variance-covariance structure. The binomial distribution is the standard distribution used to fit the binary data. If these assumptions are violated to some degree, the Maximum Likelihood Estimates are consistent, but the estimated sampling variances are underestimated for small clusters. If the estimated sample variance is greater than the theoretical variance, the situation is known as "over-dispersion" or "extra-binomial variation". This reflects a lack of independence or heterogeneity within clusters.

Residual plots for grade A showed that the normal scores of the residuals had formed a straight line which confirms the normality assumptions at centre level. The plots suggested that the standard residual is very small and constant for level 2 units. Fitting extra-binomial variation at level 1 gives a variance estimate of 54.027 with a standard error of 0.171 (using 1st Order and MQL approximation) indicating large departure from the binomial assumption. This suggested that the estimate of e_{ij} was over-dispersed due to lack of heterogeneity within some centres. However, although this finding may cast doubt on the validity of the model assumptions, but the model provided a good fit for the data which explained 96% of the variance. There are 3097 centres included in the model, 680 (22%) centres had no candidates exceeding the grade A boundary and 107 (3.5%) centres had no candidates below grade A. The 107 centres data were eliminated, for an attempt to see the effect of lack of variation within each centre on the parameter estimates. The resulting model was similar to that shown in Table 7 and the extra-binomial variation estimates went down from 54.027 to 36.238 with a standard error of 0.115.

Residual plots for grade C showed that the normal scores of the residuals had formed a straight line which confirms the normality assumptions at centre level. The plots suggested that the standard residual is very small and almost constant for level 2 units. Fitting extra-binomial variation at level 1 gives a variance estimate of 2.120 with a standard error of 0.007 (using 1st Order and MQL approximation) indicating departure from the binomial assumption. This suggested that the estimate of e_{ij} was over-dispersed due to lack of heterogeneity within some centres. There were 146 (4.7%) centres had no candidates exceeding the grade C boundary and 428 (13.8%) centres had no candidates below grade C. However, although this finding may cast doubt on the validity of the model assumptions, the model provided a good fit for the data which explained 75% of the variance.

Residual plots for grade F suggested a possible evidence of extra-binomial variation at candidate level. The variance estimate of extra-binomial variation at level 1 was 0.670 with a standard error of 0.002 suggesting under-dispersion. Within centres the candidates were less homogeneous than expected: 1450 (46.8%) centres had no candidates below grade F. However, although this finding may cast doubt on the validity of the centre level normality assumptions, the model provided a reasonable fit for the data which explained 52% of the variance.