

Occasional Paper Series  
Paper 46

When should we use multi-level modelling?

Stephen Gorard

*Cardiff University School of Social Sciences*

**The Research Team**

Executive Group:

Professor Stephen Gorard (Director)  
Professor Gareth Rees  
Professor John Furlong  
Dr Laurence Moore  
Professor Ken Prandy  
Dr Ray Crozier

Research Staff:

Dr Chris Taylor  
Karen Roberts  
Patrick White  
Helen Taylor (Project Administrator)

**Our Contact Details**

ESRC TLRP Research Capacity Building Network  
Cardiff University School of Social Sciences  
Glamorgan Building  
King Edward VII Avenue  
Cardiff  
CF10 3WT

Tel. 029 2087 5345  
Fax. 029 2087 4678  
Email. [TaylorH1@Cardiff.ac.uk](mailto:TaylorH1@Cardiff.ac.uk)

[www.cardiff.ac.uk/socsi/capacity](http://www.cardiff.ac.uk/socsi/capacity)



When should we use multi-level modelling?

Stephen Gorard

Cardiff University School of Social Sciences

Glamorgan Building

King Edward VII Avenue

CF10 3WT, UK

02920-875113

email: [gorard@cardiff.ac.uk](mailto:gorard@cardiff.ac.uk)

### **Acknowledgements**

This paper arose from work done as part of the ESRC Teaching and Learning Research Programme (grant L139251106).

## When should we use multi-level modelling?

### **Abstract**

This paper is intended to be a consideration of the role of multi-level modelling in educational research. It is not a guide on how to design or perform such an analysis. There are several references in the text to sources that teach the practicalities perfectly well, and the technique is anyway similar to other forms of regression and to analysis of variance. Rather, the paper describes what multi-level modelling is, why it is used, and what its limitations are. It does so in the hope that some readers will be enthused sufficiently to become appropriately critical 'consumers' of research using this approach, so building research capacity, and easing pressures on 'specialist' reviewers. Anyone who can read or perform standard multivariate analyses can understand, referee, or conduct a multi-level model.

### **Introduction**

Multi-level modelling (MLM hereafter) presents a particular challenge for the capacity of the UK educational research community. It is growing in use as a technique, and widely cited as something that many more would like to know about (as shown by the results of a recent skills survey, see [www.cf.ac.uk/socsi/capacity](http://www.cf.ac.uk/socsi/capacity)). Yet despite being around for twenty years now, it remains an obvious symptom of the methodological division between 'quantitative' and 'qualitative' researchers. Few people are prepared to critique papers using MLM in their own literature reviews, and even fewer are prepared to referee them for journals or grant-awarding bodies. This is not good for the health of a field, such as school effectiveness, in which cognitive/intellectual similarity could lead to a damaging cronyism (Travis and Collins 1991).

MLM is a specific advance on prior methods of multiple regression (Miles and Shevlin 2001), and one which all researchers should be aware of - if only to realise the severe limitations of its use. The purpose of this paper is to consider what MLM is, what kind of situation it is appropriate to, and why it may not be as useful for the long-term as some of its advocates would

claim. The paper begins by putting MLM in the context of clarion calls for the greater use of 'advanced quantitative techniques'. It then outlines what MLM is and why it is used in a way that I hope all researchers can follow. The next section describes the limitations of MLM and problems in its actual use. The paper concentrates on why it might be used and when it is less appropriate for use, because these appear to be the areas least addressed by current treatises which are often otherwise excellent on *how* to conduct MLM. The paper concludes with a summary assessment of the value of MLM for the future of educational research.

### **The call for more statistics**

Much has been written elsewhere about various calls to change the climate of UK educational research (in Gorard 2002a). Much that has been written about this perceived crisis has suggested that the solution is largely a technical one of increasing the role of large-scale studies and associated techniques for the analysis of complex datasets. Indeed the ESRC currently has no fewer than fourteen initiatives in place to increase use of quantitative approaches among social science researchers, for example. While such measures may be necessary to meet a need for re-skilling the educational research 'workforce', they do not over-ride the greater need for rigour whatever the methods used.

The increasing quality and availability of computer software packages for statistical analysis allows more and more complex statistical models to be built and used, so that in the end most consumers of educational research simply cannot, or would not wish to, comprehend them. Even those who work on such high level models have trouble transforming their findings into a package that does their analysis justice but also makes any sense to practitioners and policy-makers (see Goldstein et al. 2000 on the difficulties of this). This means that the 'average' consumer of research has to either implicitly accept the findings, or reject them as incomprehensible. Linked to the greater use of computers is the shotgun or dredging approach to analysis in which multiple exploratory analyses are run with the same set of data. As well as liberating us from the drudgery of multiple calculations the computer has therefore increased the frequency of the 'blind or mindless application of methods without regard to their suitability for the solution of the problem at hand, or even in the complete absence of a clearly formulated

problem' (Pedhazur 1982, p. 3). While advanced techniques themselves are not to blame for this, their complexity may compound the problem by making abuses harder to spot. Their value therefore comes at a cost. This is context in which the paper addresses multi-level modelling – the need for greater awareness among researchers of advanced statistical techniques.

### **What is multi-level modelling?**

In standard multivariate linear regression techniques (multiple regression), the analyst attempts to 'explain' the variation in a dependent variable (the outcome) in terms of independent variables (the inputs). For example, a model may be constructed in which pupil examination scores are explained in terms of prior examination scores for the same pupils, and their sex. Which of these three variables is used as the outcome is a matter of choice and custom, and the solution could be recast statistically to show prior attainment, or even sex, as the outcome or dependent variable. The choice of dependent variable depends on the causal model you adopt. In this example, we might argue that two of the variables are determined prior to the final examination, and that the arrow of causation can only run one way (one cannot change sex through performance in an examination, for example). At least implicitly therefore, all regression adopts a path analysis approach to testing the feasibility of a previously determined causal path (and it might be better for this path to be made explicit via structural equation modelling, Maruyama 1998). Regression cannot, of course, prove that the independent variables *determine* the variation in the dependent variable, but it can show whether that is possible. Whatever you do with regression, and this includes the techniques discussed below, cannot overcome this limitation.

The full set of assumptions underlying regression techniques is quite large, and therefore can be rather terrifying. It is also the subject of some dispute, both over what the assumptions really are, and over the implications for running an analysis that does not meet them (e.g. Menard 1995, Miles and Shevlin 2001, de Vaus 2002). They are presented below as thirteen separate assumptions, although some of these are clearly related.

- The measurements are from a random sample (or at least a probability-based one)

- All variables used should be real numbers (or at least the dependent variable must be, since many authorities advocate the use of dummy predictor variables)
- There are no extreme outliers
- All variables are measured without error (but when is this possible?)
- There is an approximate linear relationship between the dependent variable and the independent variables (both individually and grouped)
- The dependent variable is approximately normally distributed (or at least the next assumption is true)
- The residuals for the dependent variable (the differences between calculated and observed scores) are approximately normally distributed
- The variance of each variable is consistent across the range of values for all other variables (or at least the next assumption is true)
- The residuals for the dependent variable at each value of the independent variables have equal and constant variance
- The residuals are not correlated with the independent variables
- The residuals for the dependent variable at each value of the independent variables have a mean of zero (or they are approximately linearly related to the dependent variable)
- No independent variable is a perfect linear combination of another (not perfect 'multicollinearity').
- For any two cases the correlation between the residuals should be zero (each case is independent of the others).

In general, if these are not true for any analysis, the impact is to reduce the apparent size of any relationship uncovered. Therefore, and in general, a significant result it is still relatively safe even when some of these assumptions are not met. The assumptions are listed in full here so that the last one – the independence of cases in the analysis – can be seen in context.

This last assumption is the key one for the more complex technique called 'multi-level modelling' (MLM), and it is this that determines whether MLM is particularly appropriate for any analysis. MLM is actually the name for a range of techniques, including both fixed and random effects, developed from 'hierarchical' approaches to analysis in fields such as agriculture.

It emerged particularly in the classic paper by Aitkin and Longford (1986), and writing by Raedenbush and Bryk (1986), and Goldstein (1987). It is based on a recognition that in the social world much of the data to be analysed has an inherent structure that it may be foolish to ignore. This means that the measurements we take may routinely flout the last assumption of regression techniques. Cases are *not* independent of each other. We might expect people in one family to be more like each other than like a person in another family. We might expect the health problems of all girls to be more similar in some respects than like that of a boy. We might expect pupils in one school to perform more like each other than a pupil in another school. Technically, such cases are said to be 'auto-correlated' (although other terms such as 'intra-correlated' are also in use). MLM advocates argue that since this is the case, we are better off building these similarities into our analytical methods. MLM is therefore simply regression that allows the analyst to use both individuals and groups of individuals in the same model to avoid flouting the last assumption (of independent cases), since the standard error of any results can be affected by the clustered nature of the data.

Paterson and Goldstein (1991) use a hypothetical example based on 5,000 pupils in 100 schools to make the point about auto-correlation. If we wish to explain a test outcome using standard regression, then we would treat the 5,000 cases separately and calculate a regression equation that maximised the accuracy of our predicted test scores using the independent variables (such as sex, or prior attainment). The equation would be of the form:

*pupil score = constant + error term + (coefficient1 times predictor-variable1) + (coefficient2 times predictor-variable2)...*

or in more formal terms for pupil i:

$$y_i = b_0 + e_i + b_1X_{1i} + b_2X_{2i} \dots$$

The constant ( $b_0$ ) is the 'intercept', the suggested value of the pupil score when the value of the predictor variable(s) is zero. If we allow this constant to differ between schools, since the predicted test score may also depend on the school attended by each pupil, then we are effectively adding a 'school term' to create an equation of the form:

*pupil/school score = constant + school term + error term + (coefficient1 times predictor-variable1) + (coefficient2 times predictor-variable2)...*

or in more formal terms for pupil  $i$  in school  $j$ :

$$y_{ij} = b_0 + u_j + e_{ij} + b_1X_{1ij} + b_2X_{2ij} \dots$$

This is the simplest form of a multi-level model. The term for the constant plus school ( $b_0 + u_j$ ) is now the intercept value for each school (in the same way as above, but where it is calculated per school rather than for the whole dataset). The model can be extended to allow more variables, more levels (pupils within schools within school districts, for example), and to allow the coefficients to vary between schools. It can be used in a variety of settings, perhaps most notably in repeated measures designs, where the individuals (cases) are measured more than once. We may assume that the scores for any individual are more alike than they are between individuals (in repeated measures the person is taking the place of the school, and the repeated scores are taking the place of the pupils).

Variants of MLM have long been used in other fields, and analogues appear in techniques such as logistic regression and ANCOVA. The major change in the last fifteen years has not been so much a statistical breakthrough, but rather an advance in the specialist software available. Like most techniques we do not need to know a lot more about the calculations involved, since these will be performed for us by a computer anyway (see Plewis 1997, for example, for more on how to conduct MLM). Rather, we need to understand what the technique is about. Clearly the MLM approach is better than simply using the mean score for each school to look for differences between schools, since there would then be only 100 cases, which would make for less reliable findings, and may miss interesting variations at the pupil level. To what extent is the MLM approach genuinely better, or even different, than doing two separate analyses - one with the school means, and one with the pupil scores in our example?

## Limitations of multi-level modelling

If the first assumption for the use of regression techniques - that measurements are taken from a random sample - is met then MLM is not needed (by definition). Any apparent auto-correlation that exists in the sample will also be random (accidental) rather than the result of social structures. For example, if the population is the Year Seven pupils in all schools in the UK, then a random sample will have a national 'spread'. There may be more than one pupil from the same school, just as there will almost certainly be more than one with the same sex, prior attainment, and class and ethnic background. It may be important that all of these variables, and more, are taken into account in the regression. But this is very different to the argument for MLM which is based on a very different kind of sampling - at best cluster-random and at worst cluster-convenience. MLM is, like all regression, based on probability theory, and the p-values generated in conducting an analysis are predicated on a random sample. If the sample is not a probability sample of some kind then MLM is not appropriate. If the sample is truly random, or systematic or stratified, then MLM is not needed. Therefore the first 'intrinsic' problem is the lack of a clear area of application for MLM.

The chief argument for MLM based on auto-correlation could only apply if the sample is a cluster-random one. Note that this is not the same as any 'sample' of co-operating schools used as a cluster-opportunity sample (e.g. whereby a local network of volunteer schools is used for convenience). Nor is it the same as a sample where cases are chosen randomly within such an opportunity cluster. The argument for MLM clearly requires a defined population of clusters from which a sample *of clusters* is selected at random (with no replacement, no non-response, and no dropout). If this condition does not exist then the claimed increase in accuracy from using MLM is likely to be more than lost by the design bias (Torgerson and Torgerson 2001). Therefore, many examples of MLM in actual use are inappropriate (e.g. Lauder et al. 1999). In fact, it now becomes unclear whether MLM could *ever* be appropriate. Statistical techniques are becoming more and more complex largely to overcome deficiencies in the datasets involved. Where we attempt to make statistical adjustments for these deficiencies, then we often do not know which variables to adjust for, or do not have the necessary data at the appropriate level of aggregation, or do not have the techniques to adjust appropriately. Is MLM such a technique?

The answer, on balance, has to be no. It is important that the unit of randomisation is also the unit of treatment for results purposes (Moses 2001). For cluster samples this means that the cluster is the unit of analysis, not the individuals within them. Of course, individuals are important, but to conduct a statistical analysis for individuals generally requires the collection of data from a random sample of *individuals*. A good cluster-random sample mimics a true random sample (as does a systematic or stratified one), and insofar as it does it then allows an individual-level analysis without concern for auto-correlation at the cluster level. Put simply, auto-correlation is a deficiency of sampling not of analysis, and the appropriate solution is therefore better sampling not more complex analyses. Where deficient datasets already exist, then complex techniques may be the only way to analyse them, but in the future we may be better served by devoting more effort to the collection of better data that can be analysed more simply.

The MLM approach allows for nested hierarchies, but neither considers how high or low to aggregate these hierarchies, nor gives clear guidance on what to do in the usual situation where hierarchies are competing rather than nesting. What is done about the auto-correlation also caused by sex, social class, ethnicity, days of the week for testing, time of day for testing etc.? The hierarchies created by these clusters do not nest, so in practice they are ignored. Sex, for example, is used a simple explanatory variable at one or more levels in school effectiveness research while the school is used as a cluster for analytic purposes. There seems no clear *a priori* reason for this. In practice within school effectiveness research, MLM is used with simple nested hierarchies, often of only two levels in fact, and competing non-nested hierarchies are ignored. It would appear that the grounds for this 'decision' may be based on the requirements of MLM rather than on the social reality under investigation. Using a one-level model, on the other hand, with both individual and grouped variables allows every relevant hierarchy to be assessed in the same model. For example, a standard regression model of individual pupil examination outcomes could have individual prior attainment and mean prior attainment of school attended, as well as individual sex and mean prior attainment of same sex, as explanatory variables. This is so despite the fact that prior attainment by school and prior attainment by sex do not nest.

Paterson and Goldstein (1991) claim that 'the key technical advance of multilevel modelling is to assume that the  $u_j$  vary randomly across schools' (p.388). This constant determines the intercept for each school, and is an important part of the predicted score for each pupil. It is routinely

calculated for each school, using observed differences between pupils in those schools. The reliability and scientific safety of any school term depends therefore not on the overall sample size, nor the number of schools, but on the size and nature of the sample in each school. In their example, having rejected using the sample of 5,000 individuals because of auto-correlation caused by the clustered nature of the sample, and rejected using the aggregate scores for 100 schools as too few to be safe, Paterson and Goldstein (1991) instead advocate using the 50 cases in each school to calculate 100 versions of  $u_j$ , and relying on these to help assess a school effect. The ensuing bias is likely to be far greater than that involved in simply using 5,000 cases, and warning readers of the possible impact of auto-correlation. The medicine may be more harmful than the disease.

### **Problems in using MLM**

Advocates of MLM will argue that the technique has other advantages besides. After all, it allows us to measure the size and determinants of 'school effects'. Recall that the original statistical argument for MLM was based on auto-correlation which implies the existence of cluster 'effects'. The method is predicated on the existence of these effects, and therefore cannot be used to test for their existence. Otherwise we enter a topsy-turvy world in which the same dataset is used to both model and test scientific propositions (clearly nonsensical), and in which residual scores (the difference between best prediction and actual score) are re-labelled 'school effects' or 'value-added'. But these residuals are largely error terms (Fitz-Gibbon 2000).

MLM techniques have therefore not been above criticism, particularly where they have been used in 'school effectiveness' research (SER). However, this criticism has generally been weak and easily avoided (for example the best criticisms of SER in Slee et al. 1999 are not concerned with MLM itself). Where these criticisms have had more substance they have generally been of the way that MLM has been used in practice, and have therefore been defended by MLM advocates agreeing that the examples cited were indeed poor uses. At the AERA annual conference in New Orleans 2000 there was a public debate about the value of SER. It was 'won' hands down by the SER advocates largely because their critics showed so little understanding of the techniques underlying the research. The complexity of MLM is its main bulwark, and the

'don't do numbers' mentality of many UK educational researchers has given it a very easy ride so far.

Gibson and Asthana (1998) deplored the way in which MLM can be used to parcel out and then ignore the variance in school outcomes due to social and economic factors. Routine use in SER calculates the impact of prior attainment and socio-economic background of pupils, and notes that this is the vast proportion of the variance involved. But it then concentrates on 'partitioning' the remaining variance (often treating this remainder as a new 100%). This does a dis-service to users and policy-makers who are then given an account that appears to overemphasise the impact of teachers, school leaders, and classroom settings, while neglecting issues such as pupil background.

Coe and Fitz-Gibbon (1998) are also critical, without supporting the position of Gibson and Asthana. They argue that the usual socio-economic categories are not, in themselves, explanations. We do not know why sexes, social classes or ethnic groups attain differentially and therefore these remain pseudo-explanations. However, so are the other explanations used in standard SER. Fielding (2000) has argued that MLM is best used where experiments are not possible (perhaps for ethical reasons) in order to make up for defects in the data. In practice however, MLM is often simply used as a replacement for experiments (Fitz-Gibbon 2000), and leads to a pretence that this essentially passive approach is a kind of 'magic bullet', uncovering causation, overcoming poor design, and allowing researchers to draw robust conclusions from poor datasets (Coe and Fitz-Gibbon 1998, see also Goldthorpe 2001, Gorard 2002b). Users routinely measure the differences between schools first and then simply employ MLM to dredge for statistically 'significant' differences.

The nature of MLM techniques means that there are very few people outside what is in danger of becoming a kind of cult who can follow this kind of work, and it sometimes seems writers work to keep it that way. A clear example is the tradition of using technical variable names to report findings. It is now standard practice within this club to present findings in terms of brief acronymic names rather than descriptions. At the next level of absurdity writers try to explain the meaning of their variable names. Why not just use the description itself?

For example, in their main chapter about the relationship between school choice and school performance - the empirical guts of their book - Lauder et al. (1999) present these less than fascinating facts.. 'in the Year 11 School Certificate Study we included a Level 1 variable called FAMSTR which was not included in the Year 10 skills study' (p.116), and 'at Level 1 the variable name was MAPTITUDE' (p.117) etc. In fact their Table 7.1, which looks at first sight like a set of results, is actually just a summary of their variable names. Why should I want to know this? I would not report a t-test and point out that when entering some numbers into SPSS I referred to them as 'X' or 'VAR00001'. This is not what readers would need to know. This technical mumbo jumbo is usually presented instead of, rather than in addition to, information that I do actually need. Lauder et al. (1999) present the chief results in their Table 7.4 which contains only variable names and associated alpha levels. We are not told what the units of measurement are for each variable, so when shown the coefficients from their regression analysis we can have no idea what these values mean. The coefficients by themselves are useless information for us, and like the variable names therefore simply become rhetorical noise.

### *Representational errors*

It is generally assumed that where we are analysing school outcomes it is better to use data relating to the individual student rather than aggregated figures relating to classes, groups or schools. This is certainly the path followed in standard school effectiveness studies. However, we should all be aware that there is a considerable error component in the allocation of school outcomes. Whatever the system of moderation used, public examinations are inaccurate, so much so that they are estimated to be accurate only to within a grade or two (Nuttall 1987, Gorard 2001a). In a large aggregate analysis (e.g. at school level), we can assume that these 'random' errors are largely cancelled out, but this is not so when the analysis focuses on individuals. Raudenbush (2002) concludes that randomised studies of school reform should use the school as the unit of allocation and treatment, and therefore the analysis presumably.

Statistical analysis by computer involves very many calculations of which most of us are usually only dimly aware, and one of the dangers of this is that we cannot therefore make a reliable estimate of the 'propagation' of our measurement errors. It is a standard assumption in social science that any measurements we make are not totally accurate. We can also introduce further small errors by restricting our working to a certain number of decimal places or significant

figures. We simply do our best to take accurate readings, and include an error component in our subsequent modelling to represent these general flaws. To a large extent we behave as though the error component in our analysis remains constant, so that if we start with figures at a certain level of accuracy, we will end up with results at approximately the same level of accuracy. In some cases this behaviour may be appropriate but in others, known in extreme form as 'ill-conditioned' problems, it is not so. If we assume that all of our measurements are in error (and with most educational measures this is a safe assumption), then adding two figures also involves adding their error components. The error components may partly cancel each other out, or they may increase each other. More formally imagine two numbers whose true value is A and B, for which our measurements a and b are only approximations such that:

$$a = (A + E_a)$$

$$b = (B + E_b).$$

Where  $E_a$  is the error in our measurement of A, and  $E_b$  is the error in our measurement of B. If we add our estimates of A and B we actually reach the sum  $A+B+E_a+E_b$ . This is unlikely to be a major problem since the relative error  $(E_a+E_b)/(A+B)$  is probably not much larger than either  $E_a/A$  or  $E_b/B$  (the relative errors with which we started). Since we do not know whether  $E_a$  and  $E_b$  are positive or negative the same result occurs when we subtract A and B. If we multiply a by b we obtain  $(A + E_a).(B + E_b)$  which equals  $A.B + B.E_a + A.E_b + E_a.E_b$ . The error terms  $A.E_b$  and  $B.E_a$  could be large if A or B is very large, and in this way the original error in our measurements could propagate with every calculation we make, being added to and multiplied in turn.

In regression-type analyses we can also introduce error if we diverge from the ideal of a linear relationship between variables. MLM, like standard regression, largely misses all non-linear relationships. The error component in the residuals will swell because of any non-linearity in the relationship between the dependent variable and a basket of independent variables. To the likely measurement, transcription, and representational errors we therefore have to add potential problems from non-linearity. Where the variables involved are interval (real-numbers) in nature, as they should be according to strict assumptions, then not only can social class, school type, and sex no longer be included (for these are categorical variables) but using logarithmic

transformations to overcome non-linearity might introduce further problems (Harwell and Gatti 2001). On the other hand, where we include categorical and ordinal variables, we may violate another assumption underlying the techniques involving normality in the distribution of variables. MLM involving ordinal and dummy variables, for example, might manoeuvre round the assumption of independence of cases but flout several other assumptions in doing so.

Unless we track the potential propagation of these errors it is possible for our answer effectively to 'cancel out' the estimates we started with, and so contain a much larger proportion of error component than we started with.

Consider the simultaneous equations:

$$\begin{aligned}400 &= 200x + 200y \\201 &= 101x + 100y\end{aligned}$$

Their solution is that  $x=1$  and  $y=1$ . If a measurement in the second equation is incorrect by less than half a percent, then the true value of the first figure could be 200 (rather than 201), making the equations:

$$\begin{aligned}400 &= 200x + 200y \\200 &= 101x + 100y\end{aligned}$$

The solution now is that  $x=0$  and  $y=2$ . This is a totally different solution 'caused' by a small proportionate error in one term. Imagine the practical implications if  $x$  and  $y$  were components of an effective school as assessed by MLM. For some problems the introduction of an error component makes a large difference, and for some problems the error makes *all* of the difference.

### *Order of entry*

For any study, the regression model explaining the greatest variance in the dependent variable (e.g. exam score) will use all available independent variables. This is the model you get if you simply enter all of the variables at once. However, it is possible to create simpler models containing fewer variables but still explaining a large proportion of the variance. These models

are easier to use and understand, and so more practical. It may be that ethnicity and first language, as variables, are measuring much the same thing in terms of school outcomes. The same may be true of social class and indicators of poverty. In such cases, we are better off picking the best single indicator from a group of related measures, and using only that one. We could pick the best indicator on theoretical grounds, or in terms of availability. Both of these approaches are fine. However, the most common ground for selection of variables is the proportion of variance that they explain. If language and ethnicity are related and language is the better predictor of examination scores then we might omit ethnicity from our analysis.

In several forms of multivariate analysis, such as multi-level modelling, the order in which independent variables are entered into the explanatory model can also make a very substantial difference to the results obtained. Different explanations of social phenomena can be derived using the same technique but with only minor variations in the order of entering variables. Since many of the best-known and influential theories in education are based on precisely such models the importance of bearing this principle in mind is difficult to exaggerate. Put simply, in the absence of greater detail about the order in variables are considered, some of these theories may be less secure than previously imagined (in Gorard 2001b).

Clearly both the assumption of no measurement error, and the sensitivity of results to changes in the order of calculations, apply equally to other regression techniques as to MLM. They are rehearsed here largely to remind readers that MLM does nothing to solve these major issues, since there is a tendency for some novices to see the technique as a push-button solution to more than the relatively minor problem for which it was devised. However the scale of these problems for MLM are also greater, I would argue. Where regression is used simply to assess the R-squared (total variance explained), or even to measure robust differences between coefficients, then the introduction of representational errors and order of entry effects may be less important. MLM, however, is routinely used to partition the R-squared into prior attainment and socio-economic background (explaining between 80 and 100% of the variance, see Gorard 2000), and the residual including the error term and the school effect (explaining between 20 and 0% of the variance). The important results for SER then come from partitioning this residual variance into the determinants of value-added scores and so on. This is playing with the 'loose change' of the model, in which it is far more likely that the errors and design bias will dwarf the impact of the

variables of interest. While textbooks warn of the danger of handing over scientific control to 'stepwise' approaches in regression (e.g. Bernard 2000), descriptions of MLM seem to assume that nesting can only be in one direction and make no reference to the possibility of different results arising from different methods of partitioning variance (e.g. Plewis 1997). Perhaps it is obvious to readers that social background always precedes school effect, or that individual variance precedes classroom variance. However, I would like to see the case argued. For me, the arrow of causation is far from clear here. Are the levels in a multi-level model meant to be mutually determining or analytically separate? Why are the same variables routinely used to 'explain' variance at the individual level, and then aggregated to explain variance at the cluster level as well (e.g. Lauder et al. 1999)? Does this test for a school-mix effect, or does it take such an effect for granted?

MLM was introduced to try and overcome what is only one of around thirteen statistical assumptions underlying the use of standard regression techniques, and even that line of argument is based on accepting a deficiency of either research design, or sample quality, or both. MLM involving ordinal and dummy variables, for example, might manoeuvre round the assumption of independence of cases but flout several other assumptions in doing so. What we gain from using MLM may be more usefully attained by better design or greater use of genuine random samples, and the gain is anyway liable to be lost among the errors and complexity involved. It remains the case that anything that can be done with MLM can also be achieved by other means - perhaps most simply by conducting an analysis for each level with aggregated scores where necessary. This latter approach also has the advantage of allowing the levels to be non-nesting in a simple way (such as sex and school). MLM 'is technically an improvement over the traditional multiple regression model... but there are simpler ways, ways that do not need more than one iteration' (Kreft, p.14). When five large datasets were analysed using both standard regression and MLM the two sets of results from each correlated at around 0.99 (Fitz-Gibbon 2001). It is interesting to note how often a standard single-level regression analysis using 'clustered' variables (such as the proportion of free-school meal pupils in each school) completely replicates a multi-level analysis. Approximately the same amount of variance is explained by roughly the same variables with MLM as without. In these cases, results using MLM are less accessible to a wider readership *for no apparent reason*. They are not as parsimonious as traditional regression (requiring more parameters to be estimated), they are less generalisable (i.e. more specific to the

dataset they are fitted to), need a larger dataset, and are more complex to estimate. Therefore, 'after fifteen years of promotion of these models, some disappointment has set in' (Kreft 1996, p.1).

## **Discussion**

The decision to use, or not to use, multi-level modelling, should be as considered as any other method decision, and cannot be taken independently of other aspects of design, analysis, and sampling. MLM is clever and complex, but also difficult to explain, incompletely theorised, and loses the concept of a clear individual predicted score. Above all, how much practical benefit has been generated thereby? Put another way - what secure knowledge has it generated that would not have been possible anyway? The focus of school effectiveness studies (the main use of MLM so far) has been very technical. Its findings have not always been transformed into anything usable (recall that a considerable amount of the value-added work actually going on in schools is *not* based on these techniques). It can be seen by teachers as being imposed on them from outside, and little is actually done with it in schools once the researchers have left (Saunders 2000).

As statistical techniques evolve over time, generally becoming more sophisticated, they provide a wider choice of modelling strategies and therefore the opportunity for more realistic (i.e. 'lifelike') analyses in social science. The potential downside is twofold. By increasing the number of decisions to be made by the analyst, the newer techniques introduce new sources of bias, and by becoming more complicated they reduce the number of readers prepared to check all of the caveats in the methodology. There is therefore a danger that method 'messiahs' can peddle their own solutions, mistakenly thinking that these can be judged apart from the problems that they are used for (in Snow 2001), and being respected for it simply because their technique is complex. This leads to mono-methodic researchers (and even entire fields), who use one technique like MLM again and again. Presumably the only logical way that this behaviour could be explained is that these researchers deliberately seek out problems suitable for their one technique, and deliberately ignore problems that would require them having to learn a new one. And there are

many examples of 'suboptimal' use of such complex techniques (Maruyama 1998, p.275), some bordering on statistical fantasy.

Some authorities argue that auto-correlation anyway only leads to loss of power (Raudenbush 2002), and this could be righted more simply by increasing the sample size rather than changing our methods of analysis. Perhaps 'we don't need more complex analytic techniques, we need better data collection' (Brighton 2000, p.135).

## References

- Aitkin, M. and Longford, N. (1986) Statistical modelling in school effectiveness studies, *Journal of the Royal Statistical Society Series A*, 149, 3, 1-43
- Bernard, R. (2000) *Social Research Methods: qualitative and quantitative approaches*, London: Sage
- Brighton, M. (2000) Making our measurements count, *Evaluation and Research in Education*, 14, 3&4, 124-135
- Coe, R. and Fitz-Gibbon, C. (1998) School effectiveness research: criticisms and recommendations, *Oxford Review of Education*, 24, 4, 421-438
- de Vaus, D. (2002) *Analyzing social science data: 50 key problems in data analysis*, London: Sage
- Fielding, A. (2000) *Explanatory modelling of complex social structures with case studies in educational research*, presentation at ESRC/BERA Advanced Training Workshop, Birmingham, July 2000
- Fitz-Gibbon, C. (2000) Education: realising the potential, in Davies, H., Nutley, S. and Smith, P. (Eds.) *What works? Evidence-based policy and practice in public services*, Bristol: Policy Press
- Fitz-Gibbon, C. (2001) *Value-added for those in despair: research methods matter*, British Psychological Society
- Gibson, A. and Asthana, S. (1998) Schools, pupils and examination results: contextualising school 'performance', *British Educational Research Journal*, 24, 3, 269-282
- Goldstein, H. (1987) *Multilevel models in educational and social research*, London: Griffin

- Goldstein, H., Huiqi, P., Rath, T. and Hill, N. (2000) *The use of value-added information in judging school performance*, London: Institute of Education
- Goldthorpe, J. (2001) Causation, statistics, and sociology, *European Sociological Review*, 17, 1, 1-20
- Gorard, S. (2000) *Education and Social Justice*, Cardiff: University of Wales Press
- Gorard, S. (2001a) International comparisons of school effectiveness: a second component of the 'crisis account'?, *Comparative Education* , 37, 3, 279-296
- Gorard, S. (2001b) *Quantitative Methods in Educational Research::The role of numbers made easy*, London: Continuum
- Gorard, S. (2002a) What is research capacity building?, *Building Research Capacity*, 1, 2-3
- Gorard, S. (2002b) The role of causal models in education as a social science?, *Evaluation and Research in Education*, (forthcoming)
- Harwell, M. and Gatti, G. (2001) *Review of Educational Research*, 71, 1, 105-131
- Kreft, I. (1996) *Are multi-level techniques necessary? An overview, including simulation studies*, <http://www.calstatela.edu/faculty/ikreft/quarterly/>, Accessed 19/6/02
- Lauder, H., Hughes, D., Watson, S., Waslander, S., Thrupp, M., Strathdee, R., Simiyu, I., Dupuis, A., McGlenn, J. and Hamlin, J. (1999) *Trading in futures: Why markets in education don't work*, Buckingham: Open University Press
- Maruyama, G. (1998) *Basics of structural equation modelling*, London: Sage
- Menard, S. (1995) *Applied logistic regression analysis*, London: Sage
- Miles, J. and Shevlin, M. (2001) *Applying regression and correlation*, London: Sage
- Moses, L. (2001) *A larger role for randomized experiments in educational policy research*, presentation at AERA annual conference, Seattle, April 2001
- Nuttall, D. (1987) The validity of assessments, *European Journal of Psychology of Education*, 11, 2, 109-118
- Paterson, L. and Goldstein, H. (1991) New statistical models for analysing social structures: an introduction to multilevel models, *British Educational Research Journal*, 17, 4, 387-393
- Pedhazur, E. (1982) *Multiple Regression in Behavioural Research*, London: Holt, Rhinehart and Winston
- Plewis, I. (1997) *Statistics in Education*, London: Arnold
- Raedenbush, S. and Bryk, A. (1986) A hierarchical model for studying school effects, *Sociology of Education*, 59, 1-17

- Raudenbush, S. (2002) *New directions in the evaluation of Title I*, presentation at AERA, New Orleans April 2002
- Saunders, L. (200) Understanding schools' use of 'value-added' data: the psychology and sociology of numbers, *Research Papers in Education*, 15, 3, 241-258
- Slee, R., Weiner, G. and Tomlinson, S. (1998) *School Effectiveness for Whom? Challenges to the school effectiveness and school improvement movements*, London: Falmer Press
- Snow, C. (2001) Knowing what we know: children, teachers, researchers, *Educational Researcher*, 30, 7, 3-9
- Torgerson, C. and Torgerson, D. (2001) The need for randomised controlled trials in educational research, *British Journal of Educational Studies*, 49, 3, 316-328
- Travis, G. and Collins, H. (1991) New light on old boys: cognitive and institutional particularism in the peer-review system, *Science, Technology and Human Values*, 16, 3, 322-341