Full Report of Research Activities and Results

Background and aims

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. Such data, within the UK public exam system, are typically reported as a series of grades in each subject taken. The data are used extensively for purposes of University selection and institutional comparisons, typically by assigning scores to examination grades and averaging these scores across individual subjects. Such use of the data relies upon particular assumptions about the utility of the scoring systems and the equivalence of these scores across the subjects taken by candidates. While the main thrust of the project is methodological, it has also undertaken analyses which are of substantive interest. In particular the project has studied the use of A level data in so called 'institutional (school) effectiveness' research, which attempts to study the relationship between student and institutional factors and to compare institutions, having adjusted for exogenous factors such as 'intake' achievements (in the present case the results of GCSE examinations taken two years earlier). In terms of policy this has particular relevance for the use of performance indicators based upon examination results.
The original aims and objectives were as follows:

1. To extend existing multilevel modelling techniques to analyse institutional performance data where the response is a set of ordered categories; where there is measurement error, and where there several responses, not all of which are present.
2. To provide important substantive information about the gender differences in different subjects in A level examinations, adjusting for GCSE performance.
3. To study institutional differences in terms of A level performance, especially differential performance by subjects and student characteristics.

The *specific tasks* for 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 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 and to share the results with them.

Acquisition and preparation of the dataset.

It was anticipated that this stage of the project would be completed within the first six months since the constituent data sets were available and had been used already by DfEE in preparation of their own statistical returns and performance tables. In the event this phase took somewhat longer.
The data consisted of records for some 720,000 students in some 2800 (16-19) institutions for the years 1993 – 1997 inclusive. The extent of the data and the need to transform it into a format suitable for multilevel modelling took longer than anticipated. Further problems then occurred as follows.

- There was a paucity of written documentation about the data. The data had been coded by different people according to different specifications, perhaps inevitably given the nature of the dataset, but this did make it extremely difficult to produce a clean file.

- Specific problems were as follows: The point scoring system used for the 1993 data was different from every other year and this did not emerge until the first analyses had been completed. It emerged that the 1993 data had different distributions from the other four years: no reason for this could be found and it was therefore decided to drop this year. When merging data for the four years it transpired that some institution identification codes had changed after 1995. The DfEE was able to provide a linking file which solved the problem. Several candidates had multiple entries and time was spent sifting the data to determine which of these were genuine.

Because of these data problems the analysis stage did not start until 12 months into the project and in the second year of the project further attention had to be paid to data cleaning tasks. In large measure these data problems resulted from the use of a dataset which originated from administrative records and we are very grateful to the DfEE for making it possible for us to resolve the various problems. We do, however, think that more careful attention needs to be paid in future to the consistent coding of examination results.

Gender and institution type differences

Multilevel models of increasing complexity were fitted to the data, typically starting with a simple variance components model through models incorporating random coefficients and those which adjusted A (or a combined A and AS level) level scores or grades for the corresponding GCSE examination results. The latter were available for all the students, both in terms of individual subject results and average scores across subjects together with number of subjects taken. A detailed description and analyses using the scoring system are given in Yang and Woodhouse (2001 and Appendix A).

In the event there was less attention to gender differences than had been anticipated because the steering committee member most interested in these, Dr Elwood, left the committee half way through on taking up a post in Northern Ireland. Nevertheless, there are some interesting results that did emerge.

While the project did not undertake extensive comparisons between all curriculum subjects, a detailed analysis was carried out for Chemistry and Geography. For both subjects, but especially for geography, females had higher average performances than males but made less progress on average between GCSE and A level, most markedly for Chemistry. For both Chemistry and Geography, however, the greater progress for males is apparent only for those with average GCSE scores below about the 70th percentile, above this females increasingly make more progress. In an analysis looking at subject choice in mathematics a similar result was found, with females making more progress above about the 40th percentile for GCSE. 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 we have not studied separately any humanities/arts subjects, which constitute the majority of examination entries, and this remains an area for future work.

Institutions have been classified by whether they are independent or maintained, by whether they are selective and within the maintained sector further by whether they are grant-maintained and whether they are sixth forms, sixth form colleges or further education colleges. This yields 11 categories plus a very small number who could not be classified. Without any adjustment in the model for GCSE scores (Yang and Woodhouse, 2001) all categories of selective institutions have the highest scores on average by about three quarters of a (between-student) standard deviation above students in maintained comprehensive schools, as would be expected. 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), again as would be expected.

Once GCSE scores, together with various interactions 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 (about one point on the standard A level scoring system) lower than those from maintained comprehensive schools. For independent non-selective schools and maintained selective schools all differences are small and non-significant. For students with high and low GCSE scores, however, the pattern is somewhat different. For example, 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 30th percentile. Note that a standard deviation in this analysis is equivalent to about 5 points on the usual A level total point score scale (which ranges from 0 to approximately 50).

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. Are some institutions consistently ‘effective’ in terms of ‘value added’ or adjusted measures and is it possible to identify institutions that improve or deteriorate in effectiveness over time?

The size of the dataset and the existence of four years worth of data make this sample very suitable for studying this issue. Appendix B contains details of the analyses carried out. 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). The analysis compared the predicted performance of each institution using the data from years 1994, 1995 and 1996 with the actual (adjusted) performance in 1997 and demonstrated only a moderate prediction with a correlation of 0.51. Also, there are very few institutions that can be identified, taking account of sampling error, as consistently improving or deteriorating over time.

Scales of measurement

For this, largely methodological, part of the project the two subjects of Chemistry and Geography were chosen to make a comparison between the use of a point score and the actual grades (Yang, Fielding and Goldstein (2000), Appendix C). Since the point scoring system is based upon the grades obtained, and since other scoring systems are possible, it is important to know whether the actual system used produces similar results to the use of the original grades themselves, and also what information may be lost by summarizing the grades by numerical scores.

The model chosen for the grades is the cumulative odds model where the response consists of the ordered set of odds given by the ratio of the probability of a grade A to grades B-F, the ratio of the probability of grades A or B to grades C-F etc. Clearly such a model retains as much as possible of the information given by the grades and utilises the ordered nature of them. The probabilities are connected to a set of functions of predictor variables via a link function, in the present case logit and probit links are used.

The project studied various assumptions. The simplest one for such models is to assume that the odds ratios are all a function of a common set of predictor variables, differing only by an intercept term. This was elaborated by allowing different coefficients (interactions) for each odds ratio. At the institution level in the hierarchical structure the simplest model assumes a simple between-institution variance term and this was elaborated by allowing the intercept term for each odds ratio to vary randomly. All of these models were compared to models using point scores as outcomes.

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 also 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 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. The results from this project provide a useful starting point for such analysis and reporting.

The project has also looked at ordinal models with cross classifications at higher levels (Fielding and Yang, 1999 Appendix D). Further work in conjunction with Anthony Fielding is proceeding to fit models with multivariate ordinal responses and models with mixtures of ordinal and continuous responses.

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 (scores or grades) from different subjects will depend on the overall combination chosen. Thus the standard approach which considers the structure as one with a multivariate response where many of the responses are missing at random, is not applicable. When modelling such relationships account needs to be taken of the chosen combinations. The study of such relationships among subjects is useful in contributing towards an understanding of student performance and also ‘effectiveness’ at the level of the institution; current research into ‘school effectiveness’ is increasingly becoming concerned with individual departmental and subject performance (Goldstein and Woodhouse, 2000). The analysis undertaken in the project appears to be the first systematic analysis of this issue.

Performance in mathematics was chosen in order to study the methodological issues in modelling multivariate outcomes in the presence of informative choice of response combination. Full details are given in Appendix E. 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; the project looked at Pure maths, Applied maths and Further maths, accounting for the majority of additional entries. In terms of performance on the 'main' paper it is clear that average results are strongly related to combinations chosen, for example those taking further maths tended to have higher scores than those taking only main maths. A series of multivariate response models, adjusting for average GCSE score and GCSE maths results were fitted to the data, where adjustments were made for the combination chosen. The four responses were Main maths, Further maths, Pure maths and Applied maths. Correlations among responses (main and Further Maths, Applied and Pure Maths) at student level were relatively high (about 0.7) and higher still at institution level (about 0.9). At institution level it was also possible to estimate correlations between A and AS level results and these turned out to be low for each response (about 0.3). This weak relationship may reflect organizational and teaching structures within institutions and would be an interesting area for further research.

The project also modelled the effects associated with choice combination as varying randomly across institutions. This showed that the effect of combination choice did vary among institutions, 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. The modelling also showed that the between-student variation differed among combinations. A similar approach to modelling multiple responses, where choice is involved, may be applicable in other situations such as exam question choice. A paper describing these analyses is attached as Appendix E.

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. Woodhouse et al (1996) showed that this was also true in multilevel models. However, they only studied the variance components case and the present project aimed to extend this to the case where predictor variables measured with error have random coefficients. Appendix F describes this work in detail.

The approach to this problem is via MCMC estimation since, as Woodhouse (1998) showed, maximum likelihood and moment based estimators encounter considerable difficulties in these cases. On the basis of simulations we have determined that the procedure adopted does provide satisfactory estimates, at least for the single variable with error case. The procedures have been incorporated into a development version of MlwiN. Work is continuing to extend the models to cases where there are several variables measured with error where the errors may be correlated, to the case where a polynomial or other function of a variable measured with error is used and to the case where misclassification can take place in categorical variables. The project is collaborating closely with Dr Dougal Hutchison of the NFER who holds an ESRC research grant to investigate errors of measurement using bootstrap procedures.

The procedure has been applied to a subset of the A level dataset to ascertain the extent to which adjusting for errors of measurement affects inferences (Appendix G). It appears that the fixed effect estimates from the major analyses, reported above, 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 to 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. It is intended to investigate further educational and other data sets in the continuing ESRC funded project (R000238217 – Applications and understanding of multilevel modeling in the social sciences).

Dissemination

The steering group for the project consisted of representatives from the DfEE, from Local Education Authorities as well as academics. This represented one forum for dissemination of results. All the papers supplied in the Appendices have been submitted for publication (the paper in Appendix A has been accepted). Seminars presenting the results of the project were held at the Institute of Education in November 2000 and January 2001 attended by policy makers, school teachers and researchers. A paper on the work was presented at the British Educational Research Association in September 2000 and also to the 5th International Conference on Social Science Methodology, Cologne, October 2000. The public presentations have concentrated on the substantive findings of the project, and it is intended to present the methodological findings at statistical meetings and conferences during 2001, including a presentation to the Education section of the Royal Statistical Society. A short report appeared in the Multilevel Modelling Newsletter in August 2000 (Appendix H). A subset of the data and a copy of several papers have been placed on the Multilevel Models Project web site. The data used in the project have been deposited with the Data Archive.

General

The project has met its aims in terms of covering the areas set out in the grant application. Substantively, a number of new findings have emerged, especially on gender differences, institution type differences and the variability of institution effects over time. Because of the extended period of data cleaning, the project was not able to pilot interpretations of results with a small number of institutions as proposed in the grant application. Nevertheless, the results have been presented to several audiences, including policymakers, and feedback from these groups has gone into the various articles and reports of the project. We believe that methodologically the project has produced some important new work. It has carried out detailed work on the appropriate model to use when adjusting A level data for GCSE scores. The work on measurement errors particularly has provided an important new methodological tool and it is intended to develop this further methodologically and to apply to other data sets. The approach to handling incomplete multivariate (subject) responses where combinations are at choice, is novel and provides an important tool for use in similar datasets. The comparison of point scores with the use of ordered grade category models has demonstrated the usefulness of the latter for providing more detailed characterisations of institutional differences. The methodological work on measurement errors is being incorporated into the MlwiN software package so that it will become generally available. The methodology developed by the project is broadly relevant in many other areas of social science data analysis, where measurement errors occur, where ordinal scales are used as in surveys, and where incomplete multivariate responses are obtained.

Acknowledgements

We are most grateful to the members of the project steering committee who gave valuable advice and encouragement. We are especially grateful to Trevor Knight and Audrey Brown for their help with the data and for general support.

References

Goldstein, H. and Woodhouse, G. (2000). School effectiveness research and Educational Policy. *Oxford Review of Education* **26**: 353-363.
Woodhouse, G., Yang, M., Goldstein, H. and Rasbash, J. (1996). Adjusting for measurement error in multilevel analysis. *Journal of the Royal Statistical Society, A.* **159**: 201-12.
Woodhouse, G. (1998). *Adjustment for measurement error in multilevel analysis*. Institute of Education. London, University of London.

Appendix References

A: Yang, M., and Woodhouse, G. (2001). Progress from GCSE to A and AS level: Institutional and gender differences, and trends over time. *British Educational Research Journal* (to appear).

B: Gray, J., Goldstein, H. and Thomas, S. (2001). Predicting the future: the role of past performance in determining trends In Institutional effectiveness at A-level. *British Educational Research Journal.* (to appear).

C: Yang, M. Fielding, A. and Goldstein, H. (2001). Multilevel ordinal models for examination grades. Submitted for publication.

D: Fielding, A. and Yang, M. (2001). Ordered category responses in multilevel and cross-classified structures. Submitted for publication.

E: Yang, M. Goldstein, H., Browne, W. and Woodhouse, G. (2001). Multivariate multilevel analyses of examination results. *Journal of Royal Statistical Society, C.* (to appear).

F: Browne, W., Goldstein, H. Woodhouse, G. and Yang, M. (2001). An MCMC algorithm for adjusting for errors in variables in random slopes multilevel models.

G: Goldstein, H. (2001). Measurement errors in examination scores.

H: Multilevel Modelling Newsletter, August 2000.