Explanatory Modelling of Complex Social Structures with Case Studies in Educational Research

Antony Fielding[1]

Senior Lecturer in Social Statistics, Department of Economics, University of Birmingham

& Visiting Research Fellow, Multilevel Models Project, Institute of Education, University of London

DRAFT

ESRC/BERA Advanced Training Workshop
6-8th July 2000.
School of Education, University of Birmingham

[1] E-mail: A.Fielding@bham.ac.uk. Part of the work has been supported by the Economic and Social Research Council under an award H51944500497 of the Analysis of Large and Complex Datasets Programme.

Prologue

‘ We dance around in a ring and suppose. The secret sits in the middle and knows’ (Robert Frost)

1 Design and explanation

The possibility of explanation through causal inferences is often at the forefront as a priority in the planning of statistical investigations. It is well known that in the social sciences such investigations can rarely have the design conditions that make such inferences in physical and some biological sciences very tightly based. Classical statistical methods and inference have their roots in investigations in such areas, and in particular the stimulus given by agricultural experimentation. The notion of random allocation of experimental subjects to treatment groups or conditions has been at the heart of this tradition. Complex designs emerge as necessary to conduct investigations with sufficient statistical precision and to encompass more intricately combined treatment effects. It is of course the even greater complexity of social processes , even if random allocation of individuals to interesting conditions were possible, that makes these design conditions more difficult to establish.

Causal explanation in the social sciences is in most senses far too wide a philosophical concept for statisticians to encompass it as the preserve of themselves alone. Statisticians, however, have a role to play in designing studies which will bear the weight of substantive inferences from collected data. They also aid in developing appropriate analytical techniques to further the aims of explanation when strict design conditions are far from ideal. Inappropriate inferences are often made by inexperienced users of statistical methods , who do not respect the conditions of design under which data has been collected. We will discuss later, for instance how cluster structure of sample data either by design or by necessity needs to be explicitly considered before meaningful analyses can emerge.

In both design and analysis there must be an interplay between the statistical treatment of data and the theoretical underpinnings of the subject matter. The latter can only be informed by experts in that subject matter and not the statistician. If a causal theory is under review the statistician attempts to provide the best approaches possible given the limitations of the available design and data. He might also have informed the design of the data collection had he been consulted beforehand. Far too often inadequate or incomplete data have been collected through deficient design. The dialogue between statistician and subject specialist is a necessary pre-requisite of good explanation. The range of statistical approaches to (causal) explanation have also expanded from traditional analysis of designed experiments to statistical modelling of complex processes. Much has been achieved towards developing explanations by the modelling of data collected in ways to reflect these processes but which depart from strictly controlled and randomised designs. However, to be successful models of the process and data must be informed by theoretical and empirical knowledge of the process. Any statistical model attempts to approximate the realities of the world given the design and conditions under which the data has been collected. In uncovering explanations of phenomenon or evaluating them they are only as good as the data collected and the thought that has gone into how it is collected. The aims of statistical modelling are broad but as part of the strategy of suggesting causal explanations they have a role to play, providing that role is not imposed too rigidly. The general area of causal explanation from non-experimental data is too deep , and possibly too difficult a subject , to be fully explored here. However, in the case studies to be discussed we hope some insight may be gained into these issues. We will see that by careful statistical modelling and careful attention to what data is collected and how, we may get somewhere in the thorny problem of attempting causal explanations. Some potential explanations may be ruled out if taken in conjunction with informed substantive theory. Perhaps more importantly evidence may also be gathered in favour of others. Deterministic proof of the existence of a causal explanation can rarely or never be established but attention to adequate modelling and design may propagate evidence of its plausibility. We hope to give some indication of some of the conditions under which developing advanced statistical methodology can further this end.

However, unless there is the real constructive interplay between subject specialist and methodologist some misunderstandings about the role of modelling in explanation may arise. A worthy set of criticisms and recommendations of school effectiveness research which focuses to a large extent on explanation in statistical modelling is provided by Coe & Fitz-Gibbon (1998). Many instances are cited of where more is often read into model fitting results by statistically unaware observers than the circumstances or the model assumptions warrant. However, to anticipate the later discussion of the role of control variables in modelling, we may take on board one of their particular criticisms. Broadly this objects to the inclusion of certain types of control as explanatory variables in the statistical models used. They say ‘An example of such ungrounded modelling may be found in the use of such variables as ‘sex’ or ‘ethnic origin’ which ‘explain’ (in a statistical sense) part of the variation in outcomes, but which do not explain differential performance in any true sense—unless it is argued that effects result from purely biological differences or from unfair discrimination’ They go on to argue that this results in stereotyping of all girls, for instance, rather than focusing on some more direct explanation, e.g. gender differences in spatial visualisation skills (Hodgson (1995)). It might be noted that implicitly in this criticism is a notion that the interpretation put on the model gender effect is a causal one. No statistician would countenance this. However, it must also be recognised that there is another aspect to the use of modelling in explanatory research that has not been considered. If gender effects are established by the model then this has uncovered a phenomenon that is in search of an explanation. If enough data and attention to study design had been informed by theoretical and empirical knowledge then competing explanations could be evaluated by more extended modelling. Indeed Coe and Fitz-Gibbon echo a refrain that will be repeated in the cases studied in this paper. There is often a need for collection of more detailed and relevant information that might inform this quest for explanations in causal terms. It is, perhaps, a little unfair to focus on modelling and statistical analysis for exaggerated substantive conclusions. In any case it is the use of such variables as controls rather than as potential direct causal factors that merits their inclusion in models. Whatever the nature of the explanations for gender differences they may nevertheless exist. It is the impact of these gender differences on other modelling objectives that are of interest. To use experimental terminology the 'treatments' may be schools and differences between them are of interest. In the absence of the ability to randomly allocate pupils to schools, controls over important pupil differences in the schools must be exercised. Gender may be a proxy for many other interesting but unmeasured direct causal influences on performance, but this need not matter. If differences in performance due to schools are the object of inquiry then other potential effects on these differences must be 'ironed out'. A direct causal interpretation of gender differences need and indeed usually should not enter into the picture at all. However, its inclusion in a model may advance evidence of other sources of causal explanation of school differences

As a prelude to his introduction to multilevel models in the context of school effectiveness Goldstein (1997) draws attention to practical difficulties in many areas of educational research of experimentally manipulating and randomising over conditions of interest. Thus the strict conditions that makes causal attribution firmer are often unavailable. Much the same could be said of work in most other social sciences. Thus in the real world we must see how closely we can approach this by using data and statistical tools that happen to be available or can be developed .In planning an investigation from scratch and with more attention to the experimental ideal we may for instance design a survey that reflects the structure of a process with theoretical propositions about relationships also in mind. To take a simple example if we were interested in both individual and primary school differences in educational progress we would not simply take a national simple random sample of primary school children. There might be too few students per school to make adequate inferences about school differences. The situation requires a clustered design of some sort. We would also ensure that we collected sufficient data relevant for the context of the inquiry at both school and individual levels. We would also develop models to reflect the two level process and clustered design. In some cases we are not free to make judgements about design and data collection and must perforce rely on data collected for a wide variety of other purposes. Later an example is discussed of where profitable use can be made of routine administrative data providing the limitations of such data are recognised. At the heart of any attempts at explanation, experimentally based or otherwise, is the development of adequate statistical models, reflecting not only the available data but also how it has been collected.

Even where elements of experimental manipulation are possible there is pressure of time and resources. In observing relationships between design factors and outcomes we might try to reach a 'causal' understanding of what '*appears'* to matter (Goldstein (1997). There still remain possible explanatory factors not explicitly informing the design, though expert grounding in theory and subject matter should have considered many of them. There is also no guarantee that relationships will remain stable in the future since relationships may change with changing educational and social conditions. Putting the experimental investigation into context is often difficult. In a current climate of a concern with ‘evidence based’ policies there is a strong pre-disposition to discount all but experimental evidence. In evaluation of specific school improvement programme initiatives evaluation by such approaches may be possible. However, Coe and Fitz- Gibbon (1998) consider that much experimental evidence to date in these areas are disappointing for a variety of reasons. For instance, they claim that improvement initiatives are often carried out within changing educational conditions. This precludes the establishment of effective control groups which makes it difficult to conclude what might have happened without the initiative. Another point they stress is the short term nature of many outcomes in such experimental evaluations. These make prognostication about long term sustenance of these effects difficult to ascertain. The important issue of the effect of class sizes on pupil progress has also received much attention with the large scale randomised controlled trial (RCT) Tennessee STAR study to the forefront (Word et al (1990)). A thorough review of the difficulties of making causal inferences about class size effects even under supposedly RCT conditions is provided by Goldstein and Blatchford (1998). The issues they raise also have much broader relevance and is a good reference for further detailed study of the potentialities of RCTs.. Apart from the question of changing historical conditions previously mentioned they raise some other important issues. The first is one of selection of experimental schools. The STAR study, for instance required extensive resource and time commitments by participating schools. This may make such schools atypical in ways that also affect the relationships under investigation. Further even if randomisation were possible it is only possible within schools, and different schools have different sizes and methods of organisation , for instance. Conclusions applying to large schools may well not apply to smaller ones and so on. There is also the difficulty of designing for interactions, e.g. . how compositional effects of the class in terms of ability mediate results. Another matter is that there is no real possibility of independence of ‘treatments’ within a school since teachers and students interact and this may contaminate the relationships. Also present since trials cannot be blind is the possible contamination by expectations or ‘Hawthorne’ effects.

Thus even many attempts at experimental manipulation pose limits to potential causal explanations. The examples of limitations discussed above are due to factors which are neither controlled by design nor can be dealt with by randomisation In design. To avoid problems requires adjustment for the factors in various ways in order to reach causal understanding. This can only really be done by exploratory analytical investigation incorporating relevant adjustment controls in formal statistical models. In principle these are the same sort of analytical models employed in investigations of explanation from observational or survey studies.

Advanced extensions to such models for explanation in observational studies or to data from administrative sources are the main focus of this paper. Thus we might outline in a technical way some of the main features by which they achieve explanatory ends. The model is the mechanism by which we attempt to adjust for relevant factors. As Goldstein (1995) points out in the context of the study of school effectiveness we have no control over which children attend which schools, or are assigned to which teachers, classes or teaching methods.' The best we can do is to try to understand what factors *migh*t be responsible for assigning children to schools or teachers to ways of organising teaching' The aim is to measure accurately along all dimensions possible ways by which individual pupils differ, e.g. prior ability , and by which schools differ on factors beyond the school's control. If we could adjust for all of these we have achieved what is often regarded as the paradigm of school effectiveness research ' a comparison of like with like'

Thus the attempt is to try through data analysis to deal with limitations imposed by lack of ability to experiment. The only constructive way that many factors can be simultaneously adjusted in analysis is through a framework of a statistical model. In this sense the adjustment results in comparisons in a 'typical' constant environment. If all possible explanatory factors had been considered we might get somewhere towards achieving the ideal: the attribution of school differences to the causal influence of school effects *per se*. Through the use of formal models the effect of potential explanatory factors is handled through the model analysis rather than through ideal designs. However, a modelling framework can also at the same time evaluate whether and in what ways the differences on crucial variables between pupils and schools affect the outcomes. It has often been argued that in the absence of randomisation there may be many other factors which can explain differences and which have not been explicitly considered or measured. Three things can help here. First theory and prior research should guide detailed model investigations with such factors. Secondly a precondition for this is the availability of information on relevant variables which in turn means acquisition of a range of relevant good data. Thirdly the operation of unmeasured influences which may operate on outcome differences are assumed to not systematically affecting the main relationships of interest. The incorporation of random variables in models to reflect these is a proxy for random allocation in design. If a variable has an effect that is systematic then we should be considering it explicitly and we are back to the necessity for good theory and data. Problems may arise not from the use of modelling tools but in the lack of prior conditions in particular circumstances for developing a good model. This point is made by Coe and Fitz-Gibbon (1998) in their critique current modelling methodology to explain and assess real school differences in terms of 'value-added'. Sets of theoretically perfect controls are often unavailable. Thus they say. ' In practice, therefore, any measure of value added which we may calculate may be thought to as an attempt to measure 'pure' value added that is biased towards unadjusted raw performance.' This critique has a great deal of merit. However, the force of their arguments seem to be directed towards the modelling methodology itself rather than the real recognition of the need for informed investigation and better data. The interplay between statistician and subject area expert emphasised above needs much further development Thus Goldstein (1997) says ‘ The task facing school effectiveness research is to try to establish which factors are relevant in the sense that they differ between schools and also they may be causally associated with the outcomes being measured. In this respect most existing research is limited, contenting itself with one or two measures of academic achievement and a small number of school and other background variables, with little attempt to measure dynamically evolving factors during schooling’ . The lesson appears to be that there are now a developing range of good modelling methodologies. However, what is lacking and required is deeper thought about the processes and data to inform modelling strategies to advance explanation.

For those interested in school effectiveness or more generally student progress research the article by Coe and Fitz-gibbon (1998) offers much that is thought provoking. It offers many warnings and suggestions for ways forward in the complicated area of causal explanation in research in education and other area. For instance there is the wide practice of establishing group (school) differences in outcomes *a priori* and then *ex post facto* searching for related characteristics by which the schools differ. This is roundly condemned as an extreme form of the abuse of the general principle that ‘correlation is not causation’ However, as implied by some of our previous discussion many of their valid and relevant points seem to be very much misdirected against modelling frameworks and attendant statistical methodology. These may be at fault occasionally but pre-requisites for the choice of an appropriate models are sound educational thinking and availability of relevant information.

In developing some extensions of multilevel methodology in the present paper we indirectly address two further particular concerns mentioned in these cited references. Firstly we recognise that in modelling the outcomes are often measured inappropriately. Ordered grades which form many response variables are not intervals on some arbitrarily chosen interval measurement scale. Secondly by the use of cross-classified effects we go some way towards reflecting the complexities of processes under review.
In education it is becoming recognised that different classes within schools often have stronger effects than differences between schools. Further the effects of classes may be a combination of several interrelated effects, e.g., teacher subject, history of pupil group and so on. Basic multilevel models have provided a framework for explanatory separation of hierarchical effects, such as pupils, classes and schools. Such modelling has been cited by Goldstein (1997) as one of the minimum requirements for effective explanation. The aim will now be to disentangle effects operating within levels in a way that makes optimum use of available data. As Raudenbush (1993) has noted, these developments may be viewed as part of a long term effort to develop analytical tools that correspond to those familiar in classical experimental investigation but which have greater flexibility. Part of this flexibility is a scientific approach to isolating explanatory effects which add to the possibilities of making causal attributions. Complex structures both of process and data require fairly complex modelling and extensive relevant data.

2 Basic multilevel modelling in hierarchical social structures

In complex social structures, and not the least in education effects which we wish to evaluate, unravel and explain operate in complex ways. Not the least important of these is the hierarchical structure of these processes. It is important that investigation designs, data collected, and models used in analysis reflect this hierarchical structure. It has become recognised over the past decade that multilevel statistical models are the appropriate way of reflecting this complex structure. Such models enables the derivation of information about relationships among measurements operating at different levels simultaneously. They also provide a framework for exploring explanations of these relationships. In combination with the caveats and principles raised in the introductory section they also provide the background for evaluating causal mechanisms.

Multilevel modelling can get quite complex with a growing body of applications in diverse areas. The classic generic paper is Aitkin & Longford (1986). A thorough technical discussion of the theory, methodology and range of applicable models is Goldstein (1995). In education Bryk and Raudenbush (1992) provide a detailed account of the principles of the models and methodology but could be understood by researchers with some advanced training in statistical methods There are appearing a growing number of articles and texts dealing with the diversity of applicable subject matter (Reise & Duan (2000). Here we follow Goldstein (1995) in presenting fairly briefly the basic ideas with a minimum of statistical complexity as background for more advanced type of models in later sections.
The articles by Paterson (1990), Paterson & Goldstein (1991) and Plewis (1998) provide more thorough but basic introductions.

In one of case studies later we have results on Key Stage 1 Standard Assessment Tasks (SATS) of children in the primary schools in Birmingham Local Education Authority. For analytical purposes we might treat this as a clustered design of samples of schoolchildren within a sample of schools. Here we look at the Mathematics Test and assume it to be scored as points on an underlying interval level scale Later we will see there may be more appropriate ways of treating graded responses and we will examine the Reading test with this in mind. The most simple multilevel model attempts to relate the test outcome or response to children's attendance at different schools. It may be noted that as it stands such a model is limited as an aid to the evaluation of explanations, but it forms a base model against which further developments may be assessed. Such base models usually form the standard in most published applications.

If y_{ij} is the response of the i-th student in the j-th school the base two level model is

$$\begin{aligned} y_{ij} &= b_j + e_{ij} \\ &= b + u_j + e_{ij} \end{aligned}$$

So far this is a fairly simply understood characterisation which specifies that the response can be formed as the addition of a school contribution (level 2) to an individual student contribution (level 1). The school contribution b_j can be broken down into the overall mean b and the school 'residual' u_j. The terms e_{ij} are the deviations of the individuals' responses from their school contributions. The latter are sometimes referred to as level 1 residuals or individual model disturbances.

If we were interested only in the particular schools in the sample the above model is in effect only a re-parameterisation of the simplest of traditional one way fixed effects analyses of variance. However, to further explanation and to be able to generalise about schools at large we need to consider the selected schools as randomly drawn. This involves treating the u_j as unobserved sample drawings from some distribution. Typically we may assume that these school residuals have a normal distribution with mean zero and a variance of say , s_u^2. The student residual is also assumed normal with variance s_e^2.

With this framework in mind to further explore explanations in our model we might wish to examine the impact of other explanatory variables. It has generally been found that in educational achievement the most powerful predictor of educational achievement at some stage is a measure of achievement at earlier stages. We may want to include this measure as an explanatory variable both to assess its importance on individual achievement in its own right and also to adjust school differences in outcome for their initial differences, The latter is the controlling aspect we have discussed at some length, although there is a long way to go in making causal attributions. In our example we have a baseline measures of achievement at entry to reception classes. The model becomes

$$y_{ij} = b_0 + b_1 x_{ij} + u_j + e_{ij}$$

where x_{ij} is the baseline achievement measure and b_1 is average predicted increase in KS! Mathematics score for a unit increase in baseline achievement. The residuals u_j now represent the difference between the mean KS1 score for each school for any given baseline level and this mean for the population as a whole.

It is clearer to see from this model another reason for the multilevel model with the school term It might even be the case that we were only interested in assessing the extent to which prior ability influenced individual achievement at KS1 and wished to leave the question of school differences aside. The temptation might be to consider a simple model which ignores the clustered sample design and fit by standard ordinary least squares (OLS) a model of the form $y_i = b_0 + b_1 x_i + e_i$. Here the model is fitted to all children and their school membership is ignored. Earlier studies of educational achievement (Coleman et al (1996), Rutter et al (1979)) were of this type. Apart from the fact that little can be said about the influence of schools, another problem is poor estimation of the parameters β_o and β_1. An assumption which lies behind good estimation from OLS is the independence of the disturbances e_i. Yet we know that due to the hierarchical nature of the process, design and data the e_i in the same school will not be independent. They will be correlated due to common influence of the school effect as evidenced by the fact that we know e_i is really the sum of two components (u_j+e_{ij}). This intra-school correlation is often a phenomenon of interest I its own right It is well known that it leads to incorrect estimation of standard errors and biased inferences about the parameters, e.g. confidence intervals that are too narrow and optimistic. Standard texts in econometrics (e.g. Greene (1999)) provide a good reference to this problem. Thus even if the main interest was in regression slopes in these simple models we still need to take cognisance of the multilevel structure and develop estimation procedures accordingly. Of course if we were interested in school effects as well , as we usually are, a multilevel model becomes very much a practical necessity.

	Base Variance Components Model		**Model with reception baseline achievement controls**	
Fixed Parameters	*Estimate*	*Standard Error*	*Estimate*	*Standard Error*
Intercept	0.012	0.05	0.018	0.03
Baseline test assessments: 4-points: standardised				
Number			0.30	0.016
Algebra			0.18	0.015
Shape and Space			0.02	0.017
Data handling			0.10	0.017
Speaking and Listening			0.09	0.016
Reading			0.11	0.017
Writing			0.08	0.018
Random Parametrs				
School variance	0.19	0.027	0.17	0.030
Pupil variance	0.83	0.009	0.55	0.014
Intra-school correlation	0.20		0.22	

Table!: Multilevel models for KS!: Mathematcs: Variance component model and model with baseline assessment adjustment

Table 1 above presents results of the two basic variants of these simple models. Full analyses are given in Fielding (1999).The response variable has been scored 0-4 for the four levels and then linearly transformed to standardied to have mean zero and standard deviation unity on the whole sample. The only difference in Table 1 is that a full set of baseline measures are introduced in the control but no differences of principle are evident. In the base model a first question of interest is the size of the school variance on KS1 mathematics relative to that of students within them. This provokes further thought about how to develop the explanatory model further. It must also be noted that this represents variation in raw unadjusted achievement in a model without control variables. Thus interpretation of results in substantive terms must proceed with caution. Even small school variation In the base model should not be taken at face value as suggesting uninteresting school differences. It is possible, as Fielding (1999) shows that this variance can actually increase on introduction of control explanatory variables. This will be recognised as a possible suppression effect for those knowledgeable about causal analyses. In the simple base model and since we assume that u_j and e_{ij} vary independently the total variation is

$$Var(y_{ij}) = E(y_{ij} - b_0)^2 = s_u^2 + s_e^2 = 1.02$$

The proportion of variance in responses that is attributable to schools is $s_u^2/(s_u^2 + s_e^2)$=0.20. This is also the value of the correlation of outcomes between two pupils in the same school that we have called 'intra-school correlation'. When we introduce the prior ability variable the regression coefficients show the size of the effect of that variable and the standard errors provide a basis for classical hypothesis testing procedures about those effects. Importantly, in working towards explanations we can evaluate the changed size of the variances. Naturally the individual level 1 variance has been considerably reduced by 50% emphasising the importance of prior ability on explaining individual variation in achievement at KS1. The school variance has also changed but not by much in this model reflecting the efect of the aggregate intake ability characteristics on overall differences between schools.

We have said very little so far about methods of estimation of these and much more advanced multilevel models, other than recognising that standard OLS regression procedures are inappropriate. Estimation is achieved by a variety of sound statistical procedures available in specialised software. In practice the two most common are the HLM (Bryk & Raudenbush (1992)) and the MLwiN (Rasbash et al (1999). The latter deals with many advanced modelling procedures which are not incorporated elsewhere and in addition a flexible macro language that permits flexible adaptation. MlwiN and its macro facilities are used for the developments in this paper. The latest version also incorporates many Markov Chain Monte Carlo (MCMC) Bayesian approaches which are becoming popular and which up to have been widely available in the BUGS package (Gilks, Richardson & Spiegelhalter (1996)). The standard large statistical packages such as SPSS do not provide much but SAS, STATA and econometric packages such as LIMDEP deal wit a few very limited variants.

MlwiN also provides facilitates for estimating the residuals from a multilevel model which are of obvious direct interest. The level 2 residuals from the base model may be viewed as estimated school effects on raw performance. Residuals in models with control factors may be taken as estimates of 'adjusted school effects' and as such have been taken as a basis for 'value-added' assessments since they more nearly reflect differential progress between schools. Goldstein (1995, 1997) shows that the residuals estimated

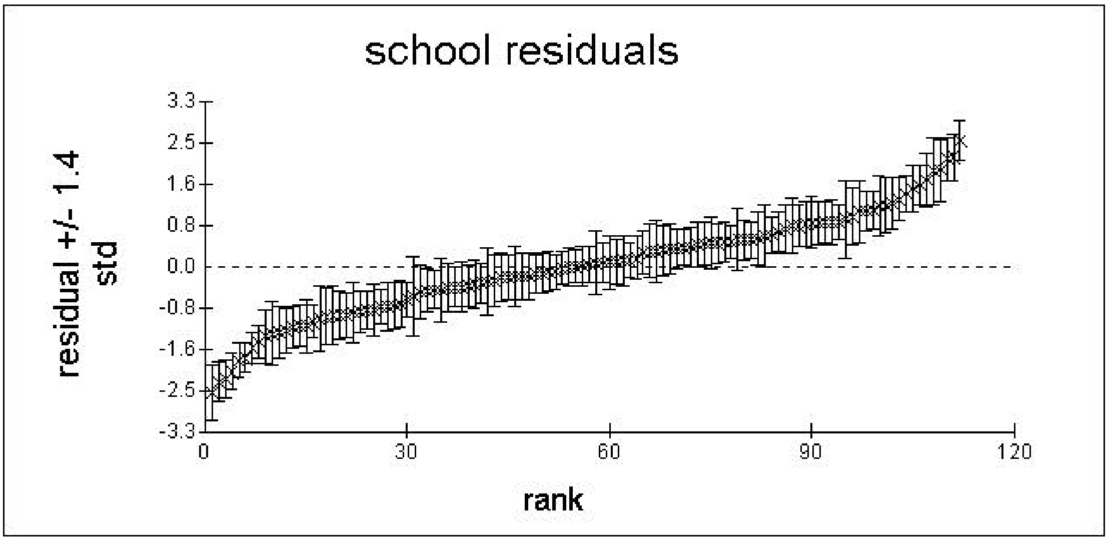


Figure 1: School residuals and uncertainty intervals   on KS1 Mathematics after adjustment for baseline assessments..

by multilevel procedures are ‘shrunken’ in the sense that information on their distribution from all schools is used to correct for the fact that otherwise estimates based on schools with a small number of observations are very imprecise. This issue of shrunken estimates will not be explored further Figure 1 below plots the school residual estimates from our basic model adjusting for baseline achievement at reception. Naturally being statistical estimates they are subject imprecision and uncertainty due to sampling errors. Their standard errors can also be estimated as part of the model fitting procedure. Thus the residuals are ordered by value and surrounded by 95% confidence bands in a ‘caterpillar diagram’, of the sort that has become familiar in literature. The explanation of the use of the multiplier 1.4 is given by Goldstein and Healy (1995). A well known feature of such caterpillar diagrams in progress research, however much control is exercised, is also exhibited in this figure. There is considerable overlap of the intervals with 50% of them covering the overall mean of zero. Thus attempts to rank or separate schools in league tables, even where there has been proper adjustment is subject to a high degree of uncertainty (Goldstein & Spiegelhalter (1996)).

Estimates of residuals at both levels also have extensive use as diagnostics in selecting appropriate models, making judgements about the adequacy of models, checking the appropriateness of assumptions, and generally to guide the details of further model exploration and development (Goldstein (1995)) Any number of potential control or explanatory variables can be added to the model above according to the context of the problem or what is understood about the process operating. These models then take their place in the armoury of equipment available in exploration. In school progress research In educational research there is considerable debate about which variables should be included according to the purpose at hand or what forms of explanation are required. The role of such individual variables as gender, ethnic origin , or socio-economic group has been discussed previously. School level factors such as type, size and organisational variables such leadership etc. may be introduced, if it was desired to see what effect they might have in explaining school differences and if adequate information was available. It is also not uncommon to include pupil level aggregates such as school level means on intake ability factors. These provide measures of school contextual features.

There are many other ways in which multilevel models may be extended. Of particular importance is to allow the relationships between the response and explanatory variables to vary at different levels. These are termed random coefficient models and in the context of school research model ‘differential effectiveness’. Formally instead of the regression coefficients being fixed parameters they may be allowed to vary at the school level. The simple adjustment model above is extended to

$$y_{ij} = b_0 + b_{1j} x_{ij} + u_{0j} + e_{ij}$$

where $b_{1j} = b_1 + u_{1j}$

The subscript j on b_{1j} indicates that each school may have a different slope and its difference from the overall average slope is represented by the random variable u_{1j}. We now have two random variables at the school level each with mean zero and a separate variances, $s^2_{u_o}$ and . $s^2_{u_1}$. In general they can also be correlated and have covariance $s_{u_o u_1}$. Thus it might be that schools with high intercept residuals have steeper slopes or vice-versa.
The between school variance now varies according to the particular value of the variable x and is the quadratic function $s^2_{u_o} + 2s_{u_o u_1} x + s^2_{u_1} x^2$. To illustrate the differential effect , the figure in Appendix A is taken from Goldstein (1997) from his example of modelling eleven year old reading scores adjusting for eight year old scores. The slopes of three schools are presented. What this example illustrates is that there is very little difference between the three schools for high ability students but for low ability ones they have considerably different effects. It is even possible for the lines to intersect for some schools so that a school which has apparently more impact than another school for higher achievers may have the reverse effect for lower ones. A fuller exploration of the importance of these differential regression coefficients ids given by Paterson (1990) and Plewis (1998).

The models are capable of being extended in a wide variety of ways. The previously cited range of references convey some idea of the complex processes that can be viewed in these ways. A key one is Goldstein (1995). For instance their can be many more than two levels. Our advances later will deal with the issue of variation between classes within schools (three level) which is becoming recognised as of prime importance. In common with familiar linear models interactions between variables may be important and these can cut across variables at different levels. There is also the possibility of modelling more complex variance at level 1. It has been observed for instance that there are not only gender differences in mean achievement

levels but there is also heterogeneity in response variation for different groups such as these. Models are also available for non-continuous variable data such as binary responses or count data. Multivariate multilevel models have also been developed to deal simultaneously with correlated responses, e.g. scores on reading and mathematics at a given educational stage.
In the next two sections we deal with two specific extensions which deal with some of the problems of explanation outlined in the first section. Firstly we consider responses which are ordered categories which up to now have often been treated by assuming continuous responses measured as arbitrary points scores. Not only are chosen outcomes in research often inappropriate (Coe and Fitz-Gibbon (1998)) but also they are measured inappropriately in operation. Multilevel modelling here requires extensions of the single level methodology of generalised linear models developed by McCullagh & Nelder (1989). Secondly we combine these generalised approaches with cross-classified random effects models to attempt the disentangling of effects operating within levels. We see that effects may also be weighted to mirror very unbalanced split plot designs in experimental research.

3 Modelling ordered category responses in a multilevel framework

Many response variables in research are ordered categorisations possessing minimal measurement properties, e.g. educational grades, Key Stage 1 achievement levels, 5-point attitude scaling. Modelling of these often proceeds by assigning them arbitrary scores and treating them through the linear multilevel models described in the previous section as if they were interval scale continuous measurements. There are a number of difficulties with this which have been widely discussed in the methodological literature but it is the norm in much empirical research. The familiarity, and ease of both access and understanding of linear regression models also encourages this routine application.

Most critiques of modelling using scores surround the questionability of many of the assumptions which is necessary to make. There a range of such objections in the literature and we briefly review them. A fuller review is provided in Fielding (1999, 2000).

Firstly there is a measurement issue arising from the interpretation of arbitrary scored scales as if they were natural, even when used as a modelling device. This is really a substantive question since regressions of the score scale may be approximately sufficient for some contexts. A typical objection on measurement grounds is exemplified by a standard econometrics text: Greene (2000) states baldly that , 'if the responses are coded 0,1,2,3, 0r 4 a linear regression would treat the difference between a 4 and a 3 in the same way as a difference between a 3 and a 2 , while in fact they are only a ranking'. Substantive meaning has been attached to rank orders as if they were units of measurement. Results based on grade scores are assumed to relate to grades as a higher level unit of measurement.

Secondly when linear models are used, be they multilevel or otherwise, questionable assumptions must be made. Do we really believe for instance that effects included in a model operate in an additive fashion on the chosen scale rather than some other arbitrarily chosen one, or perhaps a real one from which categories have been empirically formed? Points scores scales are often chosen as the response scale as a matter of convenience rather than anything else. Thus a main questionable assumption may be that the standard specification of traditional models may not be appropriate for the arbitrarily chosen scale . This very arbitrariness of a scored response scale in this respect is at the root of many objections that have been raised in the literature, which is reviewed in Fielding (1999).

A third group of objections surround the discrete nature of the observable responses which would arise even if the arbitrary scaling issue were resolved.. Distributional aspects surround the use of linear models which assume a continuous (usually normal) response. There are many technical difficulties with this. Some other problems are common to those arising even when the observed groups are formed with known cut-points from a natural unit of measurement. It is well known that in this situation standard regression procedures using category scores (possibly mid –points) can produce very biased regression parameter estimates and incorrect standard errors (Haitovsky, 1973; Stewart, 1983). Most models assume conditional disturbances and hence responses are continuously distributed (possibly normal) but this is cannot be reflected in the operational responses. This fact is largely responsible for the estimation problems with regression coefficients. Some simulations by Fielding (2000) have shown that this problem has even greater impact on variance components and residuals in multilevel models. There are other related difficulties connected to the discreteness of the response data. McCullagh and Nelder (1989) have raised the issue of ambiguity of estimation results or inferences that can arise when new responses are formed by possibly arbitrary amalgamation of old adjacent ones (or their desegregation). If scored responses are used in continuous variable regression models rules must be devised for scoring schemes to remove these ambiguities. How do we compare effects on GCSE grades for instance over time when the A* grade has been introduced. Hedeker and Gibbons (1994) also discuss the distortions that may arise due to ‘ceiling or floor effects’ if grouped scored responses are used in multilevel linear models. If, as may be supposed, the end categories reflect open intervals on some latent scale then extreme recorded category scores can rarely reflect responses to extreme effects of explanatory variables. An earlier contribution by McKelvey and Zaviona (1975) discusses the biases that can arise from such effects in traditional linear regression models. In many investigations the finding of significant polynomial terms or complex interactions in explanatory variables have been attributed to these distortions due to grouped arbitrary scores (Fielding (1999), Yang, Fielding & Goldstein (2000))

For standard single level regression , generalised linear modelling frameworks have been developed which deal with most of these questions (McCullagh & Nelder (1989)). They model grade distributions directly , impose no arbitrary scaling assumptions, and deal with the discrete nature of observed data directly. Here we discuss recent developments in multilevel ordinal response models. We may also see that from a substantive view they may offer also more useful ways of presenting results and offer more insight into the nature of explanatory effects.

Example 1: Progress in Primary School Education

To assist teachers in meeting learning needs of individual children baseline testing programmes have been set up in Local Education Authorities (LEAs) in England and Wales. Very young children are assessed on their joining the reception classes of primary schools. It has become possible to trace their progress through to the statutory Key Stage 1 (KS1) curriculum standard assessment tests (SATs) taken two years later. The data used is two level with 4444 children in 114 schools from one large LEA. The response outcome modelled is a KS1 Reading national test. A combination of four standard levels with a fine grading of level two yields six ordered grades.

In developing a multilevel model for this response there may be a variety of analytical aims. Some of these can be outlined although generally there is a wider substantive interest in factors affecting educational outcomes. Relating outcomes to prior ability measures mainly by linear models has been widely developed in the study of progress, or ‘value added’. It has also been recognised that other individual and school socio-economic control variables, though often correlated with ability, may make net contributions to progress. Understanding the nature of the combined contribution of these explanatory factors is a particular aim. Such models have informed target setting based on known individual characteristics at baseline. To facilitate such exercises the assessment of potential explanatory factors is an aim. Sometimes in the absence of prior ability measures targets have been designed controlling for only certain other characteristics. Such targets are also set at the school level. Thus implicit in national literacy targets for schools (Department for Education and Employment [DfEE], 1997) is the sole use of percentage of children eligible for free school meals, an easily available measure. Where fuller information is available and appropriately modelled, predictions may assess to what extent such procedures are possible. Indeed one of the main aims of the models to be discussed is to estimate probabilities of achieving certain grades given individual profiles on sets of explanatory characteristics. Presented pictorially these are often referred to as ‘chances’ graphs’. In the United Kingdom there is also a growing government concern with performance indicators, and in particular for schools. The estimation of school effects on progress through well-specified multilevel models , controlling for relevant variables, establishes a sound methodological basis for these. Even here, though, caution must be exercised before wild comparative conclusions are drawn (Goldstein and Spiegelhalter**,** 1996). In exemplifying the model approaches detailed substantive results in the direction of these sort of aims will not be fully explored. However, they may be seen as background motivating the methodology. A detailed substantive account of modelling the KS1 Mathematics test on the same children is given by Fielding (1999). There were sufficient and considerable differences in progress patterns to warrant the separate analysis of reading.

The overall distributions of the responses are given in the detail of rows 1-4 of Table 2. The child baseline assessment variables were in the seven areas of spelling, reading, writing, number, algebra, shape & space, and handling data. They are teacher assessments on a four-point scale. Thus they are also ordered categories. After investigation, however, it has been found satisfactory, to enter them into models as equal interval scales. They have been standardised to have mean zero and unit variance over the data.

Table 2: Distribution over KS1 Reading Test Grades and comparison with approximate marginal expectations given by various models.

	Level	0	1	2c	2b	2a	3
1	Ordered Category Number	1	2	3	4	5	6
2	Number	160	871	786	738	532	1357
3	Sample &	3.6	19.6	17.7	16.6	12.0	30.5
4	Cumulative %	3.6	23.2	40.9	57.5	69.5	100.0
5	Model AL Cumulative %	2.8	20.7	39.5	57.9	71.0	100.0
6	Model AL %	2.8	17.9	18.8	18.4	13.1	29.0
7	Model AP Cumulative %	2.2	20.6	39.5	57.6	70.6	100.0
8	Model AP %	2.2	18.4	18.9	18.1	13.0	29.4
9	Model BL at means Cumulative %	1.3	15.7	38.4	63.2	78.9	100
10	Model BL at means (%)	1.3	14.4	22.7	24.8	15.7	21.1

Definitions of the range of other explanatory variables and some summary measures given in Table 3. They are mostly level 1 variable. Separate school level data for contextual purposes, such as catchment area data, is unavailable. It will be seen , however, that aggregating child variables to the school level has formed some context factors.

A basic two level ordinal model formulation

The response is treated as a set of category indicators. If the categories are labelled s=1,2.3,4,5.6, then $p_{ij}^{(s)}$ denotes the set of probabilities that the ith child in the jth school achieves grade s on the KS1 Reading test. Since categories are ordered it is convenient to use the equivalent set of cumulative probabilities $g_{ij}^{(s)} = \sum_{h=1}^{s} p_{ij}^{(h)}$, the probability of achieving at least grade s Only (s-1)=5 of these need be explicitly considered, since by definition $g_{ij}^{(6)} = 1$. Linear multilevel models consider effects both fixed and random(school in varying individual conditional expectations on a scored response. Random variability of individuals is given by a level 1 variance, The formulation here

Table 3: Definitions of variables and summary statistics used in the models of Table 4

Variable	**Description and summary statistics**
Level 1: Pupil	
Gender	1, Male (50.3%); 2, Female (49.7%)
Free school meals	1, Eligible for free school meals (38.3%); 2, Not eligible (61.7%)
Nursery	1, Previous nursery education (66.5%); 2, Others (33.5%)
Centred Age	Age in months centred on 86 months at KS1 testing (st. dev.=3.5)
Ethnic-Language dummies	Fourteen compound categories were formed from ten ethnic groups and twelve first languages. The base for dummies is White with all languages (59.0%). Only 0.3% of children were White with languages other than English
EthLang2	1, Afro-Caribbean-English (7.1%); 0, Others
EthLang3	1, Afro-Caribbean-not English (0.4%); 0, Others
EthLang4	1, Other ethnic groups-English (5.1%); 0, Others
EthLang5	1, Pakistani–not English (16.0%); 0, Others
EthLang6	1. Indian-Hindi (0.1%), 0, Others
EthLang7	1, Indian-Punjabi (5.8%), 0, Others
EthLang8	1, Indian-other languages not English (1.0%); 0, Others
EthLang9	1, Bangladeshi-not English (4.0%); 0, Others
EthLang10	1, Arabic-not English (0.5%); 0, Others
EthLang11	1, Chinese-not English (0.2%); 0, Others
EthLang12	1, Vietnamese-not English (0.2%); 0, Others
EthLang13	1, Mixed race-not English (0.5%); 0, Others
EthLang14	1, Other ethnic groups -not English (0.2%); 0, Others
Level 1: School	
baseline aggregate	Average of seven percentages of children in school at level 2 or above on each baseline assessment
Free school meals context	Percentage of children eligible for free school meals

consider effects on this complete set of cumulative probabilities Individual variation is governed by these probabilities in a multinomial fashion. There is a possibility of extra multinomial variation but this will not be considered in detail here. (see \Fielding (2000)). In general, a set of cumulative probabilities for the ordered grades may be conceived as a scale for the ordered grade responses in that they are monotonically related to the set of grades: the increasing cumulative probabilities corresponding to increasing difficulty of achieving at least a certain grade level. It is the changing nature of this probability scale across individuals in response to fixed and random explanatory effects that we now wish to model.

The set of cumulative probabilities are constrained to lie between zero and one. It is usually desired to allow effects to operate in a linear and additive fashion akin to standard multilevel linear models. A functional monotonic transformation of the probabilities $L(g_{ij}^{(s)}) = a_{ij}^{(s)}$ to a scale occupying the whole of the real line on which effects can operate achieves this end. Technically, in general this transformation or link function can be specified by any inverse distribution function characterising a continuous random variable on the real line. In particular the logit function which is the inverse of a logistic distribution is often used :

$$L(g_{ij}^{(s)}) = \log it(g_{ij}^{(s)}) = \log(g_{ij}^{(s)} /(1-g_{ij}^{(s)})$$

We also use the probit link $\Phi^{-1}(g_{ij}^{(s)})$ which is the inverse function of a normal distribution. Thus for any individual cumulative probability scale we now have a new linear scale along the interval $(-\infty,+\infty)$. Conceptually a set of thresholds or cut-points on this scale for each individual are determined by the individual's probabilities over the grades (Bock & Lieberman, 1970). Thus $g_{ij}^{(s)}$ (s=1,2….6) under transformation correspond to intervals on the linear scale $(-\infty, a_{ij}^{(1)}], (a_{ij}^{(1)}, a_{ij}^{(2)}], (a_{ij}^{(2)}, a_{ij}^{(3)}], (a_{ij}^{(3)}, a_{ij}^{(4)}], (a_{ij}^{(4)}, a_{ij}^{(5)}], (a_{ij}^{(5)}, +\infty)$, with $a_{ij}^{(s)}$ constituting the thresholds. Changing the cumulative distributions over the ordered grades through the operation of predictor variables on the transformed scale changes these thresholds. The nature of these changes in response to effects is what we wish to model and estimate

Thus we subject these transformed probabilities to a full range of linear effect multilevel models through a linear predictor (LP) be some form of standard continuous variable response model including higher level random effects, but excluding any level 1 variance specifications. For multinomial variation there are no separate estimable variance parameter akin to the individual disturbance in continuous variable multilevel models. Level 1 variation is multinomial through the expected probabilities that we are modelling. This fact of inseparability of parametric specifications of expectations and variances is often a source of some confusion

Goldstein (1995) discusses the formulation of these models in a multilevel context. We start with the most basic of logit link models. This two level ordinal level model for ordered grade probabilities is conceptually comparable to the base variance component model we discussed in section 2.

We have for s=1.2,3,4,5:

$$\operatorname{logit}\left(g_{ij}^{(s)}\right) = \log\left(\frac{g_{ij}^{(s)}}{1-g_{ij}^{(s)}}\right) = a_{ij}^{(s)} = a^{(s)} + u_{0j}$$

The parameters $a_{ij}^{(s)}$ correspond to the cut-points for the average cumulative distributions around which those of schools vary. A fit to this model estimates only the series of marginal cut-points with a single random location effect for the distribution for the j^{th} school; that is the cumulated probabilities do not depend on any individual level characteristics denoted by the subscript i or institutional ones other than a single random effect u_{0j}. The latter is assumed

normal with zero mean and variance $s_{u_0}^{2}$. An individual's response follows a multinomial distribution determined by the cumulative proportions $g_{ij}^{(s)}$,
We extend to include explanatory variables and as in section 2 and a logit development is

$$\text{logit}\left(g_{ij}^{(s)}\right) = \log\left(\frac{g_{ij}^{(s)}}{1-g_{ij}^{(s)}}\right) = a_{ij}^{(s)} = a^{(s)} + (Xb)_{ij} + u_{0j}\,.$$

Differential school regression coefficients can be modelled so that some of the coefficients can be expressed as $b_{kj} = b_k + u_{kj}$. There will thus be many random residuals at the school level and these are assumed to have a multivariate normal distribution. This model possesses a useful property referred to as proportional odds. This can be seen by anti-logging to give the cumulative odds of achieving at least a certain level:

$$\left(\frac{g_{ij}^{(s)}}{1-g_{ij}^{(s)}}\right) = \exp(a_{ij}^{(s)}) = \exp\{a^{(s)}\}\exp\{(Xb)_{ij} + u_{0j}\}$$. Since the second

term in the expression is invariant to s changes to it in response to the effects change the set of odds across s in a proportional fashion. No such simple interpretation is available for other link functions such as the probit. Thus In the logit, since effects operate linearly on log–odds, some useful interpretations of parameter estimates are possible. For instance, in Table 4 results for a logit model BL the estimated baseline number coefficient is 0.80. This is the net effect on log-odds of a standard deviation unit change in number. Such a change will shift the entire response distribution in such a way that the set of five log-odds are all shifted by a constant 0.8. Alternatively the net effect is to multiply the set of odds proportionally by a factor exp(0.8)=2.23. Thus the odds ratios for pairs of categories remain unaffected. School random effects are also additive on the set of log-odds and operate proportionally in similar ways,

Another interesting and useful interpretation of these ordinal models is through the concept of a continuously varying though unobserved latent scale variable (*lv*) underlying the grading. It may be supposed to follow a continuous response multilevel model (*lv*)$_{ij} = (Xb)_{ij} + e_{ij}$. Note thee is no intercept since the location of this conceptual scale can be arbitrary Unknown cut-points, q_s,

Table 4. Parameter estimates for two-level ordered category models for KS1 Reading Test Estimated standard errors are in parentheses: An extra-multinomial parameter f has been fitted.

Fixed	Model AL	Model AP	Model BL	Model BP	Model CL
$\theta^{(1)}$	-3.56 (0.11)	-2.00 (0.06)	-4.35 (0.14)	-2.46 (0.08)	-4.54 (0.12)
$\theta^{(2)}$	-1.34 (0.08)	-0.82 (0.05)	-1.68 (0.12)	-1.00 (0.06)	-2.22 (0.09)
$\theta^{(3)}$	-0.43 (0.08)	-0.27 (0.05)	-0.47 (0.11)	-0.28 (0.06)	-1.24 (0.09)
$\theta^{(4)}$	0.32 (0.08)	0.19 (0.05)	0.54 (0.11)	0.30 (0.08)	-0.44 (0.09)
$\theta^{(5)}$	0.90 (0.08)	0.54 (0.05)	1.32 (0.11)	0.76 (0.06)	0.61 (0.09)
Baseline assessment:					
Spelling			0.21 (0.04)	0.12 (0.03)	
Reading			0.35 (0.04)	0.20 (0.03)	
Writing			0.21 (0.04)	0.13 (0.03)	
Number			0.80 (0.04)	0.46 (0.03)	
Algebra			0.32 (0.04)	0.18 (0.03)	
Shape and Space			0.02 (0.04)	0.01 (0.03)	
Handling data			0.28 (0.04)	0.16 (0.03)	
Gender					-0.61 (0.05)
Free School Meals					-0.71 (0.06)
Nursery					0.23 (0.06)
Centred Age					0.10 (0.07)
EthLang2					-0.01 (0.12)
EthLang3					0.14 (0.44)
EthLang4					0.27 (0.13)
EthLang5					-1.10 (0.11)
EthLang6					0.51 (0.85)
EthLang7					-0.44 (0.14)
EthLang8					-0.06 (0.27)
EthLang9					-1.28 (0.17)
EthLang10					-1.09 (0.41)
EthLang11					0.55 (0.62)
EthLang12					2.02 (1.05)
EthLang13					-0.34 (0.39)
EthLang14					0.43 (0.65)
Random effects					
School variance: $\hat{s}_u^2$	0.581 (0.086)	0.201 (0.031)	1.233 (0.177)	0.387 (0.058)	0.301 (0.052)
School % of residual variance in lv model	15.0	16.3	27.2	27.9	8.4
Approximate reduction in level 1 logistic latent variance from model AL	--	--	72%	--	71%
Approximate rescaled school variance using AL model logistic CMS scale	0.581	---	0.880	--	0.216
Extra-multinomial					
$\hat{f}$	0.968 (0.010)	0.988 (0.009)	0.956 (0.009)	1.091 (0.008)	0.960 (0.009)
-2 log-likelihood	7680.70	7168.52	-3045.7 1		

Fixed	Model DL	Model EL
$\theta^{(1)}$	-4.90 (0.16)	-4.96 (0.14)
$\theta^{(2)}$	-2.20 (0.13)	-2.27 (0.11)
$\theta^{(3)}$	-0.97 (0.13)	-1.03 (0.11)
$\theta^{(4)}$	0.06 (0.13)	-0.01 (0.10)
$\theta^{(5)}$	0.85 (0.13)	0.79 (0.10)
Baseline assessment:		
Spelling	0.21 (0.04)	0.20 (0.04)
Reading	0.35 (0.04)	0.36 (0.04)
Writing	0.15 (0.04)	0.17 (0.04)
Number	0.80 (0.04)	0.79 (0.04)
Algebra	0.30 (0.04)	0.30 (0.04)
Shape and Space	0.02 (0.04)	0.04 (0.04)
Handling data	0.28 (0.04)	0.29 (0.04)
Gender	-0.46 (0.06)	-0.45 (0.06)
Free School Meals	-0.41 (0.06)	-0.39 (0.06)
Nursery	-0.08 (0.07)	-0.08 (0.07)
Centred Age	0.01 (0.01)	0.01 (0.01)
EthLang2	-0.09 (0.12)	-0.10 (0.12)
EthLang3	0.15 (0.49)	0.21 (0.48)
EthLang4	0.22 (0.13)	0.21 (0.13)
EthLang5	-0.25 (0.13)	-0.28 (0.12)
EthLang6	1.12 (0.87)	1.15 (0.86)
EthLang7	0.07 (0.15)	0.04 (0.14)
EthLang8	-0.08 (0.29)	-0.12 (0.29)
EthLang9	-0.29 (0.19)	-0.29 (0.18)
EthLang10	-0.26 (0.42)	-0.32 (0.41)
EthLang11	1.09 (0.67)	1.11 (0.68)
EthLang12	2.78 (1.04)	2.69 (1.04)
EthLang13	0.24 (0.40)	0.22 (0.40)
EthLang14	0.04 (0.68)	0.02 (0.69)
School contexts:		
Baseline aggregate		0.83 (0.09)
Free school meals context		0.38 (0.06)
Random		
School variance: $\hat{s}_u^2$	1.112 (0.161)	0.522 (0.08)
School % of residual variance in lv model	25.2	13.7
Approximate reduction in level 1 logistic latent variance from model AL	69%	68%
Approximate rescaled school variance using AL model logistic CMS scale	0.764	0.355
Extra-multinomial		
$\hat{f}$	0.952 (0.012)	0.943 (0.009)
-2 log-likelihood	-4070.51	-4504.31

on this variable may further be supposed to form the set of ordered categories. Category 1 is observed if $-\infty < lv < q_1$. Category 2 corresponds to $q_1 < lv < q_2$ and so on to where category 6 is$q_5 < lv < \infty$. The logit model above ensues if it is assumed that the distribution of the level 1 disturbance e_{ij} is a standard logistic distribution. An ordinal probit model follows if e_{ij} ~ N(0,1). Full details of why this is so are given in the Appendix of Fielding & Yang (1999). Note the models fix the level 1 variance in standard units. By definition a latent variable is only conceptual and unobserved and has no measurement unit. The latter may be arbitrary up to constant multiplication but then with differently scaled effects. However, the ordinal model based on observed categories cannot distinguish between these arbitrary choices. Fixing the arbitrary measurement multiplier by fixing the variance of conditional level 1 disturbances in the latent model is a way of resolving the issue. With a free parameter the ordinal model would not be unidentifiable. Since the variance is being fixed it may as well be determined by standard distributions. It may be noted that by using a *lv* we are not arbitrarily deciding on a scale for which to build a linear model, which has been objected to. We are merely postulating the existence of some scale on which a linear model operates.

Model estimation

A variety of estimation procedures for specific situations have been suggested in the literature. The widely available and very flexible MlwiN software (Rasbash et. al.,1999).This is built around iterative generalised least squares procedures (Goldstein, 1995). The extensive macro facilities offer flexibility in adapting to a wide variety of complex models including the current ones. A suite of specially written macros, MULTICAT distributed with the package is under continuous development. The worksheet and model set up must be handled using the MlwiN command language in ways described in the macro manuals (Yang et. al., 1998)

.

Application to the example

Firstly, the estimated base Models AL (logit) and AP (probit) in Table 4 have no controls and establish a framework. The models are $\ell n(g_{ij}^{(s)}/(1-g_{ij}^{(s)}))=q_s+u_j$ and $\Omega^{-1}(g_{ij}^{(s)})=q_s+u_j$ respectively. School random variation operates on sets of probabilities through u_j according to the operative model. Conditional on being in a particular school, response variation for students is determined by converting these to probabilities. Modelling strategies are then usually directed towards attempted explanations of overall response variation from both sources that are evident in such base models. This will be directed towards analytical aims. It may be stressed again that individual and school variation now operates differently. In linear models sources of variation are additive , with varying u_j shifting conditional expectations and Level 1 variance separately specified.. This may be true of an underlying latent model. However, once this is cast in ordered category form a set of expectations in the form of probabilities fully express response variation. Expectations and variability are bound up with each other. Thinking 'linearly' means sometimes mistaken impressions are formed of results and wariness must be exercised..

For similar reasons, there should be care when the fixed parts of models are assumed to be marginal population expectations as if models were linear. In the logit base model AL , marginal log-odds are indeed $q^{(s)}$ since $E(u_j)=0$. However, non-linearity means $E_j(g_{ij}^{(s)})$ is not $e^{q^{(s)}}/(1+e^{q^{(s)}})$. In turn the expected value of $e^{\hat{q}^{(s)}}/(1+e^{\hat{q}^{(s)}})$ cannot be evaluated by simply substituting ,the expectations of the $\hat{b}_{\ell}$. However, for many practical purposes these may be used as reasonable approximations. With this proviso, the estimates $\hat{q}^{(s)}$ for models AL and AP can be inverted in this way and cumulative distributions converted to category probabilities. The distributions are given in rows 5-8 of Table 3 below the empirical one. Without schools random effect both models would be a simple re-parameterisation of probabilities. Both models give similar results and close to the empirical. There is a minor difference in the small lowest category. This reflects the model sensitivity to small categories in tails. If binary responses, with grade zero the focus, had been of interest , then it is likely there might be more concern about model choice. The effect of different *lv* scaling can be seen in the table 4 results. This would affect any model comparison. The results for the cut points, $\hat{q}^{(s)}$, may look entirely different on the surface. Apart from small effects of distributional differences , a factor of $p/\sqrt{3}$ brings them broadly into line. The latter is the standard deviation of the standard logistic distribution implied by the logit model. From implicit level 1 standard variances, the % of residual variance due to schools can be calculated. As table 4 shows the school contribution to overall response heterogeneity is estimated much the same (15-16%) for both base models.

Explanatory models are now developed. Models BL and BP consider introducing the baseline reception assessments into logit and probit models respectively. On comparing, by similar calculations to those above, it would be seen that these give similar impressions. Thus for other specifications logit models only are illustrated in Table 4. Logit model CL examines a range of individual background factors without prior ability controls. Model DL combines the two sets and Model EL has some school context variables found relevant and interesting.. It must not be supposed that these models emerged without detailed investigation. In general statistical model evaluation and selection strategies can be quite varied. They often involve changes in likelihoods (λ) or deviances based on $-2\log\lambda$. For these models likelihood values estimated from procedures in MlwiN are based on approximate linearisation and can be sometimes a little unreliable for formal use. Heuristically they can, however, give an indication of the extent of improvement of the explanatory power of models. They have been quoted for illustrative purposes in Table 4. For formal evaluation, however, alternatives are available in the form of the MlwiN FTEST or RTEST. These use Wald type test statistics for single or joint contrasts of fixed or random parameters using the estimated variances and covariances of the estimated parameters. The statistics have approximately a χ^2 distribution with degrees of freedom appropriate to the relevant contrast null hypothesis. If desired, confidence intervals for parameters, either singly or jointly, can also be constructed. To illustrate their uses consider comparing model BL with AL. The -2logλ change of 512 indicates that the baseline assessments are instrumental in explaining response variation. Formally a Wald statistic can examine the baseline coefficients simultaneously. There are seven contrasts, each simply involving the single baseline parameters which are hypothesised zero. The joint FTEST yields a value of 1649.6 on 7df , which naturally is highly significant. The modelling can also handle random variation across schools in the explanatory variable coefficients ('differential effectiveness). For instance If a random baseline spelling coefficient was so added to BL the estimated variance was substantively small. A RTEST yielded 3.6 on 1 df. This is not significant even at a 5% level. Other potential random variation of various coefficients in all the illustrated models yielded similar results. For these reasons, random variation at the school level is simply represented by a single intercept variation in the illustrated results

Such procedures have proved useful in model development, but this author cautions against over excessive reliance on formal classical point hypothesis testing in framing published results. Sizes of effects are often of more interest than whether they are significantly different from zero. They may also be of interest even if they are small relative to their precision. Parameter values other than zero may be plausible and may be included in models and reported on substantive grounds (with confidence intervals if there is specific interest). Again, and for other reasons it is sometimes important to avoid the publication bias in reporting only statistically significant results. Certain negative findings can sometimes be important. For instance, for these and other reasons, all baseline tests have been included in the fitted models. Other insignificant coefficients or those with small effects have also been included. These highlight a number of useful interpretations. For example, the results can illuminate properly the relative importance of various baseline tests on the response net of all others. Again, as another example, the inclusion of the insignificant nursery variable in the final models DL and EL demonstrates that its obvious influence on performance is not net of prior ability. It explains performance but is a small influence on progress. .

A combination of formal procedures and attention to such substantive considerations has informed the process of model fitting. Although limitations of their use have been discussed, it may be noted that points-scoring models have a useful role to play in model exploration. They have been used in this way. They are much quicker to implement and initially will usually identify the source of major effects, as exemplified by the discussion part of the paper by Ezzett and Whitehead (1991). Sometimes, of course, according to research direction such identification may seem to be sufficient. For more detailed analytical objectives, effects specified and adequately parameterised in subsequent generalised. models might be preferred. From exploration to refinement it has been also been importantly noted that the refined category models often result in simpler patterns of effects. This may be tentatively suggested as another of their advantages. For instance some quadratic terms in baseline measures appeared important in scoring models, possibly as a reflection of ceiling and floor effects. After investigation they proved not necessary in refining the illustrated models. Similar conclusions arose in evaluating potential interactions that had emerged as plausible in the initial exploration. Again points scoring models suggested random variation across schools for some baseline and other coefficient. Simple school variation proved sufficient for ordered category models.

Some comments may now be made about the interpretation of the results. After illustrations of just two of the many possible analytical uses will be made; individual predictions of 'chances' and evaluation of school's 'value added'.

Some researchers prefer interpretations directly in terms of ordered category probabilities and others focus on their meaning in terms of the *lv*. The preference is subjective since there is consistency, but tastes vary across disciplines. With the first focus, the fixed effects can be directly interpreted as linear contributions on the link scale of transformed probabilities. Since the baseline variables are standardised, the most important of them is number. The coefficient of 0.80 in BL is an estimate of the net marginal change in all of the set of five log-odds per standard deviation unit change in the number variable. On the lv interpretation it also represents a marginal effect of O.80 on the standard latent response variable, scaled for that particular model. A change to a more substantively meaningful marks scale is always possible by linear transformation. The recognition of re-scaling on this interpretation may , however, be seen to be crucial. Marginal effects on probabilities cannot be so simply represented. Due to the non-linearity, the partial derivative of $g_{ij}^{(s)}$ by, say, baseline number depends on all the other effects in a model and on s. Sometimes for summary purposes average distributions and average marginal effects may calculated by setting other variables and random effects at their mean values. The latter are all zero for model BL. The summary may be seen as only representative. For model BL rows 9-10 of Table 3 gives estimated average percentage distributions at these zero mean values. Here these are formed simply from antilogits of model BL estimates $\hat{q}^{(s)}$. Partial differentiation of each $g_{ij}^{(s)}$ with respect to baseline number, for example, and evaluation at the means of variables and effects may be performed straightforwardly. The results are estimates of the marginal effects of a standard deviation increase in number at the means of all variables, including number itself. They form one sort of suitable average summary. If necessary marginal effects can be estimated at a number of other points By differencing they may be converted into the marginal effects of number on category probabilities. At the means these are -0.010, -0.095, -0.083, 0.003, 0.053, and 0.133 for the six categories sequentially. They sum to zero as they should since the sum of probabilities is constrained .They show the obvious effect that increasing number level has on shifting the underlying response variable upwards. School random effects u_j are also additive on link scale. For a logit model their variance is variability in log-odds across schools. However, it may be usefully interpreted in relation to the scaled *lv* variance at level 1. Percentages of residual variance due to schools are thus given in Table 4. The similar estimates of 27.2 % and 27.9%

are given for BL and BP, and may be usefully contrasted with the 15.0% and 16.3% for base models AL and AP. After baseline control it seems there is relatively more of the residual variation that can be attributed to schools. It could be asserted that progress differences in schools relative to individual progress heterogeneity are sharper than those for raw performance at KS1. . .

With these ideas fixed , it is of obvious interest to interpret changes in different models as more effects are introduced. We expect that factors identified as influences on the response may 'explain' variation in that response at one or both levels of the hierarchy. Teasing out the nature of this is slightly more complex than for linear multilevel models. The main reason for is, of course, that the level 1 variation is characterised in terms of differently specified probabilities without a separate variance parameter. Correspondingly, and for much the same reason even under a *lv* interpretation, model changes involve changes in scale of residual variances. For instance, the introduction of certain explanatory factors may reduce heterogeneity at level 1. One manifestation of this may be seen by noting the heavier concentration in the middle categories in row 10 of Table 3 (model BL) compared to row 6 (model AL). However, although a reduction in KS1 level 1 residual variation is evident the implicit lv is rescaled to be standard logistic. This also impacts on school $\hat{s}_u^2$. Changes in this across logit models in Table 5 may be directly interpreted as changes in school variability in log-odds. It may be sufficient to note them without further qualification. However, they are a combination of changes of scale and changes in residual school response heterogeneity. If , as often, the latter is of interest some additional work is required.

Changes in the $\hat{q}^{(s)}$ are instructive in this and other work in interpreting results changes across models. However, in this connection it is useful to view the $\hat{q}^{(s)}$ from their two perspectives On the first perspective, they account for the differences in link transformed $g_{ij}^{(s)}$ across categories, independently of the values of the x_{ij} and u_i in a model. The latter shift the entire conditional level 1 ordered category response distributions and hence $g_{ij}^{(s)}$ in model specified ways. Influential explanatory variables may operate at both levels after their introduction but do so in different ways. Direct influences on level 2 variability will operate through the LP to change the dispersion in the location of the conditional distributions, much as in linear modelling.. In the logit model, if these were the only influences there would be a reduction in variation between sets of odds across schools, without necessarily affecting the proportionality factors within those sets; the relative sizes of $\hat{q}^{(s)}$ may not change. In contrast level 1 influences will result in less within school variability in the ordered response and hence a conditional level 1 distribution will have smaller dispersion. This will not be explicit in comparative model results. However, the $\hat{q}^{(s)}$ estimates characterise the spread of the conditional distributions over different response categories. If there is changed dispersion $\hat{q}^{(s)}$ will change relative to each other. Thus in the logit, proportionality amongst odds will change. Additionally as noted above there will be effects on the school variance estimate $\hat{s}_u^2$ additional to changed school heterogeneity. This is a consequence of the additivity of school effects on the link scale that is changed consequent on the other changes. That there is a scale change is best seen if the $\hat{q}^{(s)}$ are viewed from the second perspective as rescaled cut-points on the underlying lv. In either situation the examples do not consider allowing random school variation in specific $\hat{q}^{(s)}$. Sometimes this is very useful but creates additional interpretive points in studying changes that will not be pursued..

A contrast of model AL and BL on introduction of baseline controls, might clarify these two perspectives and help to interpret changes in model estimates, including $\hat{s}_u^2$. First the stretching in the $\hat{q}^{(s)}$ parameters may be noted. However, it is only changes in $\hat{q}^{(s)}$ relative to each other that have non-trivial relevance. For any model they are log-odds at zero realisations of variables and random effects, and hence where the mean of the underlying *lv* is also zero. For models AL and BL they are also realisations at the mean, which is convenient since they are then directly comparable. A stretching of log-odds is indicated straightforwardly from the comparative $\hat{q}^{(s)}$ estimates. However, for non-zero means of included variables the reference *lv* will not have mean zero at these means of variables There is an intercept shift which will move all the $\hat{q}^{(s)}$, log odds at zero values, up or down by the same amount in addition to changes in their relative size to accommodate changing proportionality. A glance at the $\hat{q}^{(s)}$ for models CL and DL where explanatory variable means are not zero should make this apparent. Comparable values of log-odds at the mean of variables would require a fairly simple intercept adjustment $\sum_{\ell=1}^{L} \hat{b}_l \bar{x}_l$. However, irrespective of this, since it is the relative changes that are of interest one way to view them is to first take ratios of the corresponding proportionality factors $e^{q^{(s)}}$ To examine the relative changes they can first be converted to. $e^{q^{(s)}}$ to which odds are proportional: These are (0.028, 0.261, 0.650, 1.377, 2.450), (0.013, 0.184, 0.625, 1.716, 3.74) and for AL and CL respectively. Taking ratios for BL to AL gives as percentages (46, 70, 96, 124, 152). This is a useful summary of one impact of baseline controls. Since more generally the $e^{q^{(s)}}$ express only the proportionality of odds, it is these ratios relative to each other that is important. It is convenient then for such comparisons to scale the ratios to make say one 100%. The reductions or increases in the proportional odds can then be seen relative to this one. This also accommodates the problem of any shifting of the intercept. From the above the index set (48, 73, 100, 129,158) is one such summary through which the relative changes in $\hat{q}^{(s)}$ can be viewed. Either summary shows the reduced level 1 conditional response dispersion, as evidenced, for example, by evaluation of probabilities at the mean in Table 3. The odds are cumulative. Thus, wherever the conditional location is, a greater concentration of the response distribution will be evidenced by such patterns in the indices. It is possible, in terms of the odds perspective, to give more detail about the extent of changes in variation at both levels. However, possibly the *lv* perspective of changes in $\hat{q}^{(s)}$ are more illuminating. The baseline controls reduce variation in response at level 1 but in moving to model BL the conceptual lv is re-standardised to the extent of this reduction. This will be accompanied by an inflation of the cut points to accommodate the new scale. Such scale comparisons are also best made conditionally at where the lv has mean zero. For models CL.DL, and EL for instance, some variables are not centred at means and a simple intercept adjustment is required to the $\hat{q}^{(s)}$ before comparisons are made. The model estimation does not operate directly on the unobserved continuous lv scale and the fit is obviously to ordered categories only. Also the fits are only attempts at good approximations by theoretical models to complex reality in the data. For these reasons in it is not expected that estimates will inflate or deflate the cut-points by exactly the same factor. Contrasting model BL with model AL , where cut-points both reflect a lv mean of zero, this factor ranges only from 1.1 to 1.42. On taking reciprocals and squaring, scale factors of these orders in a theoretical model would imply underlying level 1 variance reductions to between 50% and 82% of its former value. Given the data, this is a rough idea only. However, whatever is the appropriate figure, rescaling affect estimates of other parameters that the two models have in common in addition to other consequences of the model change. Here the only common effect parameter estimate is $\hat{s}_u^2$. If the level 1 theoretical variance changed to x% of that before, then the scale of $\hat{s}_u^2$ changes accordingly. Thus before concluding that changes in $\hat{s}_u^2$ were a reflection of changes in school heterogeneity, restoration of scale by a x% multiplication is in order. To give empirical content to this a rough and ready method for getting some approximate handle on x has been devised. Full details of this method are discussed in Fielding and Yang (1999). The fitted grouped distribution for base AL in row 6 of Table 3 can be used to construct logistic CMS. For these common scale scores, the variances are calculated for the AL distributions and that of BL in row 10. A comparison shows a level 1 variance reduction in BL to 72% of that of AL. Scaling the Level 2 variance of BL gives 0.88. Variance in the underlying response at school level is viewed as increasing but by perhaps not as much as might seem from model results. This is not uncommon in primary schools where progress may be more variable across schools than the KS1 outcome itself, and becomes apparent after baseline control . It must be stressed that these figures will be rough and they must not be given an air of spurious precision. This type of scaling relative to AL is also applied to other models and displayed as additional approximate information in Table 5

There is much of interest of real substantive content in the models presented. The main aim has been to use the example to illustrate model interpretation. However, a few comments will be made about some of this content without excessive elaboration on the educational issues. Reasons have already been given for displaying some coefficient estimates that are not significant statistically. Baseline assessments and individual background variables are quite closely associated. Model CL controls for the latter separately. Many ethnic-language dummy comparisons with white English speakers are estimated imprecisely due to small numbers. However, there is some evidence that fairly large groups of Pakistani and Bangladeshi pupils whose first language is not English are disadvantaged. Other effects are as expected and there is a gender gap in favour of girls. A month difference in age increases the proportional odds by a factor of 1.1. This corresponds to 0.06 standard deviation units on the *lv* scale. It can be noted that the factors reduce between school variation, in contrast to baseline controls, which increase it. Further this reduction is relatively larger than within school. The latter reduction is similar to that exercised by baseline. Thus although the sets of factors are related for individuals they operate quite differently at the school level and are certainly not proxy in this respect. The reason is possibly that the catchment characteristics of school make them more homogeneous in intake than on ability factors. The importance of recognising the multilevel structure is again emphasised. In model DL where the effects are combined the coefficients of the baseline variables remain much the same as BL. This is a real similarity since the approximations show scaling in the two models is not too different. Thus their impact on KS1 reading is largely independent of additional controls with which they may be related. Of course with prior ability controls other effects are net of this and may be interpreted as influences on progress. An FTEST of the extra effects in DL over BL yields a highly significant chi-square of 148.4 on 17df. Girls progress more and it is interesting that this is an opposite effect to that in KS1 Mathematics (Fielding, 1999)**.** Domestic circumstances captured by the free school meals dummy is related to progress net of ability. However, there seem to be no net advantage on progress of having a nursery education or being older. Mainly due to imprecision, the majority of ethnic-language effects are not statistically discernible for this data. so not much can be concluded. However the disadvantage of certain ethnic groups in attainment levels does not seem to carry over for progress. The fairly large net positive effect of the Vietnamese dummy relates to a very small group of 15 pupils. The importance of context effects is demonstrated in Model EL results. Irrespective of individual influences pupils seem to make more progress when their peer group in the school is more advantaged. These factors explain a substantial amount of school variation even after a wide variety of pupil controls have been imposed. The approximate figures show that school variance is reduced to the order of a half of model DL. This order is confirmed by the reduction of the portion of residual heterogeneity attributable to school from 25% to 14%. A general conclusion of these results is that the separate contributions of prior ability and other characteristics to progress. They also operated differently on school variation and a multilevel level analysis brings this out. These facts have often been ignored in policy related research. As mentioned before the DfEE, for instance suggest using only aggregate free school meals to target schools for literacy interventions.

Two particular analytical uses of the results will now be outlined. The first is concerned with individual prediction. The Value Added National Project (Fitz-Gibbon, 1996) stresses that presentation of ‘chances’ of achieving certain levels for students with different profiles is a very meaningful way of communicating predictions. Such an exercise is difficult when traditional points scores models are used. Ordered category models provide a basis for predicting the entire probability distribution over categories and serves this purpose well. There may be debate about which explanatory variables to use. For illustration, Table 5 below presents 'chances' based on Model BL. They

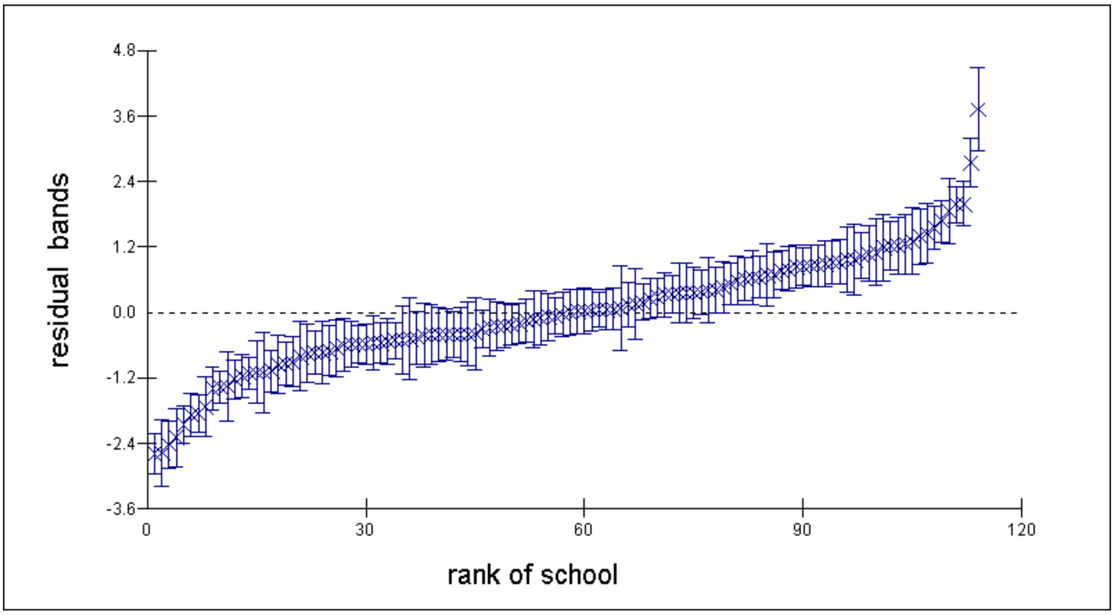
KS1 Reading Level	**Empirical Overall Percent**	**Six Baselines at Level 1 with Number at Level**			**Six Baselines at Level 2 with Number at Level**		
		0	**1**	**2**	**0**	**1**	**2**
0	3.6	8.5	3.2	1.1	1.1	0.4	0.1
1	19.6	38.9	25.6	12.3	12.6	5.0	1.8
2C	17.7	21.2	21.2	17.8	18.0	10.1	4.3
2B	16.6	15.4	17.6	17.6	17.6	15.2	8.6
2A	12.0	7.7	12.6	13.6	13.6	13.7	11.1
3	30.5	8.3	19.8	37.4	36.9	55.5	74.1

Table 5: 'Chances' distributions: Predicted percentage distributions over KS1 Reading Levels for some combinations of baseline assessment levels

might be used when only prior ability measures are available. Adjustments for expectations of non-linear functions suggested by Hedeker and Gibbons (1994) and Goldstein (1995, p79) are made before converting the fitted LP for covariate values to predictions of probabilities. The baseline values have four discrete values but experimentation has concluded that treating them linearly in the LP is fine. The top level is very rare with no more than 2.2% in the data for any test. Thus the examples set all the baseline variables except number at either level 1 or 2. The important influence of number can be seen by allowing it to vary in the table. Such distributions, perhaps with more detailed profiles, when converted to graphical displays by histograms provide a readily understood motivational device. However they are based on point estimates of model parameters. For any detailed inferences, the uncertainty of estimates should be recognised and confidence bands for the distributions could be provided. It should also be recognised that a school effect that could be included in the profile may alter these overall estimates. These fitted 'chances' are averaged over schools. In fact a pupil may be expected to do better or worse than a similar pupil does if their school effects were different.

The MlwiN procedures in MULTICAT also allow the estimation of the residual u_i and the standard errors of these estimates for a model. Generalised residuals at level 1 are a more complex matter due to the discreteness of the response and will not be pursued here (see Chesher and Irish, 1987; Fielding, 1999). A use to which estimates of school effects $\hat{u}_i$ have been put is in deriving 'value added' measures for schools. Since they result from models in which intake has been controlled there is a preference for the term 'adjusted school effects'. The details of the controls to be used and the mechanism by which they operate are an area of vigorous debate in the school effectiveness literature. A model for outcome responses is often likened to an economics production function (Woodhouse and Yang, 2000) The role of a school effect is seen as adding to the raw material characterised by other variables in the model, The model adjusts by controlling for this input quality. Official sources often eschew the use of control using socio-economic characteristics which given the above results may surely be contentious. However, for illustration Figure 2 shows the

Figure 2: School estimated residual effects for Model BL in rank order; Bands are residual +/- 1.4 estimated standard errrors



estimated school effects (residuals) for model BL using only initial ability as controls. Uncertainty must be recognised in these estimates (Goldstein and Spiegelhalter, 1996). Thus as suggested by Goldstein and Healy (1995) they are surrounded by the 1.4 standard errors giving overall 95% confidence intervals for sets of such comparisons. These caterpillar diagrams have become very familiar in school research but have as yet to make inroads into official publication. They can be routinely implemented in the graphics windows of MlwiN. The scale is that of a logistic variable and could be converted to standard deviation units of outcome on division by $p/\sqrt{3}$. Schools and others are used to standardised scores so fairly easy interpretation of such diagrams should be possible. Scales can also be converted to level units if required. The overlap between school bands, as also noted in many other published contexts, means that it is difficult to discriminate between the effectiveness of the majority of schools. For screening purposes there are some obvious extreme schools at either end. These arouse interest and could be subject to further scrutiny. A few of the schools at the top end, for instance, are particular advantaged in the context effects and socio-economic variables. These have not been controlled in the graphed results. The diagnosis and analysis of outliers is aided by the methodological developments of Langford and Lewis (1998) , which are now implemented in the graphical interface of MlwiN.

4 Cross-classified and weighted random effects multilevel models

In studies of cost-effectiveness of GCE Advanced Level provision we have become aware that costs and effectiveness variations within institutions are sometimes for more relevance than institutional differences (Belfield, Fielding & Thomas (1996)). It is with the data collected during these studies that concerns us in this section. As mentioned in the introduction there has also been a switching away in much educational effectiveness literature from interest in effective schools to effective classrooms. Hill and Rowe (1996, 1998) discuss the assertion made by Monk (1992) that ' …how much a student learns depends on the identity of the classroom to which a student is assigned ' and develop the motivation for this switch of focus. Young (1998) discusses similar views and illustrates with the work of Fraser & Tobin (1989) who stress the importance of the 'classroom learning environment'. We have already referred to the concern pf Coe and Fitz-gibbbon (1998) who make the comment, "thus school ' effects' are sought despite the fact that learning takes place primarily in classrooms and may therefore be expected to be influenced more by classroom factors". Effectiveness and scope for improvement may this have discrete classes as a proper area of concern but similar views may be echoed when we examine the resource side. In our studies of post compulsory education A level provision we have found wide disparities between resource provision between our equivalent of classrooms, subject teaching groups (Fielding et al (1998)). The interactions between effectiveness and resource constraints are thus a difficult area of direct concern. Here we will be concerned with one side of this problem, analysis of A level outcomes, suitably adjusted, at the teaching group level. The complexities of the structure of A level provision, and indeed the data we have, mean this may not be straightforwardly sought through traditional residuals from fitting well formulated *standard* hierarchical multilevel models. We will mention this structure, associated data, and the methodological problems it poses in achieving this end. We hope that some of the methods we propose and contrast will have wider relevance since similar structures are quite common. We first briefly discuss, generically, the problems connected to disentangling effects, and the levels and sources by which they may operate. These are instrumental in developing models for explanation. Although the concern is with educational achievement these issues are of general relevance.

In available literature as has been suggested educational effects operating below the level of the institution are emerging as more influential than variation between institutions. In studies of this phenomenon these conclusions are usually reached on the basis of apportionment of residual variation from (sometimes with fairly complex fitted factors) multilevel models. The evidence has not always pointed inexorably to the conclusions. However, where it does not rational explanations may be found. Luyten and de Jong (1998) conclude , for instance , that in secondary school studies that portray institutional effects as more important, prior achievement is not adequately controlled. The latter article also contains the most recent review of the many studies of this phenomenon. The general consensus that within institution effects are the more important remains fairly well substantiated.

This has addressed the problem of disentangling levels of effect and usually through multilevel modelling of nicely structured hierarchies.

However, the conclusions also beg the question of what is going on beneath the institutional level. This problem of disentangling what is happening is clearly put by Coe and Fitz-Gibbon (1998) who say, 'The combination of subject taught, teacher and pupil group is, of course, unique for each class, and effects could be attributed to all three parts of this tripartite confound'. One might even go further. Sometimes unmeasurable factors concerned with the classroom environment may operate and, of course, relevant for our broader purposes is the impact of resources. Differences in performance, suitably adjusted, in different subjects at Advanced level is well documented (e.g. Fitz-Gibbon (1996)) and is pertinent for data such as ours, where a characteristic of the teaching groups is the subject itself. We will also suggest ways in which we might address the problem of disentangling teacher effects. The way in which teachers and classes have sometimes been treated synonymously is not entirely unequivocal. This has received attention by Luyten and de Jong (1998) who use a quasi -experimental design involving 'parallel classes' as their solution. Also the fact that pupils or students being units within classes and treated hierarchically, does not always yield satisfactory analytical approaches. The pupil group is unique for each class but it is often argued that satisfactory control cannot always be exercised by taken cognisance of all relevant pupil characteristics. Yet as students cannot be treated as randomly allocated to teaching groups (classes) there may yet be important but unmeasured sources of student variation which are associated with selection into teaching groups. Any differences between groups may be partly reflecting these. A within group random student disturbance caters for unmeasured variation but may not adequately cater for these systematic effects which may be confounded with those of groups. The disentangling of many of these possible confounding factors is an important set of problems, which, in a particular context we try to go some way towards addressing here..

We firstly discuss some examples of data structures that motivate the modelling structures we propose. We will see that introducing complex cross-classified effects into multilevel models is a way forward. We will note how the complexities reflects that of the process it was gathered in. We will then elaborate in a more technical way how some of the general methodological issues raised above are specifically pertinent. In example we deal with some proposed solutions that arise out of the process of A level provision and the data we have. We apply some of these and in our results also contrast with results that might have ignored the problems. Our approach is a methodological one but we hope not to lose sight of the substantive import of our example in the study of effects in A level provision. We also hope and are optimistic that there will be broader relevance to many other areas, both in school effectiveness and outside it. An important lesson to be learned, perhaps, is that attention to data needs be given if some disentanglement of effects is to occur

Example structures and data

The ideas now to be discussed are motivated by the recognition that in certain hierarchical social structures there can be groupings of units at various levels. These groupings may themselves each form a level in a hierarchy but one that cuts across an existing level. Each grouping may contribute effects to random variation that must be disentangled. A first useful example is provided in the MlwiN manual (Rasbash et.al. (1999) dealing with attainment of children at age 16 in secondary schools. It is recognised that both these schools and the primary schools that the children previously attended can have effects. Children can be nested within both sets of schools but children within a particularly secondary school belong to many primary schools and vice-versa. This is a two-way cross-classified structure at level 2. More generally there may be multi-ways of classification and cross-classifications at many levels. Here, neighbourhood in which the child lives may be a factor making a 3-way cross at level 2. The complexity of ways and levels in which such structures can appear, and many examples, are discussed in Goldstein (1995). Good detailed applications for *continuous responses* and linear models are Goldstein and Sammons (1997) and Raudenbush (1993) and for binary responses Yang, Goldstein and Heath (2000). The data analysed by Rasbash et al (1999 are on 3,435 children attending 19 secondary schools coming from 148 primary schools in Fife, Scotland. The attainment response is graded into ten-point categories. This analysis used the points scores (1-10) and continuous response normal linear models. Comparative analyses of the same data using the ordinal models developed here are given in Fielding (2000). The second example we discuss here uses a subset of data drawn from that collected in a study of teaching group cost- effectiveness of A level provision in further education colleges in England (Belfield et. al., 1996). A similar dataset using linear points scores models has been discussed in Fielding (1998). The lowest level unit of analysis of the six-point graded outcome is an entry to a subject examination. The set used has 3683 entries nested within 314 teaching groups (classes), which in turn are nested within 6 colleges. On the surface this is a normal 3-level structure, although due to their small number colleges have been treated as a fixed effect blocking factor in later analyses. However, up to five entries are made by each of 1511 students and there is a crossing of student and group at level 2. Ignoring this crossing may result in teaching group differences being confounded with students and disentangling class such confounding effects is the concern raised above. Modelling of unbalanced cross-classified designs is a sound methodological development in this direction.

The natures of the crossing in the two examples provide contrasting types both in balance and sparseness. Both are quite unbalanced but in different ways. In the first the data is near hierarchical in the sense that each secondary school draws large numbers of students from a limited number of different sets of primary schools. A majority of the 19x148 cells in the crossing are empty, but some have relatively large numbers. In the second example the majority of cells are again empty but also there can be at most one entry per cells scattered through the crossing. In the latter sense these data are sparser. The methods to be presented are generally applicable. In general, however, such varying features of crossings will inevitably affect analytical results, particularly in the accuracy and precision of any estimates. It has also been recognised that computational efficiency may also be seriously affected (Clayton and Rasbash,1999). Future methodological investigation is diagnosed as necessary before a fuller store of knowledge is available..

The A level example will also introduce weighted random effects into the analyses. There are 145 teachers involved in the teaching groups and they may be supposed to be another specific source of heterogeneity to be disentangled and specified. Teacher effects are of obvious interest in their own right. However, without separate consideration they may confound with other features of group heterogeneity. Here each teacher usually teaches several groups. If additionally the same one teacher throughout its course had taught each group, then a three-way cross-classified structure would be a natural extension. However, except in a few isolated cases this is not so. Each group has up to six different teachers throughout its course, making their contribution in a split plot way. It is proposed to handle this by modelling the overall teacher effect on a response as a weighted average of the effects of the several teachers making a contribution. Choosing weights as proportions of the course length taught by each teacher has proved successful. Detailed timetable information in the data facilitates this. Other weighting schemes have been evaluated and results are relatively insensitive to choice. The weighting of random effects in this way is somewhat different from, but inspired in derivation by the ideas of Hill and Goldstein (1998). In *linear models only*, they apply weighting to structures where there is multiple membership of units at a level or where it is desired to attach probabilities to missing unit identifiers at levels in the data
.

Models and estimation

For six ordered categories of A level subject grade the logit model used for the cumulative distribution over grades is

$$\log it(g^{(s)}_{i(j_1,j_2)}) = q_s + \sum_{\ell=1}^{L} b_\ell x_{\ell i(j_1,j_2)} + u_{j_1} + u_{j_2} \text{ , s=1,2,......,5.}$$

The j_1 and j_2 indices range over teaching groups and students. Fixed effects dummy variables were introduced for the six colleges and are included amongst the fixed coefficients used. The model is two level with separate additive random effects for group and student at level 2.. For inference purposes these are assumed as usual to be normally and independently distributed with variances $s^2_{u_1}$ and $s^2_{u_2}$ to be estimated Level 1 observations indexed by i are lodged within cells (j_1,j_2) of the level 2 crossing. In general this model could also be elaborated in many ways, such as more crosses, more levels and random coefficients. It is also possible for the crossed effects to interact. With many colleges, their effect might have been represented as random at level 3. Further the coefficient, of a prior ability variable (the first) was considered as random at the teaching group level in model exploration. Results were uninteresting and will not be presented. However, these elaboration serve to give an example of more complex model specifications that could be entertained,

This is $$\ell n\left[\frac{g^{(s)}_{i(j_1,j_2)k}}{1-g^{(s)}_{i(j_1,j_2)k}}\right]=q_s-(b_{1j_2}x_{1i(j_1,j_2)k}+\sum_{\ell=2}^{L}b_\ell x_{\ell i(j_1,j_2)k}+v_k+u_{kj_1}+u_{kj_2})\text{, s=1,2,.....6.}$$

The elaboration of the indexing and introduction of the level 3 effect and the prior ability coefficient random across teaching groups may be noted.

An extension of the basic example to encompass weighted random teacher effects uses a model of the form

$$\ell n\left[\frac{g^{(s)}_{i(j_1,j_2)}}{1-g^{(s)}_{i(j_1,j_2)}}\right]=q_s+\sum_{\ell=1}^{L}b_\ell x_{\ell i(j_1,j_2)}+u_{j_1}+u_{j_2}+\sum_{j_3=1}^{J_3}w_{i(j_1,j_2)j_s}u_{j_3}\ .$$

The teacher random effects are denoted by u_{j_3} with variance $s^2_{u_3}$. In the weighted contribution the sum ranges across all J_3 teachers in the study,. However, for each observation most of the $w_{i(j_1,j_2)j_s}$ will be zero . The non-zero weights will be set by the teaching group of the entry observation, according to the relative contribution of the few teachers involved.

The theory of estimation of similar models for continuous response is given by Goldstein (1995) and Raudenbush (1993). We apply ordinal models. Since the crossed effects are in the linear predictor above level 1 many of the aspects of the methods that have been suggested for handling such structures carry over readily. Mostly these involve reformulating the models in various ways so that random effects can be treated hierarchically. Details are provided in the MlwiN manual, which also discusses fully the actual setting up and worksheet management. Some very detailed understanding of the complexities of the data structure is required for this. The synthesis of this with quasi- likelihood procedures for ordinal data is provide in a specially written MlwiN macro ORDCAT written by this author. This will ultimately be incorporated in the MULTICAT suite but is currently downloadable with user notes from *www.bham.ac.uk/economics/fielding*.

The example application: Subject grades at GCE Advanced Level in Six Colleges: group, student, and teacher effects.

The first two columns of Table 6 present estimates for a base and elaborated hierarchical logit models for entries within 317 subject-teaching groups. The responses are 3717 A level entries with six grades of outcome. (single subject students have been excluded as representing a special group) These models ignore the fact that responses were not independent across

Table 6 Parameter estimates for cross-classified and weighted random effects models for performance in subjects at General Certificate of Education at Advanced Level in six colleges for post compulsory school aged students. The response is the six point graded result of a subject entry. There are 3717 entries within 317 subject teaching groups from 1522 students The number of teachers involved is 145. (estimated standard errors of parameter estimates are in parentheses) The base for Subject Group dummies is Social Sciences. The base for Institution dummies is medium sized Further Education College: FEC, TC, SFC denotes Further Education, Tertiary and Sixth Form Colleges

	Base teaching group model	Teaching group model	Base model with student random effects	Model with student random effects	Base model with student and weighted teacher effects	Model with student and weighted teacher effects
Fixed effects						
q_1	-1.67 (0.07)	-1.54 (0.28)	-1.46 (0.07)	-1.33 (0.27)	-1.66 (0.08)	-1.57 (0.28)
q_2	-0.73 (0.06)	-0.47 (0.28)	-0.61 (0.06)	-0.44 (0.27)	-0.72 (0.08)	-0.51 (0.28)
q_3	0.17 (0.06)	0.59 (0.28)	0.21 (0.06)	0.45 (0.26)	0.19 (0.08)	0.53 (0.28)
q_4	1.08 (0.06)	1.67 (0.28)	1.05 (0.06)	1.35 (0.27)	1.09 (0.08)	1.58 (0.28)
q_5	2.48 (0.08)	3.32 (0.29)	2.34 (0.07)	2.75 (0.28)	2.48 (0.09)	3.20 (0.29)
STGC:GCSE score at enry to A Level Standardised		1.33 (0.05)		1.21 (0.06)		1.32 (0.05)
STGC squared		0.27 (0.02)		0.22 (0.03)		0.27 (0.02)
Female Gender		-0.12 (0.05)		-0.07 (0.08)		-0.14 (0.07)
Interaction of STGC and Gender		-0.18 (0.06)		-0.14 (0.08)		-0.20 (0.06)
SUBJECTS:						
Art, Design & Technology		-0.08 (0.20)		-0.06 (0.19)		-0.05 (0.21)
Mathematics		-0.40 (0.17)		-0.60 (0.19)		-0.17 (0.24)
Sciences		-0.38 (0.16)		-0.48 (0.15)		-0.41 (0.18)
Humanities		0.12 (0.16)		0.04 (0.15)		0.13 (0.18)
Languages		-0.49 (0.23)		-0.46 (0.21)		-0.27 (0.26)
General Studies		-0.52 (0.40)		-0.52 (0.34)		-0.44 (0.38)
COLLEGES:						
Large FEC		0.16 (0.29)		0.14 (0.29)		0.34 (0.39)
Medium sized TC		0.99 (0.30)		0.81 (0.30)		0.89 (0.31)
Small SFC		0.85 (0.31)		0.83 (0.32)		0.70 (0.34)
Medium sized SFC		-0.12 (0.29)		-0.12 (0.28)		-0.59 (0.34)
Large SFC		0.58 (0.26)		0.46 (0.27)		0.33 (0.29)
Random effects Variance						
Teaching groups % of *lv* residual variance	0.7083 (0.0785) 17.7	0.7308 (0.0807) 18.2	0.5109 (0.0607) 9.4	0.5412 (0.0620) 10.7	0.2145 (0.0711) 5.2	0.1282 (0.0614) 3.1
Students % of *lv* residual variance			1.6402 (0.0933) 30.1	1.22 (0.0766) 24.1	0.2792 (0.1164) 6.7	0.2412 (0.1138) 5.9
Teachers % of *lv* residual variance					0.3491 (0.1623) 8.4	0.4521 (0.1581) 11.0
Extra- multinomial	0.953 (0.010)	0.955 (0.010)	0.696 (0.006)	0.685 (0.006)	0.955 (0.010)	0.971 (0.010)

groups due to student effects in common. Indeed this was the type of model initially entertained in work on teaching group effectiveness before it was realised that it may have been miss-specified. However, these models provide a point of comparison for models with more elaborate specifications. The second two columns introduce the cross-classified student effect to handle this nesting of entries within 1522 students. Disentangling the group effects and the effects of the 145 teachers involved in them is attempted in the final 2 columns.

The same set of elaborating explanatory fixed effect covariates was used in each case. Those presented were a culmination of much deeper exploration of potential fixed effects for which data were available but which proved unfruitful. These included some teaching group context effects such as size and aggregate process variables such as attrition from the course. The main object of this example is to focus on the elaboration of random variation but some comment may be offered on the fixed estimates. The broad pattern of effects is similar across the differently specified models. The first four effects are for variables measured at the student level. STGC is standardised from the average of a number of GCSE subjects taken by students usually just before embarking on their two year A level courses. A quadratic term is also required for this. There is a marked ceiling to these averages and they are skewed to this ceiling. These factors may explain the quadratic effect. However, higher order polynomials are not required for the performance function as happens with linear models of the same response. The negative gender coefficients indicate that girls make less progress than males. By contrast, although not illustrated here, positive female effects emerge if STGC is not controlled and unadjusted performance is the issue. This phenomenon is also encountered in an analysis of a national 1997 cohort by Yang, Fielding and Goldstein (2000). There is also a negative interaction indicating that girls have a smaller STGC 'slope' effect. This will mean that lower ability girls will make more progress than similar boys but vice-versa at higher ability levels. A level subjects are categorised into broad groups and the dummies in Table 9 are relative to Social Science. There are some important subject effects. There is a vigorous debate in the literature about whether results such as this mean that Mathematics, Sciences and Languages can be perceived as more difficult. (Fitz-Gibbon and Vincent, 1997; Goldstein and Cresswell, 1996; Newton, 1997). This will not be pursued here. The six colleges represent a range of sizes and types found in British post school education (Belfield et al, 1996). College dummies are relative to a medium sized Further Education college. It is known that college size and type do make a difference. They have been introduced here in fixed effects as relevant block adjustment controls. There are too few in this data to draw generalisations apart from differences between specific colleges in the data.. In the tables both sets of dummies characterise the teaching group and teacher levels.

There are some differences in detail in estimates of fixed parameters across the three scenarios. They must be evaluated in the light of the relationships to extra controls that the introduction of further random effects implies. Effects on log odds mirrored in the coefficient estimates are net of random effects.. Thus we might expect some changes when student heterogeneity is introduced into the teaching group models, since they are then net of unmeasured student attributes. On a *lv* linear model interpretation there will also be consequent scale changes. The reduction in student variable coefficients is proportionately in line with changes in the cut point estimates indicative of rescaling. However, there are uneven changes in the subject and college dummies. They are not consonant with scale changes and those of the associated net log-odds. Part of the reason for this may be the clustering of student entries into certain subject groups and the attraction of some colleges for certain types of student. The mathematics and science effects are much more sharply defined. On investigation, the weighted teacher model would appear to have similar implicit conditional *lv* variability to the teaching group model. Further the cut-points and student variable coefficients have similar values. Mathematics and language effects relative to Social Sciences are no longer significant. It might be conjectured that subject effects observed in earlier models might be inextricably bound up to some extents to the type of teachers that deliver them. On introducing a teacher effect the net effects of subjects will thus change. Similar comments may be made about the changing pattern of college effects. There is quite a lot of complexity in these patterns, which might be unravelled by deeper investigation and more extensive data. The results do, however, pose some intriguing questions in the study of educational progress. They cannot be fully investigated here. As a final detailed point about the fixed estimates it may be noted that there are only minor changes in their estimated standard errors as variance specifications are refined. However, it may be pointed out that in most statistical investigations the accuracy of these estimates is sensitive to what is assumed about the specifications of variance. In general more appropriate specifications lead to better inferences.

The variance component estimates across the models raise many interesting issues of both methodological and substantive nature. In the teaching group model the covariates reduce the teaching group and entry variation proportionately. This is seen in the similar percentages (17.7 and 18.2) attributable to groups relative to standardised entry *lv* variance ($\pi^2/3$). An approximate scale calculation yields a variance reduction of the order of 30%. Introducing a student cross-classified random effect into the base model reveals two interesting features. Firstly part of the teaching group variation is now explicable by the differences of students selected into them. Students do not make an independent contribution within groups since their effects are common to certain groups. Secondly , variation amongst students is fairly high at 30% of total variation. However it is relatively much less than the 60.5% represented by the lv variance at the entry level. On this evidence there is much variation between the A level grades of subjects taken by each student. This point is conventionally recognised by some university admissions officers who specify sets of particular grade achievements for specific subjects rather than rely on aggregate points scores. For many purposes the latter hides the diversity in addition to being a dubious scaling device. In the model with student effects the greatest relative impact of the control covariates is on the student variance, which may be expected..

The weighted teacher models seem to exhibit some contrasting variance estimates. Due to its nature this model and the split plot structure it reflects has special features that need to be accounted for. Further detailed investigation of the data and the nature by which certain types of teacher associate with certain types of student and subject group would be required. Such an investigation is beyond the present illustrative purpose. However, a few important comments on the results may be made. One is that the variance contributions of teacher random effects to sampled observations are not conventionally additive. If the teacher variance estimate is $\hat{s}_T^2$ then it is $\hat{s}_T^2 \sum_{j_3=1}^{J_3} w^2_{i(j_1,j_2)j_s}$. Thus , for example a group with two equally weighted teachers would have a contribution of $\hat{s}_T^2/2$, whilst one with four equal weighted contributes $\hat{s}_T^2/4$. This shrinking of the variance contribution may be expected in that the overall teacher effect is a weighted average of several independent effects. Teacher effects may be important but their allocation to certain types of group and student may mean they alter other net random effects and may at the extreme cancel each other out. It may be asserted that observations in groups with larger number of teachers would contribute relatively more to residual entry variability. These factors may explain the apportionment of variances evident in Table 6 in a complex way. For the present purposes an examination of the weighted model in its own right reveals some useful insights. It is apparent from the base model, for instance, that on the same scale teachers exhibits more variability than either students or groups when they are jointly considered. Observed student progress and its variability would seem to have as much to do with the teachers they are exposed to as anything else. The same may apply to group variability. The control covariate model further adds to this assessment of the importance of teachers. Data on conventional teacher characteristics such as age, gender, length of service, education, and training are available. These have been tried in models with weighted fixed effects but none proved useful in explaining teacher effects. Teachers obviously matter but it is a challenge to educational research and practice to explain in what way. Some methodological tools to unravel complex effects have been provided. What is further required is more attention to study designs in relevant research and the collection of detailed data reflecting the complex structures.

Rejoinder

(1)We have stressed available methodology which is an aid to explanation even for designs which are not experimental manipulations. What is usually required, however, is more data and attention to its collection

(2) Readers requiring more familiarisation with multilevel models can teach themselves using the web based teaching resource TraMMs (2000).Details of how to access this are given in Appendix B

References

Aitkin, M., & Longford, N.T. (1986). Statistical modelling in school effectiveness studies**,** *Journal of the Royal Statistical Society , Series A*, *149, 3, 1-43.*

Belfield, C., Fielding, A., & Thomas, H. (1996). *Costs and performance of A level provision in colleges*: Research report for the Association of Principals in Sixth Form Colleges. School of Education, University of Birmingham

Bock, R.D. & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35, 179-197

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*, Newbury Park Ca, Sage

Chesher, A. D., & Irish, M. (1987). Residual analysis in the grouped and censored normal linear model. *Journal of Econometrics, 34, 33-61*

Clayton, D. J., & Rasbash, J. (1999). Estimation in large crossed random effects models by data augmentation*. Journal of the Royal Statistical Society, A, 162, 3, 425-436.*

Coe, R., & Fitz-Gibbon, C.T. (1998). School Effectiveness Research : Criticisms and Recommendations. *Oxford Review of Education, 24, 4, 421-438*

Coleman, J. S. , Campbell, E., Hobson, C., McParland, J., Mood, A., Weinfield, F., & York, R. (1966). Equality of educational opportunity, Washington D. C., Government Printing Office

Department for Education and Employment [DfEE]. (1997). *The implementation of the* national literacy strategy. London:,DfEE**.**

Ezzett, F., & Whitehead, J. (1991). A random effects model for ordinal responses from a crossover trial*. Statistics in Medicine, 10, 901-907*

Fielding, A. (1998) Methodological Innovations in the Use of Teaching Groups for Evaluating Cost- Effectiveness, *International Congress on School Effectiveness and Improvement, 1998 Proceedings*, CD Rom, ISBN 0-902252-62-3, School of Education, University of Manchester.

Fielding, A (1999). Why Use Arbitrary Points Scores? Ordered Categories in Models of Educational Progress, *Journal of the Royal Statistical Society , Series A*, *162, 3, 303-328*

Fielding, A. (2000). Ordered category responses and random effects in multilevel and other complex structures: Scored and generalised models, in *S. Reise & N. Duan, Multilevel modelling: Methodlogical advances, Issues and application*s, New Jersey, Erlbaum

Fielding, A., Belfield, C.R, & Thomas, H. (1996). *Costs and performance of A level provision in schools*: Research report for the Department for Education, School of Education University of Birmingham

Fielding, A., & Yang, M. (1999). *Random effects models for ordered category responses and complex structures in educational progress,* Discussion Paper 99-20, Department of Economics, University of Birmingham

Fitz-Gibbon C. T. (1996). *Monitoring Education: Indicators, Quality and Effectiveness*, London , Cassell

Fitz-Gibbon, C. T., & Vincent, L. (1997). Difficulties regarding subject difficulties: Developing reasonable explanations for observed data. *Oxford Review of Education, 23, 3, 291-298*

Fraser, B. J. , & Tobin, K. (1998). Student's perceptions of pyscho-social environment in classrooms of exemplary science teachers*. International Journal of Science Education, 11, 19-34*

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in practice,* London, Chapman and Hall

Goldstein, H. ((1995). *Multilevel Statistical Models*. London, Edward Arnold

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

Goldstein, H. & Blatchford, P. (1998). Class size and educational achievement: a review of methodology with particular reference to study design*. British Educational Research Journal, 24, 255-268*

Goldstein, H., & Cresswell, M. (1996). The comparability of different subjects in public examinations. *Oxford Review of Education, 22, 4, 435-442*

Goldstein, H. & Healy, M. J. R. (1995) The graphical presentation of a collection of means, *Journal of the Royal Statistical Society , Series A,158, 505-513*

Goldstein, H. & Langford, I

Goldstein, H. & 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, 2, 219-230*

Goldstein, H. & Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance, *Journal of the Royal Statistical* Society , Series A,159, 385-443

Greene, W. H. (2000). *Econometric Analysis (4th Edition),* Upper Saddle River, New Jersey, Prentice-Hall

Haitovsky, Y. (1973). Regression estimation from grouped observations. London: Griffin

Hedeker, D., & Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics, 50, 933-944*.

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

Hill, P. W., & Goldstein, H. (1998). Multilevel modelling of educational data with cross-classification and missing identification for units. *Journal of Educational and Behavioural Statistics*

Hill, P.W. and Rowe , K.J. (1996) . Multilevel Models in School Effectiveness Research, *School Effectiveness and School Improvement,* 7,1,1-33

Hill, P.W. and Rowe , K.J. (1998) . Modelling student progress in studies of educational effectiveness, *School Effectiveness and School Improvement,* 9,3,310-333

Hodgson, W. (1995). *Gender differences in mathematics and science*. Ph. D. Thesis, School of Education, University of Newcastle Upon Tyne

Langford, I. H., & Lewis, T. (1998). Outliers in multilevel data (with discussion). *Journal of the Royal Statistical Society, Series A, 161, 121-160*

Luyten, H., & de Jong, R. (1996). Parallel classes : Differences and similarities . Teacher effects and school effects in secondary schools. *School Effectiveness and School Improvement,* 9, 4, 437-473

McCullagh, P. , & Nelder, J. A. (1989). *Generalised linear models (2nd Edition),*London, Chapman and Hall

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology, 4, 103-120*

Monk, D. H. (1992). Educational productivity research: An update and assessment of its role in education finance reform*. Educational Evaluation and Policy Analysis, 14, 307-332*

Newton, P. E. (1997). Measuring Comparability of Standard Between Subjects: Why our Statistical Techniques Do Not Make the Grade. *British Educational Research Journal*, 23, 4, 433-449

Paterson, L. (1990). An Introduction to Multilevel Modelling, Chapter 2 in S W. Raudenbush and J D Willms, (Eds), *Schools, Classrooms and Pupils*, San Diego, Academic Press

Paterson, L. & Goldstein , H. (1991) New Statistical Methods for Analysing Social Structures: An Introduction to Multilevel Models. *British Educational Research Journal, 17, 4, 387-393*

Plewis, I. (1998) Multilevel Models, *Social Research Update Issue Number 23, Department of Sociology, University of Surrey*

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., & Draper, D. (1999). *A user's guide to MlwiN, Version 2.0*. Multilevel Models Project, 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, 4, 321-349.*

Reise, S & Duan, N (Eds). (2000). *Multilevel modelling: Methodological advances, Issues and*
*application*s, New Jersey, Erlbaum

Rutter, M., Maugham, B., Mortimore, P., Ouston, J., & Smith, A. (1979).*Fifteen thousand hours*, Wells, Open Books

Stewart, M. B. (1983). On least squares estimation when the dependent variable is grouped Review of Economic Studies, 50, 737-753.

TRaMMS (2000). *Training Resources and Materials for Social Scientists*. The Data Archive, University of Essex, (tramms.data-archive.ac.uk)

Word, E. R., Johnston, J., Bain, H. P., Fulton B. D., (1990). *The State of Tennessee's student/teacher achievement (STAR) project: Technical report 1985-90*, Nashville, Tennessee State University

Yang, M., Fielding, A. & Goldstein, H. (2000). *Multilevel ordinal models for examination grades,* submitted for publication , from Multilevel Models Project, Institute of Education, University of London

Yang, M., Goldstein, H., & Heath, A. (2000). Multilevel models for repeated binary outcomes: attitudes and voting over the electoral cycle. *Journal of the Royal Statistical Society, Series A, 163, 1, 49-62*

Yang, M., Rasbash, H. , & Goldstein, H. (1998). *MlwiN macros for advanced multilevel modelling*, Multilevel Models Project, Institute of Education, University of London

Yang, M. and Woodhouse, G. (2000). *Progress from GCSE to A and AS level: simple measures and complex relationships*, Multilevel Models Project, Institute of Education, University of London, submitted for publication

Young, D. Y. (1998). Rural and urban differences in student achievement in science and Mathematics: A multilevel analysis. *School Effectiveness and School Improvement, 9,4, 386-418. . .*