A Comparison of Modelling Strategies for Value-Added Analyses of Educational Data

Neil H. Spencer[a,*] Antony Fielding[b]

[a] *Department of Economics, Statistics and Decision Sciences, University of Hertfordshire, Hertford Campus, Hertford, SG13 8QF.*

[b] *Department of Economics, University of Birmingham, Birmingham, B15 2TT*

Abstract

In this paper we examine modelling strategies for value-added multilevel models and conclude that the approach offered by the BUGS software has advantages over more classical estimation methods.

Keywords: Hierarchical Modelling; Iterative Generalized Least Squares; Gibbs Sampling; Endogeneity.

1. Introduction

Since the mid 1980s, the analysis of hierarchically structured data through estimation of random effects multilevel models has become commonplace. A variety of procedures have been developed and algorithms implemented in readily available software (VARCL (Longford, 1988), HLM (Bryk et al, 1996), Proc Mixed from SAS (SAS Institute, 1992), MLn and MLwiN (Rasbash and Woodhouse, 1995; Goldstein et al, 1998)). The MLn and MLwiN software we use in this paper is developed around the Iterative Generalised Least Squares (IGLS) methods of Goldstein (1995). More recently, Markov Chain Monte Carlo (MCMC) methods of carrying out modelling in a Bayesian framework have increased in popularity, implementing Gibbs and other sampling techniques. Improvements in technology have meant that these computer intensive approaches to modelling have become more accessible to researchers, and the development of the BUGS software package (Spiegelhalter et al, 1995) has further increased interest in such modelling strategies. A good account of the methodology surrounding these procedures is provided by Draper (1998).

In this paper we discuss a multilevel model which is an example of a common situation where classical assumptions of independence of explanatory variables and the random effects no longer holds. The IGLS procedures available in MLwiN will in this case not directly yield consistent estimators. However, adaptations based on instrumental variable (IV) methods to produce a consistent estimation procedure have been suggested by Fielding and Spencer (1997). This is available as a MLwiN macro IV (http://www.bham.ac.uk/economics/staff/tony.htm). Here we discuss a case where the Bayesian modelling strategy through the MCMC methods of the BUGS software may be preferable in many regards. It should be noted, however, that the recently available update of MLn, called MLwiN, now also has the capability to carry out some Bayesian modelling. Further work comparing the BUGS and MLwiN approaches is planned.

In section 2, we present the dataset and model used in this paper, and in sections 3 and 4 we contrast results from the alternative estimation strategies. Section 5 presents results obtained using

[*] Corresponding author.

simulated datasets and section 6 considers different parameterisations of the scale matrices used in BUGS. Conclusions and ideas for further work are presented in section 7.

2. Dataset and model

We consider a hierarchical dataset consisting of data collected on 1946 pupils from 14 educational establishments in England. The response variable considered, y_{ij}, is a standardised points score in General Certificate of Education Advanced Level (A-level) for pupil j in establishment i. This is a measure of academic success over usually two, three or four examinations normally taken when aged approximately 18 years. Explanatory variables are GCSE points score (a measure of academic success over several subject areas in the British "General Certificate of Secondary Education", normally taken when aged approximately 16 years), age and gender for a pupil.

More specifically, the model we consider is

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + u_i + \varepsilon_{ij}$$

where, x_{1ij} is the standardised GCSE points score, x_{2ij} is the pupil's age and x_{3ij} is a binary variable indicating gender. The u_i parameter is a random effect varying between schools, with variance σ_u^2, and ε_{ij} is the usual pupil-level random effect, with variance σ_ε^2.

It is deliberately simple in form as the object of the work presented here is to compare modelling approaches rather than draw substantive conclusions from the data. It also serves to remind the reader that models need not be complex in order to be problematic.

Despite its simplicity, this model is of considerable interest in school effectiveness work. By controlling for the GCSE points score and other relevant factors for a pupil when analysing the A-level points score, the study can examine the "value added" by a school. This is usually evaluated by estimation of the school random effects or "residuals". This form of model is of the same genre as that used by Gray et al (1995) and Goldstein and Thomas (1996).

3. Parameter Estimation

3.1. The problem of endogeneity

Estimates of the parameters of the model shown in section 2 can be easily obtained using standard multilevel modelling packages. However, this approach ignores the fact that the GCSE point score may not be independent of the random effects in the model which such estimation requires for good properties. Indeed this variable may be regarded as endogenous and could itself be modelled using the same data:

$$x_{1ij} = \gamma_0 + \gamma_2 x_{2ij} + \gamma_3 x_{3ij} + v_i + e_{ij}$$

where x_{1ij}, x_{2ij} and x_{3ij} are as above, v_i is a random effect varying between schools and e_{ij} is the pupil- level random effect. The u_i in the model for y_{ij} and the v_i may both be regarded as school effects. It could therefore be argued that we may encounter estimation difficulties due to the correlation between x_{1ij} and the school random effect in the A-level model through the v_i. However, the extent to which this correlation exists is open to question as most of the students did not study for their A-level examinations at the same establishment at which they studied for their GCSE examinations. The v_i could then be regarded as exogenous for the A-level model. For situations where correlations at this level of the data structure may be problematic, Rice et al (1999) have developed strategies known as conditional IGLS.

Of prime concern in this paper is the correlation that exists between x_{1ij} and the random part of the A-level model through the pupil-specific random terms ε_{ij} and e_{ij}. Both these terms contain unexplained and unmeasurable pupil factors and thus cause x_{1ij} to be endogenous for the A-level model.

This endogeneity means that the assumption of independence of regressors and model error cannot be sustained and thus the parameter estimates obtained by a direct MLwiN analysis may be inconsistent. The degree to which this inconsistency affects the estimates is governed by the size of the covariance between the pupil effects at GCSE and A-level. As these covariances are not known at the time of the analysis, some adjustment of the estimation process must be undertaken to ensure that consistent estimates are produced. As the endogeneity is caused by correlations at the lowest level of the data hierarchy, the conditional IGLS methods of Rice et al (1999) are not available.

3.2. Consistent estimation using instrumental variables

A method for obtaining consistent parameter estimates for models such as those considered in this paper is given in Fielding and Spencer (1997). An instrument set is created using the regressors in the A-level model with the endogenous regressor (the GCSE score) being replaced by an instrumental variable. This instrumental variable is obtained by a multilevel modelling of the GCSE score using variables that are assumed to be independent of the school and pupil effects in the A-level model. Here, we use dummy variables that define whether the A-levels taken by the student are all science/mathematical subjects, all non science/mathematical subjects or a mixture. In order to create an instrumental variable that is a good predictor of GCSE score, it would be preferable to have a number of regressors that are relatively unrelated to the A-levels. However, further regressors of this type are not available from the study and this raises the issue of the collection of good relevant data in future studies so that this IV procedure can be carried out more efficiently.

Once the estimation of the parameters of the GCSE model has taken place, estimates of the GCSE score are obtained using just the fixed part of the model (i.e. ignoring any school and pupil effects) and these estimates are used in place of the original GCSE score in the instrument set. There is a danger of introducing correlations between the instrument set and the random part of the A-level model if any of the random effects are used to produce estimates of the GCSE score so we choose here to only use the fixed part of the model. If the regressors used in the GCSE model are truly independent of the school and pupil effects in the A-level model, then the estimate of the GCSE score from the fixed part of the model will be as well. The instrument set is then independent of the disturbance of the A-level model, and consistent estimates of the parameters of the A-level model can be obtained using standard IV techniques.

The Iterative Generalized Least Squares algorithm (used in MLwiN) can then be used to obtain estimates of the random parameters of the A-level model, with the fixed parameters being constrained to those obtained by the IV estimation. These random parameters can then be used to obtain standard errors for the IV estimates.

It could be argued that the dummy variables used to construct the instrument for the GCSE score may not be independent of the school and pupil effects in the A-level model. This problem of selecting appropriate regressors for the modelling of the GCSE score is an essential problem with the instrumental variable method of obtaining consistent parameter estimates. All that can be said regarding this problem here is that the resulting instrument is likely to be .less correlated with the disturbance of the A-level model than the original GCSE score, and the problems of inconsistency are thus likely to be less.

Table 1 shows the results of estimating the parameters of the A- level model using MLwiN with and without the IV method of estimation. It appears from these results that inconsistency of a standard analysis does not present a serious problem for this example. However, in general it is not always the case that reliable estimates are obtained and in section 5 we shall see a case where they are not. As a general rule we may prefer the IV estimators since they have a guarantee of consistency.

Table 1
Parameter estimates

Parameter	Without IV	With IV	BUGS
β_0	0·0236 (0·0551)	0·0097 (0·0667)	0·0363 (0·1241)
β_1	0·5735 (0·0210)	0·5116 (0·0903)	0·5453 (0·2188)
β_2	0·0775 (0·0196)	0·0729 (0·0268)	0·0747 (0·0339)
β_3	−0·0185 (0·0383)	−0·0180 (0·0403)	−0·0158 (0·0500)
σ_ε^2	0·6936 (0·0223)	0·6972 (0·0272)	0·7731 (0·0246)
σ_u^2	0·0261 (0·0132)	0·0236 (0·0122)	0·4833 (0·2038)

Of course, as may be seen from table 1, the standard errors obtained from the IV estimation procedure are larger, but not much larger, than those obtained from straightforward IGLS. The general problem of large standard errors and implied efficiency is well known. However, in many instances, as in these results, acceptable results can be obtained (see Spencer, 1998 for a fuller discussion of this problem).

4. Modelling with BUGS

The advantage of a Bayesian approach to the modelling of the data, as implemented in BUGS, is that the endogeneity of the GCSE points score, described in section 3.1, can be built into the model. To show this, a directed graph can be created (see figure 1). In the graph, rectangles denote constants fixed by the study design (age, gender). Nodes with circles/ovals around them are stochastic variables that are given a distribution and may be data (e.g. GCSE points score, A-level points score) or parameters (e.g. β_0, γ_0). Dashed arrows linking nodes indicate that there is a local function linking the parent nodes (at the tails of the arrows) to the child nodes (at the heads of the arrows). Solid arrows indicate that a stochastic dependence exists. As well as being an alternative way to display the model being used, this graph can be used as an aid to write the model in the BUGS language.

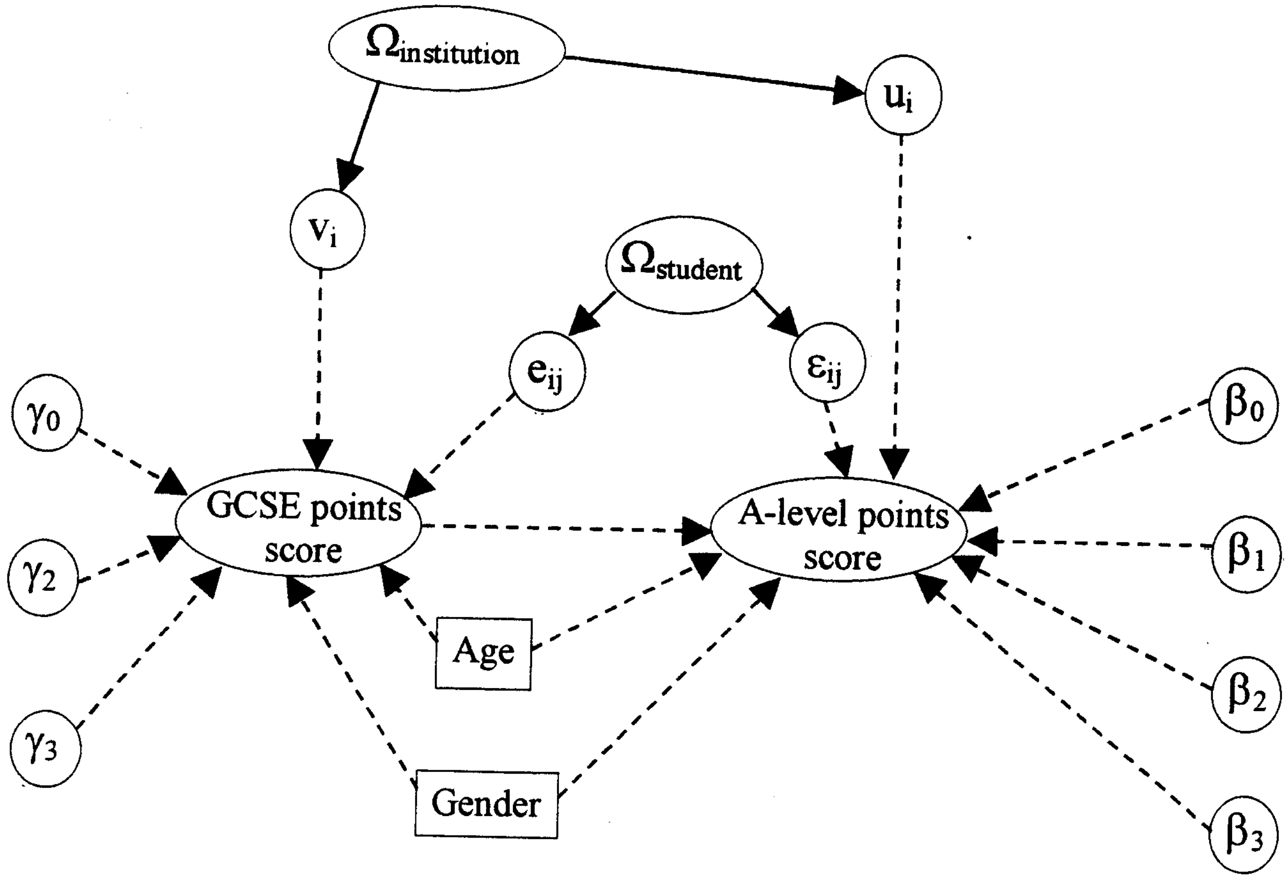

Figure 1. Directed graph

Using the directed graph, we define the A-level score to be dependent on the GCSE score, age and gender (the fixed part of the A-level model in section 2). We also simultaneously define the GCSE score to be dependent on age and gender (the fixed part of the model in section 3.1). Both the A-level score and GCSE score have school effects associated with them. To define these on our directed graph, we have created two effects (u_i associated with the A-level score and v_i associated with the GCSE score) coming from a bivariate normal distribution. To define two pupil effects, we have the A-level score and GCSE score coming from a bivariate normal distribution. It is these bivariate normal distributions (which have non-zero covariances associated with them) which induce endogeneity.

The basic idea behind the MCMC techniques is that instead of carrying out complex calculations to arrive at an exact or approximate estimate of a quantity of interest from the model, series of simulated values are obtained. The idea is that these simulated values are eventually (once the iteration process has converged) will be coming from the distribution appropriate for the quantity of interest. If this is the case then the mean, standard deviation, etc. of the posterior distribution of the quantity of interest can be examined using the simulated values. Gibbs sampling is a technique for obtaining MCMC simulations where, sequentially, each parameter of interest is simulated from its conditional distribution given the most recent updates available of the other parameters. For more details of MCMC and Gibbs sampling techniques, see texts such as Draper (1998), Smith and Roberts (1993).

4.1. Assigning priors

A frequent criticism of Bayesian approaches to statistical analyses is that concerning the need to assume prior distributions for the parameters. In the analysis presented here, we use non-informative priors. For the fixed effect coefficients, we use Normal distributions with means of zero and precisions of 0·00001. For the random parameters, we need to define two bivariate normal distributions. Each of these distributions has a vector of means associated with it and we define these to be zeros. Each distribution also has a 2 dimensional covariance matrix associated with it, and in BUGS this covariance matrix is defined as having a Wishart distribution with an associated 2 dimensional scale matrix. The difficulty comes when having to define this scale matrix and is investigated in section 6, but for now we proceed by defining an arbitrary scale matrix with the numbers 2 on the main diagonal and 1 in the off-diagonal positions. See Spiegelhalter et al (1995) for more discussion on this issue.

4.2. Burn-in iterations and convergence

In order to obtain results from BUGS, the MCMC iterations begin with pre-set starting values and a number of "burn-in" iterations are required until the parameter realisations obtained by the Gibbs sampling are not influenced by these starting values and have converged so that they come from stationary distributions. The number of burn-in iterations required is decided upon by looking at trace plots of the Gibbs sampling realisations and diagnostic statistics such as those proposed by Geweke (1992), Gelman and Rubin (1992), Raftery and Lewis (1992), Heidelberger and Welch (1983). The trace plots (see figure 2 for an example) and diagnostic statistics can be obtained from a BUGS analysis with the help of a suite of menu-driven S-Plus functions known collectively as CODA (Best et al., 1995) that accompanies the BUGS software.

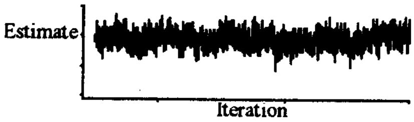

Figure 2. Trace plot

4.3. BUGS results

Table 1 shows the results obtained from a BUGS analysis of the data described in section 2 with a burn-in of 5000 iterations and 5000 monitored iterations. The results obtained again indicate that the estimates produced by the initial MLwiN analysis are acceptable. The BUGS estimates of the fixed effects are similar to those produced by MLwiN, but have higher standard errors. The estimates of the variances of the random effects are larger than those produced by MLwiN, particularly for the school level random effect. This may be in part due to difficulties in obtaining an estimate of a variance which is near the lower limit of zero.

In this example data it appears that all three methods used give comparable results. We have already commented in section 3.2 on general reasons why we might prefer IV methods to standard IGLS. However, given that MCMC modelling is more complicated to implement, it appears to convey few advantages for the case discussed, particularly in view of the lower precision of the results. In section 5 we shall see, however, that the MCMC approach can have advantages when the endogeneity issue has more impact on the standard IGLS estimation procedure.

5. Simulated data

In this section, we explore the issue further by means of simulated datasets, each with a similar structure to that introduced in section 2. The parameter values that the simulations are based on come from the BUGS analysis discussed in section 4.3: β_0 = 0·0363, β_1 = 0·5453, β_2 = 0·0747, β_3 = −0·0158, σ^2_ε = 0·7731, σ^2_u = 0·4833, γ_0 = −1·1419, γ_2 = 0·1346, γ_3 = −0·1221, σ^2_e = 0·9533, σ^2_v = 0·2252, $\sigma_{\varepsilon e}$ = 0·4462, σ_{uv} = 0·0689, where $\sigma_{\varepsilon e}$ is the covariance between ε_{ij} and e_{ij} and σ_{uv} is the covariance between u_i and v_i.

Table 2 shows the mean estimate for each parameter over the fifty datasets obtained using each estimation method, together with the associated empirical standard deviation.

Table 2
Mean parameter estimates from simulations

	Target	MLwiN		
Parameter	Values	Without IV	With IV	BUGS
β_0	0·0363	0·1210 (0·2066)	0·2250 (1·159)	0·0408 (0·2402)
β_1	0·5453	1·010 (0·0242)	0·4628 (9·592)	0·6682 (0·4770)
β_2	0·0747	0·1245 (0·0287)	0·0045 (1·422)	0·0891 (0·0631)
β_3	−0·0158	−0·0723 (0·0531)	0·2153 (2·077)	−0·0141 (0·1251)
σ^2_ε	0·7731	0·4669 (0·1753)	92·31 (403·4)	0·9640 (0·0512)
σ^2_u	0·4833	0·5571 (0·0290)	20·16 (84·33)	0·4146 (0·0717)

From table 2, we see that we have a situation where the MLwiN results without the instrumental variable procedure appear to be suffering from problems of inconsistency. The mean estimate of β_1 is well over 2 empirical standard deviations from the target value and it is this pattern of an inflated estimate for the coefficient of the endogenous variable that is typical when no adjustment for the endogeneity is made. This is because the positive correlation between the GCSE score and the random part of the A-level model causes a positive bias in the estimate of the coefficient.

When the instrumental variables procedure is employed, it is apparent that although we may well be obtaining consistent estimates, the large standard errors mean that the usefulness of the results is limited. The results obtained by BUGS are more useful with lower standard errors being produced and the mean parameter estimates being relatively near the target values. The only difficulty apparently being experienced is with the estimate of σ^2_e which is slightly higher than we would wish, with a relatively small standard error.

6. Alternative scale matrices

In section 4.1, we stated that we were able to use non-informative priors for the BUGS analysis, but still had to define scale matrices associated with the covariance matrices of the school and pupil effects. In this section, we investigate the sensitivity of the parameter estimates to changes in the scale matrices.

The first parameterisation of the scale matrix considered here has the number 1 on the main diagonal and 0·5 in the off-diagonal positions. Scale parameterisation two is the identity matrix. Scale parameterisation three has 0·7731 and 0·9533 on the main diagonal and 0·4462 in the off-diagonal positions for the pupil effects (the A-level and GCSE pupil level variances that the simulations of section 5 are based on) and for the school effects has 0·4833.and 0·2252 on the main diagonal and 0·0689 in the off-diagonal positions (the A-level and GCSE school level variances that the simulations are based on). Scale parameterisation four has 100 on the main diagonal and 10 in the off-diagonal positions for both the pupil and school effects.

Table 3 gives the results from analysing the 50 simulated datasets used in section 5. As can be seen, the mean estimates are quite stable for parameterisations 1, 2 and 3 and that used in table 2. This is a good sign in that it appears that the results are fairly robust to the choice of scale matrices. Parameterisation 4, with the extreme scale matrices, produces worse results. As in table 2, the mean estimates for the variances are less satisfactory than those for the fixed effect coefficients.

Table 3
Results from simulations with different scale parameterisations

	Target	Scale Parameterisation			
Parameter	Values	1	2	3	4
β_0	0·0363	0·0427	0·0395	0·0423	−1·426
		(0·2413)	(0·2414)	(0·2423)	(0·6555)
β_1	0·5453	0·6673	0·6512	0·6584	−1·954
		(0·4884)	(0·4856)	(0·4968)	(1·420)
β_2	0·0747	0·0893	0·0872	0·0884	−0·2168
		(0·0648)	(0·0645)	(0·0661)	(0·2038)
β_3	−0·0158	−0·0141	−0·0117	−0·0128	0·4349
		(0·1268)	(0·1270)	(0·1284)	(0·3942)
σ^2_ε	0·7731	0·9626	0·9626	0·9627	1·108
		(0·0512)	(0·0512)	(0·0512)	(0·0513)
σ^2_u	0·4833	0·3297	0·3304	0·2652	8·784
		(0·0711)	(0·0710)	(0·0707)	(0·5490)

7. Conclusions and further work

In this paper, we have shown that a BUGS analysis which has the endogeneity of the GCSE score built into the model structure can produce results which are clearly preferable to a classical analysis which ignores the endogeneity problem. The BUGS analysis can also be preferable to an analysis using instrumental variable methods which may produce estimates with unacceptably large standard errors. Also with the endogeneity being properly modelled, assumptions do not have to be made regarding the independence of the instrumental variable and the disturbance in the A-level model. It has been shown that non-informative priors can be used and that the results appear to be robust to different choices of scale matrices, providing these choices are reasonable.

More work is needed to identify those occasions, such as demonstrated in table 1, where a classical analysis can produce acceptable results despite endogeneity concerns. Additionally, further research is needed regarding the BUGS estimates of the variances in the model and how they can be improved. Nevertheless, despite these areas of research that need more investigation, it is clear from this paper that the BUGS approach to modelling can give distinct advantages over more classical approaches.

8. References

Best, N.G., Cowles, M.K. and Vines, S.K., *CODA Manual, Version 0.30* (MRC Biostatistics Unit: Cambridge, 1995).

Bryk, A.S., Raudenbush, S.W. and Congdon, R.T., *HLM: Hierarchical Linear Modelling with the HLM/2L and HLM/3L Programs* (Scientific Software International: Chicago, 1996).

Draper, D., *Bayesian Hierarchical Modeling* (http://www.bath.ac.uk/~masdd/, 1998)

Fielding, A. and Spencer, N.H., Modelling cost-effectiveness in general certificate of education A level: Some practical methodological innovations, in: Minder, C.E. and Friedl, H. (eds.), *Good Statistical Practice: Proceedings of the 12th International Workshop on Statistical Modelling, Biel/Bienne, July 7 to 11, 1997* (Schriftenreihe der Ősterreichischen Statistischen Gesellschaft: Vienna, 1997).

Gelman, A. and Rubin, D.B., Inference from iterative simulation using multiple sequences, *Statistical Science* 7 (1992) 457-472.

Geweke, J., Evaluating the accuracy of sampling-based approaches to calculating posterior moments, in: Bernardo, J.O., Berger, A.P., Dawid, A.P. and Smith, A.F.M. (eds.), *Bayesian Statistics 4* (Clarendon Press: Oxford, 1992)

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

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G. and Healy, M., *A User's Guide to MLwiN* (Institute of Education: University of London: London, 1998).

Goldstein, H. and Thomas, S., Using examination results as indicators of school and college performance, *Journal of the Royal Statistical Society, Series A* **142** (1996) 407-442.

Gray, J., Jesson, D., Goldstein, H., Hedger, K. and Rasbash, J., A multi-level analysis of school improvement: changes in schools' performance over time, *School Effectiveness and School Improvement* **6** (1995) 97-114.

Heidelberger, P. and Welch, P., Simulation run length control in the presence of an initial transient, *Operations Research* **31** (1983) 1109-1144.

Longford, N.T., *VARCL -- software for variance component analysis of data with hierarchically nested random effects (maximum likelihood)* (Education Testing Service: Princeton, NJ, 1988).

Raftery, A.L. and Lewis, S., How many iterations in the Gibbs sampler?, in: Bernardo, J.O., Berger, A.P., Dawid, A.P. and Smith, A.F.M. (eds.), *Bayesian Statistics 4* (Clarendon Press: Oxford, 1992)

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

Rice, N., Jones, A. and Goldstein, H., Multilevel models where the random effects are correlated with the fixed predictors: a conditioned iterative generalised least squares estimator (CIGLS) *Health Economics* (1999) to appear.

SAS Institute, *SAS Technical Report P-229, SAS/STAT Software: Changes and Enhancements* (SAS Institute, Inc.: Cary, NC, 1992).

Smith, A.F.M. and Roberts, G.O., Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods, *Journal of the Royal Statistical Society, Series B* **55** (1993) 3-23.

Spiegelhalter, D.J., Thomas, A., Best, N.G. and Gilks, W.R., *BUGS: Bayesian inference Using Gibbs Sampling, Version 0.50* (MRC Biostatistics Unit: Cambridge, 1995).

Spencer, N.H., Consistent parameter estimation for lagged multilevel models, *University of Hertfordshire Business School Statistics Technical Report 1, UHBS 1998:19*, 1998.