Multilevel modelling of educational data

By
Harvey Goldstein
Institute of Education, University of London

1. Fundamentals: units and levels

While this chapter concentrates on educational data, many of the issues are more generally relevant. This is especially the case when we consider what we mean by a 'unit' and a 'level'.

Traditionally, pupils or students have been viewed as well-defined units, often at the lowest level of a data hierarchy. Likewise schools have typically been considered as well-defined units with students 'nested' or grouped within them. Thus, we would ordinarily say that a particular student 'belongs to' a particular school over a period of time. We immediately see, however, that this does not always accord with reality. For example, a student may move from one school to another during a study, or a school may change by splitting into two schools or being merged with another school. Likewise, if we consider the unit of a 'class' within a school, this may vary in its form and composition frequently, and any particular student may experience several different teachers with which she or he is studying a particular subject during a year. We would regard such a student as ‘belonging to’ or having a 'multiple membership' of the set of teachers and we shall describe such structures in more detail later. The same kinds of issues arise in demographic studies (Goldstein et al., 2001), health and other human sciences. In other words our definition of a 'unit' has to incorporate temporal information (for what period does it exist in a particular form) and the relationships *between* units have to be specified in terms of membership at particular periods. This has clear implications for longitudinal studies, but also applies in many cases to purely cross sectional data. Suppose, for example, that we are comparing student achievement test scores among schools at one point in time. It is well known that the use of such data to rank order schools is problematical because it fails to take account of students’ prior ‘intake’ achievements (see also Section 3). In addition, however, it assumes that the students can be assigned unequivocally to the schools they belong to at the time of the test. Typically there will be significant amounts of mobility among schools, so that many of the students will have been in their assigned schools a relatively short time and it will be unreasonable to expect those schools to have greatly influenced the student test scores.

It is important to distinguish between actual changes in the definition of a unit and merely changes in a unit's characteristics. Thus, for example, a school may change the number of students it has or its teachers, but may still be regarded as the same unit, and of course, when modelling the effects of school size or composition such changes can be taken into account. Even so, there may be borderline cases; a school may undergo reorganisation to such an extent that it changes nearly all of its staff, has a new name and perhaps new accommodation. If, at the same time, its student composition also changes with some moving to other schools and new students entering it may be more useful to consider this as a change of unit, with the new unit only coming into existence at that time. A choice in such a case will partly depend upon the research questions being asked and also upon the modelling feasibility of different approaches which I shall be discussing later.

There is another kind of unit whose definition is closely tied in with the definition of a 'level'. This is where a collection of lower level units belongs to a higher level unit which itself is defined solely (or partly) in terms of the particular units that belong to it. Thus, for example, a friendship group is defined solely by its members and will change when any are lost or acquired. Families and households have similar properties (Goldstein et al., 2001). At any one time an individual person may belong to several friendship groups (or learning groups for example) and this can be regarded as a multiple membership structure. Because group formation can change over the period of study, it will often be the case that there will exist many more groups than there are individuals. In such a situation we may encounter modelling estimation problems in terms of separating out different effects: in terms of levels it we may regard the groups as lower level units generating measurements which are ‘nested’ within individuals, where each individual can be thought of as generating a set of friends. For example, in the case of learning groups within a classroom, if the results of a group project are being measured then the basic response is measured at the group level rather than that of the student and, so long as there is adequate movement of individuals among groups, this can be modelled as a multiple membership structure with groups at level 1 and students at level 2. This example will be discussed in more detail in Section 4. In household studies a similar situation can arise when a household characteristic, such as electricity use, is being measured over time with individuals changing households. In both these cases we can, in principle, obtain estimates of the relative contributions of groups (households) and individuals and also estimate (posterior) effects for each individual or group. The ability of such models in education to take account of group learning and performance provides a powerful tool for many situations that formerly have posed considerable analytical problems.

In order precisely to define a complex data structure of the kind I have been describing we need to specify the membership relationships among all the units involved. In fact there is just a small number of types of relationship involved and these can conveniently be set out diagrammatically for any particular structure. I shall describe how this may be done below when looking at examples.

In the next section we shall look at the specification of a basic multilevel model followed by a section on cross-classified structures, with some examples. This is followed by a section on multiple membership models, also with examples, and a final section in which I reflect on further applications and extensions.

2. The basic multilevel model.

Before describing the basic multilevel model, it is useful to reflect on why such models are necessary. For many years, educational researchers discussed the ‘units of analysis’ problem, one version of which has also been called the ‘ecological fallacy’ (see also the chapter by Diez-Roux). At one extreme, it is possible to study relationships among variables ignoring group structures. At the other extreme we can work solely with group, say school, averages in exploring relationships. Aitkin and Longford (1986) set out the statistical issues associated with various procedures. In an earlier analysis Aitkin et al (1981) reworked a well known study on teaching styles which used student level data but ignored school membership (Bennett, 1976). They showed that formerly ‘significant’ results became non-significant when a multilevel model was used. Woodhouse and Goldstein (1989) showed how the use solely of aggregate level data based upon school means could lead to unstable and misleading conclusions.

In addition to the problem of making misleading inferences, failure to model both students and schools simultaneously makes it impossible to study the extent to which school and student characteristics interact to influence the response measurement or measurements. This can only be done within the context of a multilevel model as I shall now describe.

For simplicity consider a simple data structure where a response is measured on individual students or students in a number of schools, together with one or more explanatory variables (covariates). Instead of schools we could think of households, areas, etc. We wish to model a relationship between the individual response and the explanatory variables, taking into account the possibility that this relationship may vary across schools. The response might be a continuous variable such as a test score or survival time, or a discrete variable such as an attitude towards schooling. We shall assume in what follows that we are dealing with a continuously distributed response, and for simplicity that this has a Normal distribution. Extensions to other kinds of responses follow similar lines and these are discussed by Goldstein (1995). This is a simple 2-level structure with the schools as higher level units and students as lower level units. A simple such model can be written as follows

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{ij} + u_{0j} + e_{ij} \\ \mathrm{var}(e_{ij}) &= \sigma_e^2 \\ \mathrm{var}(u_j) &= \sigma_{u0}^2 \end{aligned} \qquad (1)$$

where y_{ij} is the response and x_{ij} the value of a single explanatory variable (covariate) for the *i*-th individual in the *j*-th school. For example, the response might be an examination score measured on students at the age of 16 years and the explanatory variable a test score measured on the same students five years earlier at the age of 11 years (Figure 1 shows this schematically for three schools).

(Figure 1 here)

The slope coefficient β_1 is for the present assumed to be the same for all the schools while the random variable u_{0j} represents the departure of the *j*-th school's intercept from the overall population intercept term β_0. The first two terms on the right hand side of (1) constitute the fixed part of the model and the last two terms describe the random variation. As mentioned we shall develop the model initially assuming that the random variables have a Normal distribution

$$u_{0j} \sim N(0, \sigma_{u0}^2) \qquad e_{ij} \sim N(0, \sigma_e^2)$$

Note that we are at liberty to 'fix' one or more of the u_{0j} using an associated dummy (0,1) variable as an explanatory variable, for example if we knew that it was special and should not be considered as a member of the same population as the remainder.

This is often useful for exploring ‘outliers’ (Langford and Lewis, 1998). The key point about these random variables is that it allows us to treat the samples of units as coming from a universe or population of such units. Thus, the schools (and students) chosen are not typically the principal focus of interest; they are regarded as a random sample from a population of schools and we are concerned with making statements about that population, for example in terms of its mean and variance.

We can elaborate (1) by allowing the coefficient β_1 to vary across schools and rewrite the model in the more compact form

$$
\begin{aligned}
&y_{ij} = \beta_{0ij} x_0 + \beta_{1j} x_{1ij} \\
&\beta_{0ij} = \beta_0 + u_{0j} + e_{ij} \\
&\beta_{1j} = \beta_1 + u_{1j} \qquad (2) \\
&\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N\left(0, \Omega_u\right) \quad \Omega_{\mathrm{u}} = \begin{pmatrix} \sigma_{u0}^2 & \\ \sigma_{u01} & \sigma_{u1}^2 \end{pmatrix}, \\
&(e_{ij}) \sim N(0, \sigma_e^2), \quad x_0 = 1
\end{aligned}
$$

This model is often referred to as a ‘random coefficient model’ by virtue of the fact that the coefficients β_{0ij} and β_{1j} in the first equation of (2) are random quantities, each having a variance with a covariance between them. As more explanatory variables are introduced into the model, so we can choose to make their coefficients random at the school level thereby introducing further variances and covariances, and this will lead to models with complex covariance structures. One of the aims of multilevel modelling is to explore such potential structures and also to attempt to explain them in terms of further variables.

Having fitted such a model we can obtain estimates for the individual ‘residuals’ (u_{0j}, u_{1j}, e_{ij}) at either level by estimating their expected values (or other functions of their distributions), given the data and model estimates. Thus, for example, we can estimate $E(u_{0j}, u_{1j} | Y, \beta, \theta)$ where

$$\beta^T = \{\beta_1, \beta_2\} \quad \theta = \{\sigma_{\mathrm{u0}}^2, \sigma_{u01}, \sigma_{u1}^2, \sigma_e^2\} \qquad (3)$$

and substituting model estimates for the unknown parameters. The multilevel model is here described in non-Bayesian terms. For a full Bayesian specification of this model we would need to add prior distribution assumptions for the parameters in (3). The interested reader is referred, for example, to Rasbash et al. (2000) for details with worked examples. These procedures are all implemented in the software package *MLwiN* (Rasbash et al., 2000) and reference will be made to other features of this package in what follows.

3. Cross classified models.

Suppose a student is classified as belonging sequentially to a particular combination of primary (elementary) school and secondary (high) school and we have followed a sample of such students through each school and wish to relate measurements made at the end of secondary school to those made earlier in the primary schools. The students will be identified by a *cross classification* of primary schools and secondary schools. Note that even if we did not have prior measurements, but *did* have identification of the primary and secondary schools we could still carry out a cross classified analysis.

Raudenbush (1993) and Rasbash and Goldstein (1994) present the general structure of a model for handling such random cross classifications. In our example we have a cross classified structure which can be modelled as follows:

$$\begin{aligned} &y_{i(j_1 j_2)} = (X\beta)_{i(j_1 j_2)} + u_{j_1} + u_{j_2} + e_{i(j_1 j_2)}, \\ &j_1 = 1, \ldots J_1, \quad j_2 = 1, \ldots J_2, \quad i = 1, \ldots N \end{aligned} \tag{4}$$

in which the score of student *i*, belonging to the combination of primary school j_1 and secondary school j_2, is predicted by a linear 'regression' function denoted by $(X\beta)_{i(j_1,j_2)}$. The random part of the model is given by two level 2 residual terms, one for the primary school attended by the student (u_{j_1}) and one for the secondary school attended (u_{j_2}), together with the usual level 1 residual term for each student. We note that the latter may be further modelled to produce complex level 1 variation (Goldstein, 1995, Chapter 3), allowing for example for separate variances for males and females, etc. This applies to all our models.

As an example consider the analysis carried out by Goldstein and Sammons (1997) who fitted a series of cross classified models to a cohort of 758 students from the age 0f 8 years in primary school to the age of 16 years in secondary school. There were 48 primary schools and 116 secondary schools involved. They had achievement measures at the ages of 8, 11 and 16 years and the principal aim was to separate the effect of primary school attended from that of secondary school. The results are presented in table 1.

(Table 1 here)

Column C presents a model which includes gender and a crude measure of disadvantage (eligibility for free school meals) but no prior achievement scores, and it shows that the between-primary school variation is about three times that between secondary schools. This is partly explained by the fact that secondary schools are much larger than primary schools, but the relative importance of primary schools for achievement at the end of secondary schools is nevertheless notable. Column B shows the effect of adding in achievement measures at the start of secondary schooling so that the response effectively measures adjusted achievement or 'secondary school progress'. Both the school level variances are reduced but the ratio is relatively unaltered. In both these analyses the level 1 (student) variance has an extra term to allow for different variances for males and females. Thus, in column C the variance between females is 0.74 whereas that between males is 0.74+2 x 0.10 = 0.94 (for technical details on how this is specified see Goldstein, 1995, Chapter 3). In Analysis A in addition the between-student variance is allowed to vary also as a function of the LRT score. Thus at an extreme, high, LRT score of 2, the variation between boys is 0.50+2 x 0.092 +2 x 2 x 0.093 + 4 x 0.033 = 1.19, whereas for girls with average LRT scores it is just 0.50.

When primary school is ignored the apparent variation due to secondary school attended is estimated to be more than twice the value found in the cross classified models. The substantive importance of this for studies of schooling is that it becomes necessary to take account of achievement during periods of schooling prior to the one immediately being considered (secondary here). The researchers also carried out a bivariate response model where the 11 year achievement scores and the 16 year achievement score become responses. The simple correlation between these responses was 0.53 which reduces to 0.29 when adjustment is made for the 8 year achievements and gender (from a subsequent analysis of the data). When a cross classified model is fitted, we can decompose this correlation into a between-(primary)-school and a between-student correlation. After adjusting for the 8 year achievement score the between-student correlation was just 0.16 and the between-primary school correlation was 0.81. This suggests that in terms of progress (that is after adjusting for the 8 year prior achievement) the moderate overall correlation of 0.29 is an average of a large correlation between the effects of primary schools on these two achievements and a much smaller correlation at the student level. This further supports the finding of the persistence of the primary school effect throughout secondary schooling.

A special case is also of some interest, where some students do not belong to any units of one 'arm' of the cross classification - an incomplete cross classification. Consider a longitudinal study of children entering school for the first time in reception (kindergarten) classes. Some of these will have attended a form of pre school provision, say a nursery school. We wish to model achievement at the end of the reception year as a function of the effect of the school, together with that of the nursery where this was present. To do this we have a cross classified model for one set of students and a simple hierarchical model for the other, with a common school effect. If we thought that the school effect was different for the two groups of children then we would simply fit separate effects (with different variances etc.) and the two effects would be allowed to covary at the school level.

The examples illustrate how a cross classified model can avoid misspecification and misleading inferences as well as providing a level of detailed analysis not possible when a purely hierarchical structure is assumed. Even so, there is a further major problem with the above example. Only those students were included for analysis who had complete schooling records throughout the period and any changes in school between 8 and 11 years and between 11 and 16 years were ignored. In the next section I shall show how these problems can be overcome.

4. The multiple membership model

To illustrate this model consider just the secondary schools from the above example, and suppose that we know, for each individual, the weight w_{ij_2}, associated with the j_2-th secondary school attended by student i with $\sum_{j_2=1}^{J_2} w_{ij_2} = 1$. These weights, for example, may be proportional to the length of time a student is in a particular school during the course of the longitudinal study. Note that we allow the possibility that for some (perhaps most) students only one school is involved so that one of these probabilities is one and the remainder are zero. Note that when all level 1 units have a single non-zero weight of 1 we obtain the usual purely hierarchical model. We can write the following model for the case of membership of just two schools {1,2}:

$$
\begin{aligned}
&y_{i(1,2)} = (X\beta)_{i(1,2)} + w_{i1}u_1 + w_{i2}u_2 + e_{i(1,2)} \\
&w_{i1} + w_{i2} = 1
\end{aligned}
\qquad (5a)
$$

and more generally:

$$
\begin{aligned}
&y_{i\{j\}} = (X\beta)_{i\{j\}} + \sum_{h=1}^{J} w_{ih}u_h + e_{i\{j\}} \\
&\sum_h w_{ih} = 1, \quad var(u_h) = \sigma_u^2 \\
&var(\sum_h w_{ih}u_h) = \sigma_u^2 \sum_h w_{ih}^2,
\end{aligned}
\qquad (5b)
$$

The notation $h \in \{j\}$ means for all schools *(h)* that belong to the set of schools $\{j\}$.

In the particular case of membership of just two schools with equal weights we have $\mathrm{w}_{\mathrm{i}1} = w_{i2} = 0.5, \quad \mathrm{var}(\sum_h w_{ih}u_h) = \sigma_u^2 / 2$. In other words the contribution to the level 2 variation is just half that for a student who remains in one school, since in the former case the level 2 contribution is averaged over two (random) schools. Note that, if we ignore the multiple membership of schools and simply assign students, say, to the final school that they attend, we will underestimate the true extent of between-school variation. This is because, for those students who do attend more than one school, the true level 2 variation is less than that for students who attend a single school. In the model, however, we assume that the level 2 variation for these students is the same as that for those attending a single school, with the result that the overall level 2 variation is underestimated.

A slightly different notation to describe membership relationships is used by Browne et al. (2001). This is particularly useful when we have very complex structures involving mixtures of hierarchical, crossed and multiple membership classifications. Essentially it works by requiring just a single unique identification for each lowest level observation, in the present case a student. Each student then has a relationship with every other type of unit, that is in the present case, they are classified into primary and secondary schools, the model specifies which classifications are involved and the data structures specifies precisely which schools are involved for each student. Thus the cross classified model (4) would be written as follows

$$y_i = (X\beta)_i + u^{(2)}_{secondary(i)} + u^{(3)}_{primary(i)} + e_i,$$
$$i = 1,...N$$

where *primary(i)* and *secondary(i)* refer respectively to the primary and secondary schools attended by student *i*. The superscripts for the random variables identify the classification; where this is absent, and if there is no ambiguity, it is assumed to be the lowest level classification (1).

The multiple membership model (5b) would be written as

$$y_i = (X\beta)_i + \sum_{h \in school(i)} w_{i,h} u^{(2)}_h + e_i$$
$$\sum_{h \in school(i)} w_{i,h} = 1, \quad var(u^{(2)}_h) = \sigma^2_u \qquad (7)$$

Adoption of this notation for complex data structures is particularly useful for fitting such models within an MCMC framework which involves repeated estimation via conditional chain sampling for each unit's random effects. As Browne et al. (2001) point out, MCMC procedures become necessary when structures reach certain levels of complexity. MCMC models involve generating a chain of (correlated) values sampled from the 'posterior' distribution for each parameter of the model, including the random effects or residuals. At each step of the procedure a sample value for a parameter is drawn from the appropriate distribution conditional upon the observed data and the current values of all the other parameters in the model. This means that storage requirements are moderate. It also means that very complex structures can be defined simply by specifying how each parameter relates to the other parameters in the model.

Browne et al (2001) also use simple diagrams for representing complex structures. Thus the cross classified structure modelled in Table 1 can be represented as follows

(Figure 2 here)

The single directional lines indicate a membership relation, and here students are members of just one secondary school and one primary school. Where multiple membership is involved, two parallel lines are used (see below).

While the notation of (6) and (7) is powerful, we may often wish, as below, to retain the notation used in (4) and (5) since this may be more familiar and because it is easy to specify when structures are not highly complex.

Equations (5) and (7) specify a 2-level model where the level 2 variation among secondary schools is modelled using the set of weights for each student across all schools as explanatory variables. A similar formulation can be used to model the case where, for some students, there is no identification of the school(s) to which they belong. If we are able to assign a set of probabilities of membership among a subset of schools, however, then utilising the (square root) of these probabilities as weights (standardised to sum to 1) we can still carry out a valid analysis (Hill and Goldstein, 1998).

An extension of (5) is also possible and has important applications, for example in modelling spatial data. In this case we write

$$y_{i\{j_1\}\{j_2\}} = (X\beta)_{i\{j\}} + \sum_{h\in\{j_1\}} w_{1ih}u_{1h} + \sum_{h\in\{j_2\}} w_{2ih}u_{2h} + e_{i\{j\}}$$
$$\sum_h w_{1ih} = W_1, \quad \sum_h w_{2ih} = W_2, \quad var(u_{1h}) = \sigma^2_{u1} \quad var(u_{2h}) = \sigma^2_{u2} \qquad (8)$$
$$j = \{j_1, j_2\}$$

There are now two sets of higher level units which influence the response. In spatial models one of these sets is commonly taken to be the single area where an individual (level 1) unit lives and the other set consists of the neighbouring units that have an effect. The total weights for each set will need to be carefully chosen; in spatial models the W_1, W_2 are typically chosen each to equal 1 (see Langford et al, 1999 for an example). Another application of such a model for household data is where households share facilities, for example an address. In this case the household that an individual resides in will belong to one set and the other households at the address will belong to the other set. We can readily extend (8) to the case of multiple sets - which can be thought of as a multiple cross classification of multiple membership classification sets. This will allow us additionally to incorporate multiple spatial structures into, for example, household models.

In education we have an analogous situation to the traditional spatial correlation one. In the traditional schooling (school effectiveness) model a particular school is assumed to have an effect on student performance, attitudes etc., which is independent of any other school in the population. In practice, however, such an assumption will often be unreasonable since schools to some extent will be in 'competition' with each other for a share of resources (teachers, buildings etc.) and possibly also for certain kinds of students (see Goldstein, 2000 for a discussion of this). This leads to a 'spatial' type model, for example (see Langford et al., 1999)

$$
\begin{aligned}
&y_{i\{j_1\}\{j_2\}} = (X\beta)_{i\{j\}} + u_{1j_1} + \sum_{h \in \{j_2\}} w_{2ih} u_{2h} + e_{i\{j\}} \\
&\sum_h w_{2ih} = W_2, \quad var(u_{1h}) = \sigma^2_{u1} \quad var(u_{2h}) = \sigma^2_{u2} \qquad (9) \\
&j = \{j_1, j_2\}
\end{aligned}
$$

where the value of W_2 and the component weights will need to be carefully determined. In practice it would be advisable to carry out sensitivity analyses, trying different values for these weights. The component weights would need to take account of geographical catchment areas as well as any administrative arrangements that might affect inter-school relationships. Model (9) identifies two random effects for each school, the usual one, U_1, which is the direct effect on the students within the school and U_2, the (weighted) effect of the school on surrounding schools. Thus, we may be interested in the relative sizes of the latter among schools and also the correlation between U_1 and U_2 . We can explore other topics: for example, by fitting further explanatory variables we can study the extent to which such variables ‘explain’ the effect of each school on surrounding schools. In these ways we can address directly issues of how competition among schools operates and how the nature of the relationship between the effect of a school on the students within it and its effect on students in other schools.

We can further elaborate (9) by identifying subsets of the 'spatial' effects (U_2) corresponding, for example, to different school types which might be supposed to have differing kinds of effects and we can also of course include random coefficients for any of the classifications. Note that in the fixed part of the model, $(X\beta)_{i\{j\}}$, we can incorporate characteristics measured on the schools as well as on students. We can also incorporate ‘spatial’ effects into the fixed part of the model. For example the difference in social background between a school and the (weighted) average social background of neighbouring schools may be an important predictor of performance or behaviour, in addition to the social backgrounds of the students within a given school. If we have longitudinal data we can incorporate prior achievements of students in a given school and its neighbours, and if mobility takes place among schools then this can be incorporated by fitting a multiple membership structure for the U_1. Additionally, if we measure successive cohorts of students passing through a set of schools we can model the between-school variances as functions of time in order to obtain inferences about changing influences.

Models such as (9) can also be applied to demographic data on households, where the attitudes, opinions, consumption habits, etc. of a household, may be influenced by surrounding households. Most commonly such effects are modelled within a traditional hierarchical model where 'neighbourhoods' are defined, say using administrative regions, with households nested within them. Such models, however, have the drawback of using typically arbitrary definitions of spatial units. The advantage of a multiple membership formulation is that it requires only that a suitable 'distance function' determining the weights should be constructed. Of course, there may well be problems in defining such a function, since it may not be simply the Euclidean distance and there may be important problems to do with obtaining the necessary data. Nevertheless, as an approach it does have the potential for solving the spatial unit definition problem, especially with the advent of comprehensive Geographical Information Systems (GIS).

We see, therefore, that these models, in principle, enable us to provide adequately contextualised descriptions of schooling that can incorporate the social nature of institutions as described by their effects on each other. In particular these models are relevant to political and social debates about the effects of 'market competition' between schools and how this affects performance and other factors, especially in terms of trends over time. Existing discussions of this issue are typically carried out at aggregate (school) level and therefore are unable to obtain valid estimates of between-school variation (see for example Gorard, 2000).

To illustrate the flexibility of these models consider again the example of modelling learning groups discussed earlier. It was pointed out that where the response was modelled at the group level we had a multiple membership model where groups 'belonged' to individuals, and a model such as (7) could be used. Suppose, in addition to measuring outcomes at the group level we also have a measure of achievement or attitude at the student level. Recalling that the groups are defined as level 1 units the group response will have an individual component and this will generally be correlated with the response at the student level. We could therefore write such a model as

$$
\begin{aligned}
& y_{1i} = (X_1\beta_1)_i + \sum_{j \in group(i)} w_{i,j} u_{1j}^{(2)} + e_i^{(1)} \quad \text{(group response - } i\text{)} \\
& \sum_{j \in group(i)} w_{i,j} = 1 \\
& y_{2j} = (X_2\beta_2)_j + u_{2j}^{(2)} \quad \text{(individual response - } j\text{)} \\
& cov\ (u_{1j}, u_{2j}) \neq 0
\end{aligned}
\qquad (10)
$$

This is a bivariate response model with one response at each level. The first equation refers to a group response and, given suitable data with individuals belonging to different groups, can be used to estimate individual and group effects. The second equation models an individual student response and from the complete model we can directly estimate the correlation between a student's contribution to the group response and their individual response. Figure 3 shows the relationships using a double arrow for the multiple membership of groups within children and a dotted line joining the two child ‘effects’ to indicate a bivariate response model.

We can also identify those individuals who may be discrepant, say with low contributions to the group response but high individual effect and this might be an important diagnostic for learning potential. An alternative formulation for some purposes would be to incorporate the individual level measure as a covariate in the model for the group response. If, however, we had sequential measures on individuals then we might wish to fit trend terms with random coefficients and then the full bivariate response formulation becomes necessary (see Goldstein, 1995 Chapter 6 for such a model in the purely hierarchical case). Further elaborations can be introduced, for example by modelling classes of students (containing learning groups) within teachers and/or schools and so forth.

It is perhaps worth mentioning that multiple membership models bear a close relationship to fuzzy sets (see for example, Manton et al, 1994, for an introduction, and Haberman, 1995 for a critique) where individual units also can belong to several groups at a time, with 'membership coefficients' being equivalent to our weights. There appears to be no explicit application of fuzzy set theory, however, to general hierarchical structures.

In the final section I will explore the nature of the new kinds of knowledge that the application of such models is able to generate.

5. Types of response

We have so far considered models with a continuous (Normally distributed) univariate or multivariate response variable. Multilevel models can, however, be formulated for other response types. One of the most common is that for a binary response, for example whether or not a student passes an examination. In such a case the equivalent to (1) can be written as

$$E(y_{ij} \mid j) = \pi_{ij}, \quad y_{ij} \sim Binomial(1, \pi_{ij})$$

$$y_{ij} = \begin{cases} 0 & \text{if fail} \\ 1 & \text{if pass} \end{cases} \qquad (11)$$

$$\text{logit}(\pi_{ij}) = \log\{\pi_{ij}/(1-\pi_{ij})\} = \beta_0 + \beta_1 x_{1ij} + u_{0j}$$

$$\text{var}(u_{0j}) = \sigma^2_{u0}$$

This can then be extended in all the same ways as (1) to cross classifications etc. This is one example of a generalised multilevel linear model and another example is where we have a count as the response variable.

An important class of models is where the response is a duration. These models are often known as survival or event history models and again the multilevel formulation involves the same kinds of generalisations as with (1). So called parametric and semi-parametric models can be fitted and Goldstein (1995 Chapter 9) gives details. A particularly useful formulation for multilevel structures is the ‘piecewise’ model. Here the time interval is divided into small time segments and the occurrence of events noted. With a suitable specification we can then fit models using (11) together with all its generalisations.

All of these different kinds of response model assume Normal variation at higher levels. This implies that we can create multivariate response models where we have a mixture of different kinds of responses. For example, we nay measure the duration of tasks undertaken by children as well as their scores on a test and the responses of a group to which they belong. In principle all of these can be incorporated into a single linked model.

Finally, we can have models with underlying latent structures (see McDonald and Goldstein, 1989). These may be factor analysis models with different factor structures at each level or classification or more general structural relation models where the factors are related to each other.

6. Final thoughts about new insights

It is worth making the general point that all of these complex models require adequate, and generally large, data sources. Obtaining such data often will be costly and time consuming. Thus, as well as posing methodological challenges in terms of formulation and estimation there is the overwhelming prior condition that suitable data are available.

It would be agreed generally that the reality of education and learning involves complex processes and interactions. This does not preclude providing useful information in terms of simplified models or even simple summaries. Nevertheless, within the context of statistical modelling, I would argue that unless we can formulate models (based on appropriate data) that approach the real underlying complexity our descriptions of these processes will be incomplete and that this will hinder adequate causal explanations and understandings. For example, the models described in this paper may be suitable for the analysis of ethnographic data collected from student observations in ways that can preserve the detail with which the data are collected. Such data will often incorporate details of interactions among students and groups of students within a classroom in an attempt to understand the ways in which peer group influences may work. We may be able to begin to model such data using the spatial models and group response models that have been described. A simple example would be observations made on the tasks children are engaged on and the length of time they spend on these. Traditionally, such measurements would be assumed to be independent from child to child, i.e. the length of time spent by one child does not affect the length of time spent by a nearby child. This, however, is debateable and a spatial type model would be useful here. Furthermore, where such data are collected over a long time period, with changing group or spatial structures, the multiple membership models can be used to characterise these. By embedding such data within a formal modelling framework we can extend the range of existing analyses and construct the foundation for secure generalisations and in this way help to secure a link, that is a common framework, between existing quantitative and qualitative research traditions. As we explore deeper levels of complexity so we would expect to generate new theoretical insights and the process of developing the models themselves will often suggest new directions for data collection and theory development.

A further important class of models are those known as 'meta analyses'. These seek to combine results from several studies in order to achieve a more reliable inference. Thus, in the simplest case, each study is a level 2 unit and individuals within a study are level 1 units. A 2-level model for such studies can allow overall variability as well as variability in relation to particular coefficients. Thus, for example, if each study is concerned with the effect of class size on achievement we can see whether the relationship with class size varies from study to study, as well as estimating the average effect. We can also incorporate studies where data are available only at aggregate level, rather than at individual level. This is particularly important since many published studies quote only aggregate results. Goldstein et al (2000) discuss how such studies can be incorporated into a general multilevel framework. This is done using a device similar to that used in (10) where we have two equations, one of which models a level 2 response and one models a level 1 response. The level 1 response equation has a level 2 random effect as does the level 2 response equation. By allowing the level 2 random effects to be correlated we link the two equations and this linking allows efficient estimation of all the parameters.

While this chapter has been concerned largely with education, with some reference to demography, almost all the issues covered apply to other areas for example health, and collaboration between such disciplinary areas will be fruitful. Finally, there is a wealth of further information about applications and current developments that can be found on the website of the London based Multilevel models project (http://multilevel.ioe.ac.uk/).

References

Aitkin, M., Bonnet, S. N. and Hesketh, J. (1981). Teaching Styles and Pupil Progress: a reanalysis. *British Journal of Educational Psychology* **51**: 170-186.

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

Bennett, S. N. (1976). *Teaching styles and pupil progress*. London, Open Books:

Browne, W. J., Goldstein, H. and Rasbash, J. (2001). Fitting complex model structures to large datasets: a Monte Carlo Markov chain (MCMC) algorithm to fit multiple membership multiple classification models. *Submitted for publication*.

Goldstein, H. (1995). *Multilevel Statistical Models*. London, Edward Arnold: New York, Wiley.

Goldstein, H. and Woodhouse, G. (2000). School effectiveness research and Educational Policy. *Oxford Review of Education* **26**: 353-363.

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

Gorard, S. (2000). *Education and Social Justice*. Cardiff, University of Wales Press:

Haberman, S (1995), Review of *Statistical applications using fuzzy sets*, J. American statist. Assn., 90, 1131-1133.

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

Langford, I. and Lewis, T. (1998). Outliers in multilevel data. *Journal of the Royal Statistical Society, A.* **161**: 121-160.

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

Manton, R. G., Woodbury, M. A., and Tolley, H. D. (1994). *Statistical applications using fuzzy sets.* New York, Wiley.

McDonald, R. P. and Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British Journal of mathematical and statistical psychology* **42**: 215-232.

Rasbash, J. and Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross classified random structures using a multilevel model. *Journal of Educational and Behavioural statistics* **19**: 337-50.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., et al. (2000). *A user's guide to MlwiN (Second Edition).* London, Institute of Education.

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.

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

Woodhouse, G. and Goldstein, H. (1989). Educational Performance Indicators and LEA league tables. *Oxford Review of Education* **14**: 301-319.

Figure 1

A simple multilevel data structure for three schools with parallel slopes.

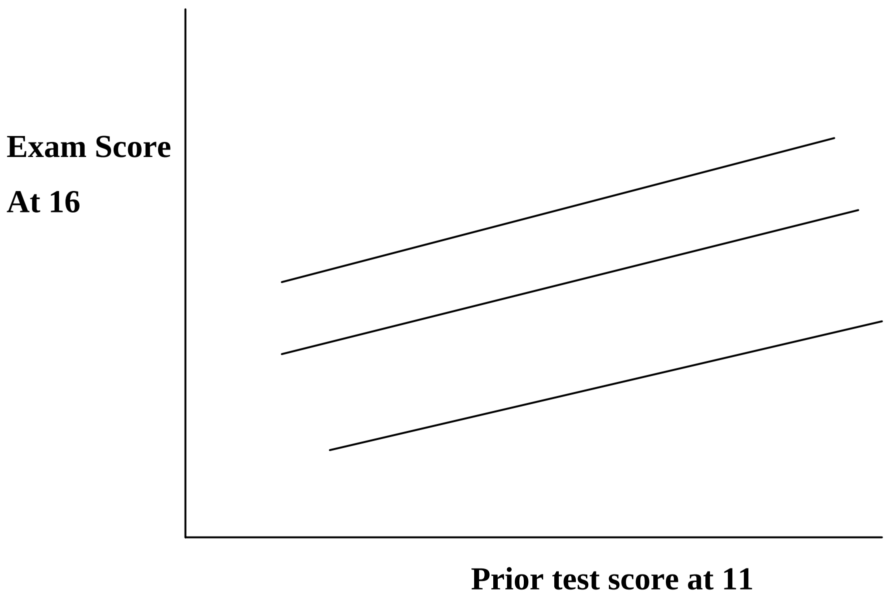

Figure 2

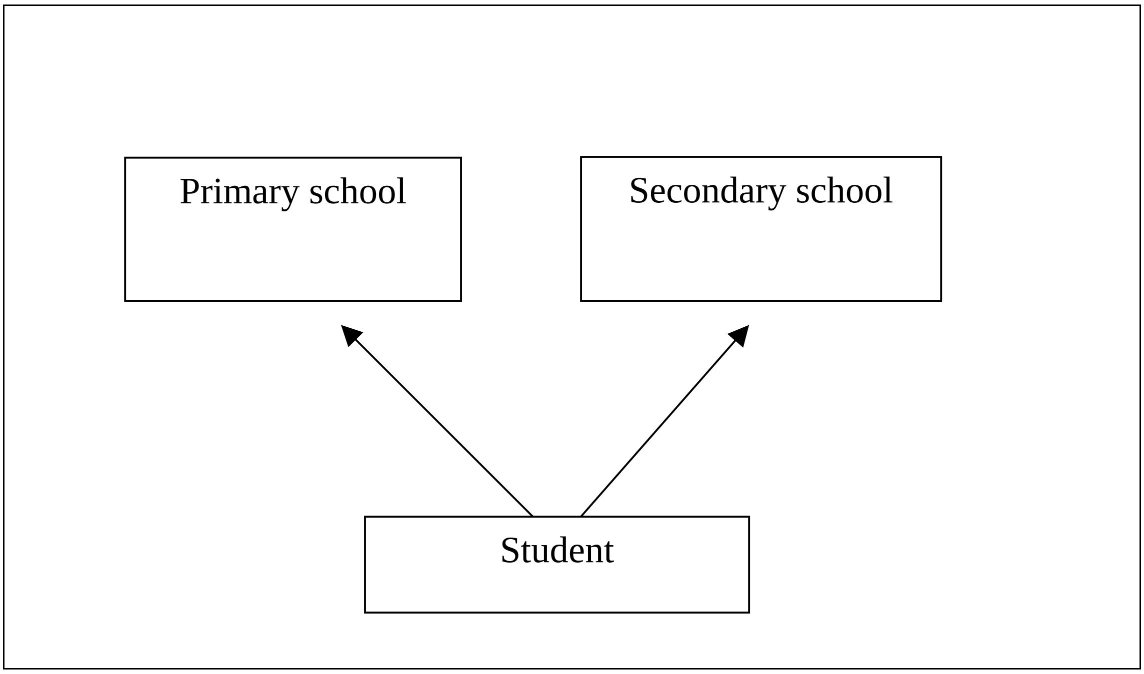

Figure 3.

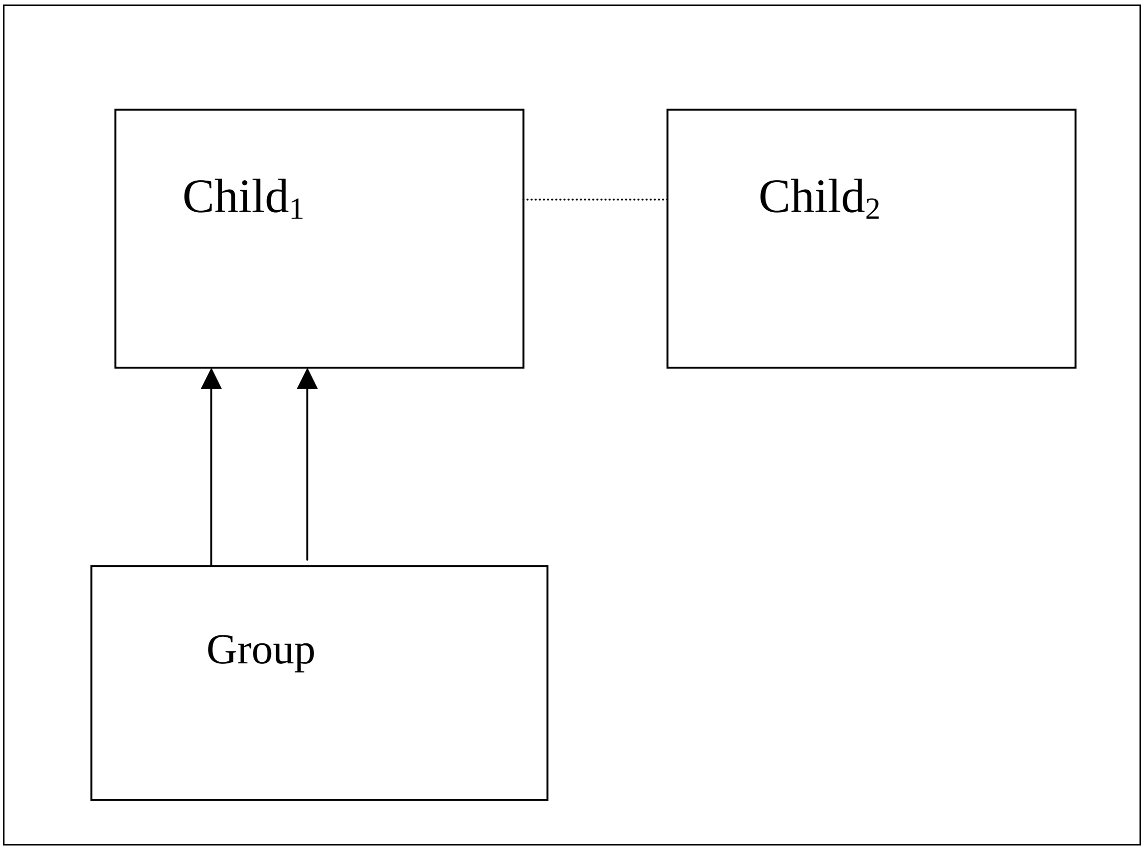

Table 1. Variance components cross-classified model for 16-year-old exam score as response. The exam score and the LRT (Reading test) score have been transformed empirically to have N(0,1) distributions. Free school meal is a binary (yes, no) variable. At level 2 the subscript 1 refers to Primary and 2 to Secondary school. At level 1 the subscript 0 refers to the intercept, 1 to males and 2 to LRT. At the age of 11 years students were allocated to a grouping based upon verbal reasoning scores: VR1 (the base category) comprises the lowest 25%, VR2 the middle 50% and VR3 the highest 25%.

	A	**B**	**C**
Fixed			
Intercept	0.51	0.50	0.25
Males	-0.21 (0.06)	-0.19 (0.06)	-0.34 (0.07)
Free school Meal	-0.22 (0.06)	-0.23 (0.06)	-0.37 (0.08)
VR2 band	-0.39 (0.08)	-0.38 (0.08)	
VR3 band	-0.71 (0.13)	-0.71 (0.13)	
LRT score	0.31 (0.04)	0.32 (0.04)	
Random			
Level 2:			
(Primary) σ^2_{u1}	0.025 (0.013)	0.036 (0.017)	0.054 (0.024)
(Secondary) $\sigma^2_{u_2}$	0.016 (0.014)	0.014 (0.014)	0.019 (0.02)
Level 1:			
σ^2_{e0}	0.50 (0.06)	0.55 (0.06)	0.74 (0.05)
σ_{e01}	0.092 (0.03)	0.06 (0.03)	0.10 (0.05)
σ_{e02}	0.093 (0.018)		
σ^2_{e2}	0.033 (0.022)		
-2Log likelihood	1848.8	1884.2	2130.3