UNIVERSITY OF BIRMINGHAM

USING CROSS-CLASSIFIED MULTILEVEL MODELS TO IMPROVE ESTIMATES OF THE DETERMINATION OF PUPIL ATTAINMENT: A SCOPING STUDY

School of Education
University of Birmingham

Department of Economics
University of Birmingham

August 2006

USING CROSS-CLASSIFIED MULTILEVEL MODELS TO IMPROVE ESTIMATES OF THE DETERMINATION OF PUPIL ATTAINMENT: A SCOPING STUDY

When citing this report please show authorship in full as follows:

Fielding, A., Thomas, H., .Steele. F., Browne W., Leyland A., Spencer, N., Davison, I.. (2006)

This report was written by Professor Antony Fielding of the Department of Economics, University of Birmingham. It arose out of a scoping project funded by the Department for Education and Science and co-directed by A. Fielding and Professor Hywel Thomas of the School of Education, University of Birmingham. Much of the modelling analysis and its computation were undertaken by various consultants; Dr Fiona Steele (Centre for Multilevel Modelling, University of Bristol), Dr Bill Browne (Mathematical Sciences, University of Nottingham), Dr Alastair Leyland (MRC Public Health Unit, University of Glasgow) and Dr Neil Spencer (Business School, University of Hertfordshire). Background research work on cross-classed structures and literature reviews were contributed by Dr Ian Davison (School of Education, University of Birmingham).

ACKNOWLEDGEMENTS

The report writer would like to thank all the members and visiting fellows of the Centre for Multilevel Modelling, Graduate School of Education, University of Bristol for the many discussions in the area which have informed the report. In particular, Professor Harvey Goldstein undertook critical reading of drafts. Gratitude is also given to Professor John Shepherd (Department of Geography, Birkbeck College, University of London) for geographical information systems advice. Finally we are very grateful to the DfES researchers and officials for their patience and invaluable guidance, and in particular Jen Helmshaw.

The views expressed in the report are the authors' and do not necessarily reflect those of the Department for Education and Skills

IBSN: 9780704426016 (International)
0704426013 (UK)

CONTENTS

Preface: Background and Acknowledgments

The starting point for this scoping study was two reports to the Department for Education and Skills. We refer to them in the current report and their contents must be regarded as pre-requisite reading. The first, which we shall refer to as 'Report 1: Vignoles' (Vignoles et al (2000)), reviewed existing research on the relationship between resource allocation and pupil attainment. It was felt that a wide variety of previous studies which failed to demonstrate such relationships (Hanushek, 1997) may suffer from methodological deficiencies. Two of these were addressed to some extent in a second report, where the outcome was attainment at Key Stage 3, and which we shall call 'Report 2: Levačić' (Levačić et al (2005)). Firstly, previous evidence had often rested on data aggregated to too high a level. What might be required was structured data from individual pupils in schools and higher level unit contexts. Secondly, statistical modelling often did not recognise the endogeneity of the resource variable. This is the process whereby some of the same factors affecting resources also affect achievement leading to a mutual intertwined dependence. The approaches to these issues were also extended to GCSE outcomes in a follow up report (Jenkins et al (2006). This issue of endogeneity in these contexts has also been discussed by Mayston (2002).

The aim of this report is to scope the possibilities for extending the specification of the educational production functions for Key Stage 3 outcomes to encompass area of residence effects of pupils in addition to their school contexts. Not only are pupils nested in schools which they attend inducing intra-school correlations but they also are clustered hierarchically by geographical areas in which they live and this must be accounted for. Further if there is variation in outcomes due to net area effects over and above other effects including resources it is important that there is adequate control for these in well specified models (Fielding 2005)). Gibbons (2002) considers that there are such real effects which though relatively small may impact on education production functions.

The approach will consider the potentialities of the use of multilevel cross-classified random effects models as a framework for these extensions. In some senses it further elaborates by incorporating crossed area effects in a multilevel model framework for school clustering suggested but not fully discussed in 'Report 2: Levačić'. The latter idea has, however, been more fully discussed in Steele et al (2006). The present report and a companion review of complex multilevel models, 'Report 3: CC Review' (Fielding (2005)), have been written by Antony Fielding (University of Birmingham, Department of Economics). However large parts of the discussion, in particular that of the experimental trial results, have been informed directly by work undertaken by our consultants. These are Professor Harvey Goldstein, Dr Fiona Steele (both of Centre for Multilevel Modelling, Institute of Education, University of Bristol), Dr William Browne (University of Nottingham, School of Mathematical Sciences), Dr Alastair Leyland (MRC Public Health Unit, University of Glasgow), and Dr Neil Spencer (University of Hertfordshire Business School). Mr Ian Davison (School of Education, University of Birmingham) worked on some of the investigative analyses forming the basis of Section 3 and Appendix 1. He also undertook some literature searches in connection with the review. The whole project has been jointly directed by Antony Fielding and Professor Hywel Thomas (School of Education, University of Birmingham). Professor Thomas has been a constant facilitator. We have also benefited from airing some of the problems arising in free ranging academic discussion at the monthly meeting of Fellows and Staff at the Centre for Multilevel Modelling, which includes many of our consultants and the writer.

1 Introduction

Using data from the Pupil Level Annual Census merged with data on schools and characteristics of social environment from other sources, 'Report 2: Levačić' considered handling the issues of endogeneity of resources and disaggregated child level data through a detailed empirical exercise. Generally this was able to show that once these two particular issues were addressed then the effect of resources was somewhat more evident than might previously have been thought. Two approaches were used. The bulk of the analyses rested on the first: Instrumental Variable (IV) estimation. This is a well known approach to endogeneity in the econometrics literature. It was further adapted to the clustered structure of the data whereby pupils are hierarchically arranged in secondary schools which are in turn nested within Local Education Authorities (LEA). Robust standard error estimation was incorporated. It is well known that unless this form of estimation is used the clustered structure of the data will overstate precision of results leading to poorer inferences. Apart from tables of results, little detail was given in the report of the use of the second approach; explicitly specified multilevel models to handle the data structure. However, it has been subsequently been reported more fully by Steele et al (2006).

In this approach an additional feature to handle resource endogeneity was by specifying linked simultaneous equations for both resources and achievement. Taking these features together this approach leads to what is referred to as multiprocess multilevel modelling. It will become apparent in what follows that this is our preferred approach to analysis and estimation of education production functions involved.

The previous models used, however, have been estimated recognising only a relatively simple one-way hierarchical structure with pupils nested within schools which were themselves nested within LEAs. As such they do not explicitly consider the possibility that the areas within which a pupils lives will also impact on their attainment. The inclusion of area effects using an additional geographical residence hierarchy may improve the specification and evaluation of other net effects in the educational production functions including school resource effects. Goldstein (2003) and 'Report 3: CC Review' discuss more fully the methodolological reasoning behind this. Such an extension will also allow assessment of the importance of net effects of residence area and then may give background insight into the relative importance of area based initiatives relative to school based ones. Thus this report, using KS3 outcomes will scope the possibilities of using a cross-classified multilevel modelling approach in various ways indicated below. It will also consider alternative ways of handling the endogeneity of resources issue within this general framework.

As a product of the research for 'Report 2: Levačić' the final database was made available for the present scoping report. The general aim of this scoping was to investigate to what extent the model of production functions used could have a further extended specification to include effects not only of school and LEA but also of areas in which the pupils lived. Since such models are likely to have more complex features a main aim was to investigate the feasibility of applying existing methodology and software to handle them. This feasibility may also be affected by the large size of the overall set of data and the range of definitions of area possible. Area of residence identifiers are available in the dataset provided to us in the form of a hierarchy of postcodes, census output area, ward, and local authority districts. This is a separate hierarchy from the educational one of schools and LEAs. The cross-cutting of these hierarchies is thus more complex to analyse than strictly hierarchical structures and must be handled by models and methods of analysis appropriate for cross-classified multilevel effects. To take an example we might regard pupils as nested within a level 2 of schools in the education hierarchy but also they

are nested within a ward level of the residence hierarchy. Further schools do not draw all their pupils from the same ward and pupils in particular wards may attend many different schools. Thus at a more complex higher level we may regard pupils as nested within particular cells in the cross-classification of wards and schools.

The first part of our scoping study has already been produced in the form of a description and literature review of the approach to such structures known as *cross-classified multilevel random effects* modelling; the report we refer to as 'Report 3: CC Review'. It also contains an extensive review of more complex extensions to cross-classified models known as *multiple membership models*. These will be required in future if research is to be undertaken on such factors as continuity of effects from longitudinal data or effects of combinations of teachers on an achievement outcome. The review also introduced in a readable way many features of statistical methodology which can be quite technically daunting. We draw on these in the current report but we would not wish too much repetition. Thus we regard the review as a necessary precursor to the understanding of some of the ideas introduced here. In this sense the review and this report should be taken together as complementary documents. A greater understanding of what we have to say here will be facilitated by a prior examination of the review.

Apart from the production of the review, and given the above background and introductory remarks the rest of our remit for this scoping study was as follows:

- We would interrogate the dataset made available to us to satisfy ourselves of its integrity and suitability for the present purpose. Any professional statistician needs to do this with second hand data before embarking on complex analyses.
- We would undertake some preliminary investigation of the nature of the crossed classifications involved in the data hierarchies in terms of balance and sparseness. These are features which govern certain statistical properties and computational feasibility of available estimation methods.
- We would investigate the possible ranges of estimation procedures available in terms of their appropriateness for the type of models and data to hand. This might include instrumental variable estimation, structural equations (multiprocess) and Monte- Carlo Markov Chain (MCMC) estimation.
- We would consider the range of software that is available for various estimation procedures and consider what computational constraints there might be that are not *a priori* apparent.
- We would investigate the appropriate specification of area effects and consider for the future models based on diffuse spatial effects. This would also include which type of area identifier and how many levels might be incorporated into models.
- We would produce some trial analyses in the light of these investigations on subsets of the data. These would highlight the potential uses and limitations of the range of methods and software investigated

To some extent we have covered certain of the essentials of the third and fourth bullet points of the above remit in writing 'Report 3: CC Review'. We will draw on these in this report but will not risk over rehearsing too much. The discussion of the potential of spatial modelling has also been explored in that review and will not be further discussed here. For the rest we now briefly describe the structure of this report and the broad issues addressed.

In Section 2 we report in some detail on our detailed investigations of the available data. We anticipated some problems of interpretation and handling which often occurs with provided data

from a secondary source. However, the problems uncovered and their resolutions were more formidable and time consuming than originally envisaged. We also had more than a little difficulty in preparing the STATA data file provided in forms suitable for other specialist software which handles multilevel models. We will conclude that this dataset or similar ones may need a lot of further attention if they are to be used in any more detailed future research.

Section 3 firstly considers some of the reasons for our choice of subsets of the data with which we later experiment. Our concern here is also to examine the structure of the cross-classes from the point of view of separation, imbalance and sparseness. These are all factors which affect statistical properties and computationally feasibility of any proposed model and methods to estimate it. Of particular importance was that we discovered a thin spreading over different areal units of a certain minority though not insignificant proportion of a school's children who lived in a different LEA than the one in which they went to school. Though we did not directly investigate this in particular, it has been pointed out to us that the borough organisation of education in London might make this phenomenon even more marked for London LEAs.

Section 4 of the report considers the possible model structures and their possible statistical and substantive interpretations. We also discuss some possible implications for methodology and software of these model frameworks and their development. Not the least of these is the user friendly properties of software for iterative model development. We also raise the possibility of using a different areal unit in modelling but which is not presently identified in the dataset: census super output area. We also discuss the way in which we require to merge onto the dataset another area identifier: the LEA area in which pupils live. Section 5 contains results of our feasibility experiments with the variety of models, estimation procedures and software. All of these previous sections inform the conclusions we reach and the recommendations we make in the final section 6.

2 Investigation and Editing Of Dataset

This scoping study made use of the dataset used in 'Report 2: Levačić'. This dataset is huge (464783 pupils and over 280 variables), assembled from different sources. The scoping revealed detailed errors and anomalies. It appears that coding decisions and variable manipulations have been made that are not fully documented in the information provided for the scoping. Consequently, it took considerable time to unpick some of these issues. More importantly, further data cleaning and greater understanding of the previous work might be required to enable construction of appropriately rigorous cross-classified multilevel models. It is clear that considerable attention would be required to the state of any database which might be provided for any future larger exercises.

2.1 The Data As Received

The database made available to the investigators consisted of a prepared STATA worksheet used for 'Report 2: Levačić'. It contained records for 464783 children at Key Stage 3 (KS3) in 2003 from the Pupils Annual School Census (PLASC). Added to PLASC data were school data derived from the Annual Schools Census and Registrar of Educational Establishments and Section 52 data on individual schools' revenue and expenditure. The pupils' postcode of residence also facilitated the addition of some socio-economic indicators from census output areas including this postcode. The range of these indicators in the database was limited and driven by what was thought relevant from exploratory analyses in 'Report 2: Levačić'. Additional sources of data which were merged with the individual pupil database appeared to be Local Government Financial Settlement Reports, Guardian Local Authority Directory, Local Elections Centre at the University of Plymouth and the geodemographic system ACORN from CACI Information Solutions. 'Report 2: Levačić' gives details of these sources.

It is clear that considerable work went into assembling this database since it required careful integration from a number of sources. However, as always with the secondary use of such databases by statisticians, unclear details of this assembly may pose many difficulties in its further use. We had such difficulties in this scoping exercise. Modelling and data organisation provide an iterative interplay and the former is assisted if secondary data sets have greater clarity of description than is the case here. In Appendix 1 we draw attention in some detail to what we regard as deficiencies in the clarity of the way this particular dataset was presented for secondary analysis. Though particular to this dataset the detail does provide some general pointers to what may be required in general for adequate secondary analyses in similar situations. For instance the, first section of the appendix, A1.1, demonstrates the lack of adequate description of several variables relating to pupil, school and LEA identifiers in the data. This is compounded by the obscurity and apparent inconsistency of missing value codes. Conventions adopted for the latter had not been detailed in any transparent documentation. Some additional detail connected with the missing value issue is also considered in Appendix 1.2. For these various reasons considerable difficulty was thus encountered in preparation for analysis.

This initial exploration of the status of key level indicators necessary for multilevel exploration proved symptomatic of other problems later uncovered. Most statisticians find it necessary to subject acquired datasets to careful investigative scrutiny before undertaking analyses. Cleaning, editing, and checking reliability and unravelling anomalies is an essential part of this process particularly when complex modelling is envisaged. This is often problematic if the data set is third or fourth hand passed on by a previous analyst's compilations from primary sources. A lesson to be drawn is the paramount importance of ensuring at a minimum the availability of

careful documentation of such data sets and accompanying codebooks of the sources, nature and content of the data. Lack of these meant that detailed scrutiny of some aspects of the database was a major initial focus of the current scoping work. The lack of codebooks meant that determining the structure of each variable, which is necessary in initial exploration, required prior analytical tabulation and listing. Some ingenuity was then required in interpreting values for some variables, in particular the representation of missing data. Another feature of lack of clarity of description meant that acronym type variable names in the data set were often confusing. In an appendix to a prior report or as side notes in the database descriptions were incomplete. Some descriptions which are given are also vague and lack the detail to pin them down precisely. Some other variable descriptions are scattered through the report but are not drawn together in an easily interpretable form. The listing of variables presented as appendix documentation also contains, rather confusingly, variable definitions which are superfluous to the database. Much of the detailed investigative work required might be avoided if proper attention was given to ways in which the database details are presented.

The statistician is also conscious that data exclusion or inclusion decisions, either of cases or variables, may have been driven by a prior analyst's main research and modelling frameworks. These may not be the same as those of investigators to whom the data has been passed. A cavalier treatment of any such acquired data is a process which any qualified statistician will seek to avoid. Treatment of excluded data is one that requires careful consideration if biased analyses and generalisations are to be avoided. In the case of the present data it was by no means certain how the final set of cases had been arrived at. The outline of the proposed investigation by DfES suggested there are over 600,000 relevant children on the PLASC database of which the current set represents a subset of below 77% of these. There are certain hints scattered throughout 'Report 2: Levačić' such as exclusion of middle schools. However, there is no definitive statement of how the final data set was arrived at. Many of the analyses presented in that report were based on fewer cases than the full subset presumably due to listwise deletion of cases with missing data on one or more included variables. It is likely that there are justifications for this process but it is by no means certain that such justifications will carry over for later more complex analyses. In this situation selectivity in choice of cases for analysis may imply biased judgements. Later we will suggest how multiple imputation or modelling missing data mechanisms may be undertaken along with analytical models to improve the reliability of results.

2.2 The Area Data

The main focus of this scoping was to extend the framework to include area of residence in a cross-classified random effects framework. Here we comment on some initial problems that arose with various area identifiers that needed resolution before this could be envisaged.

The present dataset has several area of residence identifiers in the data set which will be necessary for various forms of cross-classified multilevel model; postcode, census output area , ward and local authority district. These are hierarchically arranged but the postcode reference is on a different coding basis. Codes for the higher levels reflect the hierarchical structure, e.g. output area 00FBPE0029 is within ward 00FBPE which is within local authority district 00FB. In Appendix 1, A1.3, we detail some of the issues that arose out of our investigations of these identifiers and their potentialities in modelling. It will be evident from this detail that many similar problems as above will arise unless more detailed documentation is to be made available.

Important modelling issues also arose out of this investigation. The data set had on average only 1.45 Key Stage 3 children per postcode. Apart from substantive considerations discussed later in

the report, any treating of postcode as a higher level unit above children will yield inevitable statistical and computational difficulties. Census output areas have an average of 3.26 KS3 children per output area. For some purposes it may be possible to separate out higher level effects for such small areas but the sparsity of the clustering and the necessity for large numbers of output area random effects may render the exercise somewhat infeasible for the full data set. Ward random effects in a multilevel model and/or a crossed model are possibly computationally easier to handle due to their larger size. Wards had an average of 58.4 KS3 children. The highest area of residence identifier in the dataset was the local authority district (LAD). Some modelling may trial such higher level authority residential area if it seems there is a possibility of an effect at this level. It should be noted that although many LEAs are also the same administrative areas as LADs in general they are a layer above. For many local authority districts the responsibility for education is the responsibility of a LEA which is at a higher level than the district. In many cases such as metropolitan districts like Birmingham the two are of course the same. For reasons which will become clearer when we consider the data structure and entertain possible modelling frameworks, we would like to identify the Local Education Authority area in which the student resides in addition to the LEA of his school. The two need not be the same. Indeed it became apparent after mapping that around 9% of KS3 children lived in a different LEA area than the one in which they went to school. This will have implications for the cross-classified model frameworks to be discussed. The residence LEA identifier is not available in the data set. Also the LEA coding for schools in data is based on the DfES system and is not an integral part of the code for local districts, though the latter are hierarchically arranged beneath the LEAs. A clear prior mapping of the LADs to LEA for residence purposes would have facilitated the avoidance of some tedious data preparation. [1] .

2.3 Other Data Editing And Cleaning Issues

We have discussed above the many issues that arose because of the unclear and unsystematic treatment of missing values. More complete detail is covered in Appendix 1.2. There are some clear indications there about the sort of initial preparation that might be required for further use of this or similar databases. It will be seen that in some cases the use of the digit '0' for missing values created some modelling problems particular when dummy indicators constructed from categorised variables are entered as covariates. Without detailed further exploration it is impossible to distinguish between zero meaning 'not this category' and a missing response for the categorised variable. The point is not trivial either. For example in A1.2 (iii) we give the example of this potentiality occurring for certain ethnic group indicators. It is possible that 'missingness' may be more prevalent in certain ethnic groups and often this is substantial. Where this is prevalent, the use of derived indicators as explanatory variables in models may seriously distort analyses.

However, it is also apparent that for many variables in the dataset consistency checks which enable resolution of such issues are available in the data set itself (see A1.2 (iv)) and any thorough cleaning of the data could handle them. The question arises as to whether a secondary analyst should be expected to undertake such consistency checks, even if he has the necessary expertise to do so. This may be unlikely given relative lack of detailed knowledge of the construction of the databases.

[1] Fortunately we were able to enlist the support of Professor John Shepherd (Birkbeck College) who was able to match the district codes in our full set of cases to a corresponding higher level LEA of residence. Another slight complication is that although this dataset relates to KS3 at English schools 67 cases travelled across the border from Welsh LEAs. These are such outliers, which though small in number may impact on the analysis unless their treatment is carefully considered.

Another important matter giving rise to some concern is missing data for many school variables (see A1.2 (v)). From preliminary information it seems that the schools involved have special characteristics which may be important to analysis and these characteristics also relate to the 'missingness'. It is likely that the schools with missing information are unique and special and for reasons directly connected with the objects of the analysis and so cannot be treated as missing at random. If this is so the possibility of distortion in analyses is evident although its likely impact is difficult to assess. However, we feel that some method of imputation might ultimately be beneficial and we return to this issue later.

Other anomalies in the data are also detailed in Appendix A1.2 section (vi). Prior cleaning, editing and documentation might have resolved many of these. The necessity for thorough prior investigation as part of this process becomes even more apparent when some of the details are considered. Of particular importance, for example, for secondary analysis is the uncovering of seemingly absurd outlier values created during construction of derived variable such as ratios. As an example 24 schools in this data set are given as having ratios of fte pupils to fte unqualified teachers 02/03 of over 10,000. There may be explanations of why this should be so but such outliers give concern. The presence of such extreme outliers leads to highly skewed distributions. These may lead to distortions particularly if used in linear models unless the data features are recognised and handled. This may be missed unless detailed in codebooks and documentation.

The use of derived ratios also creates other problems which may not be immediately recognised. Quite often the denominator is zero and when subjected to software calculations will yield system missing values which may not be appropriate in analysis. For example A1.2 (vi) shows that 12% of schools have zero unqualified staff and this yields missing values for ratios of qualified to unqualified. It would not be appropriate to treat these as missing in secondary analysis yet the reason for it would not be immediately apparent without careful prior documentation.

2.4 Further Remarks On General Issues

In Appendix 1 we have gone into some detail on the outcomes of the data investigation phase of our work and problems with handling a third or even fourth hand datasets. In the above we have also summarised issues arising from this which may have broader applicability. We do these things for a number of reasons:

- With hindsight we might have envisaged that this phase might have been more difficult than we initially thought. It certainly proved time consuming and hampered the progress with other phases. We recommend that any future work with the data set provides for this and perhaps refers back to the original primary sources.
- We believe that any future more complex modelling work will require a fuller investigation and more extensive data editing and cleaning. We have only completed certain edits sufficient for our experimental subsets and variables used in analyses of them. However, the information provided will be a useful indication of the problems that might be more fully addressed by subsequent investigations.

2.5 Format Of Databases and Software Compatibility

There are many problems of having a dataset provided as a system worksheet for one particular general purpose programme such as STATA. For many other general purpose programmes such as SPSS transfer is often relatively unproblematic using the available STATtransfer programme.

However, data input is perhaps one of the less adaptable features of specialist statistical programmes for such matters as complex modelling. Such software is often very sensitive to the format of data input. We had a little difficulty, for instance preparing data for the Multilevel modelling package MLwiN which we use intensively and also WINBUGS. Many of the difficulties arose because of the issues discussed above, particularly in the recognition of missing values.

Other problems arose because these programmes are less adaptable to the use of string variables and alphanumeric identifier codes. For example the 18 digit alphanumeric code for student identifier is cumbersome for transfer to other specialist programmes. A STATA encoding command to transfer the code to sequential integers proved infeasible due to large number of unique values so EXCEL was used as an intermediary to generate this unique code *af_case_id* in an edited STATA file.

Another anomaly in the file which may also be more general for other databases is that values for string variables have occasionally been entered misspelled or with excess leading spaces. This means that the same value is often taken to be several distinct values. With careful attention, tabulation of the data variable by variable can uncover such difficulties and in previous analyses this has probably been the case. This problem also makes matching and transfer of data to other software more difficult.

SPSS was also often used for some data manipulation facilities in addressing some of these issues since it has more flexible features for some purposes. For reasons outlined above transfer of files between STATA and SPSS is relatively easy. For other purposes we had to convert the STATA data set to formatted text files which is not a straightforward task. We believe the multilevel modelling work in 'Report 2: Levačić' using MLwiN took as its input a specially prepared formatted text file. Here we had to more or less start from scratch.

The programme GENSTAT which we use in one of our scoping experiments can input files in a variety of formats including EXCEL. However, in this case the writing of special code was required to extract data for variables from the full data file for its use by GENSTAT. Intermediate data preparation programmes such as EXCEL could not handle the big file.

Data transfer and conversion problems such as we have outlined are relatively well known to experienced statistical modellers. Common FAQs, for instance, on the multilevel mailing list concern themselves with many similar problems. We hope that details of our experience in this scoping exercise will lead to a non-trivialisation of this aspect in any future modelling work with this and other data sets. It could also provide some pointers as to why careful prior preparation of databases and how they are communicated to users could be much improved processes.

3 Choice of Data Subsets And Investigation Of Crossed Structures

It was recognised that scoping the models, methods and software on the full dataset might initially prove difficult. Firstly the sheer size of the dataset might impose storage and computational restrictions which might detract from this central purpose. This was handled quite satisfactorily in 'Report 2: Levačić' for strict hierarchical models and the methods of that research but might prove too cumbersome for more complex modelling. Secondly, with crossed random effects in a model, there may be additional computational restrictions for some of specialist software and methods that we investigate. For instance, in such circumstances one approach requires extra dummy variable predictors to be created for each unit for one of the classifications in a two way crossing. The school by ward cross classification serves as an example. One particular method requires setting up dummies for 7963 wards resulting in a huge worksheet. Thirdly unless we can find discrete non overlapping blockings of groups of schools and wards matrix inversions required for many methods might prove practically infeasible.

For these reasons, and for purposes of experimentation, we decided at the outset to select purposively a certain number of school LEAs comprising about 20% of cases. Since our concern was with trialling the methods and not inference from this sample to the larger population a purposive selection was not inappropriate. We used schools from 25 LEA s consisting of 80,032 KS3 pupils which formed about 17% of the full data set. The cases are distributed over LEAs as in Table 3.1 below. This subset of the data was labelled the 'Combined Selection'. It consisted of the LEAs beginning with a B in the alphabet plus the first two C's and a few in the North West of personal but not noteworthy interest to the investigator. In fact this selection covers a wide range of different types of authority. It ranged from the cities of Birmingham and Manchester to large mixed urban/rural authorities such as Lancashire and some quite small ones such as Bracknell Forest. It will also be seen later that predictor effects estimated on this subset are not too different from those in 'Report 2: Levačić' on the full set of data.

The initial part of our investigation was concerned with examining the structure of the cross-classifications on the selected sets of data. As part of this we rapidly discovered a feature we had not initially contemplated and further implications of which receive comment in the next section. A small but substantial proportion of pupils were what we call 'out of area', i.e. they were resident in different areas than covered by their school LEA. The Combined Selection was fairly widespread geographically and such children are likely to come from adjacent LEAs not in the subset. Thus the number of area units such as wards and output area are inevitably very large. Further these children will be thinly spread over such 'out of area' units. These are two of the factors which might affect the feasibility of the methods and partly why we considered reducing the full data set in the first place. When we analyse the trials we will see that these features may indeed have led to certain difficulties with the dataset. However, we persevered with it for some trials since it gave us an idea of the impact of the large number of area random effects which might arise in extending analyses to all the data.

Table 3.1: Combined Selection data-set

Local Education Authority	**Number of Pupils**
Barking and Dagenham	1,642
Barnet	2,529
Barnsley	1,969
Bath and North Somerset	1,812
Bedfordshire	3,871
Bexley	2,760
Birmingham	9,763
Blackburn with Darwen	1,482
Blackpool	1,333
Bolton	2,957
Bournemouth	1,489
Bracknell Forest	896
Bradford	4,657
Brent	2,026
Brighton and Hove	1,912
Bristol City of	2,397
Bromley	2,987
Buckinghamshire	4,735
Bury	2,026
Calderdale	2,103
Cambridgeshire	4,535
Lancashire	12,088
Manchester	3,321
Oldham	2,545
Rochdale	2,197
Total	**80,032**

For these reasons, therefore, we also decided to experiment with a smaller more compact set of LEAs which formed contiguous areas within a specified region of the country. We call this the West Midlands dataset. The 'out of area' phenomenon still exists, but travelling across boundaries for children is more likely to be to areas which are within the LEAs of the data set. Certain structural features of the Combined Selection dataset are thus less likely to be as prevalent here. As a result anticipated difficulties in fitting models may be possibly less severe. The West Midlands data set comprised 32,579 KS3 pupils from schools in eight LEAS and is distributed as in Table 3.2 below.

Table 3.2: West Midlands data-set

Local Education Authority	**Number of Pupils**
Birmingham	9,763
Coventry	3,031
Dudley	3,481
Sandwell	2,900
Solihull	2,800
Walsall	3,078
Warwickshire	5,167
Wolverhampton	2,359
Total	**32,579**

Next we examined the structural features of our data sets from a number of perspectives. From previous experience with complex cross-classified models it was thought likely that these would play a major part in the success of the exercise. This judgement was based on intertwined statistical and computational criteria. The features are

- The number of area units at each of the levels that it might be considered using
- The extent of the imbalance of units across the cells in cross –classifications of school and areas
- The nature of the 'out of area' problem
- Sparseness of representation in certain cells of the cross-classifications.

This initial examination proved an essential guide in exploring the desirability and feasibility of the various methods and software since not only do they address the criteria but also influence ways of setting up models for analysis. We have commented on the computational restrictions that might occur due to large numbers of units at any level. The number of units at any level also determines the extent of lower level representation within them, such as KS3 pupils within output areas. This in turn affects the precision of any specific random area effects we may ultimately wish to estimate. In the next section we will suggest creative ways of accommodating the 'out of area' issue in models.

Certain features of a structure are known to affect both statistical quality of estimation and also computational feasibility. The extent of imbalance in terms of concentration of cases either marginally in certain units of each of the crossed factors or in specific groups of cells of the crossing is one important feature. A related one is sparseness with few or no cases, both in the margins for specific units in the hierarchies and amongst cells in a cross-classification. The extent of the impact of these is still a relatively under-researched phenomenon. The problems form part of an ESRC project under the direction of Dr William Browne in the School of Mathematical Sciences at the University of Nottingham which is currently getting under way. It is hoped that much more will be known in future. However, we commented generally on some aspects of the impact of these features in 'Report 3: CC Review'. Broadly we expect the structures to lack the balance of designed experiments but we would not like this to go too far, otherwise there may be confounding of separate effects in a cross classification of random effects . To avoid this there are certain general requirements. For instance there should be some reasonable spread of Level 1 units across units of one classification within each unit of the other. If at an extreme, for example there were a very large number of schools which drew their pupils from just one ward and those wards sent their pupils to these single schools then confounding

problems occur. It becomes difficult to identify separately what are school and area effects. Imbalance also affects sparseness of units and sometimes affects the precision with which particular effects can be estimated. This is likely to pose an interesting challenge in future research if for instance it was desired to specifically identify areas of low achievement. If some areas had very small numbers of children there would be much uncertainty about this identification.

However, at this stage it might be useful to look at one example of our structural investigation to further highlight these concepts and to indicate what we might learn about our selected data sets. Many of the investigations we carried out to satisfy ourselves about the structural features are in the form of large cross classification tables with many rows and columns (often thousands). As such they are impossible practically too fully illustrate here. However, they are stored in an utilisable form. Thus the example is by necessity restricted to an examination of one medium sized authority, Cambridgeshire. A few other examples and aspects of them which are feasible to illustrate in print are given in Appendix 1. We comment on these at the end of this section.

The Cambridgeshire data had 4535 pupils attending 28 schools, averaging at around 160 pupils per school ranging from a minimum of 59 pupils to a maximum of 275. Compared to the full data set a relatively small number, 123 (2.7%), had an area of residence outside the LEA as in the Table 3.3. The investigation was undertaken before LEA of residence was merged with the dataset. Hence local authority districts were used for this exercise. It may be noted that Cambridgeshire LEA itself comprises five local authority districts; Cambridge City, East Cambridgeshire, South Cambridgeshire, Fenland and Huntingdon.

Not surprisingly most of these 'out of area' pupils are from areas adjacent to Cambridgeshire. However, their wards and output area will not be represented in the Combined Selection dataset which is geographically diverse. However, even if these areas were represented in the data, their cross-classes with Cambridgeshire schools would result in sparse cells and increase the number of discrete non-overlapping blocks of schools and residence areas in the data. If these features also extended to all the Combined Selection data set, as seems likely, the total number of area units and hence random effects might not be much less than in the full dataset. These factors may affect computational feasibilities for many methods. Even within the West Midlands dataset, where such problems might be expected to be fewer, the total numbers of random effects for the chosen area levels will still be inflated considerably. Computation time in many procedures and also precision of estimates may be affected. Some of the areas in Table 3.3 seem on the surface to be absurd, e.g. Winchester or Coventry and may be due to inaccuracies in the PLASC database, or to some other inscrutable reason. It is however, just for such pupils that the area effects are likely to be important and why the out of area pupils are often an important minority of pupils. Intuition suggests that pupils who travel some way to school may have different characteristics. They might be considered to have important distinct influences and cannot often justifiably be dropped to make analytical computation of model estimates easier.

Table 3.3: Local authority district of 'out of area' pupils in Cambridgeshire schools.

Local Authority District	**No. of pupils**
St. Edmundsbury	32
King's Lynn and West Norfolk	25
Forest Heath	14
North Hertfordshire	14
Peterborough	10
Bedford	4
Uttlesford	4
Mid Bedfordshire	3
East Northamptonshire	2
Waltham Forest	2
Ashford	1
Braintree	1
Cannock Chase	1
Coventry	1
East Hertfordshire	1
Hillingdon	1
Luton	1
Newham	1
Northampton	1
South Kesteven	1
Tendring	1
Welwyn Hatfield	1
Winchester	1
Total	**123**

Restricting attention now to the 4412 pupils who live in the Cambridgeshire, it is instructive to examine the nature of the crossing of schools with both output areas and wards. There are 1471 output areas and 123 wards. With a mean of 3.0 pupils, output areas were inevitably sparsely represented in the data. 10% of areas had just one pupil and only 7% had greater than 7 pupils. We anticipated that such features might make for difficulties if output area was used as a random effect in a model. Computationally the sheer number of areas is burdensome but also estimation precision might be low due these features. To some extent our trials confirmed these beliefs. Table 3.4 below gives some summary statistics on the cross tabulations of schools and wards, and schools and output area. Detailed tables were used to examine the essential features of the structure but these summaries contain some pertinent information

In this table it is seen that schools draw their pupils from a relatively large number of both output areas and wards so this lessens the dangers of confounding. Also with a large number of output areas, and hence few children in each it was found inevitably that few schools were represented amongst the pupils in each output area. Amongst the output areas 73% had children in just one school. To some extent this represents what may be termed a near hierarchical structure .This might help with computation since it means that non overlapping blocks of schools and output areas can be more readily found[2]. By contrast with Cambridgeshire

[2] In 'Report 2: CC Review we consider reasons for this based on the efficient algorithms developed by Rasbash and Goldstein (1994).

in some other areas we have examined, such as Oldham in Appendix 1, there is less of an issue with concentration of school representation amongst output areas. This may reflect something of a difference between large county authorities and urban metropolitan ones. Of course, as elsewhere, large numbers of random effects and hence the lower precision with which they can be estimated are likely to raise further statistical and computational issues.

Table 3: 4 Summary statistics on area structure for Cambridgeshire schools and for pupils who also lived in the five local authority districts making up the Cambridgeshire LEA

School	No of pupils	No of output areas	Average no of pupils per output area represented	No of Wards	Average number of pupils per ward represented
1	171	74	2.3	13	13.2
2	152	74	2.0	17	8.9
3	201	64	3.1	9	22.3
4	92	71	1.3	17	5.4
5	162	89	1.8	11	14.7
6	59	41	1.4	7	8.4
7	113	53	2.1	12	9.4
8	94	49	1.9	12	7.8
9	158	59	2.7	16	9.9
10	138	43	3.2	8	17.3
11	59	136	2.3	13	10.5
12	30	49	1.6	4	12.3
13	102	205	2.1	12	17.1
14	88	275	3.1	11	25.0
15	64	144	2.3	8	18.0
16	67	150	2.2	10	15.0
17	93	240	2.6	17	14.1
18	92	241	2.6	15	16.1
19	97	108	1.1	47	2.3
20	35	95	2.7	7	13.6
21	91	202	2.2	10	20.2
22	59	162	2.7	10	16.2
23	81	201	2.5	23	8.7
24	67	135	2.1	14	9.6
25	87	210	2.4	14	15.1
26	68	179	2.6	10	17.9
27	91	224	2.5	15	14.9
28	34	116	3.4	6	19.3
Total	**4412**				

Wards as random effects in a cross classification would appear to present few problems. There are a much smaller number of them. Also each school in Cambridgeshire is represented by a fair number of wards as Table 3.4 indicates. However, on further examination, 20 of the 123 wards had all their children concentrated in particular schools. This should help computationally. Further problems of confounding are likely to be minimal due to those schools drawing their children from quite a number of other wards. The maximum number of schools represented per ward was 7

Another feature of the imbalance of structure is the distribution of the pupils
(Level 1) over possible cells of the crossed classifications. Table 3.5 shows that 94% of the school/output area cells are empty. Such features of sparsity do affect the statistical and computational qualities of model fits. However, given the need to distribute 4412 pupils over 41188 cells, we judge from experience with similar structures that this type of structure as manageable. Too many empty cells would verge towards confounding. Too few might add to computational demands in addition to that imposed already by the large number of output area effects. As we have mentioned very little is formally known, as yet, about the real impact of these sorts of phenomena on various methods. Table 3.6 gives similar information for the cells of the school by ward cross classification.

Table 3.5, Sparsity: Frequencies of pupils over the 41188 school by output area cells for Cambridgeshire schools and for pupils who also lived in Cambridgeshire

Number of pupils in cell	0	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	39265	883	395	257	189	90	51	32	10	6	4	2	3

Table 3.6, Sparsity: Frequencies of pupils over the 3444 school by ward area cells for Cambridgeshire schools and for pupils who also lived in Cambridgeshire

Number of pupils in cell	0	1	2	3	4	5	6	7	8	9	10	11	>11
Frequency	3076	124	35	18	16	15	11	6	5	0	5	8	125

As indicated we carried out quite a number of such structural investigations. These are not detailed here. Broad patterns emerged that satisfied us of the applicability of similar essential impacts of the features in larger data sets. However, there were some differences in detail. For instance, the information for Oldham in Appendix 1 shows a pattern implying fewer (relatively) empty cells. We hypothesis some general differences between types of authority in terms of concentration of population and number of local government divisions. This Appendix also contains some investigation of the 'out of area' issue for Birmingham LEA.

4 Modelling Frameworks

4.1 The Model Structures and Random Effects

A full potential framework for characterising the possible random effects that might be incorporated into a cross-classified multilevel model is seen by examining complex structure in the classification diagram of Figure 1.The use of classification diagrams for characterising random effects in complex models are more fully described in 'Report 3: CC Review'. It will be seen that there are two hierarchies which we shall often refer to as education (or school) and area (of residence). These hierarchies cut across one another and induce cross-classified effects. The education hierarchy leading to an examination of school and LEA effects has been the subject of 'Report 1: Vignoles'. Our scoping report is concerned with extending the modelling frameworks to additionally encompass features of the area hierarchy. It will be noted that the structure as illustrated in Figure 1 adds two classifications whose identifiers are not available in the STATA data set; super output area and LEA of residence.

To set the scene it is worth considering for the full data set the number of units at each level in the hierarchies, where these are known. These directly determine the scale of the exercise in terms of numbers of different random effects for particular models. They form constraining factors affecting the feasibility of different estimation procedures and software implementations. The STATA dataset available contains 464,783 KS3 pupils (level 1) nested within 2943 schools within 149 LEAs for the education hierarchy. The area hierarchy has postcode areas nested within 142, 597 census output areas within an unknown (but discoverable) number of super output areas (SOA). For simplicity of illustration we have indicated only one level of SOA but three hierarchically arranged layers are ultimately envisaged. [3] In turn SOAs are nested within 7964 wards from 360 local authority districts. The top level of the area hierarchy may be additionally defined in terms of the same set of units as school LEAs and we refer to this as the LEA of residence. A large number of these are the same as the local authority districts (e.g. Metropolitan Districts such as Birmingham). Thus many top level units will have single members at the lower level of district, but multilevel modelling methodology can handle this situation.[4]
The use of this 'fullish' structure, and which elements we consider in specifying random effects for particular models, is something of an open question. Without full scale research and iterative

---

[3] The Office for National Statistics website (www.statistics.gov.uk/geography/soa.asp) gives the following projected information on Super Output Areas: 'There will ultimately be three layers of Super Output Areas (SOAs), each nesting inside the layer above, with areas intermediate in size between 2001 Census Output Areas (OAs) and local authorities. This will offer a choice of scale for the collection and publication of data, and allow for the release of local data that could be disclosive if published for OAs. At present just the first two of these layers have been created. The 34,378 Lower Layer SOAs in England and Wales were generated automatically and released to the public in February 2004. The 7,193 Middle Layer SOAs were defined in a two-stage process: an initial set was generated automatically but the boundaries were then modified in consultation with local authorities and other local bodies. The final boundaries were released to the public in August 2004. The Upper Layer SOAs are expected to be created in 2006.'

[4] Since many LEAs contain several districts conceptually a random effect for district below that of LEA can be envisaged. Estimation of its variance from the conceptual distribution of random effects is made possible by these multiple occurrences. The further estimation of random district effects where these are required is also feasible since such effects are 'shrunken estimates' borrowing strength from the full distribution of effects. Full theoretical and computational aspects of this are considered by any text on multilevel modelling methodology such as Goldstein (2003).

model development we can only partly answer it. We can and will, however, give some general views based on the information above and what we judge may be worthwhile. Firstly, however, we might comment on some other features of this structure which have a bearing on these views. Although we will return to them let us ignore for the moment the top LEA levels of the twin hierarchies. The central direction of proposed models will then focus initially on the crossed effects of schools with areas of residence lower in the hierarchy. This crossing will occur at whatever level of the residence hierarchy it is chosen to focus and whichever area random effects a particular model specifies. In the classification diagram of Figure 1 we have placed schools alongside local districts to indicate that the crossing occurs at that level. As discussed in 'Report 3: CC Review' the implication is that schools are also crossed with area units lower down the hierarchy. This tautology is important in thinking about crossed hierarchies. Model formulations in terms of random effect will by necessity reflect this. Certain models for various reasons may say only include ward units as a basis for the study of area effects, or only output areas, or perhaps both. School effects will then be crossed with either or both of these in the area hierarchy.

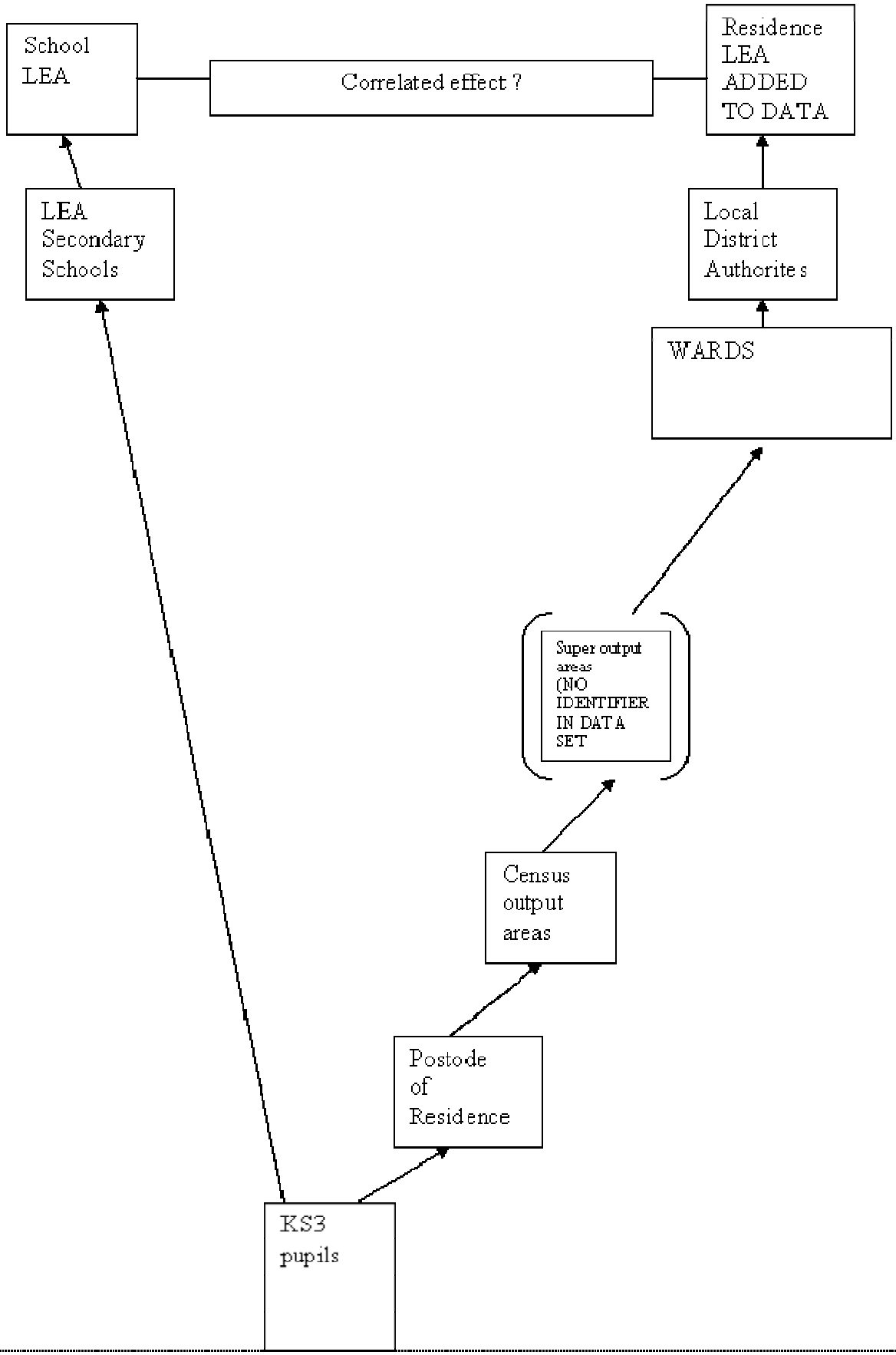

Figure 1: Classification diagram for full structure

Thus a central question to which any model development and investigation may be directed is the extent to which we reflect in analysis the full complexity of the various area effects (beneath residence LEA). In principle and if there is evidence of variation at each level then perhaps all that are feasible should be included. However, it must be recognised that for certain procedures there may be computational constraints to this. The more the complexity the less easy it is to implement a random effects model. Further iterative model development may bring greater clarity this but in this scoping study we have focused on two of the area units whose identifiers are available to us, ward and output area. Local authority districts are perhaps too close to the top LEA level to be useful if the latter are also included. The basic area identifiers in the PLASC data are postcodes of pupil residence and from this the higher area units are established. However, postcodes have an average of little over one pupil within them. Separation of random effects and variances between postcodes from those at the pupil level will inevitably be difficult. This confounding of pupil and postcode effects seems to make inclusion of postcode random effects not a very practical proposition. Even with greater representation of children the sheer number of postcode units would add to computational difficulties. However, it must be stressed that this does not mean that observable variables at postcode level would not be a useful adjunct to further research. They could easily be introduced as fixed effect predictors. No such information is available in the present data set so we must leave this issue aside. Perhaps however, if such information was merged it might be closer to the family and social circumstances that are proxied by the census output area variables utilised in 'Report 2: Levačić'. To some extent both output area and wards may be useful to include as random effects. In the models in 'Report 2: Levačić' output area fixed effect variables are included and the incorporation of such areas as a level in a multilevel model would mean that the random effects for output area reflect additional residual variation due to unobserved sources at that level. Such variation may provoke substantive interest in further ways of finding proxies for social circumstance effects. However, output areas are still very close to highly localised circumstances with on average only a little over 3 KS3 pupils within them. Our experimental analyses may show that the sheer number of output areas may impose above anything else computational constraints on model fits. The ward is of course the aggregate on which many social exclusion initiatives are based. Until recently it was also the area basis for much empirical research due to ready availability of data such as deprivation indices at ward level only. Wards are highly aggregated collections of KS3 children. However, if ward variation is of interest inclusion of wards in models are certainly more feasible technically. Some of our analyses later will utilise both wards and output areas but if it is decided that one or other is more appropriate then it is certainly easier to fit models including only one set of random effects rather than both.

Pertinent to all this is the fact that the area unit boundaries are administrative and rather artificial in reflecting the social and economic circumstances in which our KS3 children find themselves. Diffuse spatial modelling such as occurs in some health research and discussed in 'Report 3: CC Review' may be useful but is beyond our present experimental concern. The administrative nature of boundaries is a well known aspect and limitation of much policy directed research on area effects. The substantive question is really a matching of area unit boundaries on which it is proposed to implement area initiatives and the empirical research which is designed to evaluate them. Thus in the present context we could in principle use whichever area units seemed appropriate and between which there was evident variation. It is then not statistical matters but rather research and policy issues which should suggest where analyses should lead. To some extent these questions are outside our present remit. We experiment only with the area units given to us.

Evident in our classification diagram but not in the data are possibilities for the use of census Super Output Areas. We believe these may be worthy of further investigation. Substantively one

or other layer of these may form area units to which more realistic policy initiatives may be directed if real source of variation are found amongst them. This may have more impact than dilution of activity might yield if it was at ward level. It might have greater practicality than highly localised output area activity, if indeed the latter were thought feasible at all. These issues are deep ones and are geared to any research objectives set but they can inform the ends to which any analysis might be directed. A model which examined areal variation through super output areas might replace a possible joint operation of ward and output area. From an analytical and computational perspective this might certainly be easier to handle. Although not in the current dataset super output area identifiers could be merged using a geo-demographic system. Variables defined at these levels may also be derived from census information and used as potential explanatory fixed effects in multilevel education production functions. For the moment we can only leave these matters as a recommendation that they could be considered.

4.2 The Role of LEA in Both Education and Area Hierarchies.

We return now to the possibility of including higher level LEA random effects, and then we may begin to see that some new challenges arise. Our original perception, before detailed examination of the data revealed otherwise, was that a crossing of schools and areas at say ward level (if district was ignored) might be sufficient to reflect the complexity of the crossed hierarchies. This optimism was based on the initial misguided belief that the combinations of school and area of residence of all KS3 pupils would be nested entirely beneath the LEA of the school. If this was the case one set of random effects for LEA would be an adequate and relatively simple specification of this higher level random effects were it desired to include them. For the vast majority of pupils this situation arises. However, it was found that an important minority of children cross LEA boundaries to go to school. These children may also be of a qualitatively different kind from the rest so it would be important to reflect this. However, with the existing area identifiers in the data, which omits LEA of residence, the only way to handle this would have been to ignore 'residence effects' associated with higher level areas defined by the geography of LEAs, whilst recognising that areas of residence are not nested within School LEA. In terms of Figure 1 we would omit the Residence LEA box in the top right hand corner of the classification diagram and associated linked crossing to School LEA. We would then cross whatever top level area of residence was chosen, possibly ward or LAD with School LEA. In Figure 1 this top level area of residence would then move up to be horizontal to School LEA in the recast classification diagram. It is to be noted that lower level education effects such as schools and area effects such as output areas are then implicitly nested within cells of this crossing. Lower level crossings are then also implicitly taken care of within this nesting. A detailed technical explanation of why this should be so and illustrative examples are given in 'Report 3: CC Review'. After some thought and discussion it was felt that ignoring LEA of residence might not be entirely appropriate since higher level residence effects might be potentially evident. A solution was to identify the LEA areas of residence in ways we have indicated in a previous section and then merge it to the dataset. A model with higher level effects could then cross classify School LEA with Residence LEA. Within cells of this any lower level crossings are now nested. A feature of this setup is that the set of highest level units forming the residence area and the set of highest level units for the education area classifications are common. In this case there is a strong possibility that the two LEA effects may correlate; in that the LEA effect operating because a child's school was located in it may be related to the geodemographic LEA area residence effect. In our experiments we entertain this possibility but recognise the limitations of some existing software to handle it. We also pose the question of whether these effects are sufficiently important to justify the extra complexity and the greater difficulty in handling them. Of course, as in previous discussion these statistical and computational questions must be balanced against substantive research requirements.

4.3 Fixed Effect Covariates in the Model

So far we have commented only on the range of possible random effects that we might model to reflect the complex structure evidenced in Figure 1. Any analysis will also investigate and incorporate relevant and observable fixed effect covariates at any of the levels. The final choice would be a major part of any detailed analytical approach to model development in future research. 'Report 2: Levačić' conducted some such model development within the context of the hierarchical school model and arrived at sets of predictors as indicated in their reported results. There is no guarantee that this set of predictors would be fully appropriate in the new model framework. Most of them possibly would be but this feature of the models may be capable of more refinement. Other questions of interest also arise and have been raised during the discussions around this project. We might, for instance, examine such features as regional variation or differential progress for different subgroups of children. The latter might involve trying and testing interaction of the KS2 prior achievement with other child characteristics. Similar further refinement to other aspects of the specification may also reveal answers to additional interesting questions. In principle such refinements could be carried out in the current context. There is also the possibility of differential school effects on progress or impact of area being different for types of children. This would require the added complexity of random regression coefficients. Suffice it to say that for the purposes of limited experimentation we consider only the predictors and their characterisation as given in 'Report 2: Levačić'. Within this scoping we also cannot at present extend to a consideration of random regression coefficients. The latter though adding to computational demands are statistically feasible within the model frameworks we have outlined.

5 Methodology, Software and Trial Analyses

5.1 Introductory Comments

In this section we present some detail on the experimental analyses we have undertaken to guide us on the feasibility of aspects of the modelling work. In all of this we are conscious of three interrelated themes. Firstly from substantive background considerations we might be led to a consideration of a particular type of model specification to be fitted to a particular set of data. Statistical considerations might then lead us to suggest that a particular type of estimation might be most appropriate. However, we might find that software and computing constraints might make this preference infeasible. Therefore on a second theme, we might approach the issue the other way round. We ask what estimation procedures are available within obtainable and feasible computing resources and then assess their statistical quality for fitting the model at hand. Thirdly either of these first two approaches might lead us to question whether we might drop certain aspects of the model specification. To a large extent our scoping exercise and trial analyses are a matter of balancing these themes.

For reasons explained previously this experimentation was with two selected subsets that we have labelled West Midlands and Combined Selection. In addition one limited trial has been undertaken on the full dataset to explore the possibilities of the GENSTAT software.

In the previous section we outlined the full structure of potential models. However elaborate the models to be ultimately developed become, there are two distinct complexities that we must address in the present context. First is the presence of several levels in the hierarchies as well as crossing of different classifications at each level. In a previous section we have examined some aspects of the nature of the cross-classifications in terms of numbers of units at various levels, balance and sparsity. All of these will affect the statistical quality and feasibility of any projected model fitting. The latter are also conditioned by the variety of estimation procedures available. They will also affect the computational feasibility of the procedures which in turn is conditioned by availability of software and its practical implementation with computing resources available. The statistical and computational aspects are to some degree intertwined. The second complexity is the endogeneity of the resources variables which was the subject of detailed review in ‘Report 1: Vignoles’, ‘Report 2: Levačić’ and Mayston (2002). To handle this there are again a variety of statistical estimation procedures of varying quality and feasibility available. These are also conditioned by the availability of appropriate software and the feasibility of its implementation.

To some extent in our trial analyses we have separated out the two model complexities and concentrated on handling them one at a time. Also we have not always brought together all the features of the complex structure of random effects at the same time. The ultimate aim might be seen as exploring the feasibility of bringing everything together to produce model estimates with high statistical estimation quality within feasible computational frameworks. Where appropriate we will comment on where our concentration on particular aspects might ultimately bring this aim to fruition. To start with an over ambitious full set of complexities might mean running into many initial difficulties, solutions to which might be difficult to discern. The point of a scoping exercise as a learning process might then be lost. This is our general rationale for the approach to scoping we have adopted. However, we will shortly detail aspects of particular trials and specific reasons for considering them but mainly they are connected to matching and balancing particular aspects of desirability and feasibility.

Before doing this it is worthwhile commenting on availability of modelling software and some of the reasons for narrowing down our choices for experimentation to the main three we do use; MLwiN, WINBUGS and GENSTAT. For initial purposes in commenting on instrumental variable estimation we also entertain STATA used in 'Report 2: Levačić' and for which our provided data is instantly useable. STATA would possibly also be our preferred option for general purpose advanced statistics apart from multilevel modelling. 'Report 3: CC Review' commented on the wide range of software that can now handle cross-classified models and the various estimation procedures used. The reader of this report is cross-referenced to that parallel report. The possibilities are now increasing at an accelerated rate. The continuously updated review of available software on the Centre for Multilevel Modelling web page www.multilevel.ioe.ac.uk includes a crossed effects model as one of its experimental media. From this review and personal experience with similar data structures we conclude that it might be practically difficult and infeasible to contemplate using much of this reviewed software for the sort of model specifications we wish to ultimately entertain. However, although we have not been able to investigate here there certainly seems some promise in anticipated future developments of the GLAMM procedures in STATA9. Often the complexities of structure and number of levels in required models or the desired output constrains the attractiveness of most of the software. Almost all the wider range of software will also not easily allow us to build into our multilevel modelling the more refined features we might wish to explore. For instance, we might wish to consider the possibility of different area or school resource effects for certain subgroups of children. This will require us to specify regression slopes as randomly varying effects over such groups and these are difficult, if not impossible, to implement in much software. Also procedures for adapting to endogenous covariates in a multilevel setting are often unavailable.

There is also another important consideration. We wish to use software that has an attractive interface for iterative model development and model selection. The graphical interface of MLwiN makes it attractive for adapting model specification, examination of diagnostics and exploring results. MLwiN also encompasses both approaches to estimation that are widely used for linear multilevel models; maximum likelihood through iterative generalised least squares (ILGS) and Monte-Carlo Markov Chain (MCMC). These approaches to estimation are more fully explored in 'Report 3: CC Review'. MLwiN also has a very flexible macro language and facility which means that its range of procedures for a variety of situations can be adapted straightforwardly to new situations. This can be profitably used to implement our preferred approach to the endogeneity issue, the multiprocess multilevel model. Apart from its general attractiveness on these grounds MLwiN is also the most widely used specialist software for multilevel modelling in the UK and familiarity with its use and application is now fairly widespread. It is also under continuous development and some of the limitations for our purposes that we uncover in our trials may soon be remedied.

To fill the gaps where the limitations arise and learn more about statistical and computing feasibility we also use WINBUGS and GENSTAT for some trials. The WINBUGS general purpose modelling package relies entirely on the Bayesian framework of MCMC estimation. It can implement our complexities fairly readily to parallel MCMC procedures in MLwiN. However, we think it fair to say that it is not as user friendly or as fast as the MLwiN implementation. GENSTAT has a very fast performance and is quite impressive in the complexity of the multilevel, crossed and other random effects structures it can entertain. However, perhaps due to lack of familiarity we find it less user friendly in data input and procedure preparation. It does not have an equations window displaying the models symbolically which would enhance the user's ability to interact with it in model development. It is also very difficult to see at present how it could handle the endogeneity complexity which will

ultimately be necessary for most of our purposes, though it is said that it could be adapted with some suitable programming development

For our projected experimental analyses our starting points are the analyses in 'Report 2: Levačić'. Thus where we have introduced covariate explanatory variables, and for ease of comparison, we have used the same that appear in that report's Tables of results. Ultimately other covariates may be required when new complexity is desired and further model development and exploration takes place. For instance, there are a limited number of census output variables in the results and data sets. These may require expanding once area random effects are also considered. Ward level covariates may also be required. It may also be useful to seek explanatory variables for any significant LEA variation that may emerge. With changing model specification it might also be the case that different pupil level characteristics and interactions between them are required. Thus we wish to emphasise that the choice of covariates used in our trial analyses and 'Report 2: Levačić' may not entirely be the end of the story.

The bulk of analyses in 'Report 2: Levačić' do not explicitly consider multilevel specifications of the model. Instead the use of robust standard errors to adjust for school clustering has been generally adopted, though not for LEA clustering. In our review, 'Report 3: CC Review', we briefly comment that there may be some limitations to this type of model fitting and the related Generalised Estimation Equation approaches. Some references on the contrasting approaches are also given in this source. The approach to the endogenously issue has been through a standard Instrumental Variation (IV) two stage least squares (2SLS) procedure which again has its limitations. Not the least of these is the well known inefficiency and sometimes poor precision of estimates although it can be satisfactory in certain situations. In principle this approach if it was thought useful could be extended fairly easily using the STATA programme to encompass further levels of clustering induced by crossing of areas with schools. We have undertaken a repeat of some of the analyses but redefined in this way the clustering basis for the robust estimation. We do not present our detailed results. Suffice it to say we get more or less the same covariate effect estimates but now the robust standard errors as expected are different and induced by the specified additional clustering. This will affect any statistical inference on results
It has been recognised that full multilevel modelling appears preferable to encompass explicit recognition of complex level effects. However, one more thing is certain about handling the endogenously issue. At first sight it might have been thought that the available first stage predictions of resource variables which are formed from instruments in stage 1 of the original IV 2SLS estimation might just be used in multilevel specification for the achievement equation. This would be just as if the endogenously issue had been dealt with by this prior analysis. Although convenient and easy to implement this is statistically entirely inappropriate for two reasons. Firstly although this might lead to consistency in estimation of model parameters inappropriate standard error estimates for them would result. The reason for this is that information on contribution to precision of estimates arising from the instrumental variable approach is carried forward from stage 1 in the original two stage process itself and is not available from the predictions alone. Intuitively the two stages cannot be seen as entirely unrelated exercises. The process of just inputting the predictions would treat them as exogenously given and the associated usual procedures for standard error estimation are not correct. Secondly, although the resource variable is a school level variable, its model equation in terms of instruments will also need to be specified using an appropriate more elaborate multilevel model. The original resource equation from which predictions might be found does not properly recognise this.

One of our trial analyses does however investigate a similar type of IV estimation using a macro procedure for multilevel models that has been developed for MLwiN. Although this proved

rather unsuccessful from a computational perspective, if IV estimation was thought to be useful its further development might be an option. Our overall preferred option using statistical criteria for handling both structural and endogenously complexities would be a multiprocess multilevel model. This involves estimating both a resource and an achievement process equation simultaneously, whilst allowing for endogenously by having a proper structure of correlated effects and disturbances. This approach has been used in 'Report 2: Levačić ' for the strict hierarchical model involving school and LEA random effects. However, there is not much emphasis on it and details of the procedure are left somewhat unelaborated. This aspect of that work did however rely on a special MLwiN macros written by Fiona Steele to which we have access. It modifies the two equation structure in a way that can be handled by the multivariate response facilities of MLwiN. Fiona is also one of the consultants on this scoping project and the macros have been adapted for the present trialling with crossed random effects structures.

5.2 MCMC Estimation Using MLwiN and Ignoring Endogeneity

We now outline on a case by case basis some of the trials we carried out, our reasons for undertaking them, the success or otherwise with which they were implemented and some comparative results. We use the KS3 mathematics score as the dependent response[5], for brevity denoted by y, and where appropriate the expenditure per pupil averaged over the 3 years as the resource variable[6]. The feasibility results and conclusions for these will be made without loss of generality for other achievement responses and resource variables. Initially for each type of trial we were probably overambitious in what could be achieved given time and resource constraints and computational difficulties. However, important lessons were learned. After discussing the trials we will then return to recommendations regarding any future full scale work.

5.2.1 Features of trials

The trials in this section were concerned with examining the behaviour of Monte Carlo Markov Chain (MCMC) estimation in MLwiN on both our selected datasets. Accumulated experience has led the Centre for Multilevel Modeling to recommend this approach where very complex cross-classified structures and many levels are concerned. Prior to the development of the MCMC implementation the analysis of such complex structures required an adaptation of the main Iterative Generalised Least Squares (IGLS) procedure which was at the heart of MLwiN. Following Rasbash and Browne (2004) we call this adaptation the RG method. The methods are further discussed in 'Report 3: CC Review' and full technical details of the RG method are described in Rasbash and Goldstein (1994). One advantage of an MCMC approach is that certain information on structural aspects of crossings is not an integral part of the estimation process as it is with RG. For instance, one practical aspect of this in RG is that there model estimation need to use large numbers of columns of variables since explicit indicator dummy variables are required for each unit in certain classifications in the crossings. If there are a large number of units in such classifications, for example wards, demands on the worksheet size create memory restrictions. There is also the necessity in the RG method for the inversion of large matrices and this creates additional computational difficulties. Frequently model fitting crashes or takes an inordinate amount of time. Some of these problems are related to increased need for memory and the exhaustion of that which is available, but there is also a technical limitation. The RG method necessitates estimating large numbers of variances which need to be constrained to be equal. This often creates numerical instability and causes a failure of the method to converge. For simpler smaller scale models and datasets, MCMC processing may be

[5] This is labelled *k3matscr* in the results of 'Report 2: Levačić' and in the edited data sets.
[6] This is labelled *pexaav* in the results of 'Report 2: Levačić' and in the edited data sets.

much more demanding than RG due to extensive simulation, and in those cases RG may be preferred. However and conversely, there is a trade off due to the greater processing requirements and intricacies of RG when it is required to handle much complexity and larger scale models. Experience has shown that in many cases with suitable choice of Bayesian priors, the two approaches give similar results. Rasbash and Browne (2004) discuss these methods more fully and give a more detailed comparison of their relative advantages and disadvantages with example analyses.

There are two limiting features of the MCMC results to be discussed in this section which we should carefully note here:

- They ignore the endogeneity issue since our point is to examine how the estimation handles the complex data structures involved. Any computation problems will be magnified for larger datasets, since although the structures remain similar, the number of units at various levels multiplies. Thus we wish to see how the procedures manage for our subsets in this respect. The hope might be to extend the MCMC approach to the multi-process models for handling endogeneity that we have suggested. We will make some comments on the feasibility of this in discussion of one of our other trialing processes.
- The MCMC approach currently implemented in MLwiN cannot fit models where random effects from two classifications that share the same set of units are specified as correlated. As discussed previously we may wish to consider this situation for LEA of school and LEA of residence. Therefore later in this section we contrast with an approach using WINBUGS with MCMC estimation that allows such correlated effects in order to examine possible consequences of this restriction. MLwiN is also currently being developed to allow this feature.

In the discussion we need by necessity to touch on a number of technical matters, terminology, and diagnostic statistics connected with MCMC. We refer unfamiliar readers to the user's manual, Browne (2003), for detailed explanations. This includes full information on Gibbs sampling which is the simulation method used, the meaning of default prior distributions used for the model parameters, suitable choices of number of iterations after 'burn ins' and other technical concepts we may refer to. For the models considered here following the guidelines and from experience with similarly sized problems the MCMC algorithms are run for 50,000 iterations after a 5,000 iteration burn in. A desktop machine with 0.5Mb of RAM running under Windows 98 was used.

We firstly consider the West Midlands subset of the data and initially part of the fairly full random effects structure considered in the previous chapter. The structure is illustrated in the classification diagram of Figure 2. First we cross-classify school by ward. Since the geography of the cells of this specification is not hierarchically arranged beneath specific LEAs, we have separate effects for LEA of school and LEA of residence at the top levels. Thus there are the 4 higher random classifications lea(school), lea(area), school and ward, with respectively 8 , 34 , 223 and 452 units on a dataset with 32579 level 1 units (pupils). We see that there are many more lea (area) units consequent on the pupils crossing the boundaries to go to school from areas outside West Midlands LEAs. Since the data selection has been based on LEA of West Midland's schools there are very few lea (school) units. As noted in a previous chapter the pupil representation of the 26 outside lea (area) units will be rather sparse. We also note that for this subset with only 8 West Midlands lea (school) units the estimation of the variance of that random effect might be quite imprecise. An alternative might be to specify these LEA effects as fixed in a practical application. However, our concern is to examine the feasibility of the random effects structure, so for the trial we leave them specified as random. It might also be pointed out

that, as is well known in modelling treating them as fixed effects would prevent us later introducing covariates defined at the lea(school) level (see Fielding (2004) for example).

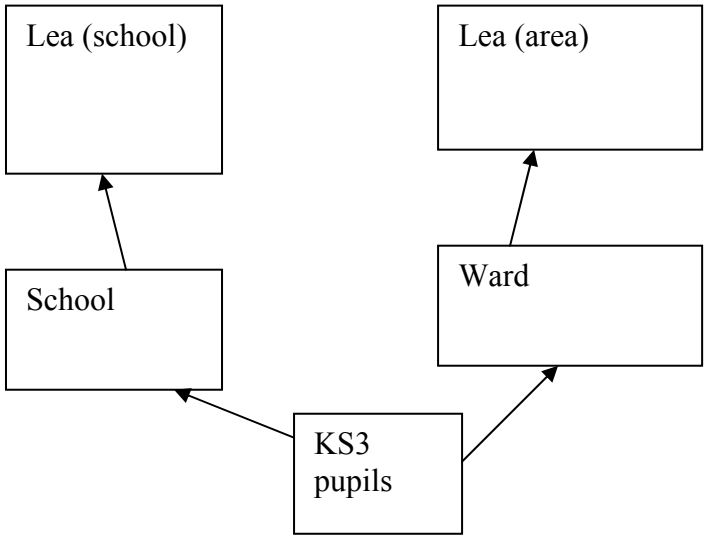

Figure 2: Classification diagram for Models 1-4 involving the ward by school crossing at level 2 and lea (school) by lea (area) at level 3

We then use the same structure but now apply it to the larger and more extensive Combined Selection data. For this the 4 higher random classifications lea(school), lea(area), school and ward have respectively 25 , 112 , 534 and 1995 units with 80,032 level 1 units (pupils). We note that due to out of area children and since the selected 25 school LEAs were geographically spread, most of the English LEAs are represented in the lea (area) classification. There are consequently a large number of wards of residence many of which will again be sparsely represented.

We will firstly consider simple variance components models in both cases with an intercept but no predictor covariates. These are labelled Model 1 and Model 2 for West Midlands and Combined Selection respectively. The variance components models establish bases by which we may judge the relative sizes of the sources of structural variation. In most substantive applications it is the convention to use such bases against which further model developments are judged. We will later look at fitting more predictors in each case. We will then in a later subsection consider using output area rather than wards to extend the scope of the experimentation. The latter will be restricted to the Combined Selection dataset.

The models we fit in this using MCMC in MlwiN are summarised in Table 5.1. A reference to this table might help the reader in keeping track of the essentials of each of the model fits as they develop. The classification diagram in Figure 2 will be appropriate for the first four models. It will be essentially the same for Model 5 and Model 6 but the ward will be replaced by output area in the right hand box at level 2.

Table 5.1: Essential features of Models 1-6

Model	Levels and crossings	Data Set	Predictor covariates
1	3: lea (area) by lea (school)	West Midlands	No
	2: ward by school		
	1: pupil		
2	3: lea (area) by lea (school)	Combined Selection	No
	2: ward by school		
	1: pupil		
3	3: lea (area) by lea (school)	West Midlands	Yes
	2: ward by school		
	1: pupil		
4	3: lea (area) by lea (school)	Combined Selection	Yes
	2: ward by school		
	1: pupil		
5	3: lea (area) by lea (school)	Combined Selection	No
	2: output area by school		
	1: pupil		
6	3: lea (area) by lea (school)	Combined Selection	Yes
	2: output area by school		
	1: pupil		

There are two possible software reasons for an analysis not to be feasible; time and memory. If the problem requires too much memory then that may make a problem totally infeasible on a particular computer, whilst if it will take 2 years to run this will also be infeasible.

Note that for each model we have chosen to only run for 50,000 iterations after a burn in of 5000 iterations as a basic test of memory usage. If we needed to run for longer then we could use thinning (see Browne (2003)), thus not requiring additional storage. In evaluating we will time the 55k iterations and look at convergence diagnostics to estimate how long we need to run for. Since this exercise is to scope analysis feasibilities we will barely comment on the substance of the empirical estimates apart from, in passing, brief remarks on a few essentials which may be of interest.

5.2.2 MCMC estimation of variance component models with four random classifications

Model 1 (West Midlands) and Model 2 (Combined Selection) have classification diagram as in Figure 2 and both are expressed symbolically as

$$y_i = \beta_0 + u^{(5)}_{0,\text{lea(school)}_i} + u^{(4)}_{0,\text{lea(area)}_i +} + u^{(3)}_{0,\text{ward}_i} + u^{(2)}_{0,\text{school}_i} + e_{0i}$$

The substantive results on variance component estimates for Model 1 and Model 2 are displayed in Table 5.2.

Table 5.2: Results for Models 1 and 2

	Model 1: WEST MIDLANDS		**Model 2: COMBINED SELECTION**	
	Estimate	Standard error	Estimate	Standard error
Fixed effects				
Intercept	5.945	0.0744	5.962	0.0541

Random effect variances				
Lea(school)	0.020	0.028	0.042	0.019
Lea (area)	0.013	0.011	0.013	0.004
Ward	0.023	0.004	0.025	0.002
School	0.426	0.043	0.357	0.023
Pupil	1.206	0.010	1.176	0.006
*MCMC Deviance**	98456.77		240110.21	

* ***Note: Here and in later tables of MCMC analyses we include the MCMC deviance statistic for completeness and for any future comparison with other analyses on the data set. It is not directly relevant to the comments we offer here. It is a goodness of fit statistic which offers one basis for model selection as models are fully developed and elaborated. Full technical details of the form and use of this statistic are found in Browne (2003).***

Model 1 took around 20 minutes to run. The larger Model 2 took just over an hour. For Model 1 the Effective Sample sizes for the 6 parameters are 211, 682, 1428, 32k, 5.3k and 48k respectively. Browne (2003) gives a technical explanation of the meaning of this concept and its role in judging the performance. The Effective Sample sizes were 176, 1017, 857, 30k, 3777, 46k for Model 2. Broadly though, this information and examination of other MCMC diagnostics such as the set of Raftery- Lewis statistics suggests that 50k stored iterations will give for the most part reasonable estimates for both models. The possible exception is for the intercept estimate in Model 2 where the diagnostics suggest that 60k iterations might be preferable. We note that the refinement of mixing of the Markov chains and hierarchical centring as discussed in detail by Browne (2004) might improve performance. We might also note as previously discussed that for the West Midlands set the LEA school effect might have been fitted as fixed effects due to there being only 8 LEAs in this subset. This might have improved computation time if models without covariates as here were all that were eventually required

Initially for substantive reasons we note that area ward variation is noticeably smaller than school. This may be due to wards being too large in aggregate to pick up area variation which might be evident for smaller discrete pockets of area effects whose socio-economic characteristics may influence educational achievement. The two LEA effects are relatively small and are also imprecisely estimated relative to their size. There may be a small element of confounding of these two effects since information on their separation is provided only by out of area children. However, we believe there are sufficient numbers of such children for us to separately identify such effects. There is scope for further research on methodology for looking at this aspect of these models.

We also might note that despite the two datasets being purposive selections with possibly different characteristics the variance component estimates are not widely dissimilar. The implication might be that subsampling of the full dataset would not necessarily be an undesirable operation in a full scale analysis. We can also see that the standard error estimates are much smaller for the larger Combined Selection set of Model 2 than they are for West Midlands. This is unsurprising since the variances for each set of random effects are based on much larger samples of units at each level.

5.2.3 MCMC estimation of models with four random classifications and with predictor variables

We next investigated adding predictor variables to the models using the same structure. For comparison the predictors are the same as the ones developed and used in the final models of 'Report 2: Levačić'. Readers are referred to that report for full definitions. These predictor covariates are a mixture of variables measured at various levels in the hierarchy. Thus we would expect some changes in the residual random effects variances away from those in the base model as predictors explained part of that variation.

There were some initial problems in fitting these models resulting in the programme crashing. On investigation there appears to be a newly discovered bug in MLwiN (now corrected) involving the missing data prior to entering the MCMC estimation code. Missing data often applied to whole units in the various lower level classifications and this creates problems since the current procedure tries to fit these units into the analysis. This bug is informative in developing the relatively new MCMC procedures and will be worked on. However a workaround is not to rely on automatic handling with missing data but to listwise delete cases on the required columns prior to analysis. This reduced the number of cases by 10% to 30910 for the West Midlands set and by 5% to 75983 for the Combined Selection. As a check on the impact of this case deletion the initial variance components models were re-run on the reduced sets. These gave similar results to before. We are thus reasonably assured that the missing data mechanism is not informative and that results are unlikely to be biased for these reasons

We then added the predictors and ran the full models; Model 3 for West Midlands and Model 4 for the Combined Selection. We note that three of the original predictors, Start lowest age 12 (school has 12 as lowest age of entry), Start lowest age 13 (school has 13 as lowest age of entry), and Jewish school could not be used in West Midlands model since they took on constant values for that subset. Similarly Start Lowest Age 12 was not used for the Combined Selection Model. . These unfitted predictors are indicated by * in the results displayed in Table 5.3.

The Model 3 run was successful and took around 78 minutes for the total 55k iterations. The Model 4 run took several hours. Model 4 as a one-off for the larger dataset, however, seems feasible. However, iterative model development which would be required in a full scale analysis would be very time consuming since each time we adopt a model change according to standard model fitting strategies we might have to wait a considerable length of time before moving to an improved changed model. For both models from an examination of the Effective Sample size statistics and other diagnostics it appeared that 50,000 stored iterations in the MCMC procedure was sufficiently long run to get precise estimates.

	Model 3: WEST MIDLANDS		**Model 4: COMBINED SELECTION**	
	Estimate	***Standard Error***	***Estimate***	***Standard Error***
Fixed effects				
Intercept	3.280	0.511	3.0488	0.167
Expenditure per pupil (averaged)	-0.0000014	0.000033	0.000013	0.00002
Female	0.0445	0.007	0.051860	0.004
Age (days from Sept 1st 1989)	0.00019	0.00003	0.00020	0.000
SEN Action/Action Plus	-0.290	0.011	-0.2936	0.007
SEN Statement	-0.331	0.034	-0.3118	0.018
Eligible for FSM	-0.104	0.009	-0.105	0.006
Ethnicity (base, white)				
Asian, Indian	0.096	0.020	0.0799	0.016
Asian, Pakistani/Bangladeshi	0.028	0.023	-0.0142	0.016
Asian, other	0.038	0.042	0.0922	0.028
Black	-0.016	0.018	-0.0144	0.013
Chinese	0.296	0.056	0.226	0.031
Mixed Ethnicity	-0.011	0.017	-0.0160	0.013
First language not English	0.038	0.018	0.0967	0.012
Key stage 2 maths adjusted	-0.126	0.023	-0.0786	0.016
Key stage 2 maths adjusted squared	0.166	0.003	0.162	0.002
School Variables:				
School has sixth form	-0.0029	0.025	-0.0100	0.017
Start lowest age 12	*		*	
Start lowest age 13	*		0.130	0.072
Gender of school (base, mixed)				
Boys' school	0.066	0.056	0.0376	0.028
Girls' school	0.146	0.052	0.0654	0.065
Type of school (base, comprehensive)				
Grammar school	0.175	0.069	0.217	0.039
Secondary modern school	0.094	0.064	0.0022	0.042
Other type of school	0.145	0.158	0.0039	0.055
Religious denomination of school (base, non-denominational)				
Roman Catholic	-0.013	0.036	-0.0206	0.020
Church of England	0.039	0.065	0.0486	0.031
Other Christian	0.053	0.093	0.0553	0.067
Jewish	*		0.101	0.111
Per cent eligible for FSM in school	-0.012	0.003	-0.00878	0.002
Per cent eligible for FSM squared	0.00013	0.00005	0.00005	0.00003
Per cent AEN in school	0.00020	0.0010	0.00156	0.001
Specialist school	0.031	0.024	0.0227	0.014
Special measures	-0.051	0.074	-0.0748	0.040
EIC or EAZ	0.095	0.057	0.0441	0.038
Beacon school	0.099	0.047	0.0331	0.024
Leading Edge Partnership	-0.090	0.082	0.0250	0.036

Leadership incentive grants	-0.061	0.053	-0.0608	0.035
Teachers' pay ratio (averaged)	0.187	0.457	0.220	0.128
Urban local authority district	-0.088	0.053	0.00012	0.021
Capacity Utilisation (averaged)	0.014	0.053	-0.0196	0.049
Census variables:				
Proportion Unemployed	-0.159	0.168	-0.176	0.122
Proportion Black Ethnicity	-0.057	0.099	0.100	0.063
Proportion Chinese Ethnicity	-0.285	0.284	-0.0936	0.157
Proportion Pakistani/Bangladeshi Ethnicity	-0.203	0.044	-0.1283	0.028
Proportion Indian Ethnicity	0.089	0.049	0.0722	0.046
Proportion Lone Parent Households	-0.099	0.033	-0.161	0.021
Proportion NVQ Level 1 or less	-0.419	0.039	-0.388	0.024
Random effects variances				
Lea(school)	0.0027	0.003	0.0021	0.00113
Lea (area)	0.0024	0.002	0.0047	0.0020
Ward	0.0039	0.001	0.0186	0.0014
School	0.022	0.003	0.0043	0.0005
Pupil	0.294	0.002	0.300	0.002
MCMC deviance	49898.88		124113.2	

Table 5.3: Results for Models 3 and 4

Note: * indicates that covariate effects could not be fitted since the covariates had constant values within the data subsets analysed

The results for fixed effect coefficients are broadly what we might expect from substantive results in 'Report 2: Levačić' so we are confident that data and other estimation problems are ironed out well. Naturally there are a few minor differences both between the above two sets of results and with those of 'Report 2: Levačić' but these may be naturally due to the nature of the purposive selections. The Combined Selection may be broader in character than some features which may be special to the West Midlands. In full scale substantive analyses the results are suggestive that effects of region and interaction with other covariates might be worthy of detailed exploration. That the results are again broadly comparable is again a pointer to designed sampling of the full data as a reasonable prospect. We take this matter up in the final section of the report. Since the point of this report is not one of substantive investigation we refrain from commenting much on the results per se. That would necessitate more in the way of an iterative model fitting strategy and full model development. In passing though we can observe that both in these models and in 'Report 2: Levačić' we can identify a few obvious variables which seem to have large effects relative to their standard errors: proportion in residence output area with NVQ Level 1, Chinese ethnicity, and dummy indicators relating to categories of statementing of Special Educational Needs. We might also note the large reduction in residual school variances over the base model which is hardly surprising since many of the explanatory covariates are defined at the school level. In more extensive model development with additional explanatory variables relating to areas of residence we might expect a larger reduction in their residual variances also.

5.2.4 MCMC estimation replacing ward by output area in the hierarchy. Variance Component model on combined selection dataset

The place of residence hierarchy in the structure of course contained further levels and we considered firstly replacing the ward level with output area in the larger Combined Selection set of data. The sample of 80,032 came from a large number of 26,294 output areas so this run is an investigation of the feasibility of a large number of random effects. The structure we investigate is as in Figure 2 but the output area replaces ward in the crossing with school at level 2. The average of about 3 children per output area is indicative of sparsity. The output area is also the level at which the census predictors are used and 'Report 2: Levačić' indicates these as proxy for family socio-economic characteristics. Here by having additional random effects for unobservable sources of small area variation we are perhaps too close to these child characteristics with small numbers of children to be able to separately unconfound these area effects from those of family. But this is part of the feasibility exercise which touches on the substantive interpretation of any such estimated effects. The run was quite successful. The results of the variance component model, Model 5, are in Table 5.4 below.

Table 5.4: Results for Model 5

Fixed effects	Estimate	Standard error
Intercept	5.9542	0.055
Random effect variances		
Lea(school)	0.0438	0.021
Lea (area)	0.0047	0.002
Census output area	0.0586	0.004
School	0.3730	0.024
Pupil	1.1371	0.007
MCMC Deviance	237402.9	

5.2.5 MCMC estimation replacing ward by output area in the hierarchy. Model with predictors on Combined Selection dataset

After listwise deleting for missing data, the model was fitted to a reduced dataset of 75,983 pupils on the Combined Selection data using the predictors. This run took over 5 hours, which is indicative that lengthy work would be involved in model development, evaluation and choice. Diagnostics revealed a satisfactory number of MCMC iterations. The results of fitting this Model 6 are in Table 5.5

We note that the output area residual variance has reduced considerably in this model over the base model and this is hardly surprising since the census predictor variables are measured at this level. However, compared to the school residual variance the unobserved sources contributing to the area effect are still quite considerable.

Results for Combined selection dataset		
Output area as lower level area of residence effect		
Fixed effects		
	Estimate	**Standard Error**
Intercept	3.056	0.165
Expenditure per pupil (averaged)	0.0000096	0.000021
Gender	0.051	0.004
Age (days from Sept 1st 1989)	0.00020	0.000019
SEN Action/Action Plus	-0.294	0.007
SEN Statement	-0.312	0.018
Eligible for FSM	-0.105	0.006
Ethnicity (base, white)		
Asian, Indian	0.080	0.016
Asian, Pakistani/Bangladeshi	-0.012	0.016
Asian, other	0.091	0.028
Black	-0.014	0.013
Chinese	0.227	0.031
Mixed Ethnicity	-0.016	0.013
First language not English	0.096	0.012
Key stage 2 maths	-0.077	0.016
Key stage 2 maths squared	0.162	0.002
School Variables:		
School has sixth form	-0.005	0.017
Stat lowest age 12	*	
Start lowest age 13	*	
Gender of school (base, mixed)		
Boys' school	0.035	0.029
Girls' school	0.063	0.028
Type of school (base, comprehensive)		
Grammar school	0.210	0.039
Secondary modern school	-0.004	0.057
Other type of school	0.002	0.055
Religious denomination of school (base, non-denominational)		
Roman Catholic	-0.022	0.021
Church of England	0.045	0.031
Other Christian	0.057	0.066
Jewish	0.093	0.114
Per cent eligible for FSM in school	-0.009	0.002
Per cent eligible for FSM squared	0.00006	0.00003
Per cent AEN in school	0.002	0.001
Specialist school	0.021	0.014
Special measures	-0.084	0.040
EIC or EAZ	0.037	0.038
Beacon school	0.029	0.025
Leading Edge Partnership	0.0260	0.037

Leadership incentive grants	-0.055	0.035
Teachers' pay ratio (averaged)	0.243	0.125
Urban local authority district	0.001	0.021
Capacity Utilisation (averaged)	-0.025	0.051
Census variables:		
Proportion Unemployed	-0.191	0.127
Proportion Black Ethnicity	0.101	0.059
Proportion Chinese Ethnicity	-0.134	0.159
Proportion Pakistani/Bangladeshi Ethnicity	-0.119	0.026
Proportion Indian Ethnicity	0.056	0.042
Proportion Lone Parent Households	-0.168	0.022
Proportion NVQ Level 1 or less	-0.396	0.024
Random effects variance		
Lea (school)	0.0020	0.0014
Lea (area)	0.0057	0.0022
School	0.0194	0.0014
Census output area	0.0098	0.0010
Pupil	0.2931	0.0018
MCMC deviance	122468	

Table 5.5: Results for Model 6

5.2.6 Testing MLwiN's MCMC procedure on the full available structure. Variance component model on Combined Selection dataset.

In the final trial with MCMC we decided to test on the Combined Selection data MLwiN's ability to cope with further higher classifications by including all of the potential residence levels outlined in the structure in the previous chapter. This structure has six higher-level classifications. For residence we have four levels; output area nested in ward nested in local authority district nested in LEA of residence. For the education hierarchy there are two; school and school LEA.

Table 5.6: Results for Model 7

Fixed effects	Estimate	Standard error
Intercept	5.958	0.055
Random effect variances		
Lea(school)	0.0411	0.0191
Lea (area)	0.0035	0.0025
Local Authority district	0.0018	0.0013
Ward	0.0227	0.0021
Census output area	0.0433	0.0036
School	0.3531	0.0231
Pupil	1.1352	0.0065

Model 7 is thus a variance component model with just a random intercept term. The results of the run are given in Table 5.6. We note again that we cannot in this implementation allow correlated effects for the two LEA classifications.

This run took some time and it would take quite a large number of hours with added predictors. It is still feasible but might prove restrictive in iterative exploratory model investigation unless the time restriction was recognised. It was concluded that the structure of the data was such that confounding of random effects does not pose serious problems.

5.3 Restricted Maximum Likelihood (REML) Estimation Using GENSTAT (Endogeneity Ignored)

GENSTAT has REML algorithms that have been specifically designed for large crossed random effect models and are reputed as computationally very efficient. In experimentation, data preparation using Excel as an intermediary proved quite difficult due to the size of dataset,. Thus this trial additionally required the programming of a short piece of code from the programming language C to get the data into the right format for GENSTAT. There is also no simple graphical user interface with GENSTAT to make model development easy. We will, however, suggest a strategy in our conclusions which might include the use of GENSTAT in a fully developed project. This might make use of its computational efficiency in a final phase of the development of a full project using cross-classified models.

Given time and resource constraints in this scoping exercise the various options cannot be fully explored. Thus experimentation is limited to the default methods for linear mixed models on intercept only variance components models.

The West Midlands subset of 30910 after listwise deletion of observations that had missing values on one or more predictors was fitted. The model with 4 higher-level classifications (LEA of school (8 possibilities), School (218), Ward (443) and LEA of residence (32)) was handled very easily by GENSTAT and was completed in a matter of seconds.

The estimates are given in the following Table 5.7 and compared with the corresponding MCMC results. Estimated standard errors are in parentheses

Table 5.7: Results of REML trial compared with MLwiN MCMC results

Parameter	Genstat REML Estimate (SE)	MLwiN MCMC Mean estimate (SE)	MLwiN MCMC mode/median
Fixed effects			
Intercept	5.949 (0.068)	5.949 (0.073)	5.949/5.949
Random effects variance			
LEA (school)	0.012 (0.018)	0.018 (0.027)	0.000/0.009
LEA (residence)	0.008 (0.006)	0.011 (0.009)	0.000/0.009
Ward	0.023 (0.004)	0.023 (0.004)	0.023/0.023
School	0.421 (0.042)	0.426 (0.043)	0.420/0.423
Pupil	1.206 (0.010)	1.206 (0.010)	1.206/1.206

We note that there is really good agreement with MCMC results. In the table the mode and median of the MCMC posterior distribution are given in addition to the mean used in the

previous results. The mode would correspond with the maximum likelihood estimate (Browne (2003)). We see that the REML estimates all lie between the MCMC mode and mean. The LEA school variance is based on only 8 LEAs and so is very small and hard to estimate precisely by all methods. This limitation might be removed on wider datasets.

For further experimentation the same model was tried on the full dataset with 464,709 observations after having removed a few with missing unit identifiers. This time we have 149 LEAs (both for school and residence), 3083 schools and 7963 wards. Impressively, although GENSTAT evidently found this dataset harder to manage this variance component model was fitted in less than 15 minutes. The estimates are given in Table 5.8

Table 5.8: REML estimates for full variance component model o full data set

Parameter	*Genstat REML*
Fixed effects	
Intercept	5.971 (0.0197)
Random effects variance	
LEA (School)	0.031 (0.006)
LEA (Residence)	0.008 (0.002)
Ward	0.028 (0.001)
School	0.267 (0.007)
Pupil	1.204 (0.003)

It should be noted that in the time available testing further models with GENSTAT on this dataset was infeasible. Problems with fitting predictor fixed effects are not envisaged. Previous experience found that GENSTAT is not very good with large sets of random effects when the response is binomial. Also there is no experience to draw on in order to comment on how good it is at fitting random coefficients or its functionality in this direction. Such extended specification of models might be required in a full project. GENSTAT could also not as yet handle multi-process models which will be required ultimately to handle the endogeneity problem. This situation may change. However, due to its relative speed it may have a role to play in a full scale project in ways we will indicate.

5.4 MCMC Estimation Using WINBUGS Allowing Correlation Of Highest Level Effects

One constraining feature of the MCMC procedures in MLwiN mentioned above is that we cannot allow correlation between random effects for two classifications sharing a common set of units. Such a correlation may prove to be an important feature. For our structure it seems reasonable to investigate a possible correlation between the lea (school) and lea (area) effects. Intuitively these pair of random effects might be thought to be subject to similar common unobserved influences. WINBUGS is a general statistical modelling with random effects package using parallel MCMC estimation, but since it does not directly use special features of multilevel structures it has proved not as efficient as the MLwiN algorithm in many previous applications. However, it can accommodate the correlated effects that we require. Another example in area effects on mortality using a similar structure has been given by Leyland and Naess (2004). Given our resources available, trials were limited to the West Midlands dataset and models tried had separate use of ward and output area crossed structures with schools. The full set of predictors are used and the model fitted with the ward effect is

$$y_i = X\beta + u^{(5)}_{0,\text{lea(school)}_i} + u^{(4)}_{0,\text{lea(area)}_i +} + u^{(3)}_{0,\text{ward}_i} + u^{(2)}_{0,\text{school}_i} + e_{0i} \ ,$$

using the same classification diagram as in Figure 2. This was as used in the MLwiN estimation above, but now correlated effects at the top level are allowed. The vector β is the parameter vector for fixed predictor effects including the intercept. Apart from variance parameters $\sigma^2_{0,\text{lea(school)}}$ and $\sigma^2_{0,\text{lea(area)}}$ for the highest level random effects we now allow them to have covariance $\sigma_{\text{lea(school),lea(area)}}$. We call this the ward model. The model was also fitted replacing the ward effect by the output area effect and we label this second model the output area model. Setting up the models for WINBUGS requires quite a different characterisation for the specification of prior distributions of parameters. This leads to somewhat different prior specifications than the default MLwiN settings used in the previous experiments. For completeness full details of the structure of the prior distributions for the ward model are given in Appendix 4. A similar characterisation follows for the output area model by replacing the ward variance by the output variance. Results are presented in Table 5.9 and are now discussed.

For estimation two chains were used with different starting values. Following a burn-in of 10k, results presented are based on chains of length 25k (i.e. based on 50k draws)[7]. For further detailed methodological technicalities connected to chaining in MCMC, Browne (2003) may be referenced.

The correlation between lea (school) and lea (area) effects was estimated to be quite positive at 0.39 in both ward and output area models. However these were estimated very imprecisely. For the ward model the 95% confidence interval is
(-0.67, 0.97). For the output area model it is (-0.71, 0.98). A main reason for the lack of precision is the restriction in this data set to only eight lea (school) units. There were many lea (area) units for which there was no matching of lea (school) units. The correlation is therefore based on quite a small sample of the top level units. It is also possible that running the chains for longer would lead to more precise estimates. We were concerned in this experiment to examine software feasibility for estimating this type of model so evidence on the substantive correlation structure may have to await full scale fitting.

[7] These methods basically work by simulating (correlated) draws from the (posterior) distribution of the unknown parameters. These draws taken sequentially form a Markov chain i.e. a chain where the value of the next draw of each parameter is only dependent on its current value and not any past values. The chain of values can then be summarised as a distribution and this can be used to calculate summary statistics for each parameter, for example it's mean and variance (estimate of standard error). The iterative process converges to a stationary distribution. Since the chain is stochastic - the current samples are conditional on the previous samples , the final values obtained are dependent on the starting values used. Running more than one chain facilitates the process of checking that the stationarity has been achieved, i.e. that the final posterior distribution is independent of the starting values used

Table 5.9: Results on variance estimates for models estimated using WINBUGSS incorporating a correlation between the top level effects *

	Ward model	**Output area model**
Random effects variances and covariances		
Lea(school)	0.0013	0.0013
Lea (area)	0.0018	0.0023
Covariance Lea(school), Lea(area)	0.0005	0.0006
Ward	0.0039	--
Output area	--	0.0121
School	0.0218	0.0228
Pupil	0.2942	0.2850
Total variance {lea (school and lea (area) the same}**	0.3239	0.3248
Total variance {lea (school and lea (area) different}	0.3230	0.3236
VPC: % of variance due to higher level lea (school) and lea(area)-same	1.2339	1.5085
VPC: % of variance due to higher level lea (school) and lea(area)-different	0.9395	1.1290

Notes: * The variances for the ward model are almost identical to the MLwiN estimation in the corresponding Model 2 above

** Pupils going to school and living in same area will have a contribution of $\sigma^2_{0,\text{lea(school)}} + \sigma^2_{0,\text{lea(school)}} + 2\sigma_{0,\text{lea(school).lea(area)}}$ from the top levels to the variance. For pupils living and going to school in different areas there will be no covariance term.

Although the correlation between the top level effects was estimated large, the actual size of the variances and covariance was small compared to the variances at other levels. Again, however, these estimates are very imprecise. Table 5.9 shows estimates (means of posteriors) of the variances and covariances of the random effects structure under the two models, together with some summary statistics. Fixed effects estimates which are not detailed were not out of line with those of previous results. Variance Partitioning Coefficients (VPC) which are the % of total residual variance attributed to levels are also given. Only about 1% of the total variation arises due to differences between lea (school) s or lea (area) s. Therefore, despite the strong correlation between the two, it is possible to question the importance of either. We initially questioned whether it was necessary to include a correlation parameter if the top level education and area effects were modelled. Further investigations possible include an examination of the relationship between residual estimates in both the MCMC framework of MLwiN and that of WINBUGS. These might have led to further insights which might govern the type of model to be used in full scale analyses. These have not been undertaken. However, to incorporate the correlation would at present require use of the more cumbersome and for our purposes time inefficient WINBUGS. Each of the above two models once set up took approximately 24 hours to run. It may also usefully noted that the machine used for this experiment had a somewhat higher memory specification than the one used for MLwiN. With the development of MCMC in

MLwiN and if such a framework was thought desirable, it could be implemented in MLwiN more easily in future. We return to this question in our conclusions.

The model fit can be assessed using the DIC Information Criterion (see Spiegelhalter (2003) for a full technical discussion and how it is interpreted.). Of relevance here is the little difference between the values of this for the two models as in Table 5.10. This and other more detailed diagnostics revealed a similarity in their goodness of fit.

Table 5.10: Model fit diagnostic

	DIC Information Criterion
Ward model	50251
Output area model	50236

A model using both ward and output area (and local authority district) is feasible for this dataset in this type of model using WINBUGS. However, it would take even much longer to run than the current trials.

5.5 Multiprocess Modelling Using RG Method In MLwiN

So far in our trial analyses we have not considered the issue of endogeneity of resources. Our concern has been to evaluate the extent to which the existing methodology and software can handle the complex structures involved in the data.
The methods we have considered have certain advantages in handling large complex datasets in ways we have outlined. These advantages are also more apparent if we require handling the top level LEA feature of the structure. This requirement must remain an open question. However, in 'Report 2: Levačić' the extension of estimation to handle endogeneity within a multilevel specification was through simultaneous equation multiprocess modelling of achievement and resources. This used the RG method and adapted straightforwardly the multivariate response features of MLwiN. However, specially written macros were required to set up the model and worksheet for the hierarchical education units though there is little detail of this in 'Report 2: Levačić' itself. In this section we consider adapting these macros to set up the more complex structures including cross-classification of schools within area units. We do not include the higher level units in this initial experimentation. In principle this could be done but the computation would be much more extensive. We will also later consider making further adaptations of this approach using MCMC estimation. However, in this connection we will outline some of the limitations of the currently developed software.

The aim of the part of the scoping study in this section then was to explore the feasibility of estimating simultaneous equations models to assess the impact of the resource variable **pupil expenditure,** denoted by *pexaav* in model formulations, on KS3 **maths** attainment (y), while taking into account the cross-classification of schools by area of residence (wards or census output areas) in the attainment equation. We did not include the full set of control covariates (predictors) since this would have added to the computation burden. The resource equation is itself defined at school level since pupils within a school have a common value so without LEA

effects it is a simple single level equation. Both the Combined selection and the West Midlands datasets are trialled

We nest pupils within a cross-classification of schools and areas. Denote by $y_{i(jk)}$ the attainment of pupil *i* in school *j* and area *k*. The multiprocess cross-classified model considered consists of the two equations of general form (1) and (2) below. These are estimated simultaneously. To emphasise the difference between the levels at which the two responses are measured it is here more convenient to use the older multilevel model notation (see 'Report 3: CC Review')

$$y_{i(jk)} = (X_y \beta)_{i(jk)} + \beta_{pexaav} pexaav_j + u_j^{(y)} + v_k^{(y)} + e_{i(jk)} \qquad (1)$$

$$pexaav_j = (X_{pexaav} \Gamma)_j + u_j^{(pexaav)} \qquad (2)$$

Here X_y is the data matrix of covariates to be applied in the attainment equation and *pexaav* is the endogenous resource variable. X_{pexaav} contains covariates for the resource equation. Subject to identifiability constraints there could be overlap between X_y and X_{pexaav} in some applications. The $u_j = (u_j^{(y)}, u_j^{(pexaav)})$ are school random effects on achievement and resources. The $v_k^{(y)}$ are area random effects; either ward or output area in the current trials. The usual pupil level disturbances are denoted by $e_{i(jk)}$.

Equations (1) and (2) define a multiprocess (simultaneous equations) model. The equations are linked by allowing for correlation between the school residuals $u_j^{(y)}$ and $u_j^{(pexaav)}$. This is the crucial way of handling the endogeneity issue. The model can also be thought of as a bivariate response model. MLwiN has special facilities for handling multivariate responses and it is to these that the MLwIN macros which need to be specially written are addressed. The MLwiN model allows the structure of the equations for multivariate response to be quite differently specified and this is a clear advantage here over some other software that might be available for multivariate responses. We should note that here the responses are defined at different levels. Attainment is at the pupil level. Expenditure is at the school level. The macro for the first trial is given in Appendix 3. This could readily be adapted for other specific models, different data sets and more complex structures.

For the purposes of this part of the scoping study X_y contains an intercept and linear and squared terms for the prior ability variable; KS2 attainment in mathematics. Including the full set of predictor variables as in previous trials would present little difficulty but time to try these models would be more extensive. The variables in X_{pexaav} used are the same set of instruments used in 'Report 2: Levačić': dummies for party political control of the local authority in 2002 with reference category Labour control, Average Standard Spending Assessment per pupil, and Full Time equivalent (FTE) number of pupils in 1999 [8]. In future work it might be appropriate to consider more extended specifications of the resource equations with additional predictors, random effects for LEAs and possibly interactions and differential LEA effects.

[8] In 'Report 2: Levačić' and in the edited data sets these variables are labelled : *yrlab02* (Labour) *yrcon02* (Conservative*), yrlib02* (Liberal*) and yrnoc02* (No overall control); *SCPAAV* (Average Standard Spending Assessment per pupil); *FTEPUP99* Full Time equivalent (FTE) number of pupils in 1999.

Given the computation times required to fit these models, only wards were considered. Due to the very large number of census output areas, models using this definition of area will take much longer to fit. As such, it may be impractical to use the RG method using output areas and more complex structural features that we have previously entertained.

LEA effects were not considered in these trials. One approach for these data sets due to their limited number might have been to use fixed effects for lea(school) but this approach would confound with LEA-level variables (see Fielding (2004)). This would be a particular problem here as some of the instruments used to identify the resource equation are at the LEA-level. This problem does not arise if the top level effects are specified as random which might be the case in future analyses.

To fit the multiprocess model in MLwiN, the data must first be restructured into bivariate response format $(y_{i(jk)}, pexaav_j)$. This is the first step in the detail of the macros. Because the expenditure variable $pexaav_j$ is at the school level, only one response per school is needed. In preparation here, $pexaav_j$ is attached to the first pupil in each school. Dummy variables ($w^{(y)}$ and $w^{(pexaav)}$) indicating the two different responses are then defined and these are interacted with covariates.

All estimation in these trials was carried out using the RG method based on IGLS, which is extremely computationally intensive given the large number of schools and areas, even for the restricted data subsets and using the larger wards rather than output areas. Estimation times can be reduced by using the XOMIT command to search for groupings and to omit cells in the school-ward cross-classification with few pupils. In this analysis, cells with fewer than certain numbers pupils were dropped in the experimentation. Whilst this enhances the feasibility and was done for the purposes of this trial of the method, future questions might arise as to the advisability of this. Many of dropped cells are likely to arise because of the 'out of area' children who travel across LEA boundaries to go to school. These pupils may have entirely different characteristics and substantive interpretations of results may be somewhat distorted.

The multiprocess cross-classified model is specified as a 3-level model. After searching for groupings in the cross-classification, the 'group' variable becomes the level 3 identifier. Schools are at level 2 and pupils at level 1. Note that although it is usually more computationally efficient to put the classification with the largest number of units at level 2 (ward, here), in this case we must put schools at level 2 since this is the classification is common to both responses. Having specified this hierarchical structure, the SETX command is used to set $w^{(y)}$ to have a random coefficient across wards. The coefficients for both $w^{(y)}$ and $w^{(pexaav)}$ are random at the school level, and the coefficient for $w^{(y)}$ is random at the pupil level. These are all essential details in the macro set up.

After dropping missing values, the Combined Selection data contains 75049 pupils, nested within 502 schools and 1969 wards. We then fitted cross-classified models after dropping cells in the cross-classification of school and ward with 9 or fewer pupils, and then searching for groupings. This led to 55 groups. The maximum number of wards per group was 117 wards. It is this latter quantity on which storage, memory requirements and hence feasibility and speed of execution of the RG method depend. The analysis dataset contains 55353 pupils, nested within 488 schools and 1001 wards, so a fair amount of data was dropped to make the operation less time consuming.

Results from this analysis are given in Table 5.11 and are also compared with a single process achievement model which ignored the endogeneity issue. We may note the expected change in the resource coefficient in the attainment equation although other parameter estimates are broadly comparable. We note that here and in some other tables of results the resource coefficient is negative. This should not worry us unduly since we are only trailing the method and have not included all the necessary control variables. Without controlling for the many other observable factors which are associated with expenditure we are probably picking up the attraction of resources to schools in circumstances associated with lower levels of attainment. Apart from the endogeneity issue the full set of predictors will control for these and better specify the direct effect of resources. Evidence of the likely impact of endogeneity which has now been controlled is shown by the relatively large 0.32 correlation between school effects. Although the multiprocess model took about an hour and a half to run this is still quite lengthy given all the computing efficiency savings that have been considered.

Table 5.11: Results from single and multiprocess cross-classified models (schools with wards) on Combined Selection data after omitting cells with ≤ 9 pupils.

	Single process model		Multiprocess model	
	Estimate	St. Error	Estimate	St. Error
KS3 maths attainment equation ***Fixed effects***				
Intercept	2.747	0.069	2.753	0.073
Expenditure per pupil (pexaav)	-0.085	0.008	-0.129	0.005
Key stage 2 maths	-0.051	0.030	-0.051	0.031
Key stage 2 maths squared	0.167	0.004	0.167	0.004
Random effects variances				
Ward	0.0095	0.0009	0.009	0.0009
School	0.040	0.0027	0.043	0.0025
Pupil	0.316	0.0091	0.316	0.009
Expenditure equation				
Fixed effects				
Constant	-	-	-6.985	1.453
Conservative in 2002	-	-	-0.054	0.043
Liberal in 2002	-	-	0.392	0.043
No Overall control in 2002	-	-	0.100	0.060
Average Standard Spending Assessment per pupil	-	-	0.003	0.0005
Full Time equivalent (FTE) number of pupils in 1999	-	-	-0.001	0.0002
Random effects variances				
School	-	-	0.800	0.110
Covariance between school residuals for attainment and expenditure	-	-	0.060	0.009
Correlation between school residuals for attainment and expenditure			0.323	
*Estimation time**	24 mins, 7 sec		1 hour, 12 mins, 33 sec	

*Estimation times are for a 2.80 GHz Pentium IV PC with 2GB RAM running Windows 2000.

The next set of results in Table 5.12 are from fitting the same cross-classified model to the Combined Selection data with less restrictive dropping of small cells. Cells in the cross-classification of school and ward with 5 or fewer pupils were dropped followed by the usual search for block groups. This led to 23 groups with now a larger maximum of up to 421 wards in each. The analysis dataset contains 62394 pupils, nested within 497 schools and 1116 wards. Some of the coefficients and parameter estimates are sufficiently different from Table 5.11 to suggest that care should be exercised in routinely dropping certain cells for computational feasibility. The advisability of doing this must be balanced in any future fuller analysis.

However, we note that the less restrictive dropping of cells has more than doubled the time taken to fit the model

Table 5.12: Results from single and multiprocess cross-classified models (schools with wards) on Combined Selection dataset after omitting cells with ≤ 5 pupils.

	Single process model		Multiprocess model	
	Estimate	St. Error	Estimate	St. Error
KS3 maths attainment equation				
Fixed effects				
Intercept	2.736	0.039	2.743	0.039
Expenditure per pupil (pexaav)	-0.089	0.008	-0.130	0.008
Key stage 2 maths	-0.032	0.018	-0.033	0.018
Key stage 2 maths squared	0.164	0.002	0.164	0.002
Ward	0.009	0.0007	0.009	0.0007
School	0.041	0.003	0.044	0.003
Pupil	0.314	0.002	0.314	0.002
Expenditure equation				
Constant	-	-	-7.071	0.717
Conservative in 2002	-	-	-0.055	0.031
Liberal in 2002	-	-	0.391	0.120
No Overall control in 2002	-	-	0.101	0.035
Average Standard Spending Assessment per pupil	-	-	0.003	0.0002
Full Time equivalent (FTE) number of pupils in 1999	-	-	-0.001	0.0001
Random effects variances				
School	-	-	0.811	0.051
Covariance between school residuals for attainment and expenditure	-	-	0.057	0.009
Correlation between school residuals for attainment and expenditure			0.302	
Estimation time	1 hour, 20 min, 46 sec		4 hours, 3 mins, 30 sec	

The model procedure was then tried on the West Midlands dataset. After listwise dropping of missing values this had 30910 pupils, nested within 217 schools and 443 wards. We analysed a cross-classified model after dropping cells in the cross-classification of school and ward but now can try a looser criterion of 3 or fewer pupils, The search for non-overlapping groups of ward and school cells led to 2 groups with 287 wards in one and 2 wards in the other. Thus the block grouping was not quite so useful here but this might be expected since the West Midlands set was based on contiguous areas and we might expect a more diverse crossing with less blocking

of groups of wards and schools. The analysis dataset contains 27641 pupils, nested within 215 schools and 289 wards. Results are given in Table 5.13 below.

Table 5.13: Results from single and multiprocess cross-classified models (schools with wards) on West Midlands dataset after omitting cells with ≤ 3 pupils.

	Single process model		Multiprocess model	
	Estimate	St. Error	Estimate	St. Error
KS3 maths attainment equation				
Fixed effects				
Intercept	2.777	0.053	2.773	0.053
Expenditure per pupil (pexaav)	-0.077	0.013	-0.060	0.013
Key stage 2 maths	-0.053	0.025	-0.052	0.025
Key stage 2 maths squared	0.166	0.003	0.166	0.003
Random effects variances				
Ward	0.009	0.001	0.009	0.001
School	0.045	0.005	0.045	0.005
Pupil	0.310	0.003	0.310	0.003
Expenditure equation				
Fixed effects				
Intercept	-	-	-9.284	1.188
Conservative in 2002	-	-	0.279	0.121
Liberal in 2002	-	-	-	-
No Overall control in 2002	-	-	0.043	0.045
Average Standard Spending Assessment per pupil	-	-	0.003	0.0003
Full Time equivalent (FTE) number of pupils in 1999	-	-	-0.0016	0.0002
Random effects variances				
School	-	-	0.606	0.058
Covariance between school residuals for attainment and expenditure	-	-	-0.024	0.012
Correlation between school residuals for attainment and expenditure			-0.145	
Estimation time	21 mins, 2 sec		41 mins, 24 sec	

Even with this smaller data set with a smaller total number of random effects the run time is still quite considerable. The results also indicate that there is now negative correlation between the two school effects. There may be a regional dimension to some of this work which has not previously been uncovered. This will be handled to some extent if higher level LEA random effects to represent unobserved LEA factors were included in models. However, this is a dimension worthy of more investigation in future fuller analyses.

The macros used and the results above have embedded within them the RG adaptation of IGLS estimation. Indications are that computing requirements even if models are feasible on fuller data sets are likely to be quite extensive. In many ways the multiprocess modelling is our preferred approach to handling endogenous variables in the complex random effect models. But it is clear that the RG method may require large amounts of resources and may indeed run into difficulties for larger datasets.

The way forward as indicated by our previous trials may be to use more efficient MCMC methods once the bivariate models have been set up. MCMC in MLwiN can handle the multivariate responses though the extra feature may mean models take much longer to run than the previous trials indicate. There is, however one further problem in advocating this way forward which may not be restrictive in the longer term. At present using MCMC it not possible to estimate bivariate response models (i.e. equations (1) and (2) jointly) where the responses are at different levels as we can using RG. Work in this direction is currently planned at the Centre for Multilevel Modelling as part of a current ESRC project. How long before this reaches fruition is an open question but any additional resource input to this development would obviously hasten the process. An alternative approach which uses a strict hierarchical modelling framework by iterating between the two hierarchies, here education and residences is suggested by Clayton and Rasbash (1999) and known as data augmentation. This approach has an element of both IGLS and Monte-Carlo simulation. It is shown to reduce estimation times for many cross –classified models with complex features. In principle the approach could be extended to multiprocess models. We have not trialled this since it would require very lengthy detailed work on writing of special macros for MLwIN but is an element of programme development that might be considered in any future full scale project.

5.6 Multilevel Instrumental Variable (IV) Estimation Using MLwiN Macros

'Report 2: Levačić' used IV estimation with the instruments for the endogenous variable as outlined previously. However this estimation did not explicitly consider a multilevel model specification but incorporated a refinement to adjust standard errors known as sandwich estimation. The latter produce robust standard errors allowing for clustering of pupils in schools. We have rehearsed some of the difficulties with IV estimation previously and in 'Report 3: CC Review'. In principle this type of estimation is possible with tighter definitions of clusters to incorporate the cross-classes envisaged and we tried these in STATA. Implementation is fairly straightforward. However, the fitting of multilevel cross-classified models is the main objective of our scoping. To add to our trialling experience we considered the potential of IV estimation within an explicitly specified multilevel random effects framework. Spencer and Fielding (2000, 2002) gave details of an MLwiN macro for such estimation of strictly hierarchical models within an IGLS approach and further considered its joint use with WINBUGS in model fitting strategies. These approaches use a two-stage idea similar to the single level approaches used in 'Report 2: Levačić' but extend to multilevel equations. Our original intention was to scope these various approaches to investigate their ease using eight trials formed from three dichotomies of two datasets, two forms of cross class (using wards or output areas) and two strategies, MLwIN macro and joint use with WINBUGS. We did not consider the higher level LEA classifications for these trials. All the trials considered require the adaptation of the macros to use the RG method for cross-classified effects. We also considered how the macros might be used in association with MCMC estimation of the instrumented achievement equation in order to improve efficiency. Most of these potential trials ran into severe difficulties for a variety of technical and computational reasons. The main ones were:

- The larger data sets ran into difficulties and crashed using the IV macro due to space issues because of large number of random effects even when only the smaller number of ward units

was used. Even with the smaller West Midlands dataset these problems arose when output areas were tried as the area classification.

- During the operation of IGLS procedure on occasion certain variances go negative during the iterative process. This would create difficulties for the matrix inversions required. In straightforward applications this is handled effectively by the automatic routine stepping in to make necessary adaptations to the iterative process by inverting a full matrix. In the special IV macro this often created insurmountable problems as MLwIN diagnostics indicated, mainly because the required matrix then became too large to cope with. This could be handled with by detailed re-writing of macros but was not possible to perform at this stage.
- In the only reasonably successful trial, that of the West Midlands data set using wards as the area classification, parameter estimates could be found. However, there was difficulty in estimating their standard errors. The technical reason for this is that the current macros struggle to cope with the lack of more than one block created at the top level by the cross-class. This is in turn due to the structure of the data with many 'out of area' pupils. Omission of certain cells with few level 1 units might aid this as with the experimental multiprocess models. However, this could not be tried with consultancy resources available. More detailed programming of the macros was also suggested as the outcome of this trial and may be possible in future.
- In the successful trial for technical reasons connected with their prevalence in certain units it proved difficult to fit the full set of predictors. This was mainly due to the constancy of their values in the limited data set used.
- The use of MCMC estimation of main instrumented equations in MLwiN, which might have overcome some of the above problems proved a non starter with the current macros. They were specifically designed for IGLS and it seems would have to be extensively re-written to handle MCMC.
- Since the joint use of the macros with WINBUGS is predicated upon successful operation of the MLwiN macros, it proved infeasible to attempt this strategy.

The results for the one trial referred to are given in Table 5.14 below along with a straight non IV estimation which did not take endogeneity into account. Where estimates are comparable results are broadly in agreement with estimates in other type of trials considered previously. We emphasise again that no standard error estimates are available for the IV estimation

Estimation using the IV macro in MLwiN for the West Midlands dataset	**Non IV estimation (ignores endogeneity of expenditure)**	**Instrumental Variable estimates**
Fixed effects	**(St error in parentheses)**	
	Estimate	***Estimate***
Intercept	3.0740 (0.4765)	1.752
Expenditure per pupil (pexaav)	0.0000387 (0.000048)	0.000394
Female	0.0469 (0.0069)	0.0724
Age (days from Sept 1st 1989)	0.0001878 (0.000030)	0.0001691
SEN Action/Action Plus	-0.2909 (0.0105)	-0.2706
SEN Statement	-0.3326 (0.0344)	-0.3499
Eligible for FSM	-0.1050 (0.0020)	-0.1077
Ethnicity (base, white)		
Asian, Indian	0.0967 (0.0198)	0.0803
Asian, Pakistani/Bangladeshi	0.0276 (0.0229)	0.0351
Asian, other	0.0364 (0.0423)	0.0223
Black	-0.0177 (0.0179)	-0.0163
Chinese	0.2957 (0.0559)	0.2705
Mixed Ethnicity	-0.0128 (0.0168)	-0.0217
First language not English	0.0396 (0.0185)	0.0703
Key stage 2 maths adjusted	-0.1270 (0.0234)	-0.1281
Key stage 2 maths adjusted squared	0.1663 (0.0028)	0.1663
School has sixth form	0.0044 (0.0231)	-0.0062
Grammar school	0.2454 (0.0545)	0.1934
Secondary modern school	0.0675 (0,0634)	0.0752
Roman Catholic	-0.0222 (0.0379)	0.0077
Per cent eligible for FSM in school	-0.0111 (0.0031)	-0.0138
Per cent eligible for FSM squared	0.0000835 (0.000050)	0.0000584
Per cent AEN in school	0.000953 (0.000938)	0.000602
Specialist school	0.0220 (0.0218)	0.0182
EIC or EAZ	0.0402 (0.0541)	0.0270
Leadership incentive grants	-0.0686 (0.0556)	-0.1174
Teachers' pay ratio (averaged)	0.3310 (0.4376)	0.5775
Urban local authority district	-0.0617 (0.0492)	-0.0574
Capacity Utilisation (averaged)	-0.0319 (0.0988)	0.1932
Proportion Unemployed	-0.1538 (0.1670)	-0.2247
Proportion Black Ethnicity	-0.1025 (0.0886)	-0.2180
Proportion Chinese Ethnicity	-0.2498 (0.2809)	-0.3298
Proportion Pakistani/Bangladeshi Ethnicity	-0.1851 (0.0373)	-02124
Proportion Indian Ethnicity	0.0883 (0.0446)	0.2086
Proportion Lone Parent Households	-0.1054 (0.0327)	-0.1582
Proportion NVQ Level 1 or less	-0.4398 (0.0368)	-0.3546
Random effects variances		
Ward	0.0038 (0.0026)	0.0040
School	0.0243 (0.0025)	0.0277
Pupil	0.2933 (0.0034)	0.2937

Table 5.14: Estimation results using the IV macro in MLwiN for the West Midlands dataset

The Instrumental Variable estimation considered in this section is nearest in idea to that used in 'Report 2: Levačić'. It may be considered as a possible approach for those familiar with this type of estimation. It is seen to be not very practical at the moment for broader studies. However, with more advanced programming of the macro, which as indicated may be required

it, could possibly be implemented for further study. This approach might appeal for those more comfortable with this type of approach to handling endogeneity.

6 Conclusions and Recommendations

6.1 General

The above discussion of our scoping exercise suggests some optimism about possibilities of undertaking a full scale exercise using cross classified models to incorporate structural area of residence effects into models of pupil attainment. However there are some limitations to these possibilities which would need to be initially addressed. Also certain questions, particular those concerning computational possibilities, could possibly only be answered by a fuller iterative model development approach to the exercise.

We are also cognisant of the broader research framework and the need for analytical models to be developed in response to sharper questions in policy contexts about the purpose of the approaches considered. Thus one caveat we might express is to what extent the identification of area differences might assist resource allocation issues when these are determined at LEA and school levels. Above all we suggest that any analytical exercise needs be very detailed and will require much more in the way of time and resources than might have originally been envisaged.

We discuss our conclusions and recommendations under a number of discrete headings, though they are inter-related. We will refer to aspects of this inter-relatedness as we outline the arguments.

6.2 Data

It will be evident from section 2 that the provided secondary dataset with which it is proposed to work may require considerable prior attention. We have outlined in some detail certain problems we uncovered and some ways of addressing them. There may also be difficulties in preparing data for use on other specialist software, which though not insurmountable, may require careful attention. In our scoping exercise we have concentrated our attention on the experimental subsets of data. The full STATA dataset must be subject to more detailed scrutiny. The preparation of a fully documented codebook of variables included and their source would be of beneficial. More generally preparation of databases and their detailed documentation must be viewed as a crucial part of any analytical process.

Although the dataset contains more of less the full set of available PLASC variables on children and schools, other potential covariates for inclusion in the achievement model are more limited. In particular the census information relates to output areas only and the variables available are restricted to those found useful in 'Report 2: Levačić'. We are aware that originally wider information on output areas was used in that report. There is no guarantee in advanced model development that the same variables might emerge as useful or indeed that the same set of fixed effect predictors will suffice. Thus access to broader sets of data on output and other areas might be beneficial.

We have considered a range of different areal units in our exercise and discuss how these enter into model frameworks below. One aim of model development might be to seek to explain how different variation in the hierarchies is explained by covariates at the appropriate levels. Thus there may be a need for data to be augmented by available measures at these levels and merged to the existing data. The rapidly developing neighbourhood statistics facility of the Office for National Statistics at *www /neighbourhood.statistics.gov.uk/* is an obvious source of potential information but there may be others that investigation may uncover.

In further discussion of model frameworks we have suggested that for many purposes additional area identifiers of one or more layers of Super Output units might be useful. In principle this could be merged into the dataset in similar ways as those we have used to merge the Local Education Authority of residence. Covariate information might also be available at this level from sources we have discussed.

6.3 Models And Structure

It will be evident from all of our discussion that our preferred way of conceptualising the situation is through multiprocess cross-classified multilevel models. The multilevel aspect of this was also considered briefly in 'Report 2: Levačić' for the education hierarchy. These types of models with more than one response in multilevel structures are also often called multilevel structural equation models. Here we have two such equations in the model, one for achievement and one for resources. We suggest that joint modelling of these be seen as the culmination of exploratory iterative model development and not as one-off exercises with pre-determined sets of effects.

For each equation there are two aspects to this which may need further consideration: which fixed predictors effects to try and which structural random effects.

We take the achievement equation first. In the experiments in this report we have gone with the fixed predictors used in 'Report 1: Vignoles' to trial the methodology. However, for reasons we have seen in commenting on the data there are others which may be fruitful and which could and should be tried. Which random effects to incorporate and reflecting different levels of the detailed structure discussed in Section 4 is a matter of trial and judgement in the process of model development. The less complex the structure to be fitted the less problematic are the statistical and computational criteria that we have considered. Balanced against this is the desire to get a full picture of areal variation at various levels. This in turn is conditioned by research questions to be addressed; caveats about which were raised above. There is, for instance, an open question of whether we need to consider the highest level of LEA variation for both the education and area hierarchies. The evidence on this is limited since the number of LEAs in our experimental data sets has been relatively few and results have inevitable low precision. However, we feel that if they are to be included, the equation should allow them to be correlated in developing the models. We feel that postcodes are of limited relevance except as a necessary identifier of other areal units. Our investigation of the data has also revealed some question of whether output areas are at all useful as a random effect given their large number and closeness to the pupil level units. Their inclusion also creates considerable computational burdens for many estimation methods. Again given their size we might question whether identification of census output area effects would be at all useful for such things as area initiatives in adding much to what we already know about the few children in each. Such initiatives may just as well be directed at individual children. However, super output areas and wards are perhaps more meaningful units to consider. They are also easier to manage analytically. The data would need to be augmented to handle them both in ways suggested in the previous section. Our general predisposition is to recommend that in model development super output areas, wards, and LEA of residence correlated with LEA of school might initially be used. This would ease any burden of computation, whilst at the same time being substantively relevant. At a later stage other effects in the area hierarchy might be explored if they were thought to be relevant for certain questions.

In some ways it is easier to make suggestions about modelling the random effects in the resource response equation. The variables in the data to be used are observed at the school level

and although applied to each pupil in the data is in fact constant for pupils in the same school. Thus, apart from school disturbances, this equation will have a simpler random effects structure including only an LEA effect if that is incorporated. Which covariates (or instruments) to include is somewhat more of an issue and can only be fully answered in the process of model development. The predictors used in 'Report 2: Levačić' were conditioned and limited to some extent by the requirements of the instrumental variable estimation used and the focus was really on improved estimation of the achievement equation. In our experimentation we have also only considered this set of instrumental predictors. However, in a simultaneous equation framework we might get a better view of the interplay of resources and achievement with a more detailed specification of predictors in the resource equation. Since the multiprocess characterisation recognises the linking of the two there is nothing to constrain resource equation specification except technicalities of whether their coefficients can be identified. It may also include many of the same variables used in the achievement equation. This is a matter of further necessary econometric investigation which could be quite detailed and somewhat outside our remit. In principle also achievement could be included in the resource equation. This might recognise the mutual dependence of resources and achievement raised as an issue in 'Report 2: Levačić'. We would recommend that such matters are fully investigated in future research.

6.4 Estimation Methodology

We have considered extensions to IV estimation in a multilevel setting in our experimentation to parallel the IV estimation in 'Report 2: Levačić' using robust standard error estimation. There are some statistical inefficiency issues arising from such estimation frameworks which in general lead us to prefer other alternatives. Apart from this our IV estimation macro would require considerable programme development were it to be considered. We leave this possibility open for further consideration if it were thought that such an approach had a more familiar appeal.

Our preference is for some way of estimating fully specified multiprocess cross classified model using the available multilevel estimation methodology. The experimentation has had some success using the RG method based on the maximum likelihood IGLS approach. However, we have noted computational limitations using this method for more complex structures and also for larger datasets. It is recognised that MCMC estimation is computationally more efficiently for cross-classified models though paradoxically not so for strictly hierarchical random effects. We have had some success in using the MCMC procedures in MLwiN which shows how they can handle quite complex random effects structures quite reasonably. Philosophic debates about the appropriateness of Bayesian procedures apart, there are also certain statistical attractions of MCMC. We get fuller information on uncertainty about parameter estimates. However, we have not as yet been able to incorporate MCMC into the multiprocess macros due to current limitations we outline below. Nonetheless, subject to these limitations being resolved we recommend MCMC estimation within the framework of MLwiN be the preferred estimation approach to model development.

6.5 Software

We have recognised that there is now a plethora of software that can now handle multilevel structures. Rather fewer are adept at cross-classified structures. The software with which we have experimented has been used to explore more fully certain model features from a statistical and computational view to enable us to judge certain feasibilities. These software items are ones with which we are familiar and are most likely to be familiar to the UK research community. This is not to suggest that other ranges of software are not capable of handling some of the

issues. We discuss this in 'Report 3: CC Review'. The programme MPLUS (Muthen and Muthen (1998)) specifically designed for multilevel simultaneous equation structures is one which might be worthy of further investigation. However, we are unable to scope all software and we believe that which we have used is likely to be most appropriate for the issues under investigation. Our preference would be to recommend handling future research using MLwiN. This software has a number of advantages apart from its facility in handling the type of model and estimation envisaged. Its windows and graphical interface enable the easy study of results and diagnostics making model investigation and development clearer and more organised. Its macro facilities also make its use adaptable to a variety of features which may not be routine in most software. In full scale model development we also envisage a greater role for random coefficients in both achievement and resource models. This would be particularly fruitful for examining differential effects for distinct subgroups of pupils, schools and/or areas. MLwiN is particularly useful for such modelling. Some limitations of MLwiN for the current exercise have, however, been uncovered, but are capable of resolution as discussed below. We can also see a place for REML estimation in GENSTAT at a late stage in any research when it is desired to fit just a particular pre-specified model quickly and efficiently. It is not software, however, with which we have a great deal of familiarity. It is also subject to some limitations as we will outline.

6.6 Software Limitations

MLwiN currently has two limitations which have informed our scoping, but which we believe can be addressed:

(i) To handle the computational constraints which our experimentation with the RG methods has shown, MCMC estimation using a multiprocess framework has been suggested. However, this estimation method in MLwiN cannot at present handle multivariate responses observed at different levels in the structure which we require.

(ii) The MCMC algorithm in MLwiN cannot at present handle correlated random effects when two classifications in the structure are the same sets of units. Thus if we desire to incorporate the LEA effects in our model development we are presently constrained. This was a main reason for experimenting with the less user friendly and slower facilities of WINBUGS.

The MLwiN programme is under further continuous development with ESRC funding and these two issues are on the agenda. How quickly they will be resolved is subject to prioritisation in this development. If future research is to follow up this scoping we recommend that part of the resources be apportioned to this development and the extension of the multiprocess macros. We make this recommendation for obvious reasons.

GENSTAT currently has limitations for simultaneous response but our consultants indicate that it may be possible to adapt it as a further aspect of programme development to which resources may be devoted. We spell out below a possible role for GENSTAT.

6.7 Full Dataset Versus Sampling

Using the full data set, even when cleaned and augmented in the ways suggested above, might present some challenges to analysis. It was used successfully in 'Report 2: Levačić'. However, in the context of more complex structures it is not so much the number of level 1 records that stretch computing feasibility as the sheer number of complex random effects generated by the full data for any suggested model. This would still be the case even if the levels are constrained

somewhat. We believe MCMC estimation could handle the full data set feasibly but the main problem would be the sheer computation time involved. This would make the process of model development quite a cumbersome business.

However the whole of statistical experience suggests that with careful attention to sample design there is no necessity to contemplate full census operations. The only difference here is that we have the full data available but would like to economise on our stretched analytical resources whilst at the same time achieving parameter estimate results which would be in broad agreement with what might emerge from a full data analysis. Thus one further recommendation we might make is that the dataset can itself be sampled according to scientific principles. One broad suggestion might be three stage epsem design of children within schools within LEAs. Stratification factors might also be introduced at each stage. We make no specific recommendation about a sampling plan but a suitable design could emerge from further investigation. The sizes of the subsets of data with which we have successfully resolved computing constraints are indicative of the sorts of final sample sizes we might expect to be more manageable. Since data gathering is not an issue here there could also be an element of cross-validation by drawing several samples using the same design. Some recent work by the PLASC user group is also considering the practical use of sampling the data. An attractive feature is that even if attention is restricted to a sample, the full data set may be drawn on for ancillary information such as that of all schools attended by sampled pupils apart from their current ones..

Although we have not explicitly discussed it in our scoping, one output from analysis to which attention might be drawn is the estimates of the residual effects at various levels in the twin hierarchies. Educational researchers are of course used to interpreting school effects in 'value added' terms and much the same interpretations might be accorded to different area effects. We might like to have these available for schools and areas in the full national dataset but which samples from the data might not fully provide. This is where we see a potential role for a suitably redeveloped GENSTAT approach where we might consider final model specifications being run quickly and efficiently as one-offs on the full set of data. However, we are unsure as yet of the capabilities of GENSTAT for handling random coefficients which may be required.

6.8 Handling and Imputation Of Missing Data

In section 2 and elsewhere we commented on handling missing data in the provided data and the potential impact of its extent. We also believe that the data set provided to us may have also had a prior deletion of many cases from the full PLASC database because of unavailable information on key variables used. In our trials we have had to perforce listwise delete of many cases when sometimes maybe only one variable used in a model was missing on a case. Such deletion of cases to handle missing information is possibly acceptable if we could regard the data as missing at random. However, there are serious caveats about whether this is so. In particular whole schools are often deleted since they have missing information on many key variables including the responses. It is becoming recognised that a cavalier treatment of missing data by deleting offending cases may lead to distortions and bias. This is particular important when reasons for missingness are often connected to the analytical purposes of the model. For instance, it may be that whole school cases missing may be in those schools with entirely different patterns of relationship of achievement to resources. Method for handling missing data in statistical modelling have been available for some time for single level models. Imputation of values using information in the data set is one approach A recent ESRC project has extended some useable imputation methods for implementation for multilevel models and in particular for use with MLwiN. The work undertaken by James Carpenter (London School of Hygiene and

Tropical Medicine) and directed by Professors Kenward, Goldstein and Molenbergs has led to the development of a web site *www.missingdata.org.com.* This amongst much else contains information on the methodology and implementable software macros.We believe that these ways of handling missing data alongside model development will vastly improve the statistical and substantive quality of any future work.

The conclusions and recommendations we make above imply a fairly large full scale exercise if the issues are all to be satisfactorily addressed.

References

Browne, W.J. (2003). *MCMC Estimation in MLwiN (Version 2.0).* Institute of Education, University of London.

Browne, W.J. (2004). An illustration of the use of reparameterisation methods for improving MCMC efficiency in crossed random effect models. *Multilevel Modelling Newsletter*, **16(1)**:13-25.

Clayton, D. and Rasbash, J. (1999). Estimation in large crossed random effect models by data augmentation. *Journal of the Royal Statistical Society, Series A,* **162**: 425-36.

Fielding, A. (2004). The role of the Hausmann test and whether higher level effects should be treated as fixed or random. *Multilevel Modelling Newsletter,* **16**, 2, 3-9. Centre for Multilevel Modelling, Institute of Education, University of London.

Fielding, A. (2005). *Cross-classified and multiple membership structures in multilevel models: an introduction and review,* Report to the Department for Education and Skills, London, October 2005 (‘Report 3: CC Review’).

Gibbons, S. (2002). *Neighbourhood effects on educational achievement: Evidence from the Census and National Child Development Study*. Discusson Paper Series 018, Centre for Economics of Education, London School of Economics.

Goldstein, H. (2003). *Multilevel Statistical Models, 3rd Edition*.London: Arnold.

Hanushek, E. A. (1997) .Assessing the effects of school resources on student performance: an update'. *Education Evaluation and Policy Analysis,* **19, (2**) 141-164.

Levačić, R., Jenkins, A., Vignoles, A., Steele. F. and Allen, R. (2005). *Estimating the relationship between school resources and pupil attainment at Key Stage 3. Research Report 679*. London: Department for Education and Skills **(**‘Report 2: Levačić’).

Jenkins, A., Levačić, R. and Vignoles, A. (2006). *Estimating the relationship between school resources and pupil attainment at GCSE. Stage 3, Research Report 727.* London: Department for Education and Skills.

Leyland A and Naess. O. (2004). *Using correlated cross-classified multilevel models to estimate area influences on health throughout the life course*. Paper presented to Royal Statistical Meeting on Multilevel Modelling: Recent Advances, October 2004.

Mayston, D. (2002) *Tackling the endogeneity problem when estimating the relationship between school spending and student outcomes, Research Report 328.* London: Department for Education and Skills.

Muthen, L.K. and Muthen, B. O. (1998). *Mplus User’s Guide*. Los Angeles, California: Muthen and Muthen.

Rasbash, J. and Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross classified random structures using a multilevel model. *Journal of Educational and Behavioral Statistics,* **19**, 337-350.

Rasbash J. and Browne W. J. (2004). Non-Hierarchical Multilevel Models. To appear in De Leeuw, J. and Kreft, I.G.G. (Eds.), *Handbook of Quantitative Multilevel Analysis* (Paper 19 on http://www.maths.nott.ac.uk/personal/pmzwjb/bill.html).

Spencer, N. H. and Fielding, A. (2000). An instrumental variable consistent estimation procedure to overcome the problem of endogenous variables in multilevel models. *Multilevel Modelling Newsletter*, **12, 1**, 4-7.

Spencer, N .H. and Fielding, A. (2002). A comparison of modelling strategies for value-added analyses of educational data. *Computational Statistics*, **17, 1**, 103-116.

Spiegelhalter, D., Best, N.G, Carlin, B.P. and van der Linde, A. (2002)). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B,*. **64**:, 683-640.

Steele, F. , Vignoles, A. and Jenkins, A. (2006). The impact of school resources on attainment: a multilevel simulataneous equation modelling approach. *Journal of the Royal Statistical Society, Series A* (forthcoming).

Vignoles, A., Levačić, R., Walker, J., Machin, S. and Reynolds, D. (2000). *The relationship between resource allocation and pupil attainment: a review, Research Report RR228.* London: Department for Education and Skills **(**'Report 1: Vignoles')**.**

Appendix 1:
Detail on the appropriateness of the database for secondary analysis

A1.1 Problems with case and level identifiers

(i) Schools:
The summative introduction to 'Report 2: Levačić' talks of 'over 3000' secondary schools and descriptive statistics reported in Table 6 of that report seem to imply that there are 3011 establishments in the dataset. However, the unique identifier in the available dataset for such establishments, *leaestab (labelled school id)*, has 3082 values, excluding a missing value for school for 17 students coded rather curiously at 0. Another variable in the dataset *k3_estab* appeared to be a code for establishments within LEAs. For this a particular code may apply to several schools across different LEAs. The code 6900 for example was applied to schools in 12 different LEAs. In principle this poses no difficulty if it is known and can be taken in conjunction with an LEA identifier. However, this was not transparent either from the report or the database itself and was only discovered by noting from tabulation that certain code values seemed to have abnormally large frequencies. If the combination with LEA information is undertaken, taking missing values on both variables into account, another derived school identifier was created by us which also has establishment values. It was then discovered, though this was not immediately transparent from documentation that such a variable, *k3-schid (and also labelled school id*) which agreed with ours had also been created in the dataset with a similar 3082 non-missing distinct codes. For this variable 16 missing code values of 0 were applied. For the most part values of *leaestab* and *k3-schid* agreed except for 9 cases where a missing value zero was applied to just one of them but not the other. Further investigation revealed that the use of missing value codes was inconsistent. Only 12 cases had common missing value codes of zero. For *leaestab, 5* missing values were given, and for *k3_schid* 4 cases, where the value of the other variable was non-missing. The reasons for this discrepancy are not apparent. The mismatch between 3082 values of the dataset and 3011 of the report may have a simple explanation but this is neither transparent nor immediately explicable to us.

Another school identifier in the dataset is *schname,* the name of the school, which may be required for some purposes although it is strictly not necessary for modelling work. There were 6841 missing values as spaces (including the 17 for which most information was missing) and these were all a subset of those cases with missing local education authority name, *leaname (*see below), and presumably left out for the similar inscrutable reasons which will be outlined. Amongst the non missing cases, 2944 distinct school names were identified. The missing cases corresponded to 69 distinct codes for *leaestab*. Allowing for some small overlap of these two groups the sum of 3013 is suspiciously similar to the 3011 schools quoted in 'Report 2: Levačić'. In any event it is different from the 3082 distinct school code numbers and there is a possibility of name duplication across authorities. These anomalies would need to be subject to detailed investigation in any full analysis of the data set. We have restricted our attention to ironing out problems for the subsets of the data with which we later experiment.

(ii) ***Local Education Authorities***:
Similar problems as with schools surround the Local Education Authority identifier. Two variables are available in the dataset pertinent to this *k3_lea* and *leaname*. *k3_lea* is apparently the DfES three digit code of LEAs. Again 17 cases were recorded as 0 (missing) and 149 distinct value codes. Comparison with DfES sources shows the absence of the code 201, City of London. Buried at the end of Appendix 1 of 'Report 2: Levačić' is a statement 'As there were only 2-3 schools in each year in this category (the LEAs with other political control were Isles of Scilly, Rutland, City of London) these LEAs have been omitted' Apart from the inscrutability of this statement which is possibly the reason for exclusion of the City of London, it is inconsistent with the inclusion of cases for Isles of Scilly and Rutland.

A tabulation of *leaname* showed 7663 cases with missing values left as *empty spaces* and 146 other distinct values. Further comparisons with *k3_lea* revealed that 374 and 22 missing values were respectively the whole subsets of cases for Rutland and Isles of Scilly respectively. Apart from 17 common missing values, the other 2250 missing names were distributed amongst 50 other LEAs which had been coded in *k3_lea*. We are not sure why these values were recorded as missing but the hint for Rutland and Scilly implies that they might be cases which were excluded from some analyses. However, for reasons explained above these reasons are not entirely transparent. For identification and later interpretation it might be advisable to have these names recorded and in updating the database we have edited them in. Apart from Isles of Scilly and Rutland, the other explanation for the number of values being 3 fewer in *leaname* was the labelling of both Hull and Kingston (Surrey) as 'Kingston Upon'. An obscure filter for this discrepancy was later discovered at the end of the data. However in an updated edit we have labelled them distinctly.

(iii) ***Student***:
The case (level 1) identifier *k3_pmr*, pupil matching reference number, appears to be satisfactorily unique with 464766 distinct codes plus the usual 17 missing spaces. However, its 18 digit alphanumeric code is cumbersome for transfer to other specialist programmes. A STATA encoding command to transfer the code to sequential integers proved infeasible due to large number of unique values so EXCEL was used as an intermediary to generate this unique code *af_case_id* in an edited STATA file. SPSS was also used for some data manipulation facilities since it has more flexible features for some purposes. Transfer of files between STATA and SPSS is relatively easy using the STATtransfer programme.

A1.2 Some other missing value issues
(i) Some difficulty was created for us initially in the conventions for missing values since nowhere are these explicitly identified. Tabulation of specific variables one by one was required to identify the nature of how they were coded. The conventions were inconsistent. Variously used were spaces, the STATA system missing value '.' and sometimes rather curiously the digit 0. The latter may not be instantly recognised as missing particularly where it is used in scaled variables such as *dobdays* (date of birth with Sept 1 1988=1) and all the KS2 and KS3 scores. The latter issue is further confounded by the statement in 'Report 2: Levačić' Appendix II that 0 also means students disapplied from the test or working towards the level. It is not clear from the commentary whether such cases had been included in the dataset, though we presume not. None of these variable values are identified as missing by the STATA data description facilities and could easily be overlooked without care and deep knowledge of this particular file construction. In STATA itself, missing values for categorised variables are not always recognised as missing values and assigned a separate category. This was one of the problems in transferring the data files to system files for specialist software which until we identified the

issue proved somewhat troublesome. We resolved this matter for the most part by recoding all missing values, once they were recognised, to something obvious like -999. These could then be reconverted to recognisable system missing values on input into other software. Many of these issues may not have proved so troublesome had an explicit codebook been available. We have not edited the full data set in these ways, being content to do it for our experimental subsets. We have also attended to the issue only for those variables and identifiers used in the trial analyses. A good deal of further cleaning and editing work on this matter may be required if the full data set is to be exploited further by other users.

(ii) The data set had 17 cases on file for which most identifying information including a pupil reference number was missing. For many specific variables used in our trial analyses the code 0 was entered unanimously into the data file. Due to unanimity (e.g. all such cases for gender were coded 0) we believe these may be missing values. Certainly no other missing value codes could be identified. The problem here is that in many cases 0 is also a code used for a specific category of a variable (e.g. for *gender*, male =0, female =1). We speculate that there may be many other cases for which information on the gender variable may be missing and yet coded as male. Without reference back to the original PLASC database it is difficult to say. Yet it is known from educational considerations that missing information often relates to children with specific characteristics which may distort analyses if there are quite a number of them and their presence not accounted for.

Similar problems arise for some other variables we have identified and wanted to use in our trial analyses such as the eligibility for free school meals indicator (*fsmpup)*, and the special educational needs variable *sen* (none=0, school action or action plus=1, statement =2), where again we believe 0 has also been used as missing. Of course the likely impact of such problems in analysis will also depend to some degree on the extent of missingness. Without further information this is difficult to assess. However, some other variables, such as First Language not English indicator (*engornot*), are quite satisfactory, as they use the explicit STATA system missing value '.'.

(iii) A related problem occurs for many variables which are dummies created from a compound categorised variable with several categories and where missing value information is not carried through to the dummies. In some cases such as *sen* this is due to possible missingness being confounded with the category coded zero as indicated above. In other cases where an explicit missing value is recognised in the original categorised variable this is not recognised in certain indicators. For instance, the dummies *ethasianpakbang* (Asian, Pakistani/Bangladeshi), ethasianind (Asian Indian), and *ethasianoth* (Asian, other) are formed from the categories of the ethnic group variable *eth_03* which contains a 21,592 case unclassified category. For some purposes these are treated as missing values but this missingness is not recognised in the indicators. Curiously it is recognised for other ethnic indicators such as *ethblack* (Black), *ethchine* (Chinese), *ethmixed* (Mixed ethnicity). This inconsistency is a source of some confusion. The point is not trivial either. It is possible that inability to classify may be more prevalent in Asian groups and since the misclassified is a relatively substantial 5% of all cases, distortions in analyses may occur. The important *ethwhite* (White) does recognise unclassified as missing but again has an extended definition to include white British, Irish, traveller of Irish origin, gypsy romas, other whites. In some analyses White is declared as a base category but in fact the base also includes the 'any other ethnic group' category. This designation may be due to the use of 0 as a code for this combined category in an alternative broader grouped ethnicity variable *ethnic_5* (major ethnic group: not white) Fortunately the 'any other ethnic' group is not large but if not recognised may lead to some misinterpretations.

(iv) One indicator used in 'Report 2: Levačić' and our trial analyses (although generally statistically insignificant) was *urban_lad* (urban/rural local authority district). This indicator was declared to have STATA system missing values for 6824 pupils in 71 schools. Examination revealed that the many of these were in local authority districts which had been assigned a value for *urban_lad* for other cases in the set. To take just two examples, out of 1867 cases in district 00CH there are 145 declared missing and out of 584 cases in 43UG, 13 are missing. Thus their declaration is rather curious unless it is due to the fact of these cases being excluded from most analyses for other undeclared reasons. A clue that this may be the reason is that the same set of cases are declared missing on *schage18* (school has pupils aged 18 or 19), taken as proxy for presence of a sixth form. Such information appeared unavailable for 71 schools in the dataset and presumably pupils and schools were dropped for analyses involving *schage18*. However, these matters are not transparent and initially led us to some perplexity in understanding until investigations were undertaken. The variable *urban_lad* is also an indication of small misclassification error in the data set. To take the two examples above, 6 of the cases in OOCH are misclassified as rural, as are 21 cases in 43UG. Such small measurement or coding errors are probably inevitable in most large data sets and this is a recognised methodological problem which may not be serious unless the impact is large. However, in this case there is a consistency check available in the data set and any thorough cleaning of the data should handle it. It may also arise for a number of other variables in the data which we have not investigated.

(v) We also find it rather curious that many schools have missing information on key variables used in analyses. The same set of 71 schools and 6824 cases are also missing information on type of governance, school gender, numbers of schools within a certain distance, and all of the wide range of special designations such as educational action zones, special classes, special measures and beacon school. This is also the case for statutory lowest age of pupils and school denomination but additionally 4 other schools and 529 cases also have missing values. A different additional set of 16 schools and 2114 cases are missing data on type of school indicators (grammar, secondary modern or comprehensive). The time series measures of school GCSE performance indicators, school size, aggregates such as percentage of pupils eligible for free school meals, staffing and expenditure are also missing on most of these schools and quite a few others. These cases have possibly been dropped from original analyses and presumably are due to deficiencies in central schools database. But this gives us some concern since there is no indication of why this should be so.

It is likely that the schools with missing information are unique and special and for reasons directly connected with the objects of the analysis and so cannot be treated as missing at random. If this is so the possibility of distortion in analyses is evident although its likely impact is difficult to assess. However, we feel that some method of imputation might ultimately be beneficial and we return to this issue later.

(vi) We have mainly concentrated on cleaning variables likely to be used in our analyses and for our data subsets both by direct checking and intuition. However, although unable to do a complete investigation we have spotted in passing a number of other anomalies. We mention three. First the variable *a5_ag_02*, the 2002 school aggregate for 5+ grade A*-G at GCSE, has been designated a percentage measure in labelling and possibly analyses; but it is in fact a proportion in the data. There may be others. The effects of such variables in analysis may be misinterpreted unless ironed out. Second, there are also some seemingly absurd outlier values in one variable we have noted. The ratio of fte pupils to fte unqualified teachers 02/03 in one school is recorded as 39,907. Indeed 24 schools covering less than 1% of cases have figures for this ratio of over 10,000. Although there are no data in the file on absolute numbers of fte

unqualified with which to compare, a clue may be found in the data for *uqtr0203*, the ratio of unqualified to qualified teachers. For example, this is 0.00061, or 6 in 10,000, in the extreme school. The only explanation for this apart from data errors is the presence of unqualified teachers working a tiny proportion of the teaching week. The only check on this would be with the original schools database but we flag such matters as of some concern. The presence of such extreme outliers leads to highly skewed distributions. These may lead to distortions particularly if used in linear models unless recognised and handled. Third there is also another problem with the use of such ratios where the denominator can be zero as it often is in this data set. The result of trying to divide by zero in routine programmed derived variable calculations usually leads to a system missing value result. Thus over 12% of schools and over 11% of cases have missing values on *uqtr0203*. Apart from some genuine missing values most of these are schools with zero unqualified staff. We are not aware that this particular variable or similar has been used in any published analyses but this is worth a flag if any such analyses were contemplated. Zero unqualified staff is after all a real proxy indicator of some concept of maximum quality and may be important. The usual way of handling this is to take the reciprocal or some other transformation which avoids the mathematically impossible division by zero.

(vii) A final anomaly with the data set is that values for string variables have occasionally been entered misspelled or with excess leading spaces so that the same value is often recognised as distinct values. With careful attention tabulation of the data variable by variable can uncover such difficulties and in previous analyses this has probably been the case. This problem also makes matching and transfer of data to other software more difficult.

A1.3 Issues connected to area units and their identifiers

- **Postcode**:
  The variable *pl_post* is the familiar six or seven digit alphanumeric code and is given for all but 15 of the usual cases. Some specialist software has difficulty with the input of such codes so we have converted to sequential numeric code by switching between programmes and using devices similar to the above treatment of student identifiers. There are 321583 distinct postcodes in the data set yielding an average of only 1.45 KS3 children per postcode.

- **Output area:** This variable, *oacode,* is a ten character alphanumeric code which for similar reasons to the above has been converted similarly to sequential numbered identification. Apart from the 14 of the usual missing cases, there are 142598 unique output areas in the data set, an average of 3.26 KS3 children per output area. For some purposes it may be possible to separate out such small area higher level effects but the sparsity of the clustering and the necessity for large numbers of output area random effects may render the exercise somewhat infeasible for the full data set.

- **Ward**: This is a six figure alphanumeric coded with 14 missing values and 7963 distinct codes, an average of 58.4 KS3 children per ward. Conversion to numeric codes in the edited data set has also been undertaken. Ward random effects in a multilevel model and a crossed model due to its larger size are possibly computationally easier to handle. Ward is also represented by its name, the variable *ward_desc* with 15 missing values and 7213 different values. The discrepancy from the number of ward codes suggests a duplication of names across different districts. If ward name were to be used as a meaningful identifier some way needs to be found to edit and re-label duplicates in

further data cleaning. We have concentrated for now on resolving anomalies for our experimental subsets used later.

- **Local authority districts**: This is the highest level of area of residence present in the data set and labelled *lad_ua* (local authority district unitary authority).As indicated above this is leading four alphanumeric characters of ward and output area codes. There are 14 missing values and 360 unique codes. Some modelling may trial such higher level authority residential area if it seems there is a possibility of an effect at this level. The local authority district is also represented by the name variable *lad_uade.* This has 73 missing values and now 356 unique names. The discrepancy of 4 distinct values has not been investigated since it would require careful matching of names to codes. Also it may be due to the larger number of missing values. This discrepancy would need to be resolved if names were to be used in any fuller set of analyses.

Appendix 2:
Further examples of structural features of the data

A.2.1 Birmingham

Table A1: Local authority district of residence of 9763 KS3 children in Birmingham schools

Birmingham		9206
Sandwell		216
Solihull		101
Walsall		76
Bromsgrove		66
Dudley		29
Lichfield		26
North Warwickshire		16
Tamworth		6
Redditch		4
Wolverhampton		3
Cannock Chase		2
Derby		2
Worcester		2
Basingstoke and Deane		1
Bexley		1
Hackney		1
Medway		1
Northampton		1
Swale		1
West Dorset		1
Wyre Forest		1
Total		**9763**

Table A.1 presents the local authority district of the 9763 children attending Birmingham achools. Of these 5.7% lived outside Birmingham. Most of theme are contained in areas adjacent to Birmingham and their wards and output area will also be represented by children in larger numbers in the West Midlands data. However, in forming cross-classes with Birmingham schools these will form sparse cells and increase the number of discrete non-overlapping blocks of schools in this data set. The numbers of random effects for areas will be increased by the areas out of this data set often with only one pupil. These will affect the computation time in many procedures and also affect precision of estimates. Some of the areas in the table also seem on the surface to be absurd, e.g. Bexley or West Dorset. This may be due to inaccuracies in the PLASC database, or some other inscrutable reason. This investigation was undertaken before LEA of residence was merged with the dataset. Hence local authority districts were used for this exercise.

A2.2: Some features of the West Midlands data set

We also investigated the 'out of area' issue for the West Midlands data-set as a whole where we now regard this in terms of areas of residence of pupils in this dataset whose area of residence was outside the total boundary of the LEAs included. Out of 32579 pupils there were only 625 (1.9%) such pupils and 588 were in nearby districts in the broader area just outside the West Midlands LEA boundaries. Thus for experimentation with the West Midlands dataset 'floating areas' are not likely to present as severe a problem as it might be for the Combined Selection set. We conducted similar exercises for the whole West Midlands dataset as for Cambridgeshire in the main body of the report and Oldham below. Details are too elaborate to present here but are held on file. We satisfied ourselves that the West Midlands dataset might be very suitable for

experimentation, though we anticipated with large number of output area effects some computational difficulties for some methods. This turned out to be the case to some extent.

A2.3: Crossing of areas by school for Oldham LEA: Summary Statistics

Oldham had 2545 KS3 pupils in 15 secondary schools. There were 304 (11.9%) 'out of area' pupils who came from 211 different output areas , within 44 wards in 8 different local authority districts. These would potentially create similar problems in terms of number of sparse cells, blocks etc as in A1.. It is of interest to know that 138 of the 304 went to just two schools , Crompton House and Bluecoat. Nearly 50% of Crompton House pupils were out of area and over 35% of Bluecoats. It is just for such schools that the area effects are likely to be important and why the out of area pupils are often a different minority of pupils. They might considered as an important influence and cannot often justifiably be dropped to make analytical computation of model estimates easier.

For the 2241 pupils who went to school and also lived in Oldham Table A2 gives for each school the numbers of children and some indication of the spread of their pupils area of residence over the 643 output areas and 20 wards.

Table A2: Summary statistics on area structure for Oldham schools and for pupils who also lived in Oldham

School	No of pupils	No of output areas	Average no of pupils per output area	No of Wards	Average number of pupils per ward
1	190	91	2.1	9	21.1
2	102	33	3.1	7	14.6
3	138	71	1.9	10	13.8
4	116	38	3.1	6	19.3
5	81	50	1,6	12	6.8
6	100	72	1.4	18	5.5
7	194	96	2.0	9	21.6
8	175	87	2.0	11	15.9
9	232	114	2.0	7	33.1
10	231	99	2.3	10	23,1
11	186	120	1.6	18	10.3
12	107	85	1.2	17	6.3
13	106	67	1.6	12	8.8
14	121	95	1.3	16	7.6
15	162	131	1.2	17	9.5
Total	**2241**				

There was an average of 3.5 pupils per output area and 102 per ward. Inevitably there was a thin spreading of each output area pupils amongst schools. Amongst the output areas 251 (37%) sent children to just one school. Together with the data on schools in Table A2 we initially thought that the issues of confounding of schools and output areas and their separate identification, when repeated elsewhere, relatively unproblematic. Of course as elsewhere, large numbers of output area random effects and the low precision with which such specific effects might be estimated is likely to raise further statistical and computational issues. Wards as a random effect in a cross-classification would appear to pose no real problems. Each school is represented by a fair

number of wards as exemplified in table A2.. Also the minimum number of schools attended by pupils for particular wards was 4. Of course the structure is quite radically unbalanced for both output area and wards as evidenced in the sparsity of Tables A3 and A4 below. However, certainly for wards, even though they are larger aggregates, their diffusion over schools is such as to potentially make their effects sufficiently easily separable from schools and identifiable.

Table A3: Sparsity: Frequencies of pupils over the 12810 school by output area cells for Oldham schools and for pupils who also lived in Oldham

Number of pupils in cell	*0	1	2	3	4	5	6	7	8	9	10	11
Frequency	8393	749	260	127	56	29	12	12	2	1	3	1

* 66% of cells are empty

Table A4: Sparsity: Frequencies of pupils over the 300 school by ward area cells for Oldham schools and for pupils who also lived in Oldham

Number of pupils in cell	*0	1	2	3	4	5	6	7	8	9	10	11	>11
Frequency	121	44	25	16	11	9	6	2	5	3	6	3	49

* 40% of cells are empty

Appendix 3:
MLwiN Macro for multiprocess cross-classified model on Combined Selection data set

MLwiN Macro for multiprocess cross-classified model on Combined Selection data set

```
NOTE: Maths expenditure analysis, treating expenditure (PEXAAV) as endogenous
NOTE: Combined selection dataset
NOTE: 2-level cross-classified model (pupils within cross of school and ward)
NOTE: stlow_12 removed as always equals zero in this subsample

NOTE: retr c:\fiona\consultancy\KS3 resource project\Tony scoping study\combined selection fiona.ws

erase c69-c295

missing -100

NOTE: omit school with missing ID
omit missing 'af_cs_schname_code' c1-c63 c65-c68 'af_cs_schname_code' c1-c63 c65-c68

note: remove missing values on response and explanatory variables
list missing c1 c2 c3 c6 c9-c68 c1 c2 c3 c6 c9-c68

coun c1 b1
put b1 1 c69
name c69 'cons'

NOTE: sort by ward ID and code wards consecutively from 1
sort 'af_ward_csid' c1-c3 c6 c9-c66 c68 'af_ward_csid' c1-c3 c6 c9-c66 c68
name c70 'new_ward_csid'
mlre 'cons' 'af_ward_csid' 'new_ward_csid'

NOTE: sort by school ID, then ward
sort 2 'af_cs_schname_code' c70 c1-c3 c6 c9-c63 c65-c69 'af_cs_schname_code' c70 c1-c3 c6 c9-c63 c65-c69

NOTE: search for small cells in cross-class of school and ward, and omit from dataset
xomit 9 c64 c70 c1-c63 c65-c69 c64 c70 c1-c63 c65-c69
xsearch c64 c70 c90 c91

note: excluding cells with <=5 leads to groups with max of 421 wards within groups
NOTE: excluding cells with <=9 leads to 82 groups with max of 117 wards within groups
NOTE: excluding cells with <=20 leads to 318 groups with max of 27 wards within groups

note: recreate pupil ID, starting from 1 in each school
mlre 'af_cs_schname_code' 'af_csidno' c71
name c71 'pupil2'

NOTE: Standardise PEXAAV, DOBDAYS, PELFSMAV (and square to get PELFSMAVSQ), PCAENAV
aver 'pexaav' b1 b2 b3
calc 'pexaav'=('pexaav'-b2)/b3
aver 'dobdays' b1 b2 b3
calc 'dobdays'=('dobdays'-b2)/b3
```

```
aver 'pelfsmav' b1 b2 b3
calc 'pelfsmav'=('pelfsmav'-b2)/b3
calc 'pelfsmavsq'='pelfsmav'*'pelfsmav'
aver 'pcaenav' b1 b2 b3
calc 'pcaenav'=('pcaenav'-b2)/b3

NOTE: sort by group, then school
sort 2 c90 c64 c1-c3 c6 c9-c63 c65-c71 c91 c90 c64 c1-c3 c6 c9-c63 c65-c71 c91

NOTE: Create data for MATHS EXPENDITURE analysis
vect 2 'k3matscr' 'pexaav' c72 c73
name c72 'resp' c73 'index'

repe 2 'pupil2' c74
repe 2 'af_cs_schname_code' c75
repe 2 'new_ward_csid' c76
repe 2 'af_csidno' c77
name c74 'pupil2_long' c75 'school_long' c76 'ward_long' c77 'csidno_long'

NOTE: covariates
repe 2 'pexaav' c78
repe 2 'k2matadj' c79
repe 2 'k2matadjsq' c80

NOTE: instruments
repe 2 'yrcon02' c81
repe 2 'yrlib02' c82
repe 2 'yrnoc02' c83
repe 2 'scpaav' c84
repe 2 'ftepup99' c85

NOTE: cross-classification variables
repe 2 c90 c90
repe 2 c91 c91

calc c86=('index'==1)
name c86 'c1'
calc c87=('index'==2)*('pupil2_long'==1)
name c87 'c2'

calc c88='c1'+'c2'
NOTE: As PEXAAV is at school level, keep only 1 response per school
NOTE: Here, select record corresponding to 1st pupil in each school
omit 0 c88 c72-c87 c90 c91 c88 c72-c87 c90 c91

NOTE: Create explanatory variables for each response, including only school-level variables for
PEXAAV
calc c200='c1'*c78
calc c201='c1'*c79
calc c202='c1'*c80
name c200 'c1_pexaav'
```

```
name c201 'c1_k2matadj'
name c202 'c1_k2matadjsq'

calc c203='c2'*c81
calc c204='c2'*c82
calc c205='c2'*c83
calc c206='c2'*c84
calc c207='c2'*c85
name c203 'c2_yrcon02'
name c204 'c2_yrlib02'
name c205 'c2_yrnoc02'
name c206 'c2_scpaav'
name c207 'c2_ftepup99'

resp 'resp'

NOTE: Explanatory variables for response 1 (KS3 score)
expl 1 c86 c200-c202

NOTE: Explanatory variables for response 2 (school-level resources)
expl 1 c87 c203-c207

NOTE: set up 2-level model with school at level 2
iden 1 'csidno_long' 2 'school_long'
setv 1 'c1'
setv 2 'c1' 'c2'

note: sort 2 c90 'school_long' c76 c72-c74 c77 c86 c87 c91 c200-c207 c90 'school_long' c76
c72-c74 c77 c86 c87 c91 c200-c207
iden 3 c90
setx 'c1' 3 c91 c300-c416 c417
rcon c417
```

Appendix 4
Structure of Prior distributions for the WINBUGS experiments

Structure of Prior distributions for the WINBUGS experiments

$\beta_k \sim U(\ )$ **Each fixed parameter has a flat prior**

$\sigma_0 \sim U(0,10)$ **Uniform prior on σ_0, the square root of total high level variance** $\sigma_0{}^2 = \sigma^2_{0,lea(school)} + \sigma^2_{0,lea(area)} + \sigma^2_{0,school} + \sigma^2_{0,ward}$

$\sigma_{0,e} \sim U(0,10)$ **Uniform prior on square root of pupil variance**

Priors are set on the ratios of school and ward variance to the summed variance at the top level $p_{school} = \dfrac{\sigma^2_{0,school}}{\sigma^2_{0,lea(school)} + \sigma^2_{0,lea(area)}}$ and $p_{ward} = \dfrac{\sigma^2_{0,ward}}{\sigma^2_{0,lea(school)} + \sigma^2_{0,lea(area)}}$. These are

$\log(p_{school}) \sim N(0,10)$ **Vague prior on log of** p_{school}

$\log(p_{ward}) \sim N(0,10)$ **Vague prior on log of** p_{ward}

Similarly a prior is set on the proportion of lea (school variance) to the summed variance at the top level

$p_{lea(school)} = \dfrac{\sigma^2_{0,lea(school)}}{\sigma^2_{0,lea(school)} + \sigma^2_{0,lea(area)}}$. This is

$p_{lea(school)} \sim U(0,1)$ **Uniform prior on** $p_{lea(school)}$

The correlation between the lea(school) and lea(ward) effects is also the parameterisation used to define a prior. Defining

$\rho = \dfrac{\sigma_{0,lea(school),lea(ward)}}{\sqrt{\sigma^2_{0,lea(school)}\sigma^2_{0,lea(ward)}}}$ We set

$\rho \sim U(-1,1)$ **Uniform prior on** ρ

We note the following relationships to the original parameters

$\dfrac{\sigma^2_{school}}{\sigma^2} = \dfrac{p_{school}}{p_{school} + p_{ward} + 1}$ Proportion of high level variance at school level

$\dfrac{\sigma^2_{ward}}{\sigma^2} = \dfrac{p_{ward}}{p_{school} + p_{ward} + 1}$ Proportion of high level variance at ward