Final report; 1000 word summary

The underlying aim of the project was to develop existing techniques for modelling hierarchically structured data in the context of Advanced level GCE examination results data and to provide also some substantive inferences from the data.
The *specific tasks* undertaken by the project were as follows:

- To compare models of examination results which use scoring systems with models that explicitly use grades.
- To study the effects on inferences of adjusting for measurement errors in predictor variables.
- To study ways of modelling efficiently multiple subject exam results where students choose different combinations of subjects.
- To provide substantive analyses of A level examination results, using data on 720,000 candidates over four years, especially with respect to gender, type of institution differences and changes in institution performance over time.
- To collaborate with the Department for Education and Employment (DfEE) in the acquisition and preparation of the data set.

Gender and institution type differences

Comparisons were carried out for the two subjects Chemistry and Geography. Females had higher average performances than males but made less progress on average between GCSE and A level. The greater progress for males is apparent only for those with average GCSE scores below about the 70^{th} percentile, above this females increasingly make more progress. Interestingly, when the overall total A/AS level point score is studied, the females tend, increasingly, to do worse than the males with increasing GCSE average score. This may be because the project did not study separately any humanities/arts subjects, which constitute the majority of examination entries.
All categories of selective institutions have the highest overall point scores by about three quarters of a (between-student) standard deviation compared to students in maintained comprehensive schools. Further Education (FE) college students have the lowest scores, by about two thirds of a standard deviation, below students in maintained comprehensive schools (equivalent to about 2 grades at A level.
Once GCSE scores have been adjusted for, only the independent selective schools and sixth form colleges show markedly higher scores than maintained comprehensives; for a student with a median GCSE score the difference is only equivalent to about a fifth of a standard deviation. Students from further education colleges, however, still fare worst, being about a fifth of a standard deviation lower than those from maintained comprehensive schools. The advantage of those in independent selective schools decreases with increasing GCSE score and the disadvantage of those in FE colleges decreases with decreasing GCSE score and is little different from those in maintained comprehensive schools below about the 30^{th} percentile.

Changes in institutional performance over time

An important practical issue when comparing the performance of institutions over time, and one which has considerable public policy relevance, concerns the stability of effectiveness measures. If results from one year to the next are studied there are very high correlations (0.88 over a three year period), but these largely reflect the fact that institutions have similar intakes over time and the adjusted (using GCSE) correlations are lower (0.55 over three years). There are very few institutions who can be identified, taking account of sampling error, as consistently improving or deteriorating over time.

Scales of measurement

The overall inferences drawn from point score and ordinal models are similar in terms of the relative sizes of effects and between-institution variation. Nevertheless, when making

inferences for individual institutions based upon estimated (posterior) residuals, some substantial differences emerge. This would be important in institutional effectiveness studies. Ordinal models convey information about differences in the way grades are distributed. Thus, for example, two institutions (or groups) can have identical average point scores but one may have a much wider spread of grades than another, and this can be illustrated for example in terms of gender differences where, for example, in Geography females exhibit less spread than males. In general it would seem more informative to report effects for institutions and groups in terms of individual grade probabilities or odds rather than mean point scores.

Multivariate models for subject choice

Candidates for public examinations choose different combinations of subjects. Such choices are purposeful, so that one can expect that the relationship between results from different subjects will depend on the overall combination chosen.

Performance in mathematics was chosen in order to study the methodological issues. At A level candidates may enter for up to four separate Mathematics papers. There is a basic 'main' paper and then some candidates enter for further papers; In terms of performance on the 'main' paper it is clear that average results are strongly related to the combinations chosen, for example those taking Further Maths tended to have higher scores than those taking only Main Maths. The major innovation of the project was to allow the effects associated with choice combination to vary randomly across institutions, showing in particular that this variation differed markedly between choice combinations. This allows estimates to be made for each institution of the specific 'effect' of each combination chosen which presumably reflects, to some extent, examination entry policies for institutions.

Measurement errors

It is well known that the presence of errors of measurement in variables in generalised linear models can lead to inferences different from those using variables from which measurement error has been removed. For multilevel models this has only been studied in the variance components case and the project has extended this to the case where predictor variables measured with error have random coefficients. It appears that the fixed effect estimates from the major analyses are affected only slightly by different amounts of measurement error, but the random effects are strongly influenced. In particular the between-institution variation increases with decreasing 'reliability' in the GCSE score. Thus, for example, at the mean GCSE score in the unadjusted analysis (that is, assuming a perfectly reliable GCSE score) the between-institution variation is just under 4% of the total and this rises to just over 5% for a reliability of 0.8 and top 14% for a reliability of 0.6. This has particular relevance for the interpretation of 'effectiveness' measures, where in previous analyses school 'effects' will tend to have been underestimated.

General

Dissemination has taken place via representatives on a steering group, seminars to policymakers and researchers, and publications.

The methodological work, especially that on measurement errors is being incorporated into the MlwiN software package so that it will become generally available.