TEACHING GROUPS AS FOCI FOR EVALUATING PERFORMANCE IN COST EFFECTUVENESS OF GCE ADVANCED LEVEL PROVISION: SOME PRACTICAL METHODOLOGICAL INNOVATIONS[1]

Antony Fielding[2]

University of Birmingham, United Kingdom

Abstract: The use of hierarchical models of educational effectiveness are reviewed. The paper also draws attention to the inadequacy of official performance indicators currently available and in particular the inattention to resource issues. Research on this issue using teaching group cost and performance has led the present author in collaboration with others to focus on cost-effectiveness issues. These issues are discussed in the paper and draw out some new methodological issues. The concern is with evaluating the performance of students on GCE Advanced Level courses at the level of the subject group with the intention of relating this to identified teaching group costs.

Since students belong to several groups standard hierarchical modelling which assumes responses at the individual level are independent are precluded. Thus models are adopted which cross classify students and groups and investigate the issue using the procedures of Rasbash and Goldstein (1994). The data arises from 6020 subject entries for 2280 candidates in 525 teaching groups from 14 institutions. The impact of ignoring the crossing on substantive results is examined. A further issue in cost-effectiveness work is differential teacher effectiveness since salaries are the largest component of costs. Hill and Rowe (1996) have attributed part of the reason for sparsity of work on teacher effectiveness to the fact that classes are exposed to several teachers. This paper approaches the issue through three way

[1] Some parts of this paper were presented at the 10th International Congress on School Effectiveness and Improvement , University of Manchester Institute of Science and Technology, and at the Royal Statistical Society Birmingham Local Group , both in January 1998. Continuation of work has been supported under the Analysis of Large and Complex Datasets programme by the ESRC through award H51944500497 whilst the author was a Visiting Research Fellow at the Multilevel Models Project , University of London Institute of Education. Acknowledgements are due to Michael Healy and David Reynolds for helpful comments on earlier drafts.

cross-classified models with random teacher effects weighted by proportions of time the class was taught by particular teachers. Teachers also cover several groups, but although the data is extremely unbalanced it proves possible to disentangle teacher variability from that associated with subject groups. Teacher variation is shown to be considerable. Traditional explanations in terms of age , experience, training, and level of qualifications are indicated as being not important in explaining this variation.

The papers conclude that teachers vary in their impact on educational progress but concrete ascribable reasons for this remain to be uncovered.

Keywords: Advanced Level, Cross- Classifed Models, GCE. GCSE, Multlevel Models, Multilple Membership Models, Teacher Effects, Disentangling Effects

[2] Contact e-mail: A.Fielding@bham.ac.uk

1 INTRODUCTION

1.1 Appropriate Models And Levels For Educational Outcomes

The methodology and substantive issues discussed in this paper cut across the concerns of school effectiveness and school improvement research and the official concern in recent years with publication of performance indicators of various sorts in the public sector of the United Kingdom. Both should require sound measurement and appropriate modelling of educational outcomes. From the indicator perspective, of particular relevance to this paper are annual institutional league tables in England and Wales of average points scores of students entering for two or more General Certificate of Education Advanced (A) Levels. These are normally taken by students after two years full time study , most often between ages 16 and 18. Entry to higher education is governed for the most part by performance on these A levels or equivalent qualifications. This factor probably heightens the interest in the raw performance tables and their wide dissemination in the media and press. Unfortunately such indicators are often perceived without caveat as indicators of effectiveness of institutions. The flaws in this perception have been widely discussed by researchers in educational effectiveness, statisticians and others .The main foci of criticism of 'league table' construction surround lack of adjustment for intake ability and other relevant background factors and inattention to the hierarchical structure of the process and data.(e.g. Gray and Hannon (1986), Willms (1992), Nuttall et al. (1989) Goldstein and Thomas (1996), Goldstein and Spiegelhalter (1996)). From the perspective of practical research these foci also parallel the criteria expounded by Goldstein (1997) and Gray et al. (1995) as requiring attention at a minimum for sound study of educational effectiveness. In the official sphere official attempts to address these issues were expounded in a government report (Department of Education and Science (1995)). However, several years later these have not yet permeated to routine and transparent publication.[3]

[3] November 1998 was to have seen the first official attempts at publication of value-added indicators but for a stated reason of concern with methodology these were withdrawn at the last minute. Cynical observers suggested it may have had more to do with strong pressure groups whose value added positions appeared less secure than their raw outcomes.

Both critical foci have led to developments in analysing effectiveness using data hierarchically structured to reflect educational processes and multilevel modelling of effects. Although work on post-compulsory education , in which A levels are placed , is less developed than in some other area, , a multilevel approach is rapidly becoming the norm in much research on effectiveness at various educational stages. The pioneering work of Aitkin and Longford (1986) focused on variance components at levels in an educational hierarchy. This was the first attempt to disentangle hierarchical effects: separating differences between institutions from the inherent variation amongst students within them. There has subsequently been a growing literature on the multilevel approach (e.g. Bryk and Raudenbush (1987), Goldstein (1995), Nuttall et al (1989), and Willms (1992)) This recognises the hierarchical nature of many educational processes such as students within classes within institutions. By modelling the variability in responses through random regression coefficients it becomes possible to understand how various effects operate at various levels.

A clear statement of many of the problems surrounding published indicators, including those mentioned above, and attempts to measure effectiveness is given by Goldstein and Spiegelhalter (1996). They also draw attention to the inherent uncertainty in any such exercise , however sophisticated in concept it may be. There has also been some controversy surrounding the ease of interpretability of results of multilevel models for those directly involved in educational practice and administration. For instance , the general conclusion of the National Value Added Indicators Project (Fitz-Gibbon (1997, 1978)) seems to be that for such purposes non-hierarchical analyses are preferable. The latter project report does, however, propose that the more sophisticated modelling may be necessary for research in educational effectiveness and improvement. Thus the controversy about appropriate methods for different contexts will inevitably continue.

Good descriptions of the theory and appropriateness of multilevel random effects models and their application to hierarchical data structures such as we find in education are Goldstein (1995), Bryk and Raudenbush (1992), and Longford (1993). A variety of multilevel estimation procedures have been suggested in this literature. The Iterative Generalised Least Squares (ILGS) methodology discussed by Goldstein (1995) and implemented in the MlwiN software (Rasbash et al (1999) is adapted for analyses presented in the present paper. Not the least attractive feature of this software is its flexible macro language which facilitates the fitting of models with a wide variety of non- standard assumptions.

1.2 Motivation: Relating Resources to Outcomes At Intermediate Levels

In work on educational effectiveness in which the methods and results of this paper form a part, we have investigated a further much neglected factor which may put the debates about educational effectiveness into context; that of resource availability, use , and allocation. There seems to be an increasing official concern with cost-effectiveness as a way of appraising public policy , and issues surrounding such concepts as efficiency' and `value for money' are all pervasive in official literature The latter criterion is also explicitly part of the remit in school and college inspections undertaken in inspections in England and Wales undertaken by the Office for Standards in Education (OFSTED). Yet this concern does not appear to have emerged as a topic of full public transparency. In earlier work Thomas (1990) and Fielding (1995.1998) have drawn attention to how assessment of cost-effectiveness, apart from issues raised in the previous section , may place any performance judgements into proper perspective. An example is the interest in institutional type differences. Judging from raw league tables there is an apparent superiority of Advanced Level provision in school sixth forms contrasted with institutions concentrating on post 16 provision. In cost- effectiveness terms this earlier work illustrates some evidence that the reverse may be the case. This evidence is even sharper when attention is given to modelling the hierarchical structures with adequate control for input.

There is a strong strand of interest in the educational economics literature on relating outcomes and effectiveness mainly through production function approaches. In reviewing the scores of such investigations , mainly American, Hanushek (1986) has stated that 'there is no systematic relationship between school expenditures and student performances'. This is a sweeping statement if it provides support to resistance to resources being made available for improvement initiatives with the aim of increasing effectiveness. Of course, there have been many explanations and implied criticisms of many studies which lead to this overarching conclusion. Amongst them are attacks on production function approaches , the technical assumptions implied by many particular models used, and very importantly the aggregate levels at which analyses have been conducted.. Many of the studies are conducted at levels above that of even institutions , in some American cases at the state wide level only. Critical discussions of these studies in the economics literature mirror the general notions that have emerged in that of educational effectiveness and improvement: lack of attention to the structure of hierarchical effects and lack of adequate input adjustment. Examples of recent work which imply that Hanushek's conclusion may not be so simply reached are provided by, for example, Card and Krueger (1997) and Figlio (1999).

We have briefly discussed this resource and outcome controversy for several reasons . Firstly , it has obvious relevance to assessments of how far schools can be expected to improve their effectiveness when , inter alia , they are subject to resource constraints. Secondly in the new paradigm of merging the traditions of school effectiveness and school improvement concerns, the availability of resources for improvement is one of a central set of issues. Lastly it puts into focus our particular current interest which is the motivation for a discussion of the problems addressed in the present paper : variations in teaching group cost effectiveness within institutions. We have seen that types of difference between institutions are not so clear cut when cost-effectiveness is the focus. However , we are also becoming aware that variations in both costs and effectiveness and by definition cost-effectiveness. within institutions are, perhaps, an issue of as much direct concern as that of institutional variation. There has 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 '. They then develop the motivation for this switch of focus. Young (1998) expresses similar views and illustrates with the work of Fraser & Tobin (1989) who stress the importance of the 'classroom learning environment'. In a comment on some directions in research Coe and Fitz-gibbbon (1998) 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 A level provision we have found wide disparities between resource provision between our equivalent of classrooms , subject teaching groups (Belfield et al. (1996a.b). The interactions between effectiveness and resource constraints at the level of a class are thus a difficult area of direct concern.

In the present paper we will not develop these arguments any further and for these reasons we will not pursue the detailed analyses of resources and costs (see Belfield et al (1996b) for a discussion of these) . However, we have discussed them as motivation for the paper's central concern. To exploit cost-effectiveness analyses at the teaching group level a prior requirement is that we have measures , suitably adjusted , of A level outcomes at the teaching group level. The latter is a component of our work we pursue further here. The complexities of the structure of A level provision mean this may not be straightforwardly sought through traditional residuals from fitting well formulated *standard* hierarchical multilevel models. We discuss this structure, associated data, and the methodological problems it poses in achieving our ends in sections to follow. We hope that the methods we propose and contrast will have wider relevance in other areas of educational effectiveness research. There, we believe, similar structures and methodological problems may be quite common.

2 DISTENTANGLING EDUCATIONAL EFFECTS

We now briefly discuss generically the problems connected to disentangling or unpicking the nature of the sources of effects operating at various levels in complex educational processes and with which educational effectiveness research is naturally concerned. It is hoped that this will put our proposed solutions to the problems faced by our specific context into further relevant perspective.

As mentioned it is becoming recognised that variations in outcomes and progress between discrete classes within institutions is becoming a central concern in the literature, because often it is viewed as more influential than variation in effectiveness between institutions. In studies of this phenomenon the conclusions are usually reached on the basis of apportionment of residual variation from (sometimes with fairly complex fitted factors) multilevel models. The evidence does not always point inexorably in this direction but where it does not rational explanations may be found. Luyten and de Jong (1998) conclude , for instance , that in some secondary school studies which portray institutional effects as more important, prior achievement is not adequately controlled. The latter article also contains a review of the many studies of this phenomenon. The general consensus that within institution effects are more important has now been well substantiated. In disentangling the effects operating at different levels in the hierarchies standard multilevel models have played a very important role.

However, important questions are also begged. What is going on beneath the institutional level and what are the nature of effects within institutions becomes of obvious interest. The problem of unpicking 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 and say that there may be many unmeasured factors concerned with the classroom environment having an effect and of course , relevant for our broader purposes, resources. Broad d differences in performance, suitably adjusted, in different subjects at Advanced level is well documented and may have an effect (e.g. Fitz-Gibbon (1996)). This is pertinent for the situation we will be considering 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 student, teacher and group effects. The not entirely unequivocal way in which teachers and classes have sometimes been treated synonymously has also received attention by Luyten and de Jong (1998). They use a quasi -experimental design involving 'parallel classes' as their solution. The fact that pupils or students are an additional level within classes does not always yield satisfactory analytical approaches through strict hierarchical modelling. The pupil group is unique for each class but it is often argued that satisfactory control for this cannot always be exercised by taken cognisance of all relevant pupil characteristics. As students cannot be treated as randomly allocated to teaching groups (classes) in situations such as ours there may yet be important but unmeasured sources of student variation which are associated with selection into teaching groups. In models these influences captured by student level disturbance terms are often assumed to operate unsystematically. Yet if they are systematically related to group differences, group effects may be partly confounded with these. A within group random student disturbance caters for unmeasured variation but may not adequately cater for these systematic effects. The disentangling of many of these possible confounding factors is an important set of problems and one which we hope to go some way towards addressing in this paper.

We firstly discuss the structure of the data we will analyse in exemplification and how its complexity reflects that of the process it was gathered in. We will then elaborate in a more technical way specific methodological problems in the general framework of those raised above with, in addition, a few newer ones. We propose some solutions which arise out of our process of A level provision and the data. We apply some of these and in our results also contrast with results which might have ignored the problems. Our thrust is methodological but we hope not to lose sight of the substantive import in the study of effects in A level provision . We are also optimistic that there will be broader relevance to other areas of research..

3 THE DATA, STRUCTURE, AND METHODOLOGICAL PROBLEMS

Many of the methodological problems we have discussed in general terms and the substantive issues to which they relate can only be addressed with attention to investigation design and good data. Often routinely collected data are only available and these have severe limitations. The data we introduce is noted by its richness at levels below that of institutions and is indicative of the detail in data requirements, that may be required. Thus our methodological discussion and substantive conclusions are also a caveat. Solutions to some issues are only useful if adequate data is available, and thus attention to this is a pre-requisite. In our early investigations of cost-effectiveness with this data (Belfield et al (1996a,b)), we contented ourselves with hierarchical formulations of students within groups within institutions. However, we became aware of some of the difficulties raised above which prompted further investigation of the specific questions raised in this paper. However, the preliminary methodological problems also raised and opportunity. Due to our focus on teaching group provision we had some detailed data available to match the proposed solutions.

3.1 The Illustrative Data

The data which we will use arise from all students enrolled for GCE Advanced (A) Level in 1993 , and entered in 1995, from nine colleges in the further education sector and seven public sector 11-18 secondary schools. In addition to data on specific A level subjects entered and grade (if any) awarded , a limited range of background factors was available on students taking them. The most important of these were ability of students at intake as measured by the number and average points score on their previous General Certificate of Secondary Education (GCSE). The latter information was unavailable for two colleges and these were excluded from later analyses. Thus we had 6020 A level entries distributed across 525 teaching groups. The latter were nested within the fourteen institutions. These teaching groups are characterised by the subject of the A level although in some institutions there are several groups for specific subjects. Although entries within groups within colleges form an hierarchy the entries must not be made synonymous with students. Each student can take up to five A level subjects and their entries are distributed across the teaching groups within their institution. Thus although students are nested within colleges they are not uniquely nested within single reaching groups. Entries which are nested within groups also have a separate nesting within students. In the data we analyse there are 2280 students with 6020 entries, an average of 2,64 per student. As part of the process of allocating real costs to teaching groups we had detailed timetable information on which teachers taught which groups, for how long and when. Thus a detailed event history of teachers teaching a group is available. Unique teacher group combinations with a single teacher undertaking the entire provision during course instruction was a rarity. Several teachers were responsible for particular groups in a variety of combinations and each teacher was usually involved in several groups. In addition to this information for the teachers involved in the study data is available on a range of their background characteristics: gender, age ,experience , level of training , qualifications, and salary levels. Substantive generalisations about institutional differences such as type effect are limited with only fifteen. However, our larger number of teaching groups , students, and teachers provides an adequate basis for some conclusions at these levels . It also indicates data needs for potential lines of enquiry into deeper institutional effectiveness

3.2 Methodological Issues

We have catalogued in a general way in earlier sections the problem of disentangling possibly confounded effects operating at various levels of the educational process. Within this framework a number of methodological problems emerged as pertinent to our data and our ultimate desire to produce adjusted performance indicators for teaching groups and to assess the role of teachers in these. In this paper we discuss just two of these and the modelling procedures we suggest to handle them.

The first problem is that responses are related across subject groups in that each candidate appears in several groups within an institution. The entries with certain groups will not be independent of entries in other groups which happen to come from the same students. This precludes our modelling of simple hierarchies in ways familiar in the literature since an assumption is that within group disturbances are independent of each other. In a multilevel model we can control for important characteristics of the entries in terms of students taking . them. However, unmeasured and uncontrolled influences remain. Incorporated in a within group disturbance term it is usual in multilevel models to assume that these operate randomly and are not systematically related to other effects. However influences on particular students will now manifest themselves in several groups and may confound with group effects. Some group differences may be due to uncontrolled influences of common sets of students selected into them. The usual aim is to use group residuals from a multilevel model to estimate group effects having adjusted for control characteristics of observations within them. This is now precluded due to the common uncontrolled influences which make entry observations across groups dependent o each other. We propose a cross-classified model to handle this, full details of which we will discuss below.

A second problem surrounds the study of teacher effects which are of obvious interest in relation to cost -effectiveness studies in that differences between teachers and their characteristics may influence both costs and performance. Teacher salaries are in any case the major source of institutional costs. Hill and Rowe (1996) have attributed part of the reason for the relative sparsity of work on teacher effects to the fact that classes are exposed to several teachers in a split-plot way. This is evident in our data but since teachers are involved in several groups we have an opportunity of seeking ways of disentangling teacher and other group effects. If teachers were uniquely involved in only one group then effects would be entirely confounded. The situation we are in has something in common with the one discussed by Hill and Goldstein (1998) in devising models to take account of different amounts of time spent by the same student in different schools. We outline a preliminary approach to the study of teacher effects by attaching weights to them according to their contribution to groups within the framework of cross-classified hierarchical models.

4 MODELS AND METHODS

4.1 Cross-Classified Models of Student With Teaching Group.

In models and analyses to follow our response variable is based on the A level grade of an entry scored according to the Universities Common Admissions Service (UCAS) tariff.(A=10, B= 8 , C=6, D=4, E=2, Fail =0). The outcome variable y used is a transformation of this to a normalised form so that overall observed y follows approximately a standard normal distribution. This has been found a useful device in adapting to standard multilevel model assumptions (Goldstein (1995), Yang and Woodhouse (2000)).

A conventional basic three level model of progress or adjusted performance widely adopted in the literature would be

$$\mathbf{y_{ijk} = (X\beta)_{ijk} + v_k + u_{jk} + e_{ijk}} \qquad (1)$$

Here y_{ijk} would be the normalised A level score for the i^{th} entry in the j^{th} teaching group within institution k. This is modelled by a set of control and explanatory variables in the data matrix **X** with effect parameters in the vector ***b*** . Of particular importance in X would be the intake ability measure. We would also have variance components for the three level effects v_k, u_{jk}, e_{ijk}. More complex variants of this are also possible (Goldstein (1995))

The first problem above arises because the entry A level responses are nested within both students and teaching groups. The assumption of independence of entry disturbances in the model above is an assumption of efficient estimation through the iterative generalised least squares procedures available to us in the software MlwiN (Rasbash et al (1999)) and other estimation methods. It is also often desired, as in our work, to estimate group residuals and we mat expect these to be contaminated by the same departure from assumptions. Correlation across entry observations is induced by the student influence. If prior information were available it might be possible to model this covariation directly but no simple characterisation suggests itself . We can note that the two separate nestings of entries cut across each other A multilevel model with entries within students such as we first entertained would confound partly group and student effects. However, we can consider a cross classification of all groups and students and this provides a way of separating group and student residual effects. Entries are now lodged within a particular combination of student and group within an institution. These combinations form cells which can now be level 2 units above entries. For a particular institution most of these cells will be empty and there is at most one entry observation per cell. Nevertheless methods have been recently developed which make possible the analysis of such sparse structures, which Raudenbush (1993) describes as 'radically unbalanced'.. Formally the type of model we entertain is

$$\boldsymbol{y_{i(j_1j_2)k} = (X\,b\,)_{i(j_1j_2)k} + v_k + u^1_{j_1k} + u^2_{j_2k} + e_{i(j_1j_2)k}} \qquad (2)$$

We thus have a three level model with cells of the student- group cross classification at level 2. within institutions at level 3 The index j_1 ranges over students and j_2 ranges over teaching groups with a particular cell denoted by (j_1j_2). The level 2 residual effect is now composed of separate additive random contributions, $u^1_{j_1k}$ and $u^2_{j_2k}$ from the student and group within institutions. These are assumed independent each normally distributed around zero with variances s^2_1 and s^2_2 respectively. The residual level 1 disturbances within cells $e_{i(j_1j_2)k}$, now free of induced correlations, are also assumed normally distributed with variance s^2_e. With this refined structure, effect and variance specification it now becomes possible to unpick the group effect and variance net of the student influences. It will be of interest to see what substantive differences to assessment of group effects this refined specification makes over the (inappropriate) hierarchical formulation.

Rasbash and Goldstein (1994) and Raudenbush (1993)) have shown how such cross-classified models may be reformulated in ingenious ways to conform to hierarchical models which can then be analysed using available methods. Details of this reformulation and how to set up the model for estimation by MlwiN are given in the User guide (Rasbash et al 1999). Fielding and Yang (1999) gives a detailed rationale for the setup instructions suitable for the type of application discussed here.. The possibilities inherent in such models are discussed in Goldstein (1995). There may be computational constraints if the number of units in either cross –classification is very large but it has proved feasible to model our data in this way

:

4.2 Models With Weighted Random Effects For Teachers.

Our second problem led to a desire to investigate ways of extending and adapting multilevel modelling procedures to disentangle effects of teachers from other group effects. The characteristics of teachers will be of importance in any study of group cost -effectiveness and our work in this area is ongoing . Here we seek to exemplify an approach to isolating teacher effects which we will further relate to teacher variables. If teaching groups had single teachers throughout their course and teachers taught more than one group modelling of teacher variation could in principle be approached by extensions of models proposed in the previous section by three way classifications at level 2. The implementation in MlwiN given that we are piecing together three sets of data relating to students, groups and teachers would require ingenuity but perhaps the main limitations of the approach would be computing considerations. The design matrix would also be very sparse and extremely unbalanced. However , the new problem is that several teachers contribute to each response. The entries formally belong to sever al teachers in a multiple membership structure (Hill and Goldstein (1998)). We handle this structure by a model which is something like a three way classification of level 2. However, instead of a single unique random teacher effect for each entry observation the teacher effect is a weighted combination of the contribution of the effects of the teachers involved. We propose that the weights be chosen in accordance with the proportion of class hours taught by the teachers of a group throughout the time of its provision. Other weighting schemes are possible but experimentation has shown that main parameter estimates are relatively insensitive to choice of weights, although the latter may effect the precision (standard error) with which they are estimated. The application of this approach within the framework of the MlwiN software is not straightforward. However, with ingenuity the MlwiN command reference is sufficiently flexible to allow setting up and estimation of the models. Fielding and Yang (1999) discusses the detailed model set up specifications required for the type of applications discussed here. Theory and some differently structured examples of weighted random effects models are discussed by Goldstein (1995).

Formally three level model is of the general form

$$y_{i(j_1j_2)k} = (Xb)_{i(j_1j_2)k} + u^1_{j_1k} + u^2_{j_2k} + \sum_1^{J_{3k}} w_{i(j_1j_2)j_3k} u^3_{j_3k} + e_{i(j_1j_2)} \qquad (3)$$

With similar interpretations to the specifications of model (2) the new feature is the introduction of weighted teachers' effect $\sum_1^{J_{3k}} w_{i(j_1j_2)j_3k} u^3_{j_3k}$. The random effect of teacher j_3 within institution **k** is $u^3_{j_3k}$. Observations are still lodged within the cell (j_1 j_2) of the students and groups but a separate disentangled effect due to teachers involved now makes a contribution. For a particular entry observation the sum of the weights (which are proportions of teaching group time taught) range over all $\mathbf{J_{3k}}$ teachers involved in institution k and . $\sum_1^{J_{3k}} w_{i(j_1j_2)j_3k} = 1$. However for each of these observations most of the weights will be zero. It may be noted that if the variance of $u^3_{j_3k}$ is denoted by s^2_3 then the contribution of teachers to the variance of an observation is $\sum_1^{J_{3k}} w^2_{i(j_1j_2)j_3k} s^2_3$. This is not as usual constant across observations and depends on the number of teachers in the group and the weight attached to them. A group with a single teacher has a teacher contribution to the variance of s_3^2 but if there are three equally weighted teachers, say , this is $s_3^2 / 9$. Averages of random variables always have smaller variance than random variables themselves. This fact should be carefully considered in interpreting results on teacher variances.

As is usual we may want to extend models to incorporate explanatory factors which may influence any observed effect variation. For the models considered student and group control factors are included in the data matrix **X** We may wish to incorporate teacher variables also. This is problematic for the present structure since there is no unique value of a teacher variable that can be associated with an observation. However, in the same way that we weighted teacher random effects in the above , we can consider weighting fixed effect teacher variables. Thus if we have observations on the l^{th} relevant teacher variable $Y_{\ell j_3 k}$ for teacher j3 in institution k we form new weighted variables $\sum_{j_3=1}^{J_{3k}} w_{i(j_1 j_2) j_3 k} Y_{l j_3 k}$ to be applied to the entry observations and these are incorporated in the data matrix **X.** We have introduced certain weighted teacher variables in this way in the illustrative examples to follow.

5 EXAMPLE RESULTS

5.1 The Student- Group Cross-Classified Model

For simplicity we take a simple ‘value added’ or adjusted performance form of model (2) in which the only fixed effect control covariates in X are the main individual student level intake variable, average GCSE score and dummies for the institutions. For detailed empirical investigation this model could be elaborated in all sorts of complex ways familiar in the educational research literature. With these simpler models we may see what might be the likely impacts of ignoring the dependence of entry observations across groups compared to the cross-classified model designed to handle it. For the former we have fitted an hierarchical model of type (1) structured as entries within groups. We have 6020 entries for 2280 candidates in 525 teaching groups within the 14 institutions. The response variable is the normalised Advanced Level grade score. The parameter estimates of interest are given in Table 1 below.

Table 1 : Parameter estimates for random effects models of normalised A level scores for entries with students and teaching groups:

Model (1): Uncrossed hierarchical model for entries within teaching groups			**Model (2): Cross-classified model with student and teaching group random effects**		
Fixed effect Parameters	Estimate	Estimated St. Error	Fixed effect Parameters	Estimate	Estimated St. Error
Intercept	-5.493	0.296	Intercept	-5.483	0.361
GCSE Average	1.842	0.047	GCSE Average	1.828	0.060
Random Parameters			Random Parameters		
Institutional variance	0.207	0.110	Institutional variance	0,206	0.117
Teaching group variance	1.365	0.129	Teaching group variance	1.511	0.127
			Student variance	2.915	0.142
Residual variance (within groups)	6.772	0.126	Residual variance	3.906	0.095

It is seen from Table 1 that here is little impact on the fixed effect estimates which me might expect since generalised least squares estimator consistency is not dependent on the model variance specification, although their estimating efficiency is . For inappropriate variance specifications standard error estimates are usually too conservative and may be biased downwards. This is reflected in the higher figures for the crossed model with a more appropriate refined specification and this may be important for inferences on the fixed parameters .The estimates of the institutional variance are much the same across models which might be expected since the crossing is below that level.. This variance is also statistically insignificant in both models. Once a student effect has been isolated in the crossed model the variance attributable to teaching groups increases and group effects seem more dispersed. Thus the failure to account for dependency on common student influences may mean underestimation of variation between teaching groups. We might expect this to also effect our adjusted measures of group effects from estimated group residuals. The correlation between estimates of group effects from the two models was 0.90 . However, this seemingly relatively high value obscures important detailed changes of rank position of groups except for extreme ones. Comparative plot, not illustrated reveals many large changes in some residuals relative to others. Also the cross classified model group effect residual estimates were estimated to be more precise. This is of real importance in our study of teaching group cost-effectiveness, where we desire to have good estimates of group output value added and we need to recognise uncertainty in these estimates (Goldstein and Spiegelhalter (1996)). The crossed model additionally enables the estimation of a student variance A large part of what was previously attributed to entry variation within groups in the hierarchical model can now be seen in the crossed model to arise from the student differences. However, a larger variance (3,906) is taken up by the residual compared to students (2.915). This can only mean that variation of grades within students is considerable and this is information that is often lost in school models which take aggregate points scores of students as the response criterion. We have also used the cross-classified model in further investigation of some factors surrounding group our initial concern with cost-effectiveness. There are subject type differences but these do not affect the broad pattern of the impressions of the random effects that we focus on here. Group size seems to little affect adjusted group performance although clearly cost-effectiveness is enhanced by lower unit costs associated with larger groups (Fielding et al (1997)) . This raises interesting results which may contribute to the class size debate.

5.2 Introducing Weighted Teacher Effects

Due to computing limitations in fitting weighted models we have analysed weighted effects models with our Further Education Funding Council (FEFC) maintained post-16 colleges only. We are experimenting with less computationally intensive procedures which will allow the encompassing of schools. The illustration on colleges is sufficient for present purposes

. We have six of our colleges from the FEFC sector for which adequate full data on teachers and students was available. These relate to 3683 entries in 314 teaching groups from 1511 students with 115 teachers administering the provision The model again uses the normalised A level grade score response . In model (3) above the institutional level was represented by a single random intercept effect v_k . However, since we have only six colleges we have decided to use fixed effect dummy variables to represent the particular effects of these colleges. The base is a small Further Education College and the other colleges are identified by size and type in tables of results. Thus the model we analyse is a restricted version of (3) and is a two level model

$$y_{i(j_1 j_2)} = (Xb)_{i(j_1 j_2)} + u^1_{j_1} + u^2_{j_2} + \sum_1^{J_3} w_{i(j_1 j_2) j_3} u^3_{j_3} + e_{i(j_1 j_2)} \qquad (4)$$

Included n X initially are the intercept , college dummies student average GCSE score. This is then extended to include some weighted teacher variables as outlined above. The proportion of the class hours for a particular entry undertaken by the teachers form the weights. The results of the implementation of the basic model on the college data set are given in Table 2. The parameter estimates, particularly the fixed ones, cannot be directly comparable to the results of our previous section since they are on a more restricted data sets. They relate to post -16 colleges which generally have lower mean raw performance and lower intake ability. The Average GCSE coefficient is slightly lower than for the models in Table 1 which may indicate differential effectiveness of the post-16 institution. Of interest are the sizes of the institutional net differences in this model. Although not presented here a three level components of variance was fitted and this showed that only 2% of residual was due to institutions. This is largely reflected in the dummy coefficients when we model institutions as fixed effects and compare them to the small FE college base. There appear to be some small substantive differences but apart from the 'small sixth form' dummy, which is marginally significant the rest are not statistically significant when compared to their standard errors. Certainly the results seem to confirm again that institutional differences are minimal compared to their internal variation. Of the level two variance not attributable to student differences and residual mainly attributable to variation within students a proportion of around 94 % is attributable to teacher differences. Once teacher effects are explicitly considered in the weighted model and thus controlled there appears small group variability due to other factors. This is considerable and demonstrates that we can be concerned with a real and substantial teacher effect. Unlike some previous studies it has been possible to isolate and unconfound the teacher contribution separate from other class effects. The evidence on why the teachers differ is limited but evaluation and possibly programmes of controlled trials with teaching methods may give answers. The methodology we have discussed and illustrated simply is capable of being applied to much more detailed planned investigations with a wider range of relevant data. As with the crossed model in Table 1 we have also fitted weighted models with other possibly explanatory fixed factors. As before , for instance, we have

Table 2: Parameter estimates for a student/ teaching group cross-classified model for A level performance incorporating weighted teacher random effects for six post-16 colleges

Fixed effect parameters	Estimate	Estimated standard error	
Intercept (base small further education college	-5.674	0.355	
College dummies:			
Large Further Education	0.344	0.410	
Medium sized Tertiary	0.503	0.426	
Small Sixth Form	1.450	0.450	
Medium sized Sixth Form	-0.163	0.408	
Large Sixth Form	0.782	0.385	
GCSE Average	1.772	0.058	
Random Parameters			Percentage of residual variance
Teaching Group variance	0.310	0.120	3.7
Teachers variance	1.642	0.267	19.2
Student variance	2.820	0.124	33.0
Residual variance	3.762	0.210	44.1

examined the subject of the teaching group which is also a characteristic of teachers and might be expected to affect their variation. There were some subject group differences but these did not affect much the conclusions about the teacher and other random effects and their variance structures.

In cost-effectiveness studies given that teacher salaries are often directly linked to observable teacher variables we have obviously been interested in the effects of teachers *per se*. Naturally we have been interested in the extent to which factors affecting their costs also affect their effectiveness. As a preliminary to explicit modelling of teacher variables we have estimated teacher residuals from the model in Table 2 and plotted them against those characteristics in our data set. These include many variables which are often postulated as possibly influentially related to teacher performance: age, years of experience, level of training, educational qualification levels, and salary levels. Apart from small effects of training and the possession of a degree results are disappointing. Most descriptive correlations or measures of association between the teacher residuals and these factors are hardly different from zero. Given pre-occupation with costs and performance related pay the salary levels seem to bear no relationship to estimated teacher effects in this set of colleges. It seems that explanations for teacher variation need to lie elsewhere than the traditional teacher background characteristics that are often the subject of much attention. .

Notwithstanding these remarks we have fitted models which include these variables as weighted teacher effects in ways described in previous sections. We seek to formally confirm the impressions gained through the exploratory descriptive plots. These models also may illuminate the marginal importance of the teacher variables used. Table 3 presents the results of a weighted effects model with dummy fixed effects fitted for the categories of gender , educational qualifications, and training.. Subject group and college dummies are also included in the model but for ease of presentation are not displayed in the results. Teacher age and experience are also included and appear in a form standardised to have mean zero and unit standard deviation over the data subset. The qualification dummies in the table are relative to a base of no degree with QUALDUM2 for teachers with a degree and QUALDUM3 a higher degree. The training dummies are relative to a base of no training with TRAINDUM2 for a Postgraduate Certificate in Education (PGCSE) and TRAINDUM3 a Certificate of Education (Cert Ed.). .

Table 3: *: Parameter estimates for a student/ teaching group cross-classified model for A level performance incorporating weighted teacher random effects and weighted teacher fixed effects for six post-16 colleges Subject group and college dummies are included in the model fitted but estimates not displayed.*

Fixed effect parameters	Estimate	Estimated standard error	
Intercept (base: male, no degree, no training, social science subject, small FE college)	-4.986	0.741	
QUALDUM2	0.034	0.513	
QUALDUM3	0.204	0.542	
TRAINDUM2	0.434	0.148	
TRAINDUM3	0.357	0.410	
Standardised teacher age	-0.014	0.017	
Standardised teacher experience	0.021	0.020	
Female teacher dummy	0.374	0.057	
Student GCSE Average	1,732	0.057	
Random Parameters			Percentage of residual variance
Teaching Group variance	0.304	0.112	3.6
Teachers variance	1.540	0.257	18.0
Student variance	2.761	0.118	32.5
Residual variance	3.910	0.201	45.9

As evidenced in Table 3 , with the exception of the PGCSE training dummy , when related to their standard errors, all teacher variables included have effects that are not statistically significant. Formal tests on these parameters can be carried out using the Wald test available in MLwiN. The PGCSE is marginally statistically significantly different from zero but the rest of the teacher fixed effects are not. There may be some substantive interest in the potential size of these effects but on the evidence available here we have no basis for asserting any effects. It seem that it is only trained graduates who make a difference and even the evidence for that is slight. The structure of the random effects and conclusions to be drawn from them are little changed by the introduction of the teacher variables. It appears that teacher effects are a major influence on performance but that explanations for teacher variation cannot be sought in the conventional career characteristics. However, in the spirit of a philosophy that regards negative findings as important it may be thought that these findings are significant in a substantive sense. Statistically significant findings are not always the most important ones to report

6 DISCUSSION

In this paper we have reviewed some of the educational effectiveness literature that pinpoints the need for the unpicking of important influences operating between the levels of the school or college and that of the student. For our particular situation we desired to disentangle the effects of teaching groups from those of students who may belong to several groups. We also desired to advance ways in which we could isolate the effects of teachers from other influences of the classroom environment. The process of GCE Advance Level provision which we study has a fairly complex structure and it is desired to use statistical models appropriate for this and the data it generates.

We have sought to discuss methodology which may be useful in other contexts and areas where similar complex structures may arise. It is apparent that although fairly sophisticated methodology can be developed it is only useful if appropriate data is available. The type of data we have used in exemplification, particularly in detailed event history of provision, is indicative of data requirements for the analysis of such complex structures. Model procedures are being made available but investigations should proceed with cognisance of complex effects operating and the need for wide relevant data to disentangle them.

Although the main strand of this paper has been methodological , we have only used a fairly simple set of explanatory variables in order to illustrate the models. The potentiality for developing these approaches further by use of more complex characterisations of influences and effects should be reasonably clear. Nevertheless we have uncovered some though provoking substantive results. Variation of subject grades within students may be as of much importance as variation between students. In the absence of a teaching group focus it may be possible to adopt an approach using multivariate responses within schools. We have not fully reported investigations of subject differences but clearly these are of relevance.. For our college provision we have also indicated ways in which teacher effects can be separately evaluated and established that these may be considerable. Teacher variation appears not to be clearly explicable in terms of standard background characteristics of teachers, their education , training, experience or remuneration. It is a challenge to educational research to establish the factors which do contribute to good teaching. Methodology to analyse complex structures is becoming available and wills the extension of the scope and design of detailed investigation of what is the process going on within schools and colleges. Adequate data within carefully designed studies is a paramount pre-condition before some of these questions can be answered.

References

Aitken, M. and Longford N.T. (1986). Statistical Modelling in School Effectiveness Studies. *Journal of the Royal Statistical Society*, A, 149, 1-43.

Belfield, C . , Thomas , H. and Fielding, A. (1996a) *Costs and Performance of A level Provision in Schools.* Research Report to the UK Department for Education , School of Education, University of Birmingham

Belfield, C . , Thomas , H. and Fielding, A. (1996a) *Costs and Performance of A level Provision n Colleges.* Research Report to the Association of Principals of Sixth Form Colleges, School of Education, University of Birmingham

Bryk, A.S. and Raudenbush, S.W. (1987). Application of Hierarchical Linear Models to Assessing Change. *Psychological Bulletin*, 101 (1), 147-158.

Bryk, A.S. and Raudenbush, S.W. (1992). *Hierarchical Linear Models*. Newbury Park: Sage

Card, D., and Krueger, A. B. (1992). Does School Quality Matter? Returns to Education and Characteristics of Public Schools in the US. *Journal of Political Economy, 100, 1-40.*

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

Department of Education and Science. (1995). *GCSE to GCE A/S Value Added.,* London: HMSO.

Fielding, A. (1995). Institutional Disparities in the Cost Effectiveness of GCE 'A' Level Provision: A Multilevel Approach, *Education Economics*, 3, 12-36.

Fielding, A. (1998), Perspectives on Performance Indicators: Cost-effectiveness in GCE

Advanced Level Provision , *School Effectiveness and School Improvement*, 9, 2, 218-231

Fielding, A., Belfield, C. R. and Thomas, H. (1997). *An Investigation of Performance and Cost-effectiveness in GCE A-level Provision in the FEFC Funded Sector*. Department of Economics Discussion Paper 97-14, University of Birmingham

Fielding, A., and Yang, M. (1999). *Random Effects Models for Ordered Category Responses and Complex Structures in Educational Progress*. Department of Economics Discussion Paper 99-20, University of Birmingham

Figlio, D. N. (1999). Functional Form and the Estimated Effects of School Resources, *Economics of Education Review*, 18, 241-252

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

Fitz-Gibbon C. T. (1997) *The Value Added National Project: Final Report* , London, Schools Curriculum and Assessment Authority

Fraser, B. J., and Tobin, K. (1989). Student Perception of Psychosocial Environment in Classrooms of Exemplary Science Teachers. *International Journal of Science Education, 11, 19-34*

Goldstein, H. (1997) Methods in School Effectiveness Research, *School Effectiveness and School Improvement*, 8,4,369-95

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

Goldstein , H. and Spiegelhalter, D. (1996) . League Tables and Their Limitations: Statistical Issues in the Comparisons of Institutional Performance. *Journal of the Royal Statistical Society, Series A*, 159, 3, 385-409

Goldstein , H. and Thomas, S. (1996) . Using Examination Results as Indicators of School and College Performance. *Journal of the Royal Statistical Society , Series A,* 159, 1, 149-163

Gray, J. and Hannon, V. (1986). H M I's Interpretation of School's Examination Results. *Journal of Educational Policy*, 1(1), 23-33

Gray, J., Jesson, D., Goldstein, H., Hedger, K., and Rasbash, J. (1995). The Statistics of School Improvement: Establishing the Agenda. In *J. Gray and B. Wilcox (eds). Good School, Bad School, Evaluating Performance and Encouraging Improvement*. Buckingham, Open University Press

Hanushek, E. A. (1986). The Economics of Schooling: Production Efficiency in Public Schools. *Journal of Economic Literature, 24, 1141-1177*

Hill, P. W. and Goldstein, H. (1998). Multilevel Modelling of Educational Data with Cross-classification and Missing Identification of Units. *Journal of Educational and Behavioral Statistics*

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*,4,310-333

Longford, N.T. (1993). *Random Coefficient Models*. Oxford: O.U.P.

Luyten, H. and de Jong, R. (1998). Parallel Classes: Differences and Similarities, Teacher Effects, and School Effects in Secondary Schools. *School Effectiveness and School Improvement, 9*,4, 437-473

Monk, D. H. (1992). Education Productivity Research: An Update and Assessment of its Role in in Educational Finance Reform. *Educational Evaluation and Policy Analysis, 14, 307-322*

Nuttall, D.L., Goldstein, H., Prosser, R., and Rasbash, J. (1989). Differential School Effectiveness. *International Journal of Educational Research*,13, 769-76.

Rasbash , J. and Goldstein, H. (1994) Efficient Analysis of Mixed Hierarchical and Cross-Classified Random Structures Using a Multilevel Model. *Journal of Educational and Behavioural Statistics, 19,4, 337-350*

Rasbash, J., Browne, W., Goldstein , H., Yang, M., Plewis, I. , Healy, M., Woodhouse, G., and 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

Thomas, H. (1990). *Education Costs and Performance: A Cost-Effectiveness Analysis* . London:Cassell.

Willms, J.D.(1992). *Monitoring School Performance: A Guide for Educators.* London: Falmer

Yang, M. and Woodhouse, G. (2000). Progress from GCSE to A and AS Level: Institution and Gender Differences , and Trends Over Time. British Journal of Educational Research (forthcoming)

Young, D.K. (1998). Rural and Urban Differences in Student Achievement in Science and Mathematics: A Multilevel Analysis. *School Effectiveness and School Improvement, 9,4,386-418*