RESEARCH

Cross-classified and Multiple Membership Structures in Multilevel Models: An Introduction and Review

Antony Fielding
University of Birmingham

Harvey Goldstein
University of Bristol

department for
education and skills
creating opportunity, releasing potential, achieving excellence

Research Report RR791

Research Report
No 791

Cross-classified and Multiple Membership Structures in Multilevel Models:

An Introduction and Review

Antony Fielding
University of Birmingham

Harvey Goldstein
University of Bristol

ISBN 1 84478 797 2

ACKNOWLEGEMENTS

The authors of this report would like to thank all the members and visiting fellows of the Centre for Multilevel Modelling, Graduate School of Education, and University of Bristol for the many discussions in the area which have informed this review. We are also extremely grateful to Professor Hywel Thomas and Dr Ian Davison of the University Of Birmingham School Of Education. The first effectively co-directed a project in which this review forms a part and was a constant source of advice. The second undertook valuable literature searches. Finally we are very grateful to the DfES researchers and officials for their patience and invaluable guidance.

CONTENTS

1 Introduction

The aim of this report is partly to introduce in a fairly readable way some of the key ideas of fairly recent statistical methodology for modelling data on complex social structures including those in education. It reviews the 'state of the art' in the development of such methodology, and its software implementation. It also considers a wide range of examples and published applications which are either drawn directly from education or suggests potentialities in that area.

Since many of the key ideas of statistical modelling of effects and the necessity for statistical control of variables are well established in traditional explanatory multiple regression this is considered first. This establishes important notions which are essential to understand as the statistical models become more complex. Section 2 then goes on to consider how data can arise from hierarchical structures such as pupils within schools and why standard regression models should be extended to encompass multilevel models. Examples from educational progress research are then considered to illustrate the applicability of such models and to further introduce major concepts such as variance components. A variety of relevant extensions and applications are then introduced to fix ideas further,

Section 3 then considers that hierarchical structures and models to handle them are only the starting point for statistical modelling of complex reality. For instance it may be seen that not only do students nest themselves within schools but may also be lodged in a parallel hierarchy of area of residence which cuts across the school hierarchy. The example of education production functions incorporating both school and area effects are given. Further examples are given and then cross-classified random effects models are introduced as an appropriate way of handling data on such structures. We then examine some of the aims of such analyses. By considering some fairly complex repeated measures designs Section 4.1 reveals even more detailed structural complexity that falls into a cross-classified model framework. To formulate and understand the statistical aspects of the models some fairly detailed structured algebraic notation is required. This is outlined in Section 4.2. The detailed examination of a published application and its results in Section 4.3 illustrates the variety of detailed answers to research questions which may be revealed. This example which crosses-classifies students by secondary school attended with their previous primary school shows that achievement at secondary school may depend not only on the secondary school but also large carry over effects of prior primary school.

Further complexity is introduced into models in Section 5 by introducing the idea of multiple membership. For instance, in an educational setting students can attend more than one institution, so that a strict hierarchy of students within institutions is no longer applicable. Effects on a response variable may thus consist of contributions from more than one unit at the institutional level. It is shown that by conceptualising these random effects as weighted contributions from these several units the multilevel modelling framework may be further extended. Classification diagrams and a more simplified notation are then discussed and together form a heuristic way of grasping the essential features of such complex structures. Section 5 then concludes by consideration of detailed examples where the multiple membership ideas are seen in practical operation. It is also seen how units with multiple memberships may also be combined with existing cross-classifications in illuminating ways. In an educational setting a set of students may be crossed with a set of teaching groups for the purposes of studying their GCE A levels. The various A level grades are nested within a cross-classification of students and teaching groups. Teachers may make contributions to several groups and also each group may be handled by several teachers during its operation. By conceptualising each grade response as being in multiple membership relation with the set

of teachers alongside a crossing of students and groups it is shown how the model framework enables the disentangling of the separate effects of individual student characteristics, group features, and teachers.

Section 6 discusses the contrasting approaches that are taken to the estimation and statistical fitting of the complex models that have been discussed. In particular the focus is on the two approaches, Maximum Likelihood (ML) and Monte-Carlo Markov Chains, which are contrasted. The wide range of statistical software which has facilities for handling the model frameworks is also briefly evaluated. Some crucial features of MLwiN which is most widely used in the UK research community are outlined. The penultimate Section 7 considers a range of quite complex applications that have appeared in the literature. This comprehensive up to date review covers the following areas; health research, survey methodology and interviewer variance, social networks, veterinary epidemiology, animal ecology, genetics, transportation, missing unit identification, generalisability theory, psychometrics and additional education applications. Where possible attention is drawn to parallel structures in education where some of the methodology of the applications may have potential. The concluding section discusses briefly the way multilevel modelling has developed in a variety of other direction that have great potential in developing comprehensive methodological approaches. A final paragraph explains the richness that multilevel modelling has brought to methodology for complex situations but also warns against its inappropriate use and the dangers of over interpreting what it tries to do.

2 Basic Multilevel Modelling in Hierarchical Social Structures

2.1 Explanatory Models Using Multiple Regression

The aim of many statistical models is to try and account for variation in some response variable by a set of one or more explanatory variables or effects. The possibility of connection to causal evaluation is a complex one and will not be pursued here. However, for a discussion of this issue in educational research the interested reader is referred to Goldstein (1997) or Fielding (2000), for example. Models used will depend on a number of considerations. One is the nature of the response. Initially in this review we will focus on those that are continuously measured, where multiple linear regression models are common. Another consideration is the structure of the available data or the design of a study. Of particular importance for present purposes will be complex structures involving hierarchies of units of observation. Multiple regression methods are then extended to accommodate these structures and this is the central aim of multilevel modelling.

We will assume that the reader has some familiarity with the idea of multiple regression. However, it may be useful to outline some central ideas through an example. Suppose a researcher is interested in a relationship between the scores on a Key Stage 1 (KS1) Standard Assessment Task (SAT) of English children and their prior ability as indicated by a baseline test administered around two years previously. We might also believe that it might be fruitful to explore, for example, differences due to gender and family background (as indicated by whether they are eligible for free school meals or not). We might be interested in these for two reasons. Firstly there may be an interest in gender or family 'effects' on the response (SAT) in their own right. However, since they may also affect the prior ability measure we will wish to include them as 'control' variables. If we did not so include them any relationship we might observe between the response and prior ability might be partly if not entirely due to the common influence of these factors on both. *At this stage we assume we have no information on the schools attended by sample data on pupils*. We label as y_i the SAT response observation for a particular pupil i, and correspondingly x_{1i} the baseline score. Both gender and free school meal eligibility are binary variables taking only two values. These are

treated by 'dummy' indicators. Thus we have x_{2i} =1 if the pupil is female and zero otherwise. Likewise we might define x_{3i} =1 if the child is eligible for free school meals and zero otherwise. The usual multiple regression model for this situation is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i .$$

The x_{1i}, x_{2i}, and x_{3i} are observed as values of 'explanatory' variables and y_i is the response or dependent variable observation. The coefficients $\beta_0, \beta_1, \beta_2\ \beta_3$ are called parameters of the model and will be estimated in some way from the data.

For later use it is also sometimes beneficial to use shorthand vector and matrix notation and alternatively write the model as

$$y_i = \mathbf{X}_i \hat{\mathbf{a}} + e_i .$$

This is similar to the form in which it appears in many standard texts in econometrics for example and is adapted in the multilevel modelling literature and our later exposition. The vector $\mathbf{X}_i$ is taken to mean a row vector $\{1, x_{1i}, x_{2i}, x_{3i}\}$ containing observations on the explanatory variables with a leading element of unity. The latter which is constant across all variables is conveniently used to act as an artificial intercept variable. The $\hat{\mathbf{a}}$ is the column vector $\{\beta_0, \beta_1, \beta_2, \beta_3\}'$ of parameters. When multiplied out using routine matrix operations

$$\mathbf{X}_i \hat{\mathbf{a}} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3ii}$$

as in the above full algebraic specification of the model.

The basic initial interest was in the size of the coefficient β_1, the 'effect' of prior ability on SAT score. Since other variables are included in the model its interpretation is of an effect net of any possible indirect effects of the other variables. Put another way it will be the average conditional effect[1] of prior ability holding gender and free school eligibility constant. Thus for a pair of specific values of these it indicates that a difference of one unit on the scale of prior ability will yield an expected difference of β_1 in the response. This point should emphasise the fact that explanatory variables in a model act as mutual controls for each other. The quantity β_0 is known as the 'intercept', interpreted as the expectation (average) of y when all the explanatory variables in the model take the value zero. We leave aside for now the role of e_i except to note that it is usually assumed to average out at zero over pupils. For explanatory variables in a model that are continuous like x_1, then the interpretation of their coefficients would follow in a similar way to that of β_1. However, in this case they are not continuous but are dummies taking on values either zero or unity. In this case their coefficients in association with the intercept take on slightly special though intuitively similar meanings. From the way the variables have been defined we see that the intercept β_0 is the average SAT score (y) for male pupils (x_2=0), not eligible for free school meals (x_3=0) when the prior ability score (x_1) is zero. To see this we can note that for these values the model specifies $y_i = \beta_0 + e_i$, and we have made the not unnatural assumption that e_i averages out at zero. For these reasons male and eligibility for free school meals are often referred to as the base or reference categories for the indicators x_2 and x_3. It may be seen then that β_2, or the net gender effect or the average difference between females and males for any pair of values of both prior ability and the free

[1] In econometrics literature this is sometimes called the 'marginal effect' in keeping with economics terminology of effects of changing an economic variable holding other contexts constant. We prefer 'conditional' since marginal in statistical models usually means the total effect.

school meals indicator. To reinforce this point a particular situation may be considered, for a child with zero prior ability score, i.e. x_1 =0, and not eligible for meals, i.e. x_3=0. Then the average y for females will be $(\beta_0 + \beta_2)$ to set against β_0 for males. Similar interpretations may be placed on the coefficient β_2 of the free school meals indicator. In regression models of this kind the coefficients of variables that have been explicitly included as explanatory in the model are often called 'fixed effects'. Bearing in mind the facts we have noted about e_i the regression $\{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}\}$ is often taken in other contexts as a predictor of y for given values of explanatory variables and is known as the 'fixed part' of the model.

We have so far said very little about the role of the term e_i in the model. This term is variously called the 'disturbance', 'residual', or 'error' and is pivotal to analyses using statistical models. It is sometimes also referred to as a 'random effect' for reasons which might become clear as we build up ideas. At its most basic level of interpretation the term is included because we can never hope to get a model to fit non-trivial empirical data exactly; however complex that model might be. We hope the model is a reasonable approximation to reality and this residual indicates the extent to which a fixed part prediction of y from the model deviates from its actual value. It can also be given interpretations in terms of measurement error in y or inherent uncertainty in the complex reality we are attempting to mirror statistically. However, another relevant interpretation in the present context is that it represents 'unobserved variability' in children's SAT responses due to effects that we either cannot observe or do not explicitly include in the model. Apart from prior ability, gender, and free school meals we recognise that that there may be many other possible influential characteristics. These characteristics vary between children and may make their response higher or lower than that expected given their set of values on the explanatory variables we have included in the model. Our best hope is that in a reasonable model these sources of unobserved variability do not operate in a systematic way. In other words over children at given levels of the explanatory variables we assume that these residual effects average out to zero. Another consideration is that we assert that these unidentified explanatory variables are uncorrelated with the variables that are included in the model. If this is the case they will not exert any appreciable controlling or systematic influence on explicit net effects modelled.

If the assumptions we have made about e_i are true, then in fitting models to data we can treat them statistically as if they were random drawings from a statistical variable with the additional property that such drawings would average out at zero in the long run. It is for this reason they are an example of 'random effects'. Later in our development we will also introduce other more complex types of random effect. In a well developed model we would also like the e_i to be small, or in other words the predictions of y from the model are judged reasonably close to the actual values. However, if there are other unaccounted for effects that are important influences in their own right, then e_i might be large. This might be so whether or not the unidentified variables systematically affect the assessment of net effects of other variables in the model. If we could observe such a variable we would then perhaps want to extend the model specification by including it as a further explanatory variable in the fixed part. We note that this process is conditioned by our ability to both know and observe what such influences might be. Poor fitting models which may also have insufficient control often arise because of the inability to do these things.

The size of all these possible combined sources reflected in the e_i is indicated by the extent of the variability of the e_i, as measured by its variance which is denoted by σ_e^2. The latter is another parameter of a model which can be estimated from the data. Indeed, although we do not explore algebraic detail it is also a basis for measures of 'goodness of fit' of a model. Good model building strategies to get well specified models will try and examine and include

as many important fixed effects as possible to get a good fit. This is part of the delicate art of statistical analysis. We could observe, for instance, in our example, whether a child had been to nursery school or not. It might be found further that in model development it had a large important net effect on SAT scores. Further this variable may be correlated with other explanatory variables and so might also indirectly influence the size of other estimated conditional effects. Without a nursery school effect included we might regard the initial model as somewhat poorly specified.

All statistical models require assumptions and the impact of their breakdown is the subject of extensive and sometimes complicated statistical theory. Apart from the requirements of a reasonable model outlined previously we must also ensure that any proposed estimation or model fitting procedures possess certain desirable statistical properties. Some minimum assumptions are usually required to ensure these. Thus Ordinary Least Squares (OLS) estimation as used in traditional multiple regression models requires certain 'classical' assumptions. As we have hinted, and these are often uncontentious if models have been carefully thought out; firstly the residual random effects should have expectation zero and secondly they should be uncorrelated with explanatory variables in the model. Thirdly whatever the observation i, the variance of e_i should be constant. Thus we should not expect variability of the random effect in an observation to be influenced by any particular characteristic of that observation. If the contrary is the case and we had what is often called 'heteroscedasticity' the standard OLS framework will need re-thinking. Fourthly, the e_i should be uncorrelated across observations. We should not, therefore, expect the size of the residual for a particular observation to be influenced in any way by sizes of residuals for other observations. As we shall see the breakdown of this latter assumption is one motivation for considering new model frameworks and estimation when we face more complex data structures. Lastly a common assumption is that each e_i is normally distributed around zero with variance σ_e^2. We denote this by $e_i \sim N(0,\sigma_e^2)$. Though strictly necessary for inferences from OLS, the procedure is relatively robust if normality is not entirely valid.

2.2 Hierarchical Data Structures And Multilevel Models

The framework considered in the previous section may be unduly simplistic since almost all kinds of social and educational data have hierarchical or clustered structures often with several levels. For example, individuals (Level 1) live in households (Level 2) which in turn are nested within neighbourhood areas (Level 3); so this is a three level hierarchy. As another example of a three level hierarchy, primary school children (Level 1) are nested within schools (Level 2) which are grouped at Level 3 by Local Education Authorities (LEAs). A snapshot representation of this three level structure is shown in what is called a 'unit' diagram in Figure 1. It is possible that children are also clustered in well defined class groups within schools so that it might become a four level model. An example we will pursue further is information on all children taking KS1 SATS within schools in Birmingham LEA in a particular year. However, we will treat this as a two level structure with children within schools as there is only one LEA, so the 3rd level is not required. Because of shared school influences children from the same school will tend to be more like each other than pupils chosen at random from the population of children at large. For instance in a particularly effective school, most children may have a higher SAT score than expected from knowing just their individual characteristics, such as we have considered above. The relationship we have been examining is then constrained or modified by the shared membership of a particular school. The weakness of standard multiple regression is that it does not take account of the hierarchical structure of the data in that it focuses too much on individual

characteristics and ignores the location of the children in these shared contexts. Multilevel modelling is a way of explicitly including these contexts.

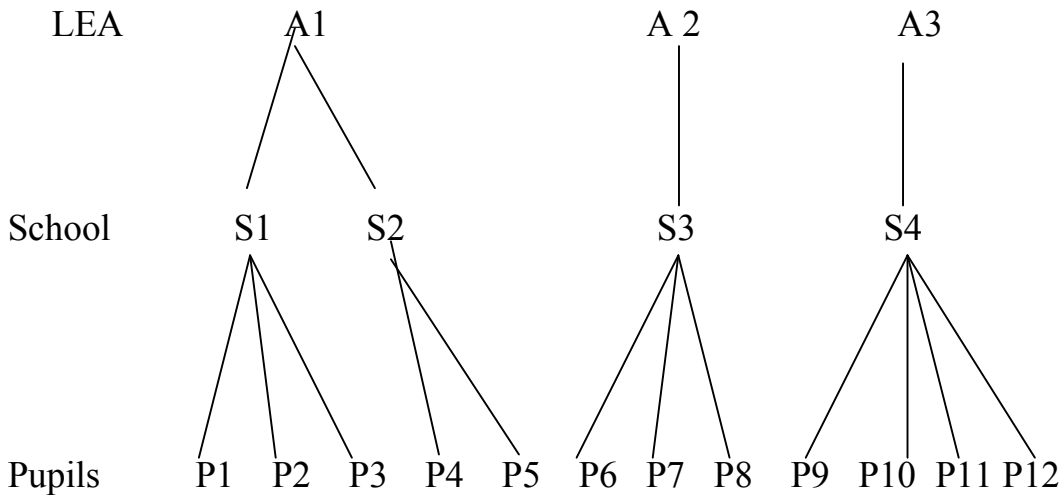


Figure 1: A unit diagram for a three level structure of pupils within schools within Local Education Authorities

The examples we have discussed are examples of hierarchies which may often be thought to arise naturally from data that is routinely collected. Many designed experiments or quasi-experiments also create hierarchies. An intervention may be carried out in several institutions with other institutions chosen as comparator controls. The data are collected on individuals within these institutions so we have a two level design (for an education intervention example see Thomas et al (2004)). Many national survey designs are cluster or multistage samples and also structure data hierarchically in this way by design. However, as Goldstein (1998) says, 'In formulating models that take account of such hierarchies we are concerned only with the fact of such hierarchies and not their provenance.' Whatever the source of the groupings by level, members of a group can both influence and be influenced by the composition of the group and its other characteristics.

Many traditional statistical analyses may be rendered invalid by ignoring these relationships and omitting the importance of group effects. To illustrate this let us consider early work in the education area by Aitkin et al. (1981). They re-analysed the data of Bennett (1976) who reported a statistically significant finding that formal styles of teaching improved the progress of children in reading. The finding had been based on traditional multiple regression analyses which treated the individual children as the only units of analysis and ignored their clustering within teachers and into classes. In the re-analysis, accounting properly for the effects of this hierarchy, the significant improvement disappeared and the progress of formally taught children was not discernibly better.

In social structures, such as arise in education, effects which we wish to evaluate, unravel and explain operate in complex ways to match the complexity of the structures. So, it is important that design of investigations, data collection, and analytical models reflect these complexities. In the rest of Section 2 we will focus on strict hierarchies as a preliminary to even more complicated structures to be considered later. It has become recognised over the past twenty years or so that multilevel statistical models provide the appropriate analytical approach. The example of Aitken et al (1981) is the first important example of multilevel analysis of social

data, though it was not labelled as such at the time. Multilevel analysis enables the derivation of information about relationships among measurements operating at different levels simultaneously.

The early classic generic paper is Aitkin & Longford (1986). Thorough technical discussion of the theory, methodology and range of applicable models is provided by Goldstein (1995, 2003). A good text which focuses essentially on educational examples is Raudenbush and Bryk (2002). Other texts requiring varying levels of statistical sophistication on the part of the reader are Kreft and de Leeuw (1998), Snijders and Bosker (1999), Hox (2002), and Heck and Thomas (2000). The articles by Paterson (1990), Paterson & Goldstein (1991), Rice and Leyland (1996), Plewis (1998) and Leyland & Groenewegen (2003) provide very readable introductions. Multilevel modelling is becoming increasingly complex and can get very technically demanding. There is an ever growing body of literature with articles and texts dealing with rapidly advancing methodology with a diversity of applicable subject matter (e.g. Duan and Reise (2000), Singer and Willett (2003)). Leyland and Goldstein (2001) edit a useful collection of articles dealing with a wide range of model developments and applications in the health area.

2.3 Basic Ideas Through An Example

Here we follow Goldstein (1997, 2003) in presenting fairly briefly the basic ideas, with a minimum of statistical complexity, as background for more advanced type of models in later sections. We will build up the idea of a multilevel model for a two level hierarchical structure in easy stages using as the motivating context the KS1 situation from the previous section. It will be somewhat artificial to start with since the aim is to communicate basic ideas. The exemplar ideas are based on the fuller analysis of Fielding (1999), where more elaborate final analytical models are presented. The responses are the KS1 results of 4444 children in a sample of 114 Birmingham LEA primary schools. This is a clustered design of schoolchildren at Level 1 within schools at Level 2. Here we might examine the Mathematics test and initially address the question of the extent to which it is influenced by measured achievement at baseline. The traditional initial approach would be to carry out a standard regression of test scores on baseline scores as outlined in the first section. In this case the model would be $y_i = \beta_0 + \beta_1 x_{1i} + e_i$. Here y represents the KS1 Mathematics score and x_1 the baseline achievement. It would likely be further developed in traditional multiple regression by extending to cover the effects of variables such as gender or ethnicity of the children. Indeed, another motivation might have been to study gender differences in progress adjusting for initial ability.

Such models do not recognise that pupils are taught in schools. They do not acknowledge that there may be shared influences from particular schools on children within those schools and that they may affect the relationships for those children. As it stands the model is to this extent incomplete. There may be influential 'school effects' which are every bit as important as other characteristics of the children. For example, it is possible that average test scores vary from school to school even after allowing for differential intake ability. We are admitting the possibility that this may be due to the school effects and an analysis that explicitly takes this into account is clearly desirable. Such an analysis has many advantages.

Firstly it enables the analyst to obtain statistically efficient estimates of regression coefficients as effects of explanatory variables. In the above discussion we discussed how omission of control effects might give misleading impressions. This is no less true if school effects are not considered in some way. Secondly, and this is a technical issue, if we use information on the clustering into schools correct standard errors, confidence intervals and significance tests are provided. Generally ignoring the clustering will indicate standard errors for the coefficients

that are too low. As in the study of Bennet (1976) the use of such standard errors will inflate, for example, traditional 't' statistics for regression coefficients so that they may be statistically significant. The use of correct standard errors will lower these values so that results may become insignificant. This impact has been known for a long time in the field of sample survey methodology. For cluster or multi-stage designs if standard errors use formulae which assume simple random sampling they will be underestimates. It is surprising how often in the not so distant past this has not been recognised for regression relationships. A third issue is that incorporating school effects enables us to explore the complexities of variation amongst schools. For example we can investigate the extent to which schools differ for different kinds of pupils. Goldstein et al (1993) found that schools show greater variation in GCSE results for students scoring higher on intake tests than they do for lower initial achievers. Fourthly an extended model may allow covariates or predictors measured at the higher levels of the hierarchy. The analysis can then investigate the extent to which variation amongst schools may be accounted for by such school factors as size of school, pupil teacher ratio, organisational practice, or concentration of pupils with certain social backgrounds. Finally in school effectiveness research there is often an interest in the relative ranking of schools based on the performances of their students after adjusting for intake characteristics and other relevant characteristics. This can be done most effectively using a multilevel model. However, Goldstein and Spiegelhalter (1996) point out the need for care with such rankings in that they may often be over-interpreted. They may be useful in identifying schools for further study which have extremely high or low rankings. However, they often have too much imprecision (very wide confidence intervals) for fine comparisons.

To clarify some of the basic issues and also the meanings of levels and units we can consider some fairly simple *hypothetical* relationships. Consider firstly the usual simple scatter diagram in Figure 2. Here KS1 results for a few pupils in *a particular school* are plotted against their Baseline Scores. Also illustrated is a standard simple regression line from a basic model which might have been fitted to pupils at that school. The variation in actual KS1 outcomes about this line is the level 1 residual variation since it relates to level 1 units (pupils) within this one school. If we were just interested in analysing this one particular school this approach might be appropriate. The residual variation would be represented as in the formulation in the previous introductory section by terms such as e_i.

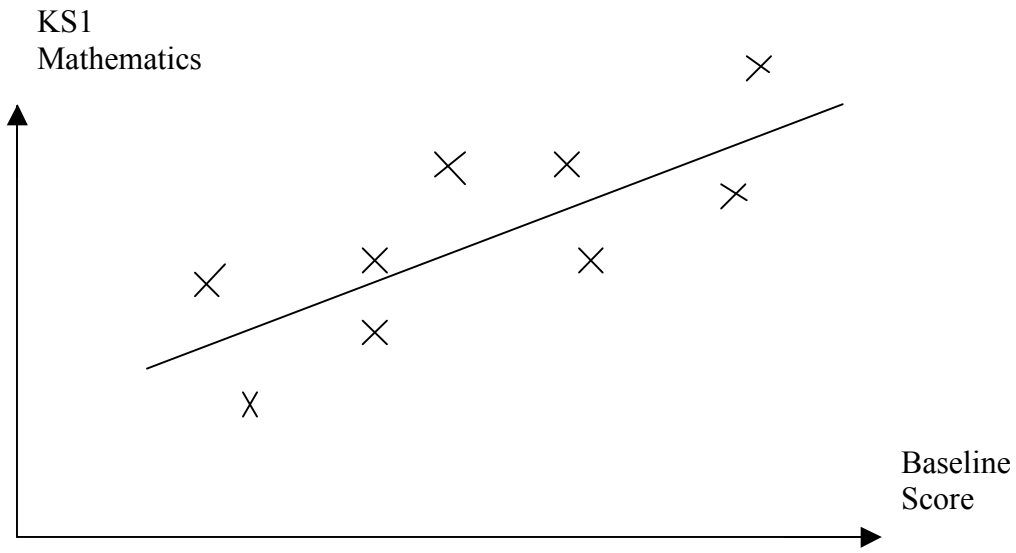


Figure 2: Level 1 Variation Amongst Pupils for One Particular School

(X indicates scatter of 9 individual pupils)

Now, however, imagine drawing several such diagrams one for each of a number of schools with *level 1 residual variation* about each of their separate lines. Figure 3 shows what hypothetically might happen if we draw separate lines for each school in our data with individual pupil data points removed.

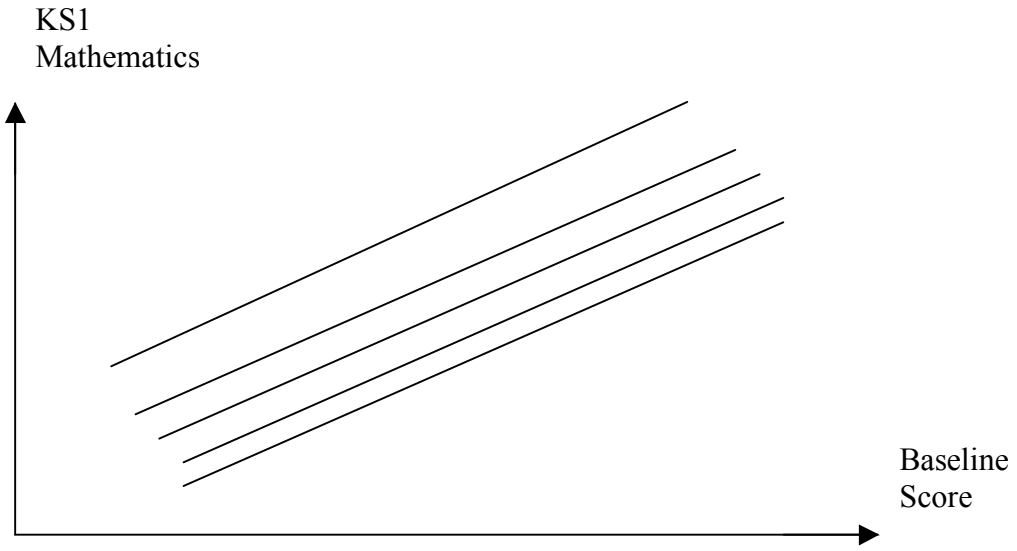


Figure 3: Regression Lines for Five Schools

(Parallel lines with same slope but different intercepts)

For simplicity of presentation we graph only five schools. We see that the school lines vary in their intercepts (the point at which they would cross the KS1 Mathematics axis). This is one particular form of *level 2 variation* and we will now focus on this situation in developing

ideas. It shows a situation where on average whatever the intake level certain schools are higher or lower than others by fixed amounts. This may be seen by noting that at any particular level on Baseline Score the difference between two schools in KS1 Mathematics levels on average is the vertical distance between the two schools above that point on the Baseline Score axis. Since the lines are parallel these distances are the same whatever point on the Baseline Score is chosen. However the net effect of differences between children's intake score is the same within all schools as evidenced by the fact that the lines are parallel and hence have the same slope. Thus a change of one unit in the Baseline Score would be expected to lead to the same change in KS1 Mathematics whichever the School.

By contrast Figure 4 shows a more complicated but more realistic situation for six schools; examples of which we will also return after developing more basic ideas.

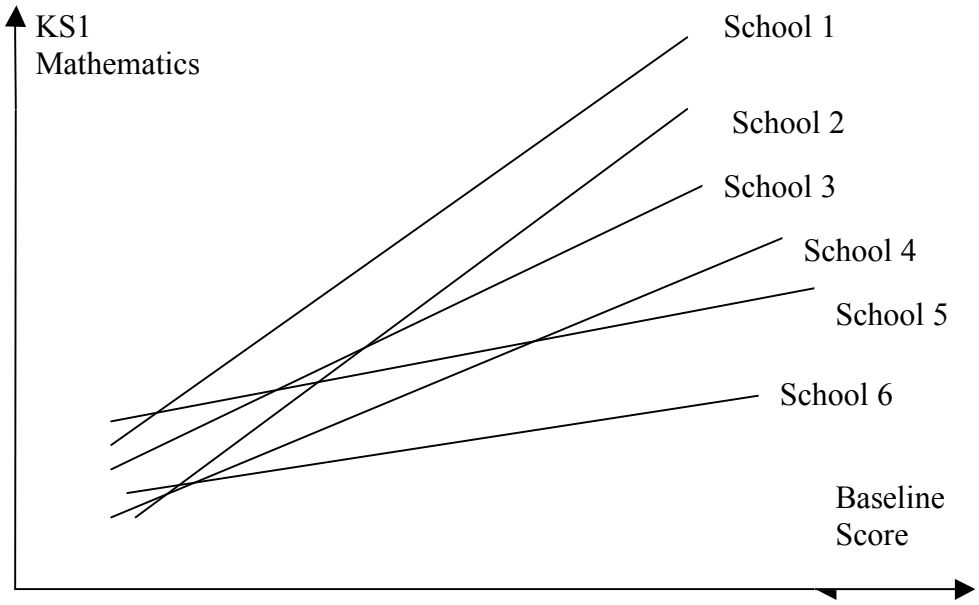


Figure 4: Level 2 Variation Amongst Six Schools in Both Intercepts and Slopes

In this Figure not only do intercepts vary but so do slopes. It is an example of multiple or complex level 2 variation since not only do schools vary in their average levels but the effect of baseline on KS1 may be different and specific for different schools. Thus for Schools 1, 2, and 3 with fairly steep slopes the effect of baseline is quite strong. In School 6 with a shallow slope it is relatively weak. Thus the conditional average difference in KS1 Mathematics for a pair of schools for pupils with a given level at baseline now varies according to that level. Vertical differences between lines are no longer constant along the length of the Baseline Score axis. There is now no single regression coefficient for baseline as we might have misleadingly analysed if we had used a traditional regression ignoring school differences. The variation in slopes is often referred to as a *differential* effect of baseline on KS1 amongst schools. If we examine the diagram we see that it also exemplifies the sort of situation envisaged by Goldstein et al (1993), as mentioned previously. At the lower end of the baseline range there is considerably less variation amongst school average KS1 scores than there is at the upper ends of the range.

For the moment we develop ideas in the context of the simpler situation of Figure 3. To accommodate different intercepts for schools in a single model we extend a regression model to reflect that situation. The appropriate model becomes

$$y_{ij} = \beta_{0j} + \beta_1 x_{1ij} + e_{0ij}.$$

We note that we now need two subscripts one for each of the two levels to indicate, for instance, that y_{ij} refers to the response of the i'th pupil in the j'th school. It might also be noted that we have added an additional zero subscript in the residual term e_{0ij}. This is for notational convenience since in later developments, as we shall see, we may introduce other random effects at the level of the pupil. We thus need a way to distinguish e_{0ij} from such terms; it also proves useful in the equation presentation in software implementations. We also see that we now have β_{0j} with a subscript j relating to a particular school rather than just β_0. This represents specific intercepts for schools which now may vary from school to school. To add greater emphasis to this a supplementary equation is often added;

$$\beta_{0j} = \beta_o + u_{oj}.$$

This equation stresses that a school's intercept varies around an overall average β_0, with u_{oj} the deviation of school j from this average. The mean of these deviations over schools is taken as zero. Substituting this back we get a model representation that is found in many texts and applications:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_{oj} + e_{0ij}.$$

We note that we have incorporated school effects into a single overall model rather than thinking about each school separately and this is a convenient way of thinking about a situation as in Figure 3. One procedure for representing this sort of school effect is a standard multiple regression model as we previously described, using a set of dummy indicator variables for schools, with one less than the number of schools. We omit a dummy indicator for one of the schools which may be arbitrarily chosen. This school becomes the base against which other schools are referenced. Thus if there are S schools indicated by s=1, 2, 3 …J and if we chose school 1 as the base this gives rise to (S-1) indicators; $x_{2,}\, x_3, \ldots\ldots x_S$ say. A typical one x_s has value unity if child i belongs to school s and zero otherwise. The model though essentially unchanged is re-formulated by dropping the u_{oj} from the above and introducing these variables and a set of associated coefficients as in the multiple regression model. Indeed it is now treated as if it were such a model. The set of dummy indicators operate as *fixed effects* in much the same way as the female indicator for gender did in the first section but there are now many more of them. The reference school 1 is treated as the reference school against which each of the other schools is compared in the same equivalent way as female was contrasted with male.

Fitting this type of model is formally equivalent to what is known as the Analysis of Covariance procedure for evaluating differences between a fixed set of schools adjusting for the covariate x_1. In some circumstances where we have just a few schools and moderately large numbers of students within each school, this might be a reasonable approach. It might also be appropriate if we are just specifically interested in making inferences about those particular schools. However, if we regard the schools as a (random) sample from a larger population of schools we might wish to make inferences about schools in general. Thus in a multilevel model in much the same way as we might regard the e_{0ij} terms as a random variable representing random unobserved effects operating at the child level we treat the u_{oj} as a random drawing from the distribution of school effects. We usually assume that

$u_{oj} \sim N(0, \sigma^2_{u_0})$ and variance term $\sigma^2_{u_o}$ sums up in one parameter the effect of school variation on the responses. With many schools this is statistically more efficient than estimating a large number of fixed school coefficients. Moreover, if our large number of schools had very few students in each, fitting a standard model with school indicators will not yield very precise estimates of these school effects. Intuitively, if we do this, we are trying to stretch somewhat limited data to tell us about a large number of unknown parameters simultaneously and in statistical parlance we run out of degrees of freedom. In a random effects model we can achieve greater precision by regarding the schools as a sample from a population and use the information from the whole sample to ultimately 'borrow strength' in studying one particular school. The technical statistical aspects of what this implies are covered in any of the introductions to multilevel modelling we have referenced. There is another major disadvantage in using fixed effect indicators; we could not develop the model further by introducing observable school characteristics in an attempt to further explain the variation amongst schools. Although this is a technical limitation, the intuitive idea is that the overall fixed effects associated with the schools sums up all that can be known about differences between schools on the average level 1 response. In modelling terms the effect of particular school variables cannot be separately identified in a model that includes school fixed effects and it would be statistically impossible to fit such models. Using random effects resolves this identification problem. There is considerable debate, particularly in the econometrics of panel data analysis, about the respective roles played by fixed effects and random effects. Fielding (2004) considers these issues more fully.

With the previous discussion in mind we can now summarise a basic multilevel model in a form similar to which it will appear in relevant literature. We write for the i'th student in the j'th school,

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_{oj} + e_{0ij}.$$

$$e_{0ij} \sim N(0, \sigma^2_{e_o})$$

$$u_{oj} \sim N(0, \sigma^2_{u_0}).$$

The term u_{oj} is the school effect and gives the additional contribution (positive or negative) that a school makes to the predicted response score (y_{ij}) given the baseline score (x_{ij}). The coefficients β_0, β_1 have the usual interpretation. The model now has two random variables specifying two random sources of variation, at the level 1 of pupils (e_{ij}) and at level 2 of schools (u_{0j}). In keeping with this such models are also often known as 'variance components' models. The two components are $\sigma^2_{e_0}, \sigma^2_{u_o}$ which need to be estimated along with β_0, β_1.

The total residual variance is $\{\sigma^2_{e_0} + \sigma^2_{u_0})$ and interest often centres on $\sigma^2_{u_0} / (\sigma^2_{e_0} + \sigma^2_{u_0})$, the proportion of the total variance that is attributable to schools, as a summary of relative importance of school effects. There is another reason for this interest. The clustering of pupils into schools induces a correlation between the responses of pairs of children who go to the same school as the previous general discussion has indicated. It can be shown that the size of this correlation, often known as the variance partitioning coefficient[2] (VPC) is indeed this same quantity $\sigma^2_{u_0} / (\sigma^2_{e_0} + \sigma^2_{u_0})$. This is one reason why traditional OLS is often inappropriate for the situation under consideration since its good behaviour depends on this correlation

[2] In sample survey design it is also known as the intra-class correlation (ICC)

being zero. Using OLS is tantamount to treating the total residual term in the model $w_{0ij} = u_{oj} + e_{0ij}$ as satisfying its assumptions when in fact within schools these terms are correlated. Methodologically, this was the main reason for the re-analysis by Aitken et al (1981) of the teaching styles data, to which we referred previously. The feature of our model that makes it multilevel is that we explicitly recognise that there are two random variables, one at each level of the data structure. Standard regression procedures which typically assume only a single random variable are inappropriate for such models. Special procedures for estimation and appropriate software are required (see the later Section 6).

2.4 An Example Of A Basic Two Level Variance Components Model

Table 1 represents results of applying two basic variants of these ideas to the data we have used as exemplar. The response variable is scored 0-4 for the various mathematics KS1 levels but standardised to have mean zero and standard deviation unity over the whole sample. In these results the only difference to the above basic model is that a full set of baseline measures will be used rather than a single score, yielding seven potential explanatory level 1 variables rather than just one. Thus we are extending the fixed effects part of the model without affecting the generality of arguments about the components of the residual variances. Also in the first part of the table we present results for a model without any explanatory variables; the base variance components model. Such a base model is of the form $y_{ij} = \beta_0 + u_{oj} + e_{0ij}$ indicating that we might be interesting in seeing how student responses differ from an overall average, and then attributing these differences separately to school effects and individual effects. The presentation of a base model is often useful as a comparative reference to see where later developments of the model might lead us. The main point of interest of the base model is that schools seem to account for 20% of the total variance in KS1 mathematics. This is interesting but it relates to raw unadjusted achievement only without control variables. Thus the interpretation of such results in substantive terms must proceed with caution. The results in the table for the 'model with reception baseline achievement controls' are after the introduction of baseline measure covariates. The estimated regression coefficients show the influence of different baseline achievements net of other achievements and the school effect. The standard errors are now estimated within a multilevel model and are thus appropriate for inference. Shape and space has a small direct effect and formally would be statistically insignificant which might suggest dropping it in further model development. The baseline tests are all on the same standardised scale so it is appropriate to compare their estimated coefficients. On this scale number seems to exert the largest net effect. The level 1 variance has reduced by 34% emphasising the importance of prior achievement of individuals in explaining quite a bit of their variation on the KS1 outcome. The school variance has reduced only slightly. This, perhaps, may reflect aggregate ability of intakes of Birmingham schools not being widely different from each other.

	Base Variance Components Model		Model with Reception Baseline Achievement Controls	
	Estimate[(a)]	Standard Error	Estimate[(a)]	Standard Error
Fixed Parameters				
Intercept	0.012	0.051	0.018	0.032
Baseline Test Assessments: 4 point scale and standardised				
Number			0.30***	0.016
Algebra			0.18***	0.015
Shape and Space			0.02	0.017
Data Handling			0.10***	0.016
Speaking and Listening			0.09***	0.017
Reading			0.11***	0.016
Writing			0.08***	0.018
Random Parameters				
School variance	0.19	0.027	0.17	0.030
Pupil variance	0.83	0.009	0.55	0.014
Intra-school correlation (VPC)	0.19		0.23	

Table 1: Multilevel Model Results for Standardised KS1 Mathematics: Variance Components Model and Model with Baseline Assessments Adjustment Control (Source: Birmingham LEA data described more fully in Fielding (1999))

Note: (a) *The *** indicates regression coefficient estimates that are significantly different from zero with p<.0.0001 using the appropriate statistical tests. On this criterion all estimates in this table are highly significant except for Shape whose net effect seems to have little predictive value when other baseline effects are included. But see footnote*[3]

Most multilevel software will also have a facility for estimating the residuals at both levels. The residuals at level 2 from the baseline controlled model may be taken as one form of 'adjusted school effects'. They or something similar are often taken as indicators of 'school value added' or 'school effects' since they reflect different progress in schools once initial ability of their intake has been accounted for. The estimates of residual are subject to all the usual uncertainty and imprecision due to sampling error in estimates and their standard errors can also be estimated from multilevel software. Thus one way of presenting residual results is a 'caterpillar' diagram as in Figure 5.

[3] In presenting the above results we have as is conventional indicated at what levels the regression coefficient estimates are significantly different from zero. There is probably little interpretative value in this beyond the information contained in the estimates and standard errors. The sample size is quite large so estimates are quite precisely estimated with relatively small standard errors. Thus the effect size estimates are the import of the results rather than whether they happened to indicate a true difference from zero. The only insignificant net effect is that of Shape and Space and that is because the estimated effect is so tiny as to be effectively zero.

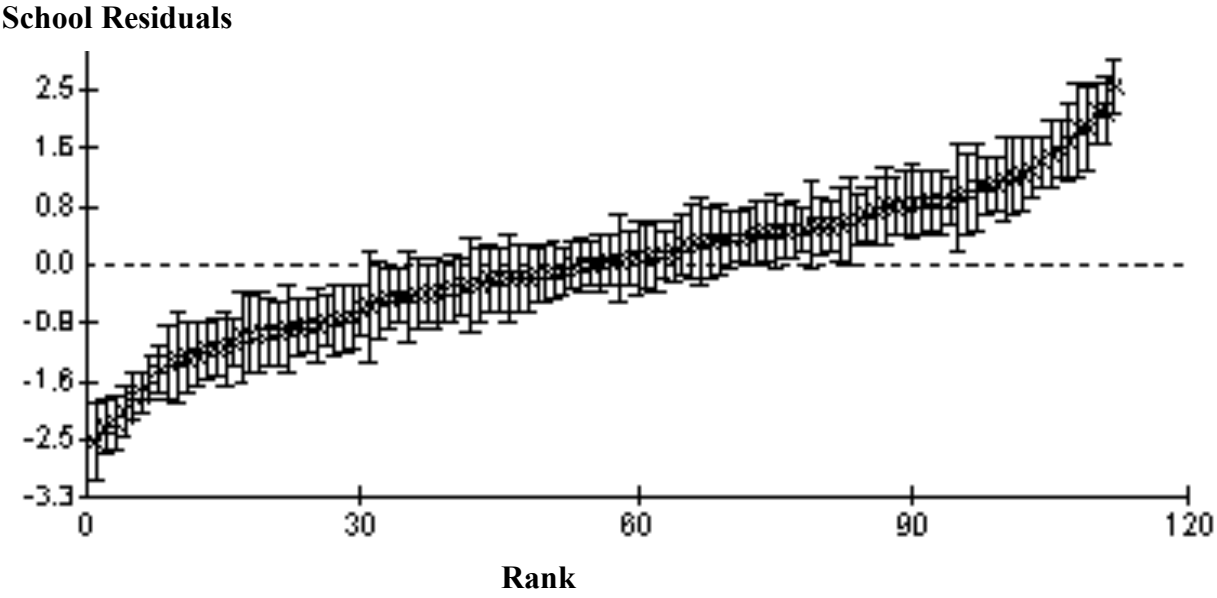


Figure 5: Confidence Bands for Adjusted School Effects by Rank Order

This diagram orders by value the residuals from the baseline adjusted model of Table 1. With their rank order on the horizontal scale, school residuals are plotted on the vertical scale surrounded by 95% confidence limits. The diagram also exhibits a well known feature similar to that in much progress research and usually present however much control is exercised; there is considerable overlap of intervals with around 50% of the intervals covering the overall mean of zero. Thus attempts to rank or separate schools in league tables, even when there has been proper adjustment, is subject to a high degree of uncertainty (Goldstein and Spiegelhalter (1996)).

2.5 Extending Hierarchical Models

In any application we would usually start as in the examples above and then introduce and examine the fixed effects of other potential explanatory characteristics. In this example, for instance, gender, nursery attendance, school meal eligibility, and certain ethnic and first language factors all proved important influences on level 1 variation. The effect of the baseline number test was also shown to be non linear necessitating a quadratic term involving the square of baseline number. One particular school level variable, the percentage of children in the school eligible for free meals (an aggregate measure of background of students) was an important context variable explaining some degree of level 2 variation.

Figure 4 above also suggests that the slopes in models can also be allowed to vary over level 2 units in assessing evidence for differential effects. Reverting to the basic model we could now also index the slope by j and write $\beta_{1j} = \beta_1 + u_{1j}$ to replace a common slope β_1. We have now allowed another random effect so that the model becomes

$$y = \beta_o + \beta_1 x_{1ij} + u_{1ij} x_{ij} + u_{oj} + e_{oij}.$$

We now have two random effects at the school level. We must usually also allow the two school effects to be correlated since it is not unusual in practice for this to arise. We now will

have three parameters to estimate to characterise school residual effect variation: $\sigma^2_{u_o}$, $\sigma^2_{u_1}$ and a covariance $\sigma_{u_o u_i}$. This type of model is often called a random coefficients model. In the example, Fielding (1999) demonstrated that the effect of the English baseline test varied over schools and that it had a correlation with the average intercept effect of -0.55. Thus in schools which in general had higher average effects there was a shallower effect of the English baseline score.

Another way in which the multilevel model can be elaborated is by relaxing the assumption that all types of children have the same degree of residual variation at level 1. We may therefore allow complex variation at level 1 by modelling a variance function by allowing the residual variance to depend on level 1 variables. Fielding (1999) in his analysis showed that level 1 variance was different for the two genders and also whether or not the child was eligible for free school meals. Different variability of outcomes amongst certain groups is often just as interesting as differences in average levels.

We can also extend the hierarchy in models to many levels. We gave examples of three and four levels in Section 2.2. In principle we can have as many as we like in addition to all the complexity of the type just discussed at any of the levels. Notation sometimes gets complicated and often has to be adapted to suit particular situations. For three level models we need three subscripts rather than two for instance. A basic three level variance components model, for example is written as

$$y_{ijk} = X_{ijk}\beta + v_{ok} + u_{ojk} + e_{oijk}$$

This relates to the i'th level 1 unit within the j'th level 2 unit within level 3 unit k. We note, for instance the use of u_{ojk} to indicate the residual for a level two unit within a particular level 3 unit. We use the generic matrix form (see Section 2.1) for the fixed part of the model since it is convenient shorthand for a model with many fixed explanatory variables.

There is also rapidly developing methodology for response variables which are not continuous. They can be binary, counts, nominal and ordered categorisations and the like. For example, Fielding et al (2004), discuss multilevel models for A level grades which treat them as sets of ordered categories rather than assuming they are points on a continuous scale using arbitrary points scores. A model for drop out of A level courses using a binary indicator dependent variable (drop out or not) in a two level structure is considered by Fielding et al (1998). We will not discuss models for these types of response in detail. However, we outline a basic 2 level model for binary response in the context of university progression studied by Draper and Gittoes (2004). Given samples of students (i) enrolling for courses in universities (j) we may record whether they progressed to second year (y_{ij}=1) or dropped out (y_{ij}=0). A well known portrayal of the way such binary responses might vary is the Bernoulli distribution, which is a special case of the Binomial with a single trial. Characterising this distribution is the probability of progression, denoted by π_{ij}. However, the probability will vary across students dependent on values of a set of explanatory variables and the university effect (and possibly in more complex models its characteristics). What we require then is a model for this dependency. For many good reasons, discussed extensively in many texts (e.g. Greene (2003)) a linear dependency model for the probability similar to linear regression is not usually appropriate. One which has found wide application and is founded on a strong theoretical and empirical base is the logit model. Rather than a linear model for the probability itself this uses a linear model for the logit; the logarithm of the odds of progression (log-odds). Odds are defined simply as the ratio of the probability of progression

to the probability of its complement, no progression. With this in mind the basic model is then written.

$$\text{logit}(\pi_{ij}) = \log_e\left[\frac{\pi_{ij}}{1-\pi_{ij}}\right] = (X_{ij}\beta + u_{oj})$$

$$y_{ij} \sim \text{Binomial}(\pi_{ij}, 1)$$

The right hand side of the first equation is for the most part the same as that for a two level model for a continuous response. There are explanatory variables in X_{ij} and Level 2 random variation is described by the continuous random effect term u_{0j}. In terms of higher level effects the interpretation is similar to linear models, but with effects operating on the log-odds of progression. However, we note there is no continuously distributed Level 1 random residual in the expression. Variation at Level 1 follows the Binomial distribution. Implicitly the binomial variance of the progression response is a function of the probability and is $\pi_{ij}(1-\pi_{ij})$. Hence there is no separate variance parameter.

Thus in summary, direct linear models as in the framework we previously discussed are inappropriate for non-continuous responses. Rather there are extensions to generalised linear models (GLM), (McCullagh and Nelder (1989))[4] to encompass multilevel random effects. A generic name for an extension involving random effects is a generalised linear mixed model (GLMM). When the random effects are hierarchical they become what are referred to as generalised linear multilevel models. Although we concentrate for ease of exposition on multilevel structures for linear models, many of the examples of complex structures we will discuss in the last section of our review are applied for non–continuous responses. Models for these usually involve transforming the responses in some way as in the example above and then only after transformation relating to a linear predictor. It is in this linear predictor that multilevel structures manifest themselves through continuously distributed random effects. Thus the ideas of these random effects as have been introduced for linear models carry over quite straightforwardly for these models. However, the estimation methodology for them is somewhat more complicated than for linear models and we will only refer to it only briefly in the discussion of this issue in Section 6.

[4] The econometrics literature often calls these limited dependent variable models (Greene (2003))

3 Cross-Classified Data Structures

3.1 The Nature Of Cross-Classifications And Their Effects

In the previous discussion we clarified the nature of explanatory statistical modelling, and discussed the difference between fixed and random effects on responses. We saw the parallels between the analysis of structures common in education and the social sciences and traditional frameworks in experimental design more common in the natural sciences. However we focussed essentially on extending standard multiple regression models to effects arising from levels in hierarchical structures. Thus, for example, in a three level structure where pupils were nested within schools which were in turn nested within Local Education Authorities we were concerned with separating out effects arising from influences at these three levels. The early development of multilevel modelling methodology and associated computer implementations, particularly in education, addressed such structures since they were quite common. However, as the work developed, it was discovered that there was often much more complexity in data structures. This led to a desire to extend the methodology to handle the analysis of effects in such structures. The separation of effects arising from what we may regard as random cross-classifications is one such extension.

Goldstein (2003) gives an illuminating example. In a basic two level multilevel structure students may be classified hierarchically by their school. In many cases, however, such units may be grouped along more than one dimension and collected data may reflect this. Students may be nested in schools they attend and also in the neighbourhood where they live. We can represent this diagrammatically in Figure 6 for just four schools and three neighbourhoods and thirty three students with groups of between one and six students nested within the cells of the various school/neighbourhood combinations. The cross classification is at level 2 with students at level 1. In this sense it is still a two level model but the level 2 units are now combinations of particular schools and neighbourhood.

	School 1	**School 2**	**School 3**	**School 4**
Neighbourhood 1	X X X X	X X	X	X
Neighbourhood 2	X	X X X X X	X X X	XX
Neighbourhood 3	X X	XX	X X X X	XXXXXX

Figure 6: A Random Cross-Classification of Students by School and Neighbourhood at Level 2

To further clarify this type of structure we can draw a ‘unit diagram’ as in Figure 7. Such diagrams, however, can become far too elaborate for some of the more complex structures we will shortly consider, and may become more confusing than illuminating. Later in this report we will, however, consider alternative pictorial representations, which are simpler in content. However, as will be seen these also require to be read alongside some algebraic detail. We eschew this for the moment until we have further developed the reasoning behind such detail. For the reasons just discussed and so the picture does not become too cluttered Figure 7 is only schematic and does not represent all the students in Figure 6. Only twelve of the thirty three students are considered. For instance, Pupils 1 and 2 attend the same School 1 but come from different areas, whilst Pupils 6 and 10 come from the same area but attend different schools

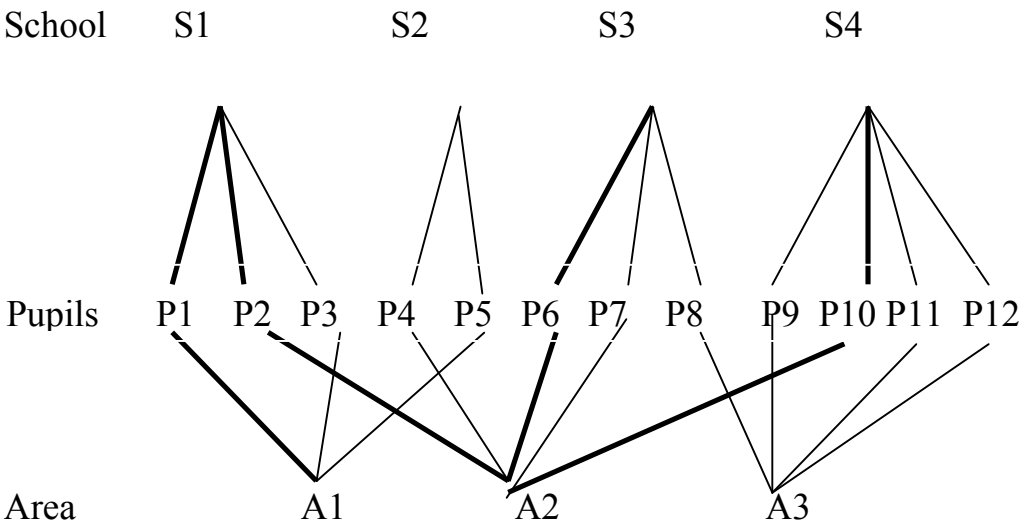


Figure 7: Twelve Students at Level 1 Nested Within an Area and School Cross-Classification at Level 2

Extending methodology for such structures means recognition that level 2 effects are now more complex and may arise from two cross cutting hierarchies. Further we desire to separate out in some way the effects of both schools and areas on whatever student outcome we are studying. This is particularly important if there is a degree of association between the areas the student lives in and the school they attend. This will usually be the case in observational and routine data. If area effects are important and they are left unspecified in a model the school effect may 'draw to itself' some effects attributable to its associated areas. This stands in contrast to the balanced designs we often see in experimental studies where such association is 'designed out'. In many ways the data collected by Pupil Level Annual School Census (PLASC) has this structure which although focussing on data collected from schools also enables classification of students by whatever area unit seems appropriate through information on their postcode. A relevant full example for the 2545 students at Key Stage 3 (KS3) in 2003 attending the 15 schools in the one Local Education Authority of Oldham is given in Figure A of the Appendix. This detailed table if it were inspected carefully would several interesting and important features which are typical of such educational structures. The 15 schools vary in size and have between 88 and 236 KS3 students. Of the 67 electoral wards of residence of these students only 20 are within the boundary of Oldham LEA. The other 47 wards, which are not all geographically quite proximate to Oldham, cover the areas of residence of a sizeable number 304 (12%) of the children in Oldham schools. We call this the 'out of area' factor. Apart from one or two wards quite close to Oldham the number of children in Oldham schools from each of them is relatively small.

These particular features present something of a challenge in applying multilevel models, which though not insurmountable may make for some daunting computation. Firstly if we were analysing the Oldham data alone we would need a fairly large number of area effects in the model apart from Oldham wards and this has computational consequences. Secondly due

to small numbers in most of these areas there are consequences, which will not be detailed here on the precision of model estimates. Thirdly, the Oldham data exemplified may be part of a larger set of many more LEAs, possibly all of them. The schools, as one factor in the cross-classification, are nested within LEAs. Taken alone so are electoral wards. However, due to the 'out of area' factor the cross-classified level 2 units will no longer be arranged hierarchically beneath LEA. Thus attempting to separate out an LEA effect by specifying it as Level 3 in a strictly hierarchical model will no longer be possible. The structure thus becomes a little more complicated. There are other features we might note about the Oldham data and a broader set of all LEAs of which it might be drawn. Each of the schools has a concentration of its students in fairly few wards, which presumably are its main catchment areas. This means there are a large number of cells in the cross-classification, even those which correspond to Oldham wards, which are empty or at most have just a few students. This sparse structure is typical of situations which Raudenbush (1993) has described as radically unbalanced. Later we will return to such further issues complexity in cross-classified structures and their resolution.

The use of cross-classified structures, and the 'building' into a statistical model the effects of its separate factors, is one of the essences of good research strategy. If we did not recognise that there are some area effects in addition to school effects on the student outcomes we would be dealing with what is often called a 'underspecified model'. The theory of statistical modelling suggests that estimation of other effects might be poor. This is true both for effects of explanatory variables we are examining and also the random effects of factors in our structure. Looked at another way we might say there is inadequate control in the model for important and possibly confounding effects. If we noted that a particular school had poor outcomes this may be due in part to the fact that this school drew large numbers of students from areas having characteristics which were detrimental to student outcomes. Conversely if we had a model which examined just area differences and did not include school effects, some areas may have higher achievement amongst its students than what might be expected. However, this may partly be due to 'school' effects of schools which these students largely attend. What we might like to study is the effects of particular schools net of any area effects and vice versa. It is cross-classified structures such as in the examples together with adequate data to support their analysis that enables us to do this. The example structures typify what might be required as essential initial conditions necessary to separate out these effects. A given area will often have its students going to a number of different schools and a given school will usually draw its students from a variety of areas with different characteristics. In Oldham LEA, for example, School F does have a concentration of 58% of its students in three wards but it draws the rest thinly spread from as many as 27 of them. Similarly a particular ward which has code 00FPBA has residing students who go to seven different schools. We can then study area effect variation controlling for school and school variation controlling for area. Fortunately an extreme situation, which would not facilitate this mutual control, only rarely arises in practice. Suppose each school drew all its students from one particular area unit and all students in an area went to one particular school. In a cross classification table like the example data there would be only one non empty cell per row and per column. The units formed by the cross-classification could still be level 2 units and variation in the effects of these units could be studied. However, we could not attribute the effects to either schools or areas. These effects are said to be completely confounded. Sometimes, certain structures, although not as extreme, approach this situation and then care must then be exercised in the attribution of effects.

In a parallel study to this review we are also considering the extension of education production functions to consider not only achievement of individual students being partly

dependent on the school they attend and the resources attracted by that school but also some 'personal resources' reflected partly in some way by the areas in which the students live. It should be clear from our discussion of the inter-relationships of effects that such effects if they are important might be included in a more elaborately specified model. Without isolating and controlling for such effects for which the cross-classified structural models are appropriate, we might not be able to disentangle such effects from others of particular interest, say the school effect of pupil teacher ratio or spending per pupil. Indeed such investigation if they yield important net area effects might suggest ways in which financial resources might be devoted to say ameliorating features of areas which have been observed to have detrimental effects. Such model results on observational data should not be over interpreted to yield firm causal explanations. However, they might suggest ways in which interventions might be designed which can then be trialled in a designed framework to yield more firmly based interpretations. Unfortunately the evidence base for area targeted interventions on children's achievement in Britain is currently sparse (Gibbons (2002)). The evidence of Gibbons (2002) on neighbourhood effects suggests that they can be relevant. However, as the wide range of research on school effects also suggests individual and hence family characteristics appear paramount. There is clearly, though much potential for more detailed research on such area effects and complex multilevel random effects models seem an appropriate tool to use.

3.2 Some Objectives Of Analysing Cross-Classified Multilevel Models

In the study of such cross-classified data structures there might be a variety of aims in building a multilevel statistical model which incorporates random effects for both factors in the level 2 structure. We conduct the discussion in the context of an area by school classification though, of course, these aims are more generally applicable for even more complex structures.

3.2.1 Improving the quality of estimates of explanatory variable effects.

Firstly we might seek to improve the estimation of effects of explanatory variables in the model by ensuring the random disturbance structure is properly specified and incorporates in a careful way essential features of important random effects. The impact of not so doing is perhaps most highly developed in the econometrics literature (see for example Greene (2003)) and is the subject of highly developed statistical theory. However, a main impact is that the fixed effect of coefficients in the model, though estimated with statistical consistency, usually report understated estimates of their precision through standard errors. The latter, play a role in statistical testing and often can lead to the conclusion that explanatory variable effects are statistically significant, when in a correctly specified model they would not be. This is true of ignoring multilevel effects in general, as we saw in the introduction. In our examples recognising school effects in a two level model might go someway to resolving this difficulty. However, if clustering of children into areas was additionally important due to real area effects then recognising this may improve further the quality of such inferences. The recent report by Levačić et al (2005) on effects of resources on individual student progress at KS3 properly recognised this grouping of students into schools[5]. However, if area effects are also

[5] In many econometric applications interest usually focuses mainly on the improved estimation of the fixed effect explanatory variable coefficients in a model. Sometimes then multilevel or crossed random effects are treated as nuisance factors. The fixed effect coefficients themselves are often estimated consistently though not entirely efficiently by standard methods which ignore the higher level random effects. However, since the standard methods produce inappropriate standard errors they are adjusted for the multilevel clustering by use of a variety of robust methods of standard error estimation ('sandwich ' estimators). This approach was used for the most part in adjusting for school effects in Levačić et al (2005), though some full scale multilevel models were

important there is the possibility that inferences may be further improved by recognising this additional random effect through the further classification of students.

3.2.2 Identifying components of variance in the outcomes.

In a cross-classified model we might initially and before introducing explanatory variables seek to examine how variation in outcomes might be attributable to differences between say schools, between areas and between individual students after controlling for area and school effects. This will give a base for then extending the model to identify which area, school and student characteristics might explain some part of these components of variance. After introducing such variables we might then estimate the residual components of variance. These will then give us an idea of the extent to which variation in outcomes might be attributable to unobserved influences operating at the level of each of the three types of unit in the model[6].

3.2.3 The study of differential effect.

We might want to investigate whether associations between student characteristics vary over schools and over areas. For instance is the effect of initial ability at entrance to a secondary school more important in some schools than in others? We might also want to see if effects of school characteristics vary across areas or if area characteristics operate differently for different schools. The effect of students going to a selective school, for example, might be greater for students from some areas than others. These types of influence are often known as differential effects[7].

3.2.4 Estimating level 2 effects

Most multilevel and other random effects model structures allow estimate of unique effects ('random effects') associated with particular schools or areas after we have adjusted for appropriate explanatory characteristics[8]. This is the approach adopted in much school effectiveness literature on 'value added' using hierarchical models (for example O'Donoghue et al (1997)). It may also be used as a screening device to identify schools with extreme adjusted outcomes for further investigation (Goldstein and Spiegelhalter (1996)). In a cross-classified model there is a further advantage in that there has been additional control for area effects which might otherwise have forced certain schools with strong areal associations to the extremes.

---

also trialled. Since econometric interest is usually on the coefficients this may be satisfactory. However, often more useful insight is often gained by explicitly modelling the higher level effects

[6] If area effect was important and had not been included in an under-specified 2 level model with just school as a random effect the component of variance due to schools might be unreliably attributed to school effects since part of it may be due to areal differences in the school intake. Rasbash and Browne (2001) give an example in health research, where we might desire to assess the relative importance of general practices and hospitals on patient outcomes. Due to the crossed nature of the two units building separate models one for patients within hospitals and patients within general practice is insufficient and may be misleading. What is required is an assessment of hospital variation net of GP effects and vice versa. This can only be done by estimating the components of variance in a cross-classified model.

[7] Technically these are handled in models by allowing random coefficients for the explanatory characteristics

[8] This is another good reason for adopting a fully specified model with explicit random effects. Treating them as 'nuisances' and adjusting for cluster effects on standard errors in a standard regression framework will not allow for this elaboration.

4 Further Examples of Cross-Classified Structures And Their Analysis

4.1 Some Examples In Education And Repeated Measures Studies

Other slightly more complicated examples occur in repeated measures studies. In such studies as tracing progress in say reading attainment the measures at different occasions over time can be regarded as nested within particular pupils. Changes over occasions are an obvious focus of enquiry. However, in an extended multilevel framework we may also want to examine and control for the effects of say teachers, different classes, schools, or raters[9]. Matters can get complicated because for particular pupils any of these either singly or in combination may change over time and sometimes several times. However, repeated measures designs are instructive in illustrating the richness of structures that can be formulated using the idea of cross-classified effects. We might note right at the outset that in multilevel analysis of these effects it is not necessary to have complete and balanced data on all occasions or sets of measurements. This is one advantage of a multilevel approach to repeated measures studies and is discussed extensively in Goldstein (2003).

We now give typical examples to demonstrate the structural ideas. Suppose a basic design has measurements at up to 7 different occasions with 2 measurements by the same 2 raters on all occasions. However, for various reasons measurements are not always complete and sometimes only one rater's measurement is taken. For a *particular student* the situation may be as shown in Figure 8 where X indicates the measurements that are taken.

	Occasion						
Rater	**1**	**2**	**3**	**4**	**5**	**6**	**7**
A	X	X			X	X	X
B	X		X	X	X	X	

Figure 8: A Cross-classification of Raters and Occasions for One Student

Below the student the structure can be conceptualised as a cross classification of a pair of raters with occasions. The various cells evident are units at level 2 with measurements at Level 1 nested within the rater-occasion combinations. It may be noted that it is also a special case of a level 2 cross classification with at most only one unit per cell. This situation can be handled by multilevel analysis although there may be some implications which we later examine.

Of course measurements will be taken on many students. Suppose for the moment that each student has its own set of raters not shared with other students so that, for instance, rater A and B are unique to one student. With this scenario the cell units that have measurements are hierarchically arranged within students at level 3. This situation is represented more fully by the diagram in Figure 9 for three particular students. We can note that with the unique allocation of pairs of raters to student and with students at level 3, each student has separable blocks of rater and occasion crossings (shaded in the diagram). This feature can have advantages in easing the estimation of corresponding models (Rasbash et al 2004). We might also note a further lack of balance in that it possible for some students to have no measurements at all on some occasions. Student 3 in the illustration has measurements only

[9] Here the term rater is used as in the educational testing literature as a generic term for such words as assessor, examiner, judge, grader or marker.

on three of the seven occasions. These types of structures can be handled quite effectively with modern methodology.

	Student 1							**Student 2**						**Student 3**		
Occasion	1	2	3	4	5	6	7	1	2	3	4	5	6	1	4	7
Rater A	X	X			X	X	X									
Rater B	X		X	X	X	X										
Rater C								X	X	X	X	X				
Rater D								X	X	X	X	X	X			
Rater E														X	X	X
Rater F																X

Figure 9: Measurements in a Cross Classification at Level 2 with only One Unit Per Cell Within Students at Level 3

Suppose now and perhaps more commonly we have the same set of raters involved with all the children in the study so that it is possible for each rater to assess more than one child. Separable blocks as in the Figure 9 structure will no longer be evident. For simplicity of exposition we now assume just one measurement per occasion rather than two, so that measurements and occasions have the same logical status in the structure. We also now consider taking measurements at up to five occasions only. An extract from the design structure which shows three of the students with which three of the raters A, B and C are involved is shown in Figure 10. The level 1 units are as before measurements but since there is only one measurement per occasion they are also conceived of as measurement occasions. In this structure the raters are crossed with students at level 2 within which the measurement occasions are nested. Formally the structure of Figure 10 is a special case of the structure of Figure 6 and could be represented by a ‘unit diagram’ similar to Figure 7 but there are now many empty cells and at most one observation per cell.

	STUDENT 1				STUDENT 2		STUDENT 3				
Occasion	1	2	3	4	1	2	1	2	3	4	5
Rater A	X	X	X		X						X
Rater B				X			X	X			
Rater C						X			X	X	

Figure 10: A Cross –Classification of Raters with Student at Level 2 with Occasion Observations at Level 1

An extension is possible if for example there were a number of measurements per occasion and each was undertaken by the same rater. In this situation each X in the diagram would now represent several measurements (level 1) within a level 2 of occasions. The crossing of student and raters would now move up to be level 3.

Similar ideas of cross classification structures also occur commonly when a simple hierarchical structure breaks down. Consider, for example, a basic repeated measures design which follows a sample of students who are observed over time, say over three school terms, within a set of classes for a single school. A three level strict hierarchical structure would ensue; occasions nested within students who are in turn nested within class. Suppose, however, and this is not entirely unrealistic, that students change classes during the course of the study. Students can no longer be regarded as nested within particular classes. For three students, three classes and up to three occasions we might have the pattern of Figure 11. Formally this is the same structural type as in Figure 10. We now have a two level structure but with cross-classification of classes by students at level 2 and level 1 observational occasions with single measurements nested within them.

	STUDENT 1			**STUDENT 2**		**STUDENT 3**		
YEAR	**1**	**2**	**3**	**1**	**2**	**1**	**2**	**3**
Class A	X	X		X				X
Class B			X					
Class C					X	X	X	

Figure 11: Students Changing Classes

Outside the field of repeated measures Fielding (2002, 2005) has examples of the same formal structure as Figures 10 and 11. The level 1 unit there are GCE Advanced Level entries from students in a number of post-16 colleges and school sixth forms. Each student enters for a number of a A-level subjects each of which is taught in separate teaching groups. Students are thus crossed with teaching groups. A level entries are nested within students but are also nested within teaching groups. To incorporate both in a combined structure we have level 1 entries within a level 2 crossing of student and subject teaching group (class). Institutions are a level 3 above this in which the student and teaching group combinations are nested. This

structure facilitates the separation of teaching group effects from those of the students selected into them. An earlier analysis which had entries hierarchically arranged within teaching groups was shown to distort teaching group comparisons since it failed to recognise that such entries were not independent across groups, since groups had students in common. Put another way the student effect was not adequately controlled.

Let us consider now a bit more complexity to add to that of Figure 11 and include schools as another higher level. If students stayed at the same school these will be classified as level 3 units beneath which the student and class crossing is nested. However, students may also change schools during the course of an observational study. Students must then be crossed with schools at the top level 3. Classes are nested beneath at level 2 with occasions as the level 1 unit. The students have moved from being crossed with classes to being crossed with schools. Note that since students are crossed at level 3 with schools they are also automatically crossed with any units nested within schools so that we do not need separately to specify the crossing of classes with students to accommodate changes of classes within the same school. This important feature receives detailed stress in Goldstein (2003).

We now consider an example which in some senses extends the ideas of Figure 11 but now with predictable regular changes of class by the students. The situation is one where class groupings of students are re-formed for each period. We might now wish to think of teacher effects rather than class group effects since we assume each class is taught by the same teacher throughout a period and where we might suppose each new class group at each period has different teachers[10]. Such a structure with four different teachers over just two periods for three students is given in Figure 12. We keep to the situation of one school for relative simplicity of illustration, though with more elaboration it could form part of a larger structure in which schools are included at a higher level and also complexities of changing schools considered. We now have a cross classification of teachers by students at level 2 with one measurement on occasions as the level 1 unit. We note again a structure where most of the cells are empty and that there is at most one level 1 unit per cell. Raudenbush (1993) gives a research example of just such a structural design separating out the classroom teacher effects from those of student characteristics in a study of growth in mathematics achievements. Raudenbush and Bryk (2002) also elaborate this example further to illustrate the flexibility of design in such studies. They wish to study the cumulative effect on the growth of combinations of teachers as students' progress. They show how this can effectively be done by introducing a further cross classification of the set of teachers on one occasion with themselves on other occasions.

[10] We note that class units in many structures are taught by the same teacher and this teacher has only one class. In this situation the set of classes and teachers are the same. If this was the situation in Figure 11 then we might call the units class/teacher. The disentangling of class and teachers effects is not possible in this situation but has been a pre-occupation in research. We give an example of a design in which this could be tackled later in the report. Although in Figure 12 classes and teachers are again in one to one relation we label the units teachers since whole class groupings move from teacher to teacher and it becomes possible to envisage this as the effect of changing teachers.

		STUDENT 1		STUDENT 2		STUDENT 3	
	Period						
TEACHER A	1	X		X			
TEACHER B	1					X	
TEACHER C	2		X				X
TEACHER D	2				X		

Figure 12: Students Changing Teachers / Groups for Each Period

We have presented just a few examples in the context of repeated measures to illustrate some elements of complexity. We have focussed on these designs since they are very apt in exemplifying the complexity of structures which can arise in may research contexts. Similar designs will occur also frequently in panel or longitudinal surveys of individuals who move from one locality to another, or workers who change their place of employment. Although there is complexity enough in the examples they do not entirely reach the end of the story. Some further detailed accounts of even more complicated designs are offered by Goldstein (2003) and Raudenbush and Bryk (2002). However, by applying similar ideas as we have discussed a variety of scenarios with many elaborate features that can with some imaginative creativity be cast into meaningful structures using hierarchies and cross-classifications.

4.2 Some Notation For Cross-Classified Models

In the early sections we introduced some basic notation for models of hierarchical structures. In a basic two level model we denoted a level two random intercept effect for level 2 unit j by u_{0j} and its variance by $\sigma^2_{u_0}$. Level 2 effects are now more complex and we require extending the notation. Firstly we need to consider the cell combinations of the two factors. If we use j_1 to indicate a particular unit of the first classification and similarly j_2 for the second factor, we identify particular level 2 units which are now combinations of two units on each factor by (j_1, j_2). In basic models the level 2 effect is the sum of two separate random effects which we now denote as $\{ u^{(1)}_{0j_1} + u^{(2)}_{0j_2} \}$. Using a similar formulation as before but using the cell notation the model for level unit i within the level 2 unit (j_1, j_2) can now be written as

$$y_{i(j_1,j_2)} = X_{i(j_1,j_2)}\beta + u^{(1)}_{0j_1} + u^{(2)}_{0j_2} + e_{0i(j_1,j_2)}.$$

We note that this model is still valid even if there is no more than one observation in a cell. We still refer to basic level 1 variance by $\sigma^2_{e_0}$ but the level two variance is now $(\sigma^2_{u_0^{(1)}} + \sigma^2_{u_0^{(2)}})$, the sum of two additive components one from each factor in the cross-classification. If we have models with random coefficients for either or both classifications of the crossing or more elaborate complex level 1 variance functions then analogous model structures and notation can be extended. Further ways of classification and more levels in the hierarchy, with possibly cross classifications at higher levels also extend it further. However, the more complex the model the more elaborate are the notational conventions. Rasbash and Browne (2001) give the complex rules and some detailed examples. Later we will introduce a more general notation

which can be used in conjunction with a particular newer form of estimation known as MCMC.

Although we will not consider this in any detail it is also possible if an application warrants it to characterise variation due to cross-classified levels by the incorporation of an 'interaction term' in ways which will be familiar to those versed in two-way analysis of variance procedures. In our examples above we might for instance suppose that the marginal effect of residence in a particular area might differ according to which school the student went to or vice versa. In other words there is something about particular combinations of areas and schools which might make the additive contribution of an area effect and school effect for a particular cell unduly simplistic. If this is the situation the cell effect is now characterised as the sum of three additive components

$$\{ u^{(1)}_{0j_1} + u^{(2)}_{0j_2} + u^{(3)}_{0(j_1j_2)} \}.$$

The usual meaning of interacting effects follows. For example, the effect of belonging to j_2 no longer makes a straight added contribution to that of j_1 whatever the unit j_1 the student is in, as it would if $u^{(2)}_{0j_2}$ were just added to $u^{(1)}_{0j_1}$. Instead the added contribution to $u^{(1)}_{0j_1}$ for unit j_2 is now $\{ u^{(2)}_{0j_2} + u^{(3)}_{0(j_1j_2)} \}$ which depends on the unit j_1. The corresponding cross-classified level variance is the sum of three components is $(\sigma^2_{u^{(1)}_0} + \sigma^2_{u^{(2)}_0} + \sigma^2_{u^{(3)}_0})$ where $\sigma^2_{u^{(3)}_0}$ the variance of the interaction effect is known as the 'interaction' variance. It should be said that although interaction terms allow for greater flexibility and specificity not many applications arise in the literature, possibly because the additive characterisation has proved adequate for most purposes.

4.3 An Example Analysis: Sixteen Year Examination Performance

Goldstein and Sammons (1993) examine data on GCSE results on students of the Junior School Project (Mortimore et al (1988)) in order to develop further insight into the continuity of school effects. In particular, it was desired to see what carry over effects the primary school attended might have on their progress at secondary school. It uses a cohort of 758 students in 48 junior schools that went on to 116 different secondary schools. The essential feature is that level 2 was a cross-classification of the different combinations of junior school and secondary school that the students (Level 1) attended.

For illustrative purposes Table 2 only presents results for a selection of four different model fits the researchers discuss.[11]

[11] This table draws directly on the Goldstein and Sammons (1997) source and in keeping with the preference of some researchers does not incorporate a starred convention to indicate significance of results as is common in some areas particularly in econometric regression results. We have not felt at liberty to embellish the table in this way nor would we really wish to do so. In Footnote 3 we have commented on this practice of starring results and explaining why it is often not as informative as it might be given that inferential content is fully covered by the estimate values and their estimated standard errors. In the context of the models in which they are presented all fixed effect estimates, apart from the 8 year scores are more than 3 times their their standard errors. As such if it were desired to refer them to the appropriate test null distribution they would all be significantly different from zero beyond the 1% level.

	Model A	Model B	Model C	Model D: No Junior random effect
Fixed effects				
Intercept	0.25	0.50	0.15	0.50
Male	-0.34 (0.07)	-0.19 (0.06)	-0.22 (0.06)	-0. 22 (0.06)
Free school meal	-0.37 (0.08)	-0.23 (0.06)	-0.22 (0.06)	-0.22 (0.06)
VR2 Band		-0.38 (0.08)	-0.36 (0.09)	-0.37 (0.09)
VR3 Band		-0.71 (0.13)	-0.66 (0.14)	-0.65 (0.13)
LRT Score		0.32 (0.04)	0.29 (0.05)	0.32 (0.04)
8-year English score			0.00016 (0.0020)	
8-year Maths score			0.0058 (0.0056)	
Random (variances)				
Junior Schools	0.054 (0.024)	0.036 (0.013)	0.025 (0.014)	
Secondary Schools	0.019 (0.054	0.014 (0.014)	0.016 (0.014)	0.028 (0.015)
Level 1 (student):				
Males	0.940 (0.05)	0.682 (0.04)	0.686 (0.04)	0.680 (0.04)
Females	0.740 (0.05)	0.554 (0.06)	0.500 (0.04)	0.520 (0.04)
LRT			0.031 (0.021)	0.030 (0.02)
Covariance of LRT and intercept			0.093 (0.018)	0.10 (0.02)

Table 2: Analysis of 16 Year Examination Results (GCSE score) with Cross-Classified Random Effects for Secondary School and Junior School (Estimated Standard Error of Effect Estimates in Parentheses). Source: Goldstein and Sammons (1997)

Note**:** The outcome GCSE score and the London Reading Test score have been empirically transformed to approximate standard normal distributions. Free school meal (FSM) is binary (yes/no) for each pupil as is the Male dummy for gender. Scores on a verbal reasoning test at the end of primary school are banded into three groups VR1 (highest), VR2, VR3. In the model there are dummy indicators for VR2 and VR3 with VR1 as the reference category.

Model A is a very basic model including only gender and eligibility for free school meals (FSM) as explanatory variables. It will be noted that as a refinement to a usual basic model different variances are allowed for males and females[12]. The level two random effect is characterised as the sum of a primary school effect and a secondary school effect. We see that the estimated variance between effects of Junior Schools is higher than that of Secondary Schools by a factor of around three. One reason for this may be that secondary schools are on average far larger than primary schools so that the sampling variance is smaller. Such an effect will often be observed where one classification has far fewer units than another, for example where a small number of schools are crossed with a large number of smallish areas

[12] Allowing level 1 variance to be complex in this way often adds useful and interesting insights. Here we note that residual variation due to unobserved characteristics amongst boys is higher than that amongst girls. This has been regarded as of interest in its own right. The dependence of variance on individual characteristics is often known in the statistical literature as heteroscedasticity.

of residence. In such circumstances we need to be careful about the interpretation of the relative sizes of these variances. However, this caveat apart, we can see that primary schools attended can exert some noticeable effect on later educational achievement in addition to that of the secondary schools themselves. To further study the issue of size of unit we could make the two between school variances a function of their sizes in a similar way as we have allowed level 1 variance to depend on gender.

In Model B further explanatory variables, LRT score and verbal reasoning group indicators are added. These reflect measured achievement at the end of primary schooling and may be regarded as input controls on the later secondary school outcomes. As expected these variables explain relatively more of the Junior School variance. From one point of view if our aim was to judge the relative effects of Secondary and Primary School on secondary school performance we should not adjust for the variables added to model B since they may be regarded as primary school outcomes and hence are primary school observed effects. They will obviously absorb a large amount of variance attributable to primary schools which were previously unaccounted for. Thus from this point of view such comparisons might be made based on models similar to A. From another point of view Model B might be more useful if the interest was in 'value added' by secondary schools after adjusting for intake performance and other unobserved effects of Junior School attended (as captured by the residual Junior School random effect). Value added issues aside, however, using Model B to directly compare the effects of Secondary and Junior on the secondary outcome might be inappropriate; the Junior School variance has been explicitly deflated by control for the Junior School attainment outcomes. These outcomes may be considered to be part of the Junior School effects on secondary progress but are not reflected in the Junior School variance in Model B which only summarises the influence of other unobserved Junior School factors.

This example is elaborated in Goldstein and Sammons (1997) where further refinements and complexities are considered. The possibility of a differential effect of LRT between secondary schools in particular is considered by making the coefficient of LRT random across schools in ways outlined in Section 2.5. This proved uninformative in an extended model specification where additionally the level 1 variance which was allowed to depend on LRT score.[13] The latter idea is incorporated into Model C of Table 2 which also adds as potential explanatory variables Maths and English scores obtained at entry to Junior school at aged 8. The latter effects are small and insignificant once the Junior School outcomes are controlled. A Model C type formulation is concluded by Goldstein and Sammons (1997) to be most appropriate for value added measurement for Secondary schools since it effectively controls for influences prior to entry to secondary schools. However, in this situation the eight year scores might be dropped without appreciable effect.

---

[13] It may be noted that in Model C and Model D where a variance term is added for LRT at Level 1, an additional covariance term is added for possible correlation between the intercept disturbance and that due to different LRT scores. The complex variance at Level 1 then allows heteroscedasticity as a quadratic function of LRT which is often observed empirically. Thus using standard statistical formulae for linear combinations of random variables we have, for example for boys in Model C the variance function :

Variance of intercept +2LRT *(covariance of intercept and LRT random effects)+ LRT^2 * Variance of LRT effect= 0.686 + 0.186LRT +0.031LRT^2. Similarly the variance for girls has the same linear and quadratic coefficients but is shifted downwards to 0.500 + 0.186LRT +0.031LRT^2, reflecting the smaller variance on average for girls.

We note as reported by Goldstein (2003) that these results, though slightly more elaborate, are broadly in line with similar results for Scottish data from schools in Fife (Paterson (1991)). Here, firstly an analysis removing the Junior School cross classification is also reported. This is a two level analysis with secondary schools only at level 2 and with a prior verbal reasoning score at entry as pupil level control. Secondly this is contrasted with a cross-classified analysis similar in type to Model A in Table 2 without the prior verbal reasoning control. The between secondary school variance is smaller but by not much in the first case than it is in the second. Thus the impact on secondary school effect variation of individual level initial ability control is similar to control exercised through primary school overall effects. However, a third analysis adds prior verbal reasoning at the individual level to the cross-classified model and here the secondary school effect variance becomes quite small. On this evidence some tentative conclusions were made about standard analyses such as the first one, which is typical of many school effectiveness studies which control for initial achievement. These analyses are often used to highlight apparent differences in progress effects between secondary schools. However, from the second type of analysis similar to Model B in Table 2 we might note that many of these differences may be due to many aspects of the primary school experience of the pupils, which were formerly left uncontrolled. A natural question arises as to what adjustments should be made in typical 'value added studies' before they can be concluded to truly reflect the contribution of the secondary school to progress. The examples discussed illustrate that adjusting for achievement at a single previous time period and usually at entry may not be entirely adequate. Possibly other aspects of the child's former experience including the primary school attended should be controlled. However, Snijders and Bosker (1999) report on a similarly structured example for a Belgian data set of Opodenaaker and van Damme (1997) using mathematics test score taken at the end of second grade of secondary schools. Here the effect of primary schools on secondary mathematics is not so marked.

To further elaborate the present example in a similar way, Model D in Table 2 retains the main specifications of Model C, except 8-year score effects, but excludes the Junior School classification. Adding 8 year scores to Model D as in Model C has little effect on this analysis. Model D is thus a 'standard' secondary school effectiveness model. It may be seen that ignoring heterogeneous effects between Junior Schools inflates the Secondary School effect. Since Secondary School variance in Model D is 75% higher than that of Model C, stronger differences between secondary schools in progress may be asserted than are really warranted. Secondary school residual estimates from models are usually taken as indicators of effectiveness. Thus the two models would not normally rank these school effects in the same order either unless primary school intake across secondary schools is homogeneous, which is unlikely.

5 More Complex Structures: Multiple Membership

5.1 The Idea of Multiple Membership

Multiple memberships are another type of complex structure where we may desire to disentangle effects and is closely related to cross-classified effect modelling. It arises in situations where lower level units in a hierarchy can be members of more than one higher level unit simultaneously. In an education system, for example, students can attend more than one institution. Suppose we are studying progress in secondary school from age 11 to Key Stage 3. It might evident that the whole of the educational experience from age 11 has relevant effects. For a student who was in the same school for the whole of that time the school effect might be unequivocal. However, a student who changed schools might have

been exposed to the 'effects' of more than one school. If in addition the level 2 in the hierarchy was a cross classification of area of residence and school the student might also have changed areas of residence and exposed to more than one area effect. There is evidence in the educational research literature that such movers have different progress profiles (Yang et al (1999)). Ignoring such multiple membership characteristics of certain students, for example by allocating them only to the units they belonged to at the response occasion, might distort analyses.

We might note that we are still concerned with disentangling and controlling for the higher level effects of particular schools (and possibly areas) in the hierarchy. However, the question arises as to how we might model these effects for observations where more than one of these school effects might be making contributions. A basic necessity is to assume a priori that for each higher level unit to which a lower level unit belongs there is a known weight (usually summing to unity for each lower level unit) to apply to the school effects. This may represent, for example the proportion of time spent in each school by the student. The choice of weights is to some extent subjective but can be important. For instance it might be thought that more recent school experience has a greater impact and might counteract to some extent the time experience. On the basis school of length of time at schools the proportion of time in the current school out of five years might be 0.2 representing the fact that the student spent only one year in his current school in the run up to KS3. Another secondary school which the student attended from age 11 to his transfer might have on this basis a weight of 0.8. Subjectively though it might be thought preferable to attach alternative weights of 0.4 and 0.6 to account for possible greater impact of more recent experience. In practice a sensitivity analysis might be carried out to determine how alternative choices of weights affect model results and inferences, if at all. Fielding (2002), for instance considered teacher effects on GCE Advanced level performance. The teaching groups in the sample data extend over two years of a course and it is the norm for more than one teacher (often two or three) to handle it. The weighting scheme adopted on the basis of timetable information was to weight each teacher involved in a group by the proportion of the course they taught. However, experimentation with different weighting schemes were carried out in exploratory work; in particular giving greater weight to teachers in the second part of the course. It turned out that the main results were relatively robust to choices of scheme except for extreme ones such as ignoring the multiple membership of responses to teachers and allocating them only to the most recent teacher. This strict hierarchical model is equivalent to giving the most recent teacher a weight of unity and the rest zero.

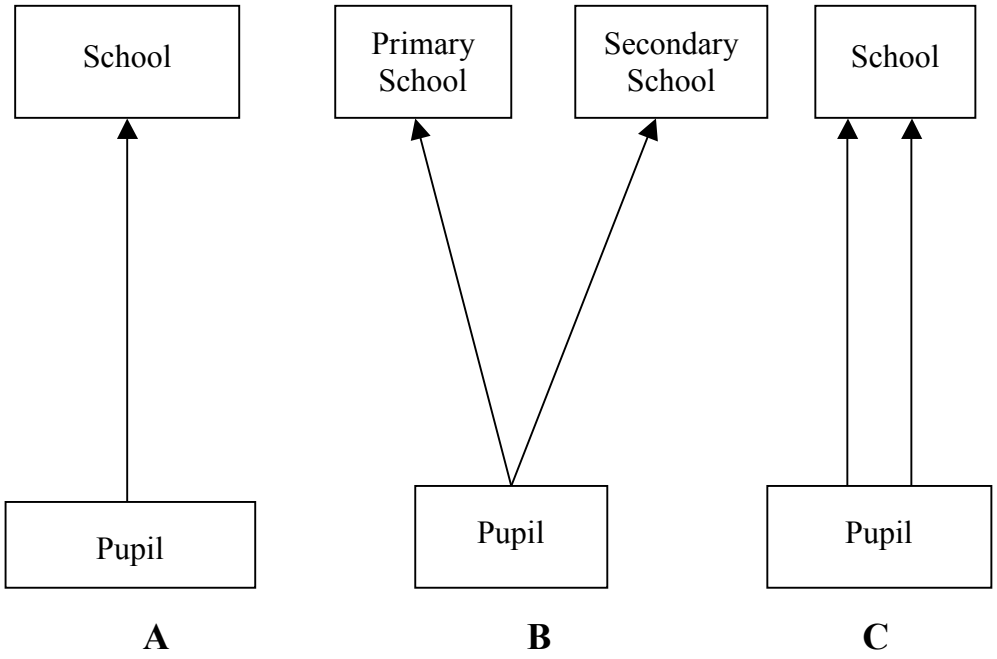
A similar situation which might be conceived as a multiple membership structure is found in data in which there is some uncertainty about which higher level units some or all lower level units belong to. Occasionally there may be ancillary information which narrows this uncertainty down a little. In a student survey, for example, information about smaller neighbourhoods of residence such as census output areas may not be available. We might, however have other information such as the ward they live in which may make it possible to assign weights to the output areas within that ward according to probabilities of belonging to each conditional on this information. In the absence of further information equal probabilities for those areas might seem sensible. As a further example in model cross-classifying schools by areas we might have missing information on area of residence for some students in a school. Aggregate school information on the relative frequency distribution of its students over areas might be a basis for assigning area probabilities to students whose area is not known. Such structures are referred to as a missing identification structures. They can be handled by similar procedures as those for multiple membership structures. Hill and Goldstein

(1998) discuss more fully both multiple membership and missing identification models with further examples. We take up this example again in Section 7.6.

5.2 Classification Diagrams And A More General Notation For Cross-Classified And Multiple Membership Structures

The notation and formal representations of models introduced earlier was that which was first introduced in the early development of multilevel methodology. They explicitly assist in gaining an intuition about the structure of hierarchical effects. They are also used in many published applications and it is mainly for this reason that they are considered in this review. Anyone making detailed examination of some of the applications reviewed will benefit from this prior exposure to these formalities.

However, as models and structure become more complex this notational framework though still explicit becomes somewhat cumbersome and sometimes difficult even for those with more than a passing acquaintance with statistical modelling. For these reasons Browne et al (2001) introduced a more simplified notation for model description. This made certain structural features of the models implicit but which could then be understood by association with 'classification diagrams', which were introduced as a schematic representation of structures. Taken together the description and diagram facilitate an understanding of the nature of complex structures. Figure 13 sets out the simplest types of such diagrams



Key:

A: 2-level hierarchical model

B: 2-level cross-classification at level 2

C: 2-level multiple membership model

Figure 13: Classification Diagrams for Two Level Hierarchical, Cross-Classified and Multiple Membership Structures

Such diagrams enable us to classify data structures as hierarchical or crossed or combinations of these at various levels or to see how multiple membership fits into the frameworks. Boxes

represent unit classifiers at various levels and those at the same horizontal level in cross-classifications are at the same conceptual level in the hierarchy. Double arrows indicate multiple membership relationships. Thus in diagram C in Figure 13 we have possible multiple membership of schools by students. We should note that this diagram does not imply that all students are in a multiple membership relation. It admits the possibility for some students. Many if not most students may be hierarchically arranged beneath just one school.

The newer model notation does not involve complicated multiple subscripts but to understand the structure involved should be read in alongside its corresponding classification diagram. In this new notation the basic two level variance component hierarchical structure associated with diagram A in Figure 13 is written as

$$y_i = (X\beta)_i + u^{(2)}_{school(i)} + u^{(1)}_{student(i)}$$
$$school(i) \in (1, 2, ...J)$$
$$student(i) \in (1, 2, ...N)$$

$$u^{(2)}_{school(i)} \sim N(0, \sigma^2_{u(2)})$$
$$u^{(1)}_{studentl(i)} \sim N(0, \sigma^2_{u(1)})$$

We have two of what are now termed classifications, with students as classification 1 and schools as classification 2 as indicated by the superscript on the random effects. The subscript i is attached to lowest level unit classifier, in this example the student, and there are as many units in this classification as there are responses in the data. This subscript uniquely identifies every measurement and random effect. The use of the set notation $student(i) \in (1, 2, ...N)$ is a reminder that a student labelled i is one of the N level 1 units and hence observations in the data. The set symbol $\in$ means 'is an element of' or 'belongs to'. The index i can thus range over possible values 1 through to N. The school (i) identifier is taken to mean the school that student i belongs to. Formally school (i) is called a classification function that maps the lowest level units, students, onto schools. The set relation $school(i) \in (1, 2, ...J)$ simply means that the result of this assignment is one of the J possible schools in the data. We note again that it is also the usual convention that we specify the assumptions we are making about the distribution of the various random effects. Normality is usually assumed, although estimation and inference procedures are often robust to departures from this. Thus in the above formal representation, $u^{(2)}_{school(i)} \sim N(0, \sigma^2_{u(2)})$ means that the level 2 school random effect is distributed normally around zero with the variance component $\sigma^2_{u(2)}$. In keeping with the notational consistency the lowest level student variance within school is denoted by $\sigma^2_{u(1)}$. Taken together with the supplementary diagram A of Figure 13 the above formalisation serves to completely specify the model and its structure.

Together with diagram B in Figure 13 the basic cross-classified model under consideration can be written as

$$y_i = X_i\beta + u^{(3)}_{primary\ school(i)} + u^{(2)}_{secondary\ school(i)} + u^{(1)}_{student(i)}$$

$$primary\ school(i) \in (1, 2, ..., J_3)$$

$$secondary\ school(i) \in (1, 2, ..., J_2)$$

$$student(i) \in (1, 2, ..., N)$$

$$u^{(3)}_{primary\ school(i)} \sim N(0, \sigma^2_{u(3)})$$

$$u^{(2)}_{secondary\ school(i)} \sim N(0, \sigma^2_{u(2)})$$

$$u^{(2)}_{student(i)} \sim N(0, \sigma^2_{u(1)})$$

Since the only subscript used in such representations is that for the lowest unit writing down of models in this notation can be extended indefinitely for very complex structures involving any number of crossed or hierarchical sets of any units at many levels. Browne et al (2001) give a comprehensive treatment of the representations. It will be noted that the equations themselves do not explicitly show the nestings and crossings. A modern estimation procedure Monte Carlo Markov Chain (MCMC) which we review later also does not require knowing the exact nesting structure providing there are unique identifiers in the data for each classification. Thus this sits very easily with the notation. However, associating each model with a classification diagram will reveal the structure of effects for easier communication and interpretation.

The basic structure of the two level classification diagram C of Figure 13, in which pupils are multiple members of schools, can also use this new notation. However, for the sake of completeness, and because it may clarify certain features when we move to the new notation, we first of all write the model in the explicit if somewhat tricky older notation. This will also be useful if further reference is required from this review to the original detail of published applications which until very recently used this older notation. In this, following but slightly adapting Hill and Goldstein (1998) and Rasbash and Browne (2001) the model is written as

$$y_{i\{j\}} = X_{i\{j\}} + \sum_{h=1}^{J} u_{0h}\pi_h + e_{0i\{j}$$

$$u_{0h} \sim N(0, \sigma^2_{u_o})$$

$$e_{0i\{j\}} \sim N(0, \sigma^2_{e_o})$$

$$with \sum_{h=1}^{J} \pi_h = 1 \ \ for\ each\ level\ 1\ unit\ (pupil)$$

We note that {j} now means the full set of school {1, 2,……J} and it is included as a subscript in the various quantities to emphasise the two level multiple member nature. The level 1 units are again indexed uniquely by i and may be members of some (or even all) of the schools in {j}. The index h uniquely indexes schools. The π_{ih} are the pre-assigned weights using criteria we have discussed. We emphasise again that for each pupil they usually sum to unity. For example suppose we have identified that a particular pupil 3, say has attended two schools, numbers 7 and 59 with weights 0.6 and 0.4 respectively, the model expression becomes

$$y_{3\{j\}} = X_{3\{j\}}\beta + 0.6u_{0,7} + 0.4u_{0,9} + e_{0,3\{j\}}.$$

We note that the weights attached to other schools in the set {j} are all zero. In most practical applications this will be the case and most of the weights entering into

$\sum_{h=1}^{J} u_{0h}\pi_{ih}$ and $\sum_{h=1}^{J} \pi_{ih} = 1$ will be zero. Also the majority of pupils may not be in a multiple member relation and attend only one school. For such a pupil the weight will be unity for this school and zero for the rest. Thus suppose pupil 7 attends only school 6 then the expression reduces to one familiar for strict hierarchical structures:

$y_{7\{j\}} = X_{7(j)}\beta + u_{0,6} + e_{0,7\{j\}}$.

However, in classification notation providing we interpret in association with diagram C of Figure 13 we can write the full model as

$$y_i = X_i\beta + \sum_{h \in school(i)} u_h^{(2)}\pi_{ih} + u_{student(i)}^{(1)}$$

$$\sum_{h=1}^{J} \pi_{ih} = 1 \quad for\ each\ level\ 1\ unit\ i\ (pupil)$$

$$student(i) \in (1, 2, ..., N)$$

$$school(i) \subset (1, 2, ..., J)$$

$$u_h \sim N(0, \sigma_{u(2)}^2) \quad for\ h \in \{j\} = \{1,2,.....J\}$$

$$u_{student(i)}^{(2)} \sim N(0, \sigma_{u(1)}^2)$$

Most of the components of this characterisation follow from our previous discussion of the notation. However, we can note the use of the set symbol $\subset$ in the definition of the classification function *school (i)* rather than as previously $\in$. The latter would uniquely assign a school to case i. However the function can now assign more than one school in the set (1,2,3…….J). The symbol $\subset$ means 'is a subset of ' and can be interpreted as *school (i)* assigning possibly more than one of the elements of (1,2,3…….J) to the case. Browne et al (2001) also consider how the diagrams and model specifications can be extended to more general models with more levels, combinations of hierarchical, cross-classified or multiple membership structures and also regression coefficients random across various classifications.

5.3 Examples Of Application Of Multiple Membership And More Complex Structures

5.3.1 Teachers, teaching groups and students in GCE Advanced Level Results (Fielding (2002)

The data came from a wider study of the cost-effectiveness of GCE Advanced Level teaching groups in six colleges in England, which at the time of data collection were funded by the Further Education Funding Council. Part of this study was designed to highlight various effects on outcomes before attributing per capita costs to student provision. The basic response was the A level points score on 3683 level 1 units, subject entries. These entries were hierarchically nested within 314 different subject teaching groups within the six colleges. There was a nesting of groups within colleges to effectively form a three level hierarchy. However, since there were so few colleges these were handled by college fixed effect indicators rather than random effects. Thus the structure will be effectively two level for modelling purposes. A total of 1511 students were involved in the entries with students taking several subjects and being contained in several teaching groups.

Thus entries are also nested within students. It was desired to control for student effects before identifying group effects which could then be related to costs in the wider study. We thus form a cross-classification of student by teaching group at Level 2 in which entries are nested. In addition it was desired to control for teacher effects not only to properly adjust the unobserved teaching group effects but also to study teacher effects in their own right. However entries were in a multiple member relation to a total of 115 teachers since several teachers may be involved in the particular subject group provision throughout the two years of its teaching. A majority of teachers were also involved in more than one subject group so the possibility of confounding of teacher and group effects was minimal. Thus we have an additional classification at level 2 by teacher to make a three way classification. The weights chosen in the analysis presented were the proportion of time of the teaching provision for the A level entry from each teacher handling the group. As mentioned above the actual choice of weighting scheme proved relatively robust. A classification diagram for this situation is given in Figure 14.

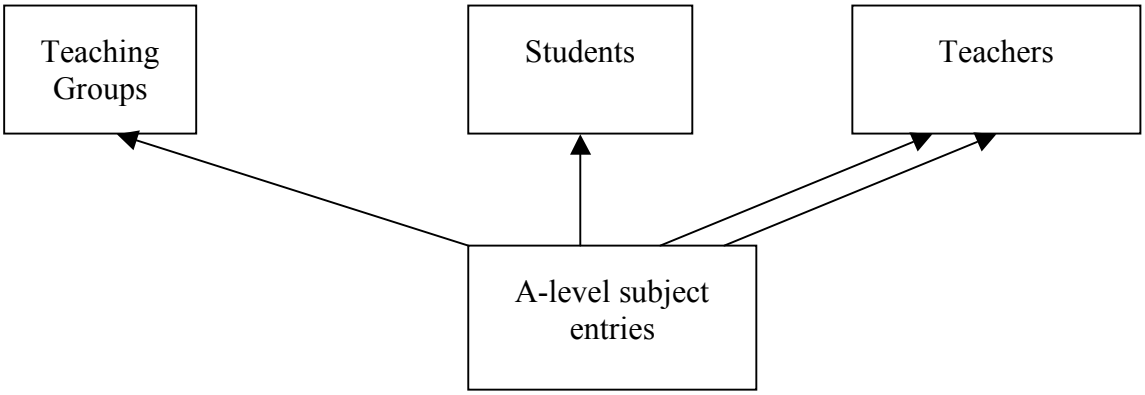


Figure 14: A Three Way Cross-Classification of A Level Entries by Groups, Students and Teachers at Level 2 with Multiple Membership of Entries Across Teachers

A brief model specification to be read in conjunction with Figure 14 is

$$y_i = X_i\beta + u^{(2)}_{student(i)} + u^{(3)}_{group(i)} + \sum_{h \in teacher(i)} u^{(4)}_h \pi_{ih} + u^{(1)}_{entry(i)}$$

$entry(i) \in (1, 2, ...,3683)$
$student(i) \in (1, 2, ...,1511)$
$group(i) \in (1, 2, ...,314)$
$teacher(i) \subset (1,2,.....115)$

In the course of development various specifications of fixed effect explanatory variables in X_i were used and models estimated by the variant of iterative generalised least squares, to be explained in a later section. The results of one such analysis are displayed in Table 3.

Fixed effect coefficients (β)	**Estimate**	**Estimated standard error**
Intercept (Base : small FE college 1, social science subject, male teacher, teacher has no degree, teacher has no teacher training)	-4.986	
Student GCSE Average (prior ability control)	1.732***	0.057
Standardised teacher age	-0.014	0.017
Standardised teacher experience (years of service)	0.021	0.020
Teacher has a degree but no higher degree	0.034	0.513
Teacher has a higher degree	0.204	0.542
Teacher has a Cert. Ed	0.357	0.410
Teacher has a Postgraduate Certificate of Education	0.434**	0.148
Female Teacher	0.374	0.267
Random effects variance		
Level 2: Students	2.761	0.118
Level 2: Teaching groups	0.304	0.112
Level 2: Teachers	1.540	0.257
Level 1 Residual for entries	3.910	0.201

Table 3: A Level Entry Performance in Six Post-16 Colleges. A Model Cross Classifying Student and Teaching Group Random Effects and Multiple Membership Across Teachers.

*** Significantly different from zero with p-value: $p<0.001$

** Significantly different from zero with p-value: $0.0001<p<0.01$

Note: Dummy variable indicators were included for college fixed effects and for eight broad groupings of subject discipline type (see Fielding (2002) for full details) but for brevity these are not displayed. The teacher characteristics as fixed effects were formed as weighted averages of the values for individual teachers contributing to an outcome using the same weights as in forming the multiple member random effect. Teacher age and teacher experience were then standardised across the data to have mean zero and standard deviation unity.

Although there is a full discussion in the original source a few important points may be mentioned. Firstly even after controlling for their original ability there is still a large amount of effect variation due to unobserved characteristics of students. It was thus important to control for these as the model does before going on to estimate teaching group residuals as a basis for adjusted group effects. Similar comments apply for control of teacher fixed and random effects. Indeed after control for teacher and student the teaching group variance is now relatively small. A stark conclusion of this analysis is that teachers do matter. However trying to explain teacher influence in terms of standard background characteristics such as age, experience, training or educational level is inconclusive on this evidence. Apart from the Postgraduate Training indicator, none of the teacher fixed effect characteristics are statistically significant and hence discernable in this data. The large variance of the residual random teacher effect indicate the importance of as yet unobserved or undiscovered teacher characteristics worthy of more detailed research.

We must, however, in such multiple members relations be careful not to over-interpret the relative sizes of the residual variance components at level 2. At first sight on converting to percentages it might appear that the relative residual contributions of students, groups, and teachers random effects to grade variation are respectively 60%, 7%, and 33%. This might be an appropriate assessment for cases where there is only one teacher involved. However, using basic statistical theory the teacher contribution to the model variance at level 2 will be $\{\sigma^2_{u(4)} \sum_{h \in teacher(i)} \pi^2_{hi}\}$ where $\sigma^2_{u(4)}$ is the residual variance amongst teachers. Thus for example if there are two teachers involved with equal weights 0.5 say the contribution to the overall level 2 variance in such cases will be $1.54*(0.5^2 + 0.5^2)= 0.77$. The relative contributions of students, groups and teachers are now 71%, 9%, and 20%. The combined teacher effect is now less influential. This accords with intuition. We might expect that the more teachers there are involved the more there will be dilution of effects of particular ones and high beneficial effects of some teachers will be counteracted by weaker effects of others.

5.3.2 Spatial models using multiple membership relations

In studying area effects, measurements on individuals within an area may be supposed to be influenced by both by an effect of that area and the effects of surrounding contiguous areas. Intuitively, this idea might be considered reasonable. Areas used in applications, such as electoral wards or postcodes are administrative constructions and may be somewhat artificial for assessing spatial impact on the process under study. Neighbouring areas which are close to one another geographically may share social or economic factors influencing outcomes in the area of interest. In studying area effects on educational progress or health status it may be realistic then to include not only administrative area of residence but to effects of surrounding areas. One way of doing this is to include a random effect for area of residence but also separately a multiple membership effect term for surrounding neighbour areas. We are here modelling two separate random effects for area of residence (classification 2) and neighbouring areas (classification 3). Figure 15 gives a basic classification diagram for this situation.

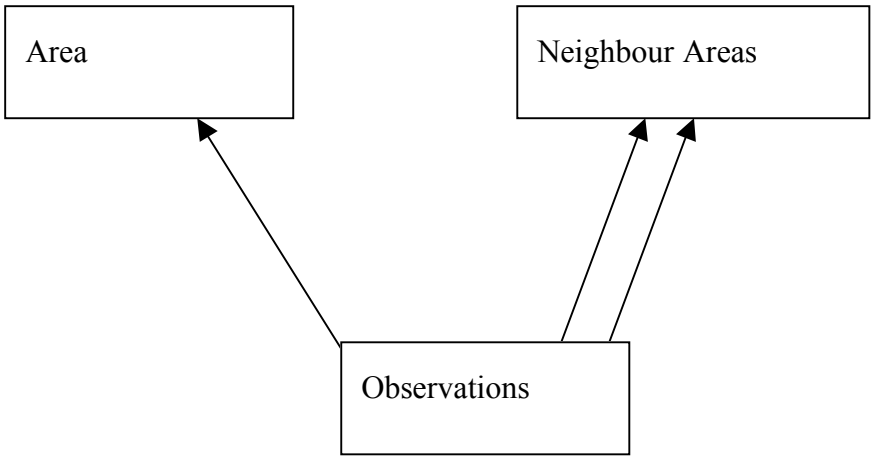


Figure 15: A Cross Classification of Area of Residence and Neighbouring Area in a Spatial Model, with Multiple Membership of Several Contiguous Neighbour Areas.

The corresponding model is expressed as

$$y_i = X_i\beta + u^{(2)}_{area(i)} + \sum_{h \in neighbour\ area(i)} u^{(3)}_h \pi_{ih} + u^{(1)}_{level1\ unit(i)}$$

$$level1\ unit(i) \in (1, 2, ..., N)$$
$$area(i) \in (1, 2, ..., A)$$
$$neighbour\ area(i) \subset (1, 2, ..., A_N)$$

The choices of equal weights for neighbours suggest itself unless there are some good reasons to the contrary. Thus if area(i) has n_i neighbours so that there are n_i terms in the sum for the multiple member effect then the weighted effects expression becomes $\sum_{h \in neighbour\ area(i)} u^{(3)}_h / n_i$. In some cases all the areas in the data are neighbours to other areas so that the set $(1, 2, ..., A)$ is the same as the set $(1, 2, ..., A_N)$. Thus there are two contributory random effects in the model for each particular area though they arise from two conceptually distinct classification sources. We might, however expect these two effects to be correlated so we can in such cases extend the model to allow for this by incorporating covariances for the two effects. Using equal neighbour weights Langford et al (1999) do this for a model of mortality using postcode sectors in 1993 for the Greater Glasgow Health Board. Table 4 below is adapted from the results in this source. The correlation between an area's direct effect and its neighbour spatial effect is 0.774 suggesting that areas have similar relative effects when they are contributing as neighbours and directly as area of residence. It might be noted that the neighbour effect has a larger variance than the direct effect. However, we should remember the caveat of the previous section The total level 2 variance for mortality in an area(i) is dependent on the number of neighbours n_i and with equal weights is given by

$\sigma^2_{u(2)} + \frac{\sigma^2_{u(3)}}{n_i}$. As Langford et al (1999) point out, 'the mean number of neighbours per postcode sector in the Greater Glasgow Health Board is 5.4; this implies a total variance of 0.0333 (*on average*), of which 49% arise from the spatial neighbour effects'

Variances and covariances in mortality	Estimate	Estimated standard error
Area of residence effect ($\sigma^2_{u(2)}$)	0.0174	0.0054
Effect of area on neighbouring area's mortality ($\sigma^2_{u(3)}$)	0.0865	0.0085
Covariance of area effect and neighbour effect (($\sigma_{u(23)}$)	0.0300	0.0334
Correlation of an area direct effect and its neighbour effect	$(0.03)/\sqrt{(0.0174)*(0.0865)}$ =0.774	

Table 4: Postcode Area Components of Variance and Covariance for Glasgow Mortality Data

Such spatial models may have potential use in the education field where area effects to proxy student background effects may be under consideration. Goldstein (2003) also raises the possibility that they may be used to model interacting schools. Thus for example the effects of schooling may come not only from a school a student attends but also from neighbouring schools that may be competing for resources, or share a governing body, etc. A model similar to the above might be appropriate for this situation possibly also with more complexity by adding other classifications such as area of residence together with a spatial model for neighbourhood areas. Leyland (2001) considers variants of these spatial models for the incidence of lip cancer in Scotland over the period 1975-1980. This example is also mentioned by Browne et al (2001) in the context of current MCMC implementations which are constrained in that they do not permit correlations for effects arising from different classifications.

Such models can be complicated even further if for instance we are studying the continuity of effects. We may for instance note that a student changes area of residence over the course of his time at a school. In this case we might want to weight areas in some way so that instead of a single term $u^{(2)}_{area(i)}$ and *area (i)* being a one to one mapping, we have a weighted term for several areas of residence. Of course if we have also included neighbouring areas as an effect these will also change and their number under consideration may obviously increase. For example suppose an individual has two areas of residence with weights 0.3 and 0.7, attaching perhaps more weight to current area. The neighbours of both areas will also now enter into the weighted multiple member term for neighbour and we would have to calculate adjusted weights for these to reflect the changes. In educational settings we could also model the dynamics of school mobility and effects of neighbouring schools in this type of model in a similar flexible way.

6 Estimation Methodology And Software issues

6.1 Introduction

At this stage it is worth discussing the distinction between multilevel models and their estimation and a body of methodology called Generalised Estimation Equations (GEE) for so called 'marginal models' (Zeger et al (1988), Liang et al (1992)). When dealing with data these latter models start with a formulation for the covariance structure for the non -fixed effects part of models. For example, but not necessarily, they may be based upon the structure of multilevel random effects such as we have been examining. However, they aim to provide estimates with acceptable properties only for the fixed parameters of the model. They treat the existence of any random effects and associated variance parameters as a necessary nuisance and without providing any explicit estimates for them. More generally GEE techniques for marginal models have useful properties for large samples when the exact nature of the random structure is unknown. Thus if interest lies mainly in the fixed parameter coefficients this approach may be useful. Even here, though they may be statistically inefficient if an assumed random structure is inappropriate. However the central substantive distinction is that marginal models seek to answer a different range of research questions. From a multilevel perspective the fact of not modelling random effects explicitly means they do not offer information on sources of variability that potentially may be even more important knowledge than average or conditional effects of explanatory variables. Thus in modelling student progress knowledge of variation between schools and how this variation depends on school factors will be important data. Although the structure may inform assumptions about the residual covariances in GEE it does not explicitly consider such information. Much the same considerations apply to estimation procedures we noted previously which use robust estimators for standard errors which recognise clustered data structures but the main model estimation still operates within the framework of marginal model ideas. Lindsay and Lambert (1998) discuss further the limitations of marginal models. The glossary by Diez Roux (2002) is also useful in clarifying the distinctions.

6.2 Approaches To Estimating Complex Multilevel Models

There are two broad approaches to estimation of *linear* multilevel models for complex structures. Up until fairly recently it has been mostly done within the general Maximum Likelihood (ML) framework though technical details of its implementations which often require approximations to make it computationally feasible vary across different softwares. The method is based on the idea of choosing estimates of parameters that maximise the probability (produce the ML) of observing the data that are actually observed given the model. More recently Bayesian estimation methods such as Monte Carlo Markov Chain (MCMC) estimation for have been developed. The actual technicalities of this approach to estimation and its philosophy are quite complex and beyond the scope of this review. However, it might briefly be noted that Bayes procedures adopt a slightly different approach than the standard 'frequentist' one to statistical inference in what is often referred to as the 'Fisherian' tradition. The essential idea is to regard the unknown model parameters to be estimated not to have fixed but unknown values but to express uncertainty about them by supposing their values are themselves governed by a probability model. Prior information on this uncertainty is then used along with collected data to update the knowledge of this uncertainty. The Bayesian method was popularised by a paper of Lindley and Smith (1972). Within the Bayesian framework the MCMC designation refers to the particular way this updating of knowledge of parameter uncertainty operates; by simulating the complex probability models for them. The method has become feasible for complex statistical models due to rapid advances in the computational environment which such a simulation approach requires. Modern books such as Congdon (2001) and Lawson et al (2003), with software guides such as Spiegelhalter et al (1997) and Browne (2002) start with very readable

introductions and many useful examples. There are philosophical and methodological features which might make a Bayesian approach attractive but which are still hotly debated amongst statisticians. However aside from this some writers have suggested other motivations for their use. One arises from problems that may arise in using other approaches due to size and complexity of the data (Browne et al (2004)). For instance, in some of the complex structures we have discussed the sheer number of random effects may make implementation of ML need large amounts of memory. It may often become computationally infeasible whatever software implementation is used. By contrast MCMC methods may become feasible but to set against this they may be computationally very demanding in terms of processor time. Further the simulation procedures used must be carefully monitored to ensure they are working properly. In these sense they are not routine once a model has been carefully specified.

The implementation and performance issues in Bayesian and likelihood fitting of multilevel models are explored in some detail by Browne and Draper (2000). Browne (2004) also introduces some newer adaptations of MCMC methodology for crossed random effects model which may make then statistically and computationally more efficient. Clayton and Rasbash (1999) using data augmentation develop a procedure for crossed random which is in some senses a hybrid of ML and Bayesian methods but this is not routinely available in software and requires some programming of macros. Hox (2000), Chapter 3, gives a fairly non-technical discussion of the range of estimation procedures available and contrasts them. Goldstein (2003) gives a full theoretical appraisal and technical discussion of the range and variety of estimation algorithms within these broad frameworks, including the supplementary use of such approaches as bootstrapping which may improve estimation. An examination of these sources will reveal that a full understanding of the ramifications of estimation methodology and the vagaries of its computer implementation is daunting to even those with technical experience and skills in statistical modelling. Thus we will not attempt here a discussion of Bayesian inference and MCMC in particular. However, with statistical guidance and some familiarity with the modelling background a seasoned researcher might be able to use the wide range of software available and be able to interpret analytical results. Indeed as Hox (2002) says, ‘ Software does not make a statistician but the advent of powerful and user friendly software for multilevel modelling has had a large impact in many research fields…..’

6.3 Software

The specialist multilevel software most familiar to users in the UK is arguably MLwiN (Rasbash et al (2004)). This software has been developed for fitting large and complex models using both frequentist likelihood and MCMC approaches (Browne (2002). MLwiN has been under development since the late 1980’s first as a DOS based programme, MLn and since 1998 in a fully fledged windows version, currently in release 2.02. It is produced by the Centre for Multilevel Modelling at the Institute of Education, University of London but now moved in 2005 to the University of Bristol. The software development has largely been funded by the Economic and Social Research Council alongside development of advanced statistical methodology and research applications, some of which is the subject of this review. At the heart of the 'frequentist' approach in MLwiN is the Iterative Generalised Least Squares (ILGS) procedure introduced by Goldstein (1986) and its implementation as discussed by Goldstein and Rasbash (1992). These articles discuss fully how certain features of a multilevel model lend themselves in the implementation to the design of computationally efficient algorithms for the required matrix operations. However, it has also been adapted successfully by efficient model formulation to encompass the more complex structures that have been under discussion. Goldstein (2003) discusses how it is formally equivalent to the maximum likelihood ‘Fisher scoring algorithm’ of Longford (1987), which was another early attempt at implementing the iterative estimation. The necessity for an iterative approach to

maximum likelihood is stressed by all developers since analytical formulae are nigh on impossible with such complex models. The MCMC estimation procedures in the latest release of MLwiN are recent but they are rapidly developing and now encompass possibilities for a wide range of complex model structures such as we have discussed. A further statistical advantage of MCMC that has been widely noted is that they yield exact inferences in situations where maximum likelihood only provide approximations; in small samples for instance.

MLwiN has many other advanced features not available in other packages and a flexible macro language which enables methodological development for even more advanced modelling. It also has a very user friendly symbolic and graphical interface so that it is designed for fully interactive use. The researcher is easily able to explore modify and develop from his results in a beneficial way. Some other software packages which may just give output from fitted models do not so easily allow this. The other main specialist software used by multilevel modellers mainly in the US is HLM (Raudenbush and Bryk (2002)) which uses the alternative EM algorithm for iterative ML also developed for random effects models by Laird and Ware (1982). This approach has been incorporated into many software packages mainly because of its computational simplicity. Zhou et al (1999) and Sullivan (1999) review and discuss the use of HLM and some other packages. HLM has also been widely used for educational data arguably since that is the substantive interest of its originators and many original applications were in this area. By similar adaptations to those used in MLwiN it can also handle the cross-classified effects, but it is restricted in the number of levels in the data one can use. The most recent version is HLM5 (Raudenbush et al (2005)).

Judging by the reference and correspondence on the LISTSERV multilevel mailing list a third major package in frequent use for multilevel linear models is the procedure PROC MIXED for random effects embedded in the general purpose SAS software (SAS/Stat (2000)). It can handle crossed structures up to two levels. It uses iterative maximum likelihood by a combination of Fisher scoring and Newton- Raphson approximations. To researchers familiar with the concepts of multilevel structures the interface can be a little tricky since levels are not explicitly recognised for the general random effects models it considers. They have to be introduced by specially arranged input and parameter files. However, Singer (1998) gives very useful details on how this is done and instrumental examples. Another useful tutorial for designed experiments is Spilke et al (2005). De Leeuw and Kreft (2001) argue that the use of SAS may not be efficient for these implementation reasons; also because one has to carry the enormous overheads in computing capacity that the whole SAS system carries. The trade off is that many use SAS as their all purpose statistical system and may be familiar with its syntax and interface. For very large data sets and complex structures with many random effects, when other software may be limited computationally, the software GENSTAT (Payne et al (2005) holds considerable promise. Recent versions include a very efficient implementation of restricted maximum likelihood procedures pioneered by Patterson and Thompson (1971) and further developed by Gilmour et al (1995) and Pan and Thompson (2000). However, the user interface is to some not very transparent or friendly.

Procedures for fitting linear and other multilevel models are now also available embedded in a range of major general statistical software packages. STATA (Stata Corp (2005) uses an iterative ML method known as Gauss-Hermite quadrature can handle nested and crossed effects up to two levels. There are also some recent possibilities in SPSS, which is the general software most widely used in the social science research. In the past few years there has also been a rapid development of the range of software that has features for carrying out basic multilevel modelling or can with ingenuity fit more complex or specialised models using macro languages. This is just about keeping up with the rapidly growing methodological

literature and literature on ever more diverse ranges of applications. Goldstein (2003), Chapter 15 lists eighteen packages which have some facilities and he gives references and notes on each one's capabilities. Useful comparative reviews of some of the packages are de Leeuw and Kreft (2001), Zhou et al (1999). Fein and Lissitz (2000) compares the specialist software MLwiN and HLM. Since Goldstein's list the range has been extended as may be seen by the maintenance of the series of reviews on the Centre for Multilevel Modelling (CMM) webpage, www.multilevel.ioe.ac.uk. This set of reviews contains detailed evaluation of how various pieces of software copes with a range of advanced models using common example data sets. Of particular relevance here are the comparisons of handling crossed – effects.

6.4 Brief Comments on Generalised Models For Discrete Responses

The last few sections have focussed in the main on methodology and software for linear models of continuous responses. For discrete response (e.g. binary) generalised linear models maximum likelihood is not straightforward. Beyond fairly basic two level hierarchical models they may involve considerable computational load involving extensive numerical integration techniques (see for example Hedeker and Gibbons (1994)). Some of the software mentioned above and other pieces do have facilities and possibilities are further evaluated in the reviews by CMM. The suite of programmes MIXOR and MIXNO specifically designed for non-continuous response is widely used. (Hedeker (1999), Hedeker and Gibbons (1996a, 1996b)). Within STATA there are advances in the use of the GLAMM procedures (Rabe-Hesketh et al (2004), Skrondal and Rabe-Hesketh (2004). These use an improved method of numerical integration called adaptive quadrature. Practical illustrations of this use of STATA are provided by Rabe-Hesketh and Skrondal (2005). The main drawbacks of such applications might manifest themselves in very complex structures where integration over many random effects is necessary. For these reasons a variety of approximation methods have been used which are called quasi-likelihood. MLwiN adapts its IGLS procedures in a variety of ways to do this and makes possible the fitting of complex structures. The GLIMMX and NLMIX macros in SAS also use an approximate approach but the type of complexity possible is a little more limited. Brown and Prescott (1999) discuss these methods and practical examples for random effects though their focus is not particularly multilevel. The main practical difficulties with quasi –likelihood approaches is that they can break down due to non-convergence for certain types of data a structure and they are still quite computationally intensive.

The MCMC approaches in MLwiN have also been applied successfully to non–continuous data in the presence of complex random effects. The general purpose WINBUGS package (Spiegelhalter et al (1997)) also uses a Bayesian approach to estimation and can be adapted quite successfully to a range of complex multilevel models of either linear or generalised types. Lawson et al (2003) compares the use of MCMC in MLwiN and WINBUGS and explores their differing potentialities through examples.

7 More Applications In The Literature And Potentiality For Similar Approaches in Education Research

In main body of this review we have concentrated on applications to exemplify the key features. This section reviews some areas of the literature where further successful applications have been used. They may be useful in the development of statistical models in many areas of research related to education where similar structures can arise. Although some of the application areas we discuss have obvious and direct connections with education, some may appear remoter. However, we stress that the design features of these applications can have direct links with the advancement of educational research with, for example, the further attention to intervention designs or increasing availability of integrated data with a great deal of structural complexity. Lessons may be learned from other areas about how disentangling of effects may be handled in the context of these complexities.

7.1 Health Research

Health statistics is an area where there has been good deal of work involving complex structures. Some of these structures involve units that parallel those familiar in education with patients instead of students, doctors or nurses instead of teachers within institutional contexts such as GP practices or hospitals. Many of the complex situations we discussed previously arise with patients consulting different health service personnel on different occasions. We also have potential for different sorts of cross-classifications such as GPs by hospitals or consultants working in different hospitals. Multiple membership relations also arise such as when patients see several doctors. Ecob et al (2004) in psychiatry studies health outcome scores in a multilevel model with crossed random effects for assessors referral sources (GPs).

Area effects are also important in the study of disease as we saw previously in our discussion of spatial models. Subramanian (2004) is a good review of their role in cross-classified model contexts. Clayton and Rasbash (1999) consider a complex cross-classified structure first addressed as a possibility by Echoard and Clayton (1998). The paper is also methodologically important since it introduces a variant of estimation known as data augmentation designed to overcome the computational difficulties first encountered with such large complex models (see Section 6.2). The application concerns artificial insemination by donor and at level 1 measurements are made on recipient women at each ovulatory cycle at which the response conception occurs. The data consist of 1901 women and 279 donors. Each donor made multiple donations and there were 1328 donations in all. A single donation is used for several inseminations. Figure 16 below is a classification diagram for this relatively complex situation. There are two crossed hierarchies; for women we have cycles within attempts within women; for donors we have cycles within donations within donors. The top level crossing is between woman and donors. Within each cell of this, which as we have seen induces a crossing of attempts by donor, we have also the possibility of a crossing of attempts by donations since donations can be used for multiple inseminations. There are now two random effects components at level 3, women and donors, and two at level 2, attempts within women and donations within donor.

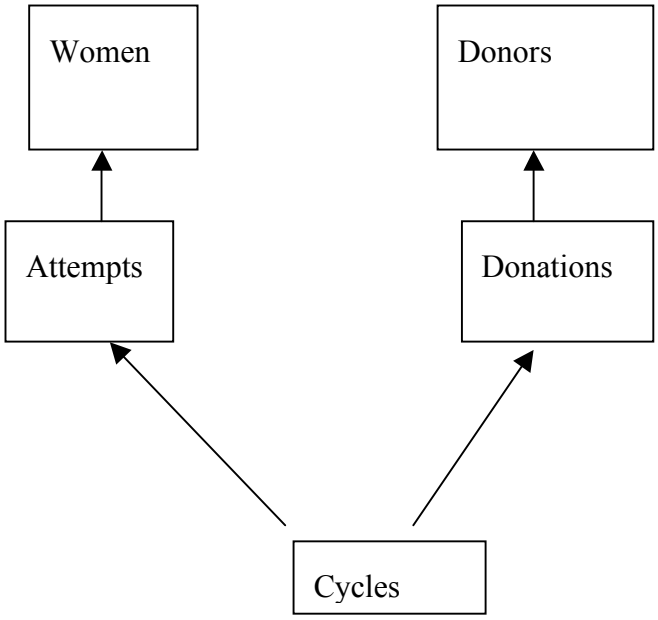


Figure 16: Cross classifications at two levels for artificial insemination

The interesting conclusion in this study is that after control for important covariates on such things as age of women and sperm count of donor, there is considerably more variation in the chance of successful insemination attributable to the woman hierarchy than that of the donor.

In the context of educational achievement particular in higher education repeated attempts at an examination under different teachers, or possibly also in different institution, provide similar structure*s*.

Crossed random effects models also arise in human health epidemiological studies. Coull et al (2001), for example, consider a study where presence of a series of observations on abnormalities are nested within a three way crossing of new born infants, location of abnormality, and type of drug exposure. Rivellini and Zaccarin (2000) discuss factors affecting fertility in Italy using crossed random effects of place of residence and place of childhood residence.

Zaslavsky et al (2004) uses a three level complex survey design on a range of consumer assessments of Medicare health plans in the US. There are three levels with sample respondent at level 1. Level 2 is a cross-classification of enrolment plan by Metropolitan Statistical Area within a state level 3 and a region level 4. The survey also was undertaken for five years and the effect of these years was allowed to be also random over the various classifications. The main focus was on the variance components though a small number of fixed effect covariates characterising the respondents were used.

7.2 Survey Methodology And Interviewer Response Variance.

The subject of response measurement error due to interviewer effects in surveys has been the subject of much methodological investigation over the years. Hox et al (1991) and Wiggins et al (1992) have studied hierarchical models of respondents within interviewers. An example of a cross-classified structure might be where we have a design clustered by areas such as parliamentary constituency and/or ward. Interviewers are assigned to different areas. Cross classified models may then simultaneously study effects of areas and interviewers on survey

respondents. O'Muircheartaigh and Campanelli (1998) found that the effect of interviewers on non response was greater than that of the areas in which they were working. They extend this work in attempting to separate cluster design effects from interviewer effects in survey precision (O'Muircheartaigh and Campanelli (1999)). Other similar examples arise in longitudinal or panel research interviewer studies where there may be different interviewers on two or more occasions. For example with just two occasions the set of interviewers on the first occasion may be crossed with the set of interviewers on the second occasion. The panel members are then nested within cells of this cross-classification. Pickery and Looseveldt (1998) and Pickery et al (2001) deal with this type of context.

7.3 Social Networks

Crossed random effects structures have also received some attention in the field of social networks. Snijders (1999) gives some detail on how level 1 units may be dyads or a directed relationship between a pair of actors. He gives an example of what a particular teacher (A) reports in a communication within another teacher (B) about individual pupils. Variables defined on that communication, such as how well the student is doing, form responses at the student level. The level 2 dyads in which such responses are nested involves not only two teachers but also the direction in which the report is made. Thus the dyads may be considered a cross classification of teachers by themselves, whether as the initiator in the dyad or the receptor. The level 2 dyads may in turn be nested hierarchically within schools. Thus we have two random effects for initiator and receptor. It is also considered that a reciprocity effect is useful which is akin to the use of an additional interaction random effect (see Section 4.2) Gender fixed effects are of interest so that three covariate dummies for the genders of each actor and whether they are the same are used. Van Duijn et al (1999) consider situations where individuals rank each others popularity. Some further examples of such type of social relation modelling are given by Snijders et al (1995), Snijders and Kenny (1999), and Snijders and Baerveldt (2003).

7.4 Veterinary Epidemiology, Animal Ecology and Genetics

An example of a complex multiple membership model is presence of salmonella infection in flocks of Danish chickens between 1995 and 1997. The example is considered in various ways in Goldstein (2003), Browne et al (2001) and Rasbash and Browne (2001). Each flock is kept in a house within a farm; a hierarchy. However, we wish to consider the effect of parentage of the flock which can be from a mixture of up to six parent flocks in a multiple member relation. These parent flocks are crossed with house within the farms. The flocks are also hatched within one of four hatcheries but this is handled by three dummy indicator fixed effects. The classification diagram is as Figure 17. Some important features that the reported analyses of this structure revealed were that most of the variation in salmonella infection was attributable to farms and parent flocks with some large hatchery differences. Residual estimates enabled extreme farms and parent flocks to be identified and scrutinised and this proved very useful in locating units which might be further examined. This idea of screening extreme and unusual units is one of the benefits of the ability to generate residual estimates.

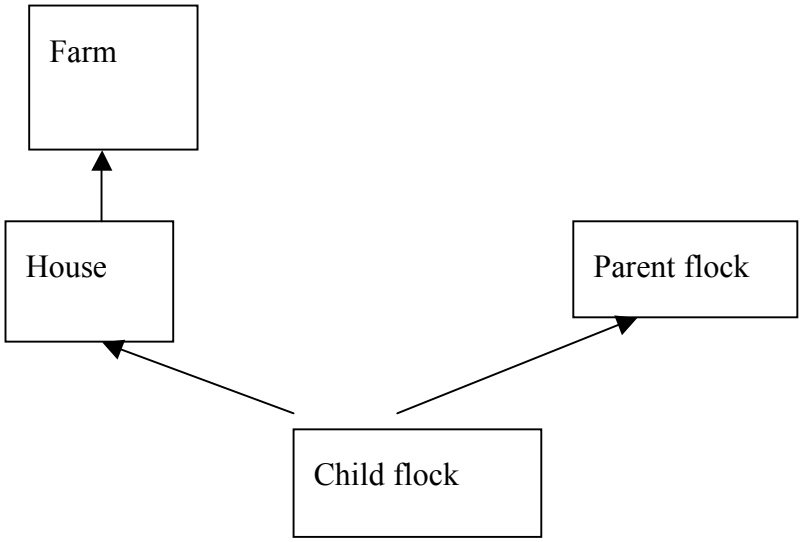


Figure 17: The Salmonella Infection Multiple Membership Structure

There is a large and old literature on the analysis of unbalanced designed experiments in agricultural and biological sciences, sometimes but not frequently with random covariate effects. Nearer to education they figure frequently in social and educational psychology. These are often formally equivalent to some of the cross-classified structures we have considered. An example, using the methods of Engel (1990), is Gengler et al (2004) which examined the crossed effects on cow's milk yield of herd, test day and milking frequency in a larger study of the heritability of yields.

There is some important recent work in psychology of depression which deals with human genetics incorporating complex crossed and multiple membership family effects (Jenkins et al (2003), Rasbash (2004)). A recent important example in bird ecology demonstrates the use of random effects modelling using quite complex arrangements of fixed and random effects and also extends to multivariate responses. The latter are clutch size, lay date, nestling mass, male survival, and female survival amongst 17 years of observations on great tits. The classifications involved are year, nest box, male parent and female parent all of which may affect the response.

The examples in this section, though taken from a variety of areas demonstrate the flexibility of designs which can be handled by the complex cross-classifications. They may be pointers to situations in educational progress and other educational research which go beyond the examples discussed in Section 4 and which have a range of tangled effects of interest which need to be unravelled.

7.5 Transportation research

In human ecology and transport research there is considerable interest in patterns of residence and factors affecting where people live and work and multilevel modelling has been applied quite widely in such areas (Bullen et al (1997), Jones and Duncan (1995), Jones and Bullen (1994)). Bhat (2000) considers a discrete choice dependent variable, mode of travel to work, and related it to such fixed effects as travel cost, ethnicity, income, and number of vehicles possessed for crossed random effects of location of work and location of residence. Such discrete responses, even directly mode of travel to school, are quite common as responses on students and these ideas are useful for showing how they might be handled in the context of, for example, school and area of residence effects. This authors claim that to the best of their knowledge this formulation of the model for discrete responses is a first. This claim is clearly incorrect. Methods for fitting such models using MLn, the precursor of MLwiN, were available by the mid nineties (Rasbash and Woodhouse (1995)).

7.6 Missing identification of units

Missing identification of which higher level certain lower units belong to is a common problem in multilevel data. However, we would not wish to drop such cases since this might seriously distort results. In school studies we might not know which school some of our students came from. This can be handled within a multiple membership relationship framework by using weights based on what we might know or assume about the probabilities that such units are members of each of the set of the higher level units. We usually have some information in the data that enables us to choose these reasonably. However, there is a slight variation from standard formulations which might have used probabilities directly as weights. Here if the probability of student i coming from school h am p_{ih} such that $\sum_{h=1}^{J} p_{ih} = 1$ we choose as weights $w_{ih} = \sqrt{p_{ih}}$ so that the weights no longer sum to unity. To see why we should do so in the missing identifier situation it is relevant to note that a student actually belongs to one and only one school even though we do not know which it is. This stands in contrast to a direct multiple relation where students actually belong to several schools and there is a weighted contribution from each. So for each unit with a missing identifier, since only one school affects the response, we would like the variance of this random effect to be school variance $\sigma^2_{u(2)}$ rather than averaging and hence diluting this variability over several weighted school effects as happens with true multiple membership relations. If we do use the w_{ih} above we can see with a bit of basic statistical algebra that the school variance contribution is

$$\text{Variance}\{\sum_{h\in\text{school}(i)} w_{ih} u_h^{(2)}\} = \sigma^2_{u(2)} \sum_{h\in\text{school}(i)} w_{ih}^2 = \sigma^2_{u(2)} \sum_{h\in\text{school}(i)} p_{ih} = \sigma^2_{u(2)},$$

which is what we would like it to be.

Hill and Goldstein (1998) consider some practical applications of this idea including extensions to mixtures of cross–classified, multiple membership and missing identification structures. In progress studies, for instance, we might wish to associate students in a multiple membership relation with a number of schools they might have attended, including their current school, over the course of a period up to the present point at which their achievement is recorded. We would like to assign weights to schools in ways that might appropriately reflect in our judgement the extent of the effects of these schools to current achievement. Suppose we are just considering school currently with school last year and decide it is appropriate to give weights 0.7 and 0.3 respectively. With full historical data on school

attended a straight multiple membership model might be analysed. However, we might not have such historical data for many students but for such students may have ancillary information that enables assigning probabilities and hence weights w_{ih} for last year's school. In a multiple membership model the current school would receive a weight of 0.7 and other schools a weight of $0.3w_{ih}$. Of course this might be extended to also accommodate students whose current school is also unknown. Goldstein (2003) also considers how this idea might also be exploited to cover missing location identities in the release of census data and which often occurs to preserve confidentiality.

7.7 Generalisability theory

The approach known as 'generalisability theory' (Cronbach and Webb (1975)) can also be formulated as a cross-classified random effects model (Schroeder and Hakstian (1990)). A test having a battery of items such as ratings is administered to a set of individuals who may be further nested within other units such as schools. Item variance is studied in combination with student effects in a cross-classification of students and items. Since we have only one unit per cell we cannot here separate out an interaction variance from the residual. Hox (2002) comments that such an approach can be adapted to examine reliability of student grades (such as degree classifications) which are the result of combining many marks or grades where variation of grades is affected by class groups, students, teachers, exercises within modules etc. These can be conceived of as random effects arising from a many way classifications (Chronbach et al (1972)). Crossley et al (2002) uses the generalisability theory idea in assessment of professionals' performance (surgeons) in medical education where there is a crossing of variety of situation types, surgeons, and assessors. There is a close link between this approach and that of Jayasinghe et al (2003) discussed below.

7.8 Psychometrics

A discussion and illustration of cross-classified models with examples in psychometrics is given by Van den Noortgate et al (2003), and which have obvious relevance in educational psychology and educational assessment and test measurement.

7.9 Further examples in education

Jayasinghe et al (2003) study assessor ratings in the peer review process of large grant proposals to the Australian Research Council. The methods may be more widely applicable to other forms of peer review such as publications, staff appraisals, job interviews, or elections to learned societies. The situation is that proposals are rated by more than one assessor and each assessor can review several proposals. This forms a cross-classification of level 1 rating by proposal and assessor at level 2 nested within a level 3 of disciplinary fields. A large number of fixed effect covariates representing characteristics of the proposal, its proposer, the assessor and interactions between them are considered. Some very interesting conclusions were reached about these covariates but an additional finding was that even after such a determined attempt at covariate control there was still a large amount of unexplained variance between assessors.

Simonite and Browne (2003) study modular degree courses in which grades are nested within a cross-classification of students and modules. This example is also very thorough in that it investigates a wide variety of covariates and interactions, allows random regression coefficients across students (e.g. year of study for student) and complex variance for the module grades (project or not, for instance). This study was also very important methodologically. It showed that there were considerable differences in some important conclusions when hierarchical models nesting module entries with students and ignoring the clustering of students within modules was used. This structure is similar to the example of

Fielding (2002a) and Fielding and Yang (2005) discussed above where ignoring the fact that A level students belonged to several teaching groups had similar less than benign consequences.

McCaffrey et al (2004) have also used cross-classified effects in modeling teacher value-added. Bell (2003) analyses student degree performance dependency on university entrance qualifications within a structure of university crossed with school attended. Teitler and Weiss (2000) use the third wave of the Philadelphia Teen Survey to estimate cross-classified two level models to see how much census area and between school variability exists in the timing of youth's initiation to sexual intercourse. Attempts are made to assess the extent to which school variation is attributable to race and normative environments of schools. May et al (2004) in a study of an educational intervention known as America's Choice recognised the importance of mobile students and cross-classified students by school attended for those attending more than one school during the span of the intervention.

8 Conclusion and Additional Comments

This review has concentrated in the main on linear models for continuous responses and to see how the random effects structure can be complex in various ways. We have also touched a little on generalised linear models for discrete and category responses. In conclusion we might mention developments for a few other issues which may also incorporate possible complex random structures.

- In a multivariate response situation we may have observations on several dependent variables observed simultaneously for each level 1 unit. These can also then vary and covary at higher levels. In addition to studying the structure of the variance over levels for each variable we will want also to examine how the covariances and correlations are structured in this way. A special case is where some measures are made at level 1 and some at higher levels so that responses for level 1 units within the higher level units are constant for these latter. We might for instance have each child's rating on some aspect of the class environment together with the class teachers. Such special cases can be accommodated by IGLS estimation in MLwiN but the current MCMC implementation requires the response variables involved to both be at level 1. Yang and Woodhouse (2001) conduct a two level multivariate multilevel analysis for various combinations of GCE A level mathematics that students can take. This analysis will also show that it is also not necessary for measurements of the response variables to be present for all observed units as would be the case for traditional multivariate analyses. Some measurements may be missing randomly or by design but all level 1 units can still be included efficiently in the analysis. This feature also means that there are procedures for the efficient design and modelling of complex rotation and matrix sample survey investigations (Goldstein (2003), Chapter 4). The multivariate idea also provides a general model for meta analysis where several studies (level 2) units are involved for some of which responses are available only in summary form at level 2 and for others detailed level 1 responses are available (Goldstein et al (2000b))
- The ability to model multivariate responses can be adapted to cover a number of other situations which on the surface have different features. An example is the extension to multilevel situations of the sample selection idea of Heckman (1979). The situation is that the response observation is missing for some units in ways that are related to variables in the model. If we can construct a supplementary model for the probability

of being missing then Goldstein (2003) shows how this can be set up so that the two equations can be simultaneously estimated efficiently in the form of a bivariate response multilevel model. A similar idea can be applied to multi-process models. Here there is a target model of interest and equations for related auxiliary processes that can, for instance, be necessary for a specification of the situation. A common situation is 'endogeneity' of an explanatory variable in the target model, where this variable is related to the random effects in this target model. In the educational production function equations of Levačić et al (2005), which related Key Stage 3 achievement to a number of explanatory variables, the resource and expenditure variables are examples. The multilevel analyses conducted in that report set up auxiliary equation for the expenditure process using specially written macros and efficiently estimated them within the framework of a bivariate response model. Steele (2001) gives another example of this idea and is currently developing advanced methodology on multi-process models under the ESRC Research Methods programme. Details may be found at www.multilevel.ioe.ac.uk. Endogeneity in single level models have often been handled by instrumental variable (IV) methods, particularly in the econometrics literature. For the most part such an approach was adopted by Levačić et al (2005). Instrumental variable approaches have been developed for multilevel models by Spencer and Fielding (2000, 2002). The main problem with IV methods is that they can be quite inefficient with low precision for estimates unless some detailed care is exercised. Another multilevel approach where no explicit second process is specified is the Conditional Iterative Least Squares estimation of Rice et al (1998)

- Modelling time spent in various states is important in many research contexts. In medicine we have survival time, or length of hospital stay. In economics duration of employment is of interest. In education students may spend different lengths of time on gaining a degree Associated with this are whether they progress satisfactorily or not at the end of each year of study, or the 'state' they are in a certain length of time after the process began. Such 'event history' or 'survival' processes can often take place within more complex structures. For example, individuals (Level 2) repeatedly pass through various periods of measurement of employment status (Level 1). Students may be measured within a number of different universities. Steele et al (2003) discuss a range of models of such situations with examples and show how they can be fitted using multilevel procedures. The ESRC project referred to above is also addressing such multi-state event history models in tandem with multi-processes.

- The closely related areas of structural equation modelling, latent variable analysis and factor analysis within multilevel structures are also recently receiving close attention, e.g., traditional factor analyses of sets of attitudes of students as indicators of underlying constructs. Variability in such latent constructs over students may also be affected by the multilevel structures in which they are placed. Details of many new developments in these areas are provided by Skrondal and Rabe-Hesketh (2004). Muthen (1994) is an early review of these ideas and has also provided specialist software MPLUS for multilevel structural equation modelling (Muthen and Muthen (2003)). Goldstein and Browne (2002) introduce methods for factor analysis within the MCMC estimation procedures of MLwiN.

- Measurement error is a problem that has received wide attention in the statistical theory and methodological literature. It is well known that when statistical models contain large components of such error statistical inferences may be misleading unless recognised and possibly adjusted for. The same is true of multilevel models. However,

the problem is possibly more complex in that such errors can occur at many levels. A discussion of multilevel approaches is given by Woodhouse et al (1996). MCMC methods are particular useful for models incorporating measurement error, particularly where the variables with error have random slopes (Browne et al (2001b). Distributions for errors can be incorporated in the simulation framework

- A programme of methodological development for handling missing data in multilevel models has been undertaken by the Centre for Multilevel Modelling and others and is continuing (the methods for missing response observations considered above is a special case). A recent project carried out by James Carpenter under the auspices of the ESRC Research Methods Programme has undertaken a wide investigation of this topic and has provided an extremely valuable research resource in the website www.missingdata.org.uk.

We finally concluded with some caveats expressed in more detail by Goldstein (2003). The application of multilevel modelling has already begun to make statistical analysis reflect the complexity of much social reality and to yield important new insights in many areas. As software is developed and becomes more widely available and user friendly the application of multilevel ideas should become more widespread and even routine. This is welcome, yet despite their usefulness, multilevel models should never become an unthinking panacea. Some researchers even doubt their value at all but naturally this will not be our view. The latter cynicism is often based on misconceptions of the nature of statistical inference or lack of understanding of the implications of uncertainty. Plewis and Fielding (2003) critically discuss some of the issues surrounding this position by some researchers. However, it is true that in certain circumstances where there is little structural complexity multilevel models may hardly be necessary. Descriptive statistics and traditional single level regression summaries may then be adequate for analysis, presentation and interpretation. On the other hand multilevel analyses can bring extra and necessary insight into the complex matters of explanation and causality. For instance, they bring efficient precision into comparisons between universities by utilising in a properly specified way what is going on inside them through data on individual students. The models are not, however, substitutes for well grounded substantive theory. They may though, through grounded empiricism help to evaluate those theories and make a contribution to their development. Part of the difficulty may be that by introducing more complexity they extend but do not necessarily simplify interpretation. But substantive theories are themselves complex and there can be no pretence that simple statistical methods, devoid of the need for caveats and detailed interpretation, are always the empirical way of cutting through that complexity. Multilevel analysis is a set of tools that are becoming invaluable in empirical research but they must be used with care and statistical understanding.

Appendix

Figure A1: Cross Classification of KS3 Students in Oldham LEA Schools by Ward of Residence and School

	OLDHAM LEA SECONDARY SCHOOLS															
WARD CODE **Wards in Oldham LEA**	A	B	C	D	E	F	G	H	I	J	K	L	M	N	0	**TOTALS**
00BPFA	8	1	66	24	1	5	0	5	0	0	3	2	0	17	2	134
00BPFB	0	1	0	0	0	3	0	1	0	49	32	10	4	3	12	115
00BPFC	0	2	0	0	0	3	0	0	1	73	8	3	1	1	16	108
00BPFD	0	0	0	0	0	40	0	4	0	0	46	1	0	10	6	107
00BPFE	1	86	3	0	1	5	2	1	0	2	26	3	0	3	4	137
00BPFF	0	0	0	0	0	1	63	0	0	10	1	0	38	0	12	125
00BPFG	0	0	1	0	5	3	0	69	0	0	1	11	0	10	0	100
00BPFH	0	0	0	0	2	4	0	64	0	0	2	2	0	13	1	88
00BPFJ	1	0	7	0	51	6	1	22	0	0	3	0	0	12	0	103
00BPFK	9	0	5	11	2	2	0	0	33	1	3	5	1	8	2	82
00BPFL	0	0	0	0	1	1	12	0	0	57	1	3	11	3	29	118
00BPFM	0	0	0	0	2	2	33	1	0	36	2	2	4	0	24	106
00BPFN	0	0	0	0	0	0	0	0	88	0	0	17	1	0	0	106
00BPFP	1	0	0	0	0	0	0	0	88	0	0	20	4	1	2	116
00BPFQ	57	0	1	0	0	2	6	0	4	1	1	5	2	1	14	94
00BPFR	27	1	4	68	2	3	14	1	5	1	6	2	0	8	4	146
00BPFS	1	1	30	2	7	2	0	6	0	0	10	4	1	14	3	81
00BPFT	0	0	0	0	1	2	62	0	0	1	1	0	34	0	13	114
00BPFU	85	0	1	9	0	2	1	0	13	0	1	14	5	9	14	154
00BPFW	0	10	20	2	6	14	0	1	0	0	39	3	0	8	4	107
Wards outside Oldham LEA																
00BNFC	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
00BNFF	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	2
00BNFK	0	0	0	0	1	0	0	5	0	0	0	0	0	0	0	6
00BNFL	0	0	0	0	0	2	0	0	0	0	1	1	0	0	0	4
00BNFP	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
00BNFU	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	2
00BNFY	0	0	1	0	0	2	0	3	0	0	4	1	0	0	0	11
00BNGB	0	0	0	0	2	24	0	20	0	0	9	1	1	0	8	65
00BNGC	0	0	1	0	4	1	0	32	0	0	1	1	0	2	0	42
00BQFA	0	0	1	0	0	1	3	0	0	2	1	0	14	0	4	26
00BQFB	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
00BQFC	0	0	0	0	0	0	0	0	0	0	0	2	6	0	0	8

00BQFE	0	0	0	0	0	0	0	0	0	0	0	1	4	0	0	5
00BQFF	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
00BQFG	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	3
00BQFH	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
00BQFJ	0	0	0	0	0	1	0	0	0	0	0	1	4	0	0	6
00BQFK	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	2
00BQFL	0	0	0	0	0	1	0	0	0	1	1	1	2	0	0	6
00BQFM	1	0	0	0	0	0	0	0	0	0	3	0	0	1	0	5
00BQFN	0	0	0	0	0	1	0	0	0	0	0	7	17	0	0	25
00BQFP	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
00BQFQ	0	0	0	0	0	0	0	0	0	0	0	4	10	0	1	15
00BQFR	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	2
00BQFS	0	0	0	0	0	0	0	0	0	0	0	3	10	0	0	13
00BQFU	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	3
00BQFW	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
00BTFA	0	0	0	0	0	0	0	0	1	0	0	3	0	0	0	4
00BTFB	0	0	1	0	0	0	0	0	0	0	0	2	0	0	1	4
00BTFC	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
00BTFD	0	0	0	0	0	0	0	0	1	0	0	3	0	0	0	4
00BTFG	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	2
00BTFK	0	0	0	0	0	0	0	1	0	0	0	2	0	0	0	3
00BTFM	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
00BTFN	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
00BTFP	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
00BTFQ	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	4
00BTFS	0	0	0	0	0	0	0	0	2	0	0	3	0	1	0	6
00BTFT	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
00BTFU	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	6
00BUFT	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
00CYFR	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
17UHGY	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	2
17UHHB	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
30UDHF	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
30UMFR	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
30UMFX	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
TOTAL	191	102	143	116	88	135	198	237	236	234	209	167	184	126	179	2545

References

Aitkin, M. and Longford, N. (1986). Statistical modelling in school effectiveness studies (with discussion). *Journal of the Royal Statistical Society, A.* **149**: 1-43.

Aitkin, M., Anderson, D. and Hinde, J. (1981). Statistical modelling of data on teaching styles (with discussion). *Journal of the Royal Statistical Society, A.* **149**: 148-61.

Bell, J. F. (2003) Analysing student progress in higher education using cross-classified multilevel logistic models, *4th International Amsterdam Conference on Multilevel Analysis*, (http://www.ucles.org.uk/assessmentdirectorate/articles/ confproceedingsetc/MA2003JB).

Bennett, N. (1976). Teaching Styles and Pupil Progress. London, Open Books.

Bhat, C. R. (2000). A multi-level cross-classified model for discrete response variables. *Transportation Research Part B*, **34**, 567-582.

Brown, H. and Prescott. R. (1999). *Applied Mixed Models in Medicine*. Chichester, Wiley.

Browne, W. J. (2002). *MCMC Estimation in MLwiN.* London, 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.

Browne, W. J. and Draper, D. (2000). Implementation and performance issues in Bayesian and likelihood fitting of multilevel models. *Computational Statistics*, **15**, 391-420.

Browne, W. J., Goldstein, H., and Rasbash, H. (2001a). Multiple membership multiple classification (MMMC) models. *Statistical Modelling*, **1**, 103-24.

Browne, W. J., Goldstein, H., Woodhouse, G. and Yang, M. (2001b). An MCMC algorithm for adjusting for errors in variables in random slopes multilevel models. *Multilevel Modelling Newsletter*, **13**, 1, 4-9.

Browne, W. J., McCleery, R. H. Sheldon, B. C. and Pettifor, R. A. (2004). Using cross-classified multivariate mixed response models with applications to life history traits in great tits (parus major). *University of Nottingham Statistics Research Papers 03-18* (http://www.maths.nott.ac.uk/personal/pmzwjb/bill.html).

Bullen, N. Jones, K. and Duncan C. (1997). Modelling complexity: analysing between individual and between place variation –a multilevel tutorial. *Environment and Planning*. **29**, 585-609.

Chronbach, L. J, and Webb, N. (1975).Between class and within class effects in a repeated aptitude x treament interaction: re-analysis of a study by G. L. Anderson. *Journal of Educational Psychology*, **67**, 717-724.

Chronbach, L.J., Gleser, G. C., Nanda, H. and Rajratnam, N (1972). *The Dependability of Behavioral Measures.* New York, Wiley.

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

Congdon, P. (2001). *Bayesian Statistical Modelling*. Chichester, Wiley.

Coull, B. A., Hobert, J. P., Ryan, L. and Holmes, L. (2001). Crossed random effect models for multiple outcomes in a study of teratogenesis. *Journal of the American Statistical Association* **96**: 1194-1204.

Crossley, J., Davies, H., Humphris, G. and Jolly, B. (2000). Generalisability: a key to unlock professional assessment. *Medical Education,* **36**, 972-78.

De Leeuw, J. and Kreft, I. G. G. (2001). Software for multilevel analysis. *Multilevel Modelling of Health Statistics.* A. Leyland and H. Goldstein. Chichester, Wiley.

Diez Roux, A. V. (2002). A glossary for multilevel analysis. *Journal of Epidemiology and Community Health*, **56**, 588-594.

Draper, D, and Gittoes, M. (2004). Statistical analysis of performance indicators in UK higher education. *Journal of the Royal Statistical Society, Series A***, 167, 3**, 449-474.

Duan, N and Reise, S. (2002) *Multilevel Modelling: Methodological Advances, Issues and Applications*. New York, Erlbaum.

Ecochard, R. and Clayton, D. G. (1998). Multilevel modelling of conception in artificial insemination by donor. *Statistics in Medicine* **17**: 1137-1156.

Engel, B. (1990). The analysis of unbalanced linear models with variance components. *Statistica Neerlandica*, 44: 195-219.

Fein, M. and Lissitz, R. W. (2000). *Comparison of HLM and MLwiN Multilevel Analysis Software Packages: A Monte Carlo Investigation into the Equality of the Estimates*. University of Maryland.

Fielding, A. (2000). Explanatory modelling of complex social structures with case studies in educational research. *ESRC/ BERA Advanced Training Workshop.* School of Education, University of Birmingham (http://www.economics.bham.ac.uk/people/fielding/fielding3.pdf).

Fielding, A. (2002a). Teaching groups as foci for evaluating performance in cost-effectiveness of GCE advanced level provision: some practical methodological innovations. *School effectiveness and school improvement.* **13**: 225-246.

Fielding, A. (2002b). Ordered category responses and random effects in multilevel and other complex structures: scored and generalised models. *Multilevel Modelling: Methodological Advances, Issues and Applications*. N. Duan and S.Reise. New York, Erlbaum.

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

Fielding, A., Belfield C. and Thomas H. (1998). Rhe consequences of drop-outs on the cost-effectiveness of 16-19 colleges. *Oxford Review of Education*, **24**, 4, 487-511.

Fielding, A., Yang, M. and Goldstein, H. (2004). Multilevel modelling of ordinal grades. *Statistical Modelling*, **3**, 127-153.

Fielding, A. and Yang, M. (2005). Generalised linear mixed models for ordered responses in complex multilevel structures: effects beneath the school or college in education. *Journal of the Royal Statistical Society, A*. **168**: 159-184.

Gengler, N., Wiggins, G. R. and Gillon, A. (2004). Estimated heterogeneity of phenotypic variance of test-day yield with a structural variance model. *Journal of Dairy Science,* 87. 1908-16.

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.

Gilmour, A. R., Thompson, R. and Cullis, B. (1995). AIREML, an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, **51**, 1440-50.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, **73**, 43-56.

Goldstein, H. (1987). Multilevel covariance component models. *Biometrika* **74**: 430-431.

Goldstein, H. (1991). *Computational Algorithms for Random Cross Classifications.*

Goldstein, H. (1994). Multilevel cross-classified models. *Sociological Methods & Research* **22**: 364-375.

Goldstein, H. (1995). *New Directions in the Statistical Modelling of Hierarchically Structured Data.* SSS 95, Utrecht, ProGamma.

Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement,* **8**, 4, 369-395.

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

Goldstein, H. and Browne, W.J. (2002). Multilevel factor analysis modeling using Markov Chain Monte Carlo estimation. In *Latent Variable and Latent Structure Models* edited by G. Marcoulides and I. Moustaki, London, Lawrence Erlbaum, 225-44.

Goldstein, H. and Rasbash, J. (1992). Efficient computational procedures for the estimation of parameters in multilevel models based on iterative generalized least squares. *Computational Statistics and Data Analysis*, **13,** 63-71.

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D. and Thomas, S. (1993). A multilevel analysis of school examination results. *Oxford Review of Education,* **19**, 4, 425-33.

Goldstein, H. and Sammons, P. (1997). The influence of secondary and junior schools on sixteen year examination performance: a cross-classified multilevel analysis. *School Effectiveness and School Improvement,* **8**: 219-230.

Goldstein, H. and Spiegelhalter, D. (1996). League tables and their limitations: statistical issues in comparisons of educational performance (with discussion). *Journal of the Royal Statistical Society, A*. **159**: 385-409.

Goldstein, H., Rasbash, J., Browne, W., Woodhouse, G. and Poulain, M. (2000a). Multilevel models in the study of dynamic household structures. *European Journal of Population,* **16**, 373-387.

Goldstein, H., Yang, M., Omar, R. and Turner, R. (2000b). Meta analysis using multilevel models with an application to the study of class size effects. *Journal of the Royal Statistical Society, C*, **49**, 399-412.

Greene, W. H. (2003). *Econometric Analysis, 5th edition*. Upper Saddle River NJ, Prentice-Hall.

Heck, R. H. and Thomas, S.L. (2000). *An Introduction to Multilevel Modelling Techniques*. Mahwah NJ, Lawrence Erlbaum.

Hedeker, D. (1999). MIXNO: A computer program for mixed effects logistic regression. *Journal of Statistical Software*, 1-92.

Hedeker, D. and Gibbons, R. D. (1994). A random effects ordinal regression for multilevel analysis. *Biometrics*, **50**, 933-44.

Hedeker, D. and Gibbons, R. D. (1996a). MIXOR: A computer program for mixed effects ordinal probit and logistic regression analysis. *Computer Methods and Programs in Biomedicine*, **49**, 157-176.

Hedeker, D. and Gibbons, R. D. (1996b). MIXREG: A computer program for mixed effects regression analysis with autocorrelated errors. *Computer Methods and Programs in Biomedicine*, **49**, 229-252.

Hill, P. W. and Goldstein, H. (1998). Multilevel modelling of educational data with cross classification and missing identification of units. *Journal of Educational and Behavioral statistics* **23**: 117-128.

Hox, J. (2002). *Multilevel Analysis: Techniques and Applications.* London, Lawrence Erlbaum.

Hox, J., de Leeuw, E. D. and Kreft, I. G. G. (1991). The effect of interviewer and respondent characteristics on the quality of survey data: a multilevel model. In *Measurement Errors in Surveys,* P, P, Biemer, R. M. Groves, L. E. Lyberg, N. A Mathiowetz and S. Sudman (Eds.). New York, Wiley.

Jayasinghe, U. W., Marsh, H. W. and Bond, N. (2003). A multilevel cross classified modelling approach to peer review of grant proposals: the effects of assessor and researcher attributes on assessor ratings. *Journal of the Royal Statistical Society, A*. **166**: 279-300.

Jenkins, J., Rasbash, J., and O'Connor, T. G. (2003). The Role of the Shared Family Context in Differential Parenting. *Developmental Psychology*, **39,** 1**.**

Jones, K. and Duncan, C (1995). Individuals and their ecologies: analysing the geography of chronic illness within a multilevel framework. *Health and Place,* **1**, 27-40.

Jones, K. and Bullen, N. (1994). Contextual models of urban home prices: a comparison of fixed and random coefficient models developed by expansion. *Economic Geography*, **70**, 252-272.

Kreft, I.G.G., de Leeuw, J., and van der Leeden, R. (1994). A review of five analysis programs: BMDP-5V, GENMOD, HLM, ML3, VARCL. *American Statistician*, **48**, 324-335.

Kreft, I. and de Leeuw J. (1998). *Introducing Multilevel Modelling*. London, Sage.

Laird, N.M and Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics*, **38**, 963-974.

Langford, I.H., Leyland, A.H. Rasbash, J. and Goldstein H. (1999). Multilevel modelling of the geographical distribution of diseases. *Journal of the Royal Statistical Society, C, 48, 2, 253-268.*

Lawson, A. B., Browne, W.J. and Vidal Rodeiro, C. L. (2003). *Disease Mapping with WINBUGS and MLwiN*, Chichester, Wiley.

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 Number 679, London, Department for Education and Skills.

Leyland, A. H. (2001). Spatial analysis. In *Multilevel Modelling of Health Statistics.* A. Leyland and H. Goldstein (Eds.). Chichester, Wiley.

Leyland, A. H. and Goldstein, H. (2001). *Multilevel Modelling of Health Statistics.* Chichester, Wiley.

Leyland, A. H. and Groenewegen, P. P. (2003). Multilevel modelling and public health policy. *Scandinavian Journal of Public Health*, **31,** 267-274.

Liang, K., Zeger, S.L. and Qaquish, B. (1992). Multivariate regression analysis for categorical data. *Journal of the Royal Statistical Society, Series B,* **54,** 3-40.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for linear models (with discussion). *Journal of the Royal Statistical Society, Series B,* **34,** 1-41.

Lindsay, J.K. and Lambert, P. (1998). On the appropriateness of marginal models for repeated measurements in clinical trials. *Statistics in Medicine,* **1**7, 447-469.

Longford, N. A fast scoring algorithm for maximum likelihood estimation in unbalanced linear models with nested random effects. *Biometrika,* **74**, 817-27.

May, H., Supowitz, J. A. and Perda, D. (2004). *A longitudinal study of the impact of America's Choice on student performance in Rochester, New York 1998-2003*. Consortium for Policy Research in Education, University of Pennsylvania. (www.cpre.org/publications/AC10.pdf).

Mortimore, P., Sammons, P., Stoll, L., Lewis, D. and Ecob, R. (1988). *School Matters*. Wells, Open Books.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., et al. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics* **29**: 67-102.

McCullagh, P. and Nelder, J. A. (1989). *Generalised Linear Models, 2nd edition*. London, Chapman and Hall..

Muthen, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research,* **22,** 376-98.

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

O'Donoghue, C., Thomas, S., Goldstein, H., and Knight, T. (1997). *1996 Study of Value Added for 16-18 Year Olds in England*. London, Department for Education and Employment.

O'Muircheartaigh, C. and Campanelli, P. (1998). The relative impact of interviewer effects and sample design effects on survey precision. *Journal of the Royal Statistical Society, A*. **161**: 63-78.

O'Muircheartaigh, C. and Campanelli, P. (1999). A multilevel exploration of the role of interviewers in survey non-response. *Journal of the Royal Statistical Society, A*. **162**: 437-446.

Pan, J. X. and Thompson, R. (2000). Generalised linear mixed models: an improved estimation procedure. *COMPSTAT: Proceedings in Computational Statistics* edited by J. G. Bethlehem and P. G. M van der Heijden, Physica Verlag, 373-378.

Paterson, L. (1990). An introduction to multilevel modelling. *Schools, Classrooms and Pupils.* S. W. Raudenbush and J.D Willms, San Diego, Academic Press.

Paterson, L. (1991). Socio-economic status and educational attainment: a multidimensional and multilevel study. *Evaluation and Research in Education*, **5**, 97-121.

Paterson, L and Goldstein, H. (1991). New statistical methods for analysing social structures: an introduction to multilevel models. *British Educational Research Journal*, **17,** 4, 387-393.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika,* **58**, 545-54.

Payne, R., Murray, D., Harding, S., Baird, D. and Soutar, D. (2005). *GenStat for Windows (8th Edition): Introduction*. VSN International.

Pickery, J. and Looseveldt, G. (1998). The impact of interviewer and respondent characteristics on number of 'no opinion' answers. *Quality and Quantity,* **32**, 31-45.

Pickery, J., Looseveldt, G. and Carton, A. (2001). The effect of interviewer and respondent characteristics on response behaviour in panel surveys: a multilevel approach. *Sociological Methods and Research,* **29**, 4, 509-523.

Plewis, I (1998). *Multilevel models*. Social Research Update No 23. Department of Sociology, University of Surrey.

Plewis, I. and Fielding, A. (2003). What is multilevel modelling for? A critical response to Gorard (2003). *British Journal of Educational Studies*, **51**, 4, 408-19.

Rabe-Hesketh, S., Pickles, A. and Skrondal, A. (2004). *Multilevel and Structural Equation Modelling of Continuous, Categorical and Event Data.* College Station, Texas, Stata Press.

Rabe-Hesketh, S. and Skrondal, A. (2005). *Multilevel and Longitudinal Modeling Using Stata.* College Station, Texas, Stata Press.

Rasbash, J. (2004). *Including Genetic Effects in Multilevel Models.* Centre for Multilevel Modelling, Institute of Education, University of London.

Rasbash, J. and Browne, W. (2001). Non hierarchical multilevel models. *Multilevel Modelling of Health Statistics.* A. Leyland and H. Goldstein. Chichester, Wiley.

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**: 337-50.

Rasbash, J. and Woodhouse, G. (1995). *MLn Command Reference Version 1.0*. Multilevel Models Project, Institute of Education, University of London.

Rasbash, J., Steele, F, Browne, W. and Prosser, B. (2004). *A User's Guide to MLwiN Version 2.0.* Centre for Multilevel Modelling, Institute of Education, University of London..

Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross sectional and longitudinal research. *Journal of Educational Statistics* **18**: 321-349.

Raudenbush, S. W., Bryk, A.S., Cheong, Y, F., Fai, Y. and Congdon, R. (2001). *HLM5: Hierarchical Linear and Non-linear Modelling*. Lincolnwood Illinois, Scientific Software International.

Raudenbush, S. W. and Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, *2nd Edition*. Thousand Oaks, Sage.

Rice, N. and Leyland A. (1996). Multilevel models: applications to health data. J*ournal of Health Services Research Policy*, **1**, 154-164.

Rice, N., Jones, N. and Goldstein, H. (1998). *Multilevel Models where the Random Effects are Correlated with Fixed Predictors. A Conditional Iterative Least Squares Estimator* (CIGLS). York, University of York, Centre for Health Economics.

Rivellini, G. and Zaccarin, S. (2000). *Multilevel Analysis in Social Research: Some Critical Issues*. Research Paper No 70m Dipartimento di Scienze Economiche e Statistiche, University di Trieste.

SAS/Stat. (2000). *SAS/ Stat Users Guide Version 8*. Cary, NC.

Schroeder, M. L. and Hakstian, A. R. (1990). Inferential procedures for multifaceted coefficients of generalisability. *Psychometrika,* 55: 429-447.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance components.* New York, Wiley:

Simonite, V. and Browne, W. J. (2003). Estimation of a large cross-classified multilevel model to study academic achievement in a modular degree course. *Journal of the Royal Statistical Society, A*. **166**: 119-134.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Research*, **25**, 323-355.

Singer, J. D. and Wllett, J. B. (2003). Applied longitudinal data analysis: Modeling change and event occurrence. New York, Oxford University Press.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalised Latent Variable Modelling: Multilevel, Longitudinal and Structural Equation Models*. Boca Raton, Florida, Chapman and Hall.

Snijders, T. A. B. (1995). Multilevel models for family data. *Advances in family research.* J. J. Hox, B. E. Van der Meulen, J. M. A. M. Janssesns, J. J. F. Laak.

Snijders, T. A. B., Spreen, M. and Zwaagstra, R. (1995). The use of multilevel modelling for analysing personal networks: networks of cocaine users in an urban area. *Journal of Quantitative Anthropology*, **5**, 85-105.

Snijders, T. A. B and Baerveldt R. J. (2003). A multilevel network study of the effects of delinquent behavior on friendship evolution. *Journal of Mathematical Sociology*, **27**, 123-151.

Snijders, T. A. B and Bosker, R. J. (1999). *Multilevel Analysis: An Introduction to Basic and Multilevel Modelling.* London, Sage.

Snijders, T. A. B and Kenny, D. A. (1999). The social relations model for family data: a multilevel approach. *Personal Relationships*, **6**, 471-486.

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.J., Thomas, A., Best, N. G. and Gilks, N. R. (1997). *BUGS: Bayesian inference using Gibbs sampling, Version 0.60.* Cambridge, Medical Research Council, Biostatistics Unit.

Spilke, J. Piepho, H. P. and Hu, X. (2005). Analysis of unbalanced data by mixed linear models using the MIXED procedure of the SAS system. *Journal of Agronomy and Crop Science*, **191**, 1, 47-53.

Stata Corp (2005). *Stata Statistical Software Release 9*. College Station, Texas.

Steele, F. (2001). A multiprocess multilevel model to allow for selection effects of source of contraceptive supply in an analysis of contraceptive discontinuance in Morocco, *Bulletin of the Institute of Statistics*, 53rd Session of the ISI, Seoul.

Steele, F., Goldstein, H., and Browne, W. J. (2004). A general multilevel multistate competing risks model for event history data, with an application to a study of contraceptive use dynamics. *Statistical Modelling*, **4**, 145-159.

Subramanian, S. V. (2004). The relevance of multilevel statistical models for identifying causal neighborhood effects. *Social Science and Medicine*, **58**, 1961-7.

Sullivan, L. M., Dukes, K. A. and Losina, A. (1999). Tutorial in biostatistics: an introduction to hierarchical linear modeling. *Statistics in Medicine*, **18**, 855-888.

Teitler, J. O. and Weiss, C. C. (2000). Effects of neighborhood and school environments on transitions to first sexual intercourse. *Sociology of Education*, **73**, 2, 112-32.

Thomas, H., Brown, C., Butt, G., Fielding, A., Foster, J., Gunter, H., Lance, A. Potts, L., Powers, S., Rayner, S., Rutherford, D., Selwoood, I. and Szwed, C. (2004). *The Evaluation of the Transforming the School Workforce Pathfinder Project*. RR541, London, Department for Education and Skills.

Van Duijn, M. A. J., van Busschbach, J. T. and Snijders, T. A. B. (1999). Multilevel analysis of personal networks as dependent variables. *Social Networks*, **21**, 187-209.

Van den Noortgate W., de Boeck, P. and Meukders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, **28**, 369-386.

Wiggins, R. D, Longford, N. and O'Muircheartaigh, C. A. O. (1992). *A variance component approach to interviewer effects. Survey and Statistical Computing*, A. Westlake, R Banks, C. D. Payne and T Orchard. Amsterdam, North-Holland.

Woodhouse, G., Yang, M., Goldstein, H. and Rasbash, J. (1996). Adjusting for measurement error in multilevel analysis. *Journal of the Royal Statistical Society, A*, **159,** 201-12.

Yang, M., Goldstein, H., Rath, T. and Hill N. (1999). The use of assessment data for school improvement purposes. *Oxford Review of Education***, 25**, 4, 469-483.

Yang, M. and Woodhouse, G. (2001). Progress from GCSE to A and AS level: institutional and gender differences and trends over time. *British Journal of Educational Research*, **27**, 2, 245-267.

Zaslavsky, A. M., Zaborski, L. B. and Cleary, P. D. (2004). Plan, geographical, and temporal variation of consumer assessments of ambulatory health care. *Health Services Research,* **39**, 5, 1467-84.

Zeger, S.L., Liang, K. and Albert, P. (1988). Models for longitudinal data. A generalized estimation equation approach. *Biometrics*, **44**, 1049-1060.

Zhou, X., Perkins, A. J. and Hsui, S. L. Comparison of software packages for generalized linear multilevel models. *American Statistician*, **53**, 282-290.

Copies of this publication can be obtained from:

DfES Publications
P.O. Box 5050
Sherwood Park
Annesley
Nottingham
NG15 0DJ

Tel: 0845 60 222 60
Fax: 0845 60 333 60
Minicom: 0845 60 555 60
Online: www.dfespublications.gov.uk

Produced by the Department for Education and Skills

ISBN 1 84478 797 2
Ref No: RR791
www.dfes.go.uk/research