Award Number: R000237394

Date: 22 November 2001

Professor Harvey Goldstein
Dept of Mathematical Sciences
Institute of Education
University of London
20 Bedford Way
London
WC1H 0AL

Dear Professor Goldstein

APPLICATION OF ADVANCED MULTILEVEL MODELLING METHODS FOR THE ANALYSIS OF EXAMINATION DATA

I am writing to report the outcome of the evaluation of your research project. An overall grade on the scale Outstanding, Good, Problematic or Unacceptable is assigned to each Report. Your project has been graded as Outstanding: 'High quality research making an important contribution to the development of the subject. An Outstanding grade indicates that a project has fully met its objectives and has provided an exceptional research contribution well above average or very high in relation to the level of award.'

I have enclosed the unattributed comments of the rapporteurs together with their grades, which we hope you will find constructive and useful.

The ESRC does not publish the details of project evaluations and they remain confidential to the Council and its Boards. We may, however, publish reports containing general details of the outcome of completed research awards together with their grades; these reports are likely to be made available to a wider audience. We will lodge your End of Award Report with the British Library where it will be available for public access, as a record of the research you have undertaken with ESRC funds.

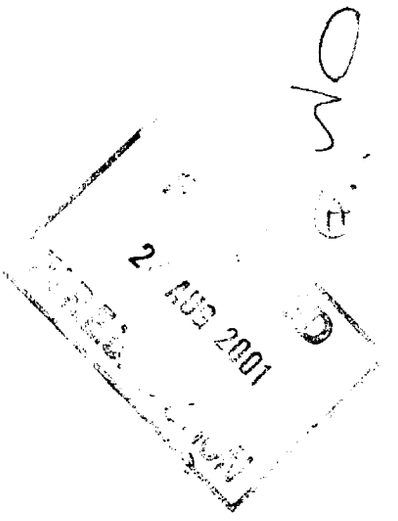
We are keen to maintain records of the output from ESRC-funded research. The staff of our publication's database will be contacting you periodically to enable you to keep us up to date on the published output of your research. We would appreciate your co-operation in this matter.

POLARIS HOUSE
NORTH STAR AVENUE
SWINDON SN2 1UJ
TELEPHONE 01793 413000
FACSIMILE 01793 413001
GTN 1434
http://www.esrc.ac.uk

I must apologise for the delay in feeding back to you on your grade and thank you for your co-operation with ESRC's evaluation. I hope that it has been of use to you and your work. If there are any points you wish to raise in response to the evaluation, please put them in writing to me not later than four weeks from the date of this letter.

Yours Sincerely

Suzanne Tanner
Policy and Evaluation Division
Tel: 01793 413112
Fax: 01793 413128
suzanne.tanner@esrc.ac.uk



Report on; Application of advanced MLMs for the analysis of examination data. #R000237394

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

The report and the accompanying papers show that significant advances have been made in each of the stated areas of research. The project involved the manipulation and analysis of a very large data set provided by DfEE. Cleaning the data set turned out to be more time consuming than anticipated . This highlights one of the problems of using routinely collected administrative data, namely that it is rarely suitable for immediate analysis. The ESRC should note this and ensure that research projects aiming to use such data are given sufficient time and resources to clean the data before undertaking the methodological and substantive studies proposed.
The substantive outcomes are summarised in the report and in the MLM Newsletter paper by Yang and Woodhouse. They show The importance of the MLM framework for the analysis of educational data. Complex models are needed to capture institutional types and the variation of institutions within type. Also subject combinations are not chosen at random and so summary measures need to take into account the subject combinations deliberately chosen and the covariances between subjects. In the UK there is the added complexity of A and AS levels, of different exam boards and of the choices of papers, especially in maths, within boards. This highly complex structure cannot be summarised by simple institutional averages. The models must capture the complexity and the analyses must reflect this. The substantive findings show this to be necessary.
The importance of value-added analyses is convincingly demonstared. THe conlusions are changed when A level results are adjusted for GCSE results. However, the analses also show that value- added analysis is not simple, it has to take into account the complexity of the system. The papers show how this may be achieved and represent a significant advance in the methods of analysis. Of particular interest are the results on gender effects, some of which appear to show that males make better progress than females between GCSE and A level, contrary to current perceived wisdom.
Another major substantive finding is that relating to trends in institutionl performance over time. Very few institutions can demonsrate consistent improvement and there is clear evidence of the well-known phenomenon of regression to the mean. It would have been interesting to see a non-parametric analysis of trends based on run lengths and on permutation tests. It may be that the vast majority of institutions, when results are adjusted for value added, behave like a set of independent random variables. If this is true then the concept of a league table is a nonsense.
The project also looked at the problems of measurement. It is demonstrated that measurement errors can have a significant effect on results. It is also shown that the results can be analysed as ordered categories, ther is no need to make the implicit assumption of the normality of the underlying scores. However, the substantive results seem to show that in practice

there is little difference between the conclusions from the two types of analysis. Despite this an analysis based on correct assumptions , namely ordered categories, should be preferred to one based on incorrect assumptions even if the results are similar. Once software is available the correct analyses should become the norm.
In summary the MLM team continue to produce research of the highest standard. They challenge perceived wisdom by analysing data within its true social context. But above all they provide the methods and the software that enable others to replicate their analses and to advance them if necessary.
I grade this project as OUTSTANDING.

========

Research quality

How would you rate the overall quality of this research?

(Please tick once only)

☑ **Outstanding**

High quality research making an important contribution to the development of the subject.

An **Outstanding** grade indicates a project has fully met its objectives and has provided an exceptional research contribution well above average or very high in relation to the level of award.

☐ **Good**

Good quality research making a useful contribution to the development of the subject.

A **Good** grade indicates a project whose research activities and contribution is fully commensurate with the level of award, approach and subject area, and which has addressed its major objectives.

☐ **Problematic**

Acceptable research but with problems or weaknesses in the design, method, analysis or outcomes.

A **Problematic** grade indicates a project which has failed to address one or more of its major objectives, has encountered significant difficulties in the execution of the project, has incomplete work, or has achieved substantially less than expected for the level of the award, the approach or the subject area.

☐ **Unacceptable**

Research poorly conducted with unreliable results, or report with insufficient details on which to base a satisfactory judgement.

An **Unacceptable** grade indicates a project which has failed to conduct the work as agreed at the time of the award (and any subsequent agreed changes to the work plan); for example, failure to conduct agreed surveys or analyses, or failure to address most of the major objectives.

RAPPORTEUR'S COMMENTS ON END OF AWARD REPORT FORM R000237394 (GOLDSTEIN ET AL)

1. Activities and achievements

This project has fulfilled its aims to an outstanding level, in the following most important respects:

i It has developed innovative new mathematical theory and converted that into useable computer software, dealing with multilevel approaches to measurement errors, ordered categorical data, and non-randomly missing data. The software is an outstanding research product. This work is not only first class in a UK context; it would rank among the leading methodological work for quantitative social science internationally.

ii The quality of the resulting data analysis is also excellent. The researchers manage to distil clear messages from the data while also paying proper attention to checking validity, and to variation. One of the main reasons why they have not allowed variation to stand in the way of clear results is that they have had a lot to say about variation itself: in many ways, the study of variation is what this project (and its predecessors) have been all about. The researchers have therefore shown that quantitative research can be highly subtle, and need lose none of the sensitivity to local context that would more normally characterise qualitative research. However, see comments on further research below (point v).

iii The data collected are of a very high quality for the purposes in hand. A better data set for facilitating methodological development could not be envisaged, consisting as it does essentially of census-type material for a period of four years. The researchers spent a great deal of time ensuring the quality of the data, and they have adhered to the usual ethical rules concerning survey data in their use of it.

iv The project has contributed to our applied knowledge about the stability of school effects, about gender differences in school attainment, about the relevance of patterns of subject choice in public examinations to examination attainment, and about differences between institutions in their examination performance. These topics all relate to the important policy issue of measuring the performance of educational institutions.

v The work made use of the excellent network of research and policy contacts which the Institute of Education team already had. This is one of the best-established research networks in quantitative social science in the UK, and maintaining it is of lasting benefit in ways that will not have been explicitly recorded here (eg in exposing PhD students and junior researchers to discussion and to draft papers).

2. Highlights

The main highlight is the methodological development, especially in respect of a multilevel treatment of ordered categorical variables and of measurement error. The consequential development of the software is significant because it will make these important mathematical advances available to other researchers.

The main policy impact will be in debates about measuring the contribution which educational institutions make to their students' progress. No-one reading the publications from this project could continue to believe that there is a simple way of measuring 'school effectiveness'. The researchers have offered technically valid and feasible ways around some of the dilemmas. Their conclusions are often that most institutions do not differ from each other in their effects, and that even fewer reliably differ from each other over a period of years.

This kind of conclusion does bring into question the ESRC's rather crude notion of 'impact on policy makers', since no amount of good-quality research is likely to deter

policy makers at present from seeking to measure schools statistically. The relevant impact of research of the present quality, then, is not so much on policy makers as on the quality of the debate which teachers and others can engage in with those policy makers who try to evaluate their institution.

3/4 Dissemination and audiences

That last point raises the issue of dissemination. The source of the data meant that the project remained close to people in the DfEE, but, because of the unavoidable delay in producing a good-quality data set, the stage of disseminating results to schools, colleges and LEAs has had to be postponed. I think this last stage, however, is of great importance, not only for the reasons outlined under 2 above, but also because it represents an excellent opportunity for the wider dissemination of good-quality research: teachers and LEA policy makers will make a receptive and easily accessible audience.

The dissemination strategy to academic outlets has been very good, and indeed I had already come across two of the project's papers (not directly from the researchers) before I received the request to write this report. For a three-year project with one full year spent on data collection that is highly commendable.

5. Further research

I would identify five areas where the work could be taken forward. Saying this is not at all a criticism of the present project: it has achieved an impressive amount in three years. It is in fact the quality of what it has achieved that raises questions about further work.

i Make the new software available through the standard issues of MLWin. From comments made in the end of award report, and from the track record of these researchers, I have no doubt that this is already in hand.

ii Extend the analysis to more stages of school education. Progress from GCSE to A level is interesting, but in terms of young people's overall experience of pre-18 education is relatively unimportant. For example, we know from quite a lot of research that most of the social class differences in attainment are already evident by early in primary school (most recently from John Bynner's work using the BHPS, and also from ESRC-funded work at the Centre for the Analysis of Social Exclusion at the LSE). Of course, achieving data links back to data at earlier stages (eg data on Key Stage assessments) would be an even more difficult task than that encountered here, and would also presumably involve problems of confidentiality.

iii It would also be interesting to relate the findings here to progress in stages of education beyond age 18, notably in higher education. An earlier ESRC-funded project by Andrew McPherson and Chris Robertson ('Schools' effects on attainment in school and higher education', 1994) found that students from schools that had higher value-added did relatively less well in first year at university than similarly qualified students from schools that appeared to be less 'effective'. This raises questions about how we should 'control' for prior attainment: perhaps the rate of progress in earlier stages should be controlled for as well as level reached. It could be that the present project's work on measurement error and on subject presentations could give new insights into this kind of finding. Might apparently standardised public examinations actually be subject to more measurement error than is commonly supposed, in the sense of 'error' induced by particular school policies on presentations, or by schools' encouragement of 'cramming' for examinations?

iv Extend the range of variables. This would not be feasible with the present data set, and may be possible only by working with LEAs (as the present researchers have done on previous occasions). But the present project's findings on how the differences between institutions diminished when GCSE performance was controlled for raises interesting questions about what would happen if further controls were introduced (eg social class, ethnicity) or if progress through earlier

stages were introduced as a control for progress between GCSE and A level. This work would allow the research to address questions of social and educational theory which could not really be dealt with using the kinds of data available to the present project.

v The other variables might include not only further individual-level data, but also information about school and LEA policies. For example, might it be more educationally effective for a school to put most of its effort into making progress between intake and GCSE? Such a school would probably have a relatively poor 'effectiveness' measure in the present data set, but might overall be of relatively high effectiveness.