ORDERED CATEGORY RESPONSES IN MULTILEVEL AND CROSS-CLASSIFIED STRUCTURES

Antony Fielding[1]
University of Birmingham

Min Yang
Institute of Education, University of London

Ordered category responses are often scored and subjected to multilevel models appropriate for continuous data. This may have some statistical and substantive implications. Models are considered which model the category distribution directly, and methods for these reviewed.. They are extended to cover cross-classified random effects at levels within the hierarchical structure. A macro suitable for implementation in the multilevel software MlwiN is introduced. The motivation is discussed through an example where it is desired to disentangle teacher, student, and subject effects in performance at the General Certificate of Education Advanced Level in England and Wales. This involves a sparse unbalanced structure at the level of the subject teaching group. Simulation results are discussed which evaluate the quasi-likelihood methods used. Some promising improvements in estimation involving simulation of latent variables underlying the response are suggested as an area for further research

Keywords: Crossed effects, Educational grades, Extra-multinomial, GCE Advanced Level, Latent variables, MLwiN, MULTICAT, ORDCAT, Ordinal data, Quasi-likelihood, Sparse data, Unbalanced designs.

[1] A.Fielding@bham.ac.uk. The author acknowledges support from the UK Economic and Social Research Council in a Visiting Fellowship award number H51944500497 of the Analysis of Large and Complex Datasets programme

1 INTRODUCTION AND MOTIVATION

In the social sciences ordered category response variables are often converted to arbitrary points scores. They are then frequently modelled by methods appropriate for continuous interval scale dependent variables. Up to recently this has certainly been the norm in official and research literature on educational achievement grades. In particular much analysis of hierarchically structured data with normal theory linear multilevel models proceeds in this way. Statisticians and others have for some time pointed out the methodological and substantive deficiencies of modelling by the imposition of arbitrary measurement scales on such grouped ranked responses. These are particularly evident when the number of categories is small, as is common, for instance, in educational gradings or in opinions and attitudes

The limitations of linear modelling of category scores that have been raised in the literature) are many and varied. These are rarely spelled out in detail but are reviewed in Fielding (1999). There are measurement and substantive considerations about the scaling implied by any sets of arbitrary scores, particularly when there is an attempt to do 'funny arithmetic' on tied rank orders. Questions arise in linear modelling about whether effects operate additively on a particular points scale rather than some other arbitrarily chosen one. Continuous distribution properties applied to discrete measurements lead to difficulties. Of prime importance, even if the scaling issues where resolved, is estimation bias when grouped data is observed. This is well recognised in traditional single level modelling (Stewart, 1983). Simulation studies by Fielding (2000a)[2] have shown that estimation of variance components, which are central to multilevel modelling, is affected even more seriously than that of fixed parameters

The popularity of scoring models for ordinal data possibly owes much to the familiarity with standard linear multilevel models and accessible software. However, it should be apparent there is a motivation to develop procedures which model the ordinal categorised responses directly, and that further make no arbitrary scaling decisions. Generalised linear models for single level structures have been developed for some time (McCullagh and Nelder, 1989). Only recently are generalised linear mixed models (GLMMs) for ordered categories, and appropriate for multilevel and other random effects, becoming known and more widely available (McCulloch, 1999). After describing modelling

frameworks, this paper will review the availability and uses of some of the existing methods and software. We then focus on the multilevel modelling software MlwiN (Rasbash et. al. 1999) and add to knowledge by suggesting the use of the specially written MULTICAT macros for categorical responses (Yang, et. al. 1998, 2000a). The latter are under continuous review and the latest version offers procedures for cross-classified and weighted random effects at various levels in the structure. The latter are the focus of the applications to be discussed.

Apart from statistical considerations discussed, it is becoming apparent that ordinal data models have many practical advantages. Fielding (1999), for example, has shown that in hierarchical models of primary school progress, simpler patterns of explanatory fixed and random effects emerge than for points models. A specific instance is that it no longer becomes necessary to introduce higher order polynomial terms in intake ability controls. In linear models the latter have been suggested as being required to handle distortions due to 'ceiling' and 'floor' effects which cannot be evidenced by observed grouped grades on the response at the top and bottom of the scale. Yang et al. (2000b), in another educational example, have also commented on the greater flexibility, ease of interpretation, and practical usefulness of generalised models. For instance, the ability to directly predict outcome 'chances' or probabilities of achieving certain grades given a profile of effect values, is a major advantage. This paper also demonstrates that more interesting comparisons can be made between level 2 school random effects. Variations between progress in achievement of certain grade thresholds may be of as much practical concern as average levels of adjusted outcome.

A particular motivation for extending models in the present work is informed by the increasing awareness in school effectiveness work that effects operating at any level in educational structures are complex and cross-classified (Coe and Fitz-Gibbon, 1998). We desire modelling frameworks for ordinal data which attempt to disentangle and unconfound these effects. Two similar examples of cross-classified effects in this area are provided by Goldstein and Sammons (1997) and in the MlwiN user guide (Rasbash et. al, 1999), although the analyses use linear models and points scores for educational grades. In both the issue addressed is of the continuity of school effects at different stages in childrens' careers. The question arises as to what effect does the primary school attended have on later performance

---

[2] The author s Departmental discussion papers and other unpublished papers referred to are available from his web page under

at secondary school. Students are nested within secondary schools but they are also nested within the primary schools they came from. Thus at a level above the child there is an unbalanced cross-classification of primary and secondary schools and models to estimate the joint contributions of these are required. Using the generalised model for ordered data to be to be presented here Fielding (2000b) re-analyses the data set used in the second example. In this paper we will discuss an example structure where effects may be even more complex. Cost-effectiveness studies of teaching groups in English colleges in the post-compulsory education sector, providing instruction for the General Certificate of Education at Advanced Level, have been undertaken by Belfield, Fielding and Thomas (1996). The data are not hierarchical in that students may be present in several teaching groups. Unless it is disentangled teaching group effects may be confounded with the students who select a particular combination of subject groups. Cross–classification of students and teaching groups in a model facilitates the separate estimation of their effects. In attempting to unravel the tri-partite confound (Coe and Fitz-Gibbon, 1998) we are also interested in teacher effects separate from those of the teaching group. If groups were taught by the same teacher throughout the course and teachers taught several groups we could extend the structure and models to deal with three way classifications. However single teacher provision is the rare exception and up to five different teachers may be involved in a split-plot way. This motivates our suggestions for extending our models to encompass weighting teacher contributory effects. Unlike the school crossing examples , the structures are very sparse. For example there can be only one observation on a crossing of students and groups and the majority of cells are empty. This may affect the quality of the estimators proposed and this matter is considered in the last section of the paper

2 MODELS AND ESTIMATION

2.1 Hierarchical Odered Category Models

Traditional normal theory linear regression models relate conditional expectations (and variances) of a continuous response to linear predictors (LP) involving covariates. As mentioned in the introduction the assumption that arbitrary point scores applied to ordered

category responses follow such models is somewhat untenable. By contrast standard single level generalised linear models (GLMs) deal directly with the conditional probability distribution over the set of ordered categories. They do this by relating the conditional cumulative distribution over the categories through a 'link' transformation to a linear predictor (LP). Each of the set of cumulative probabilities is by definition constrained to lie between zero and one. Usually, though not always, when a link function is applied to them the results of the transformation will occupy the whole of the real line. There are then no constraints on the feasible values arising from effects of covariates in the LP. For these and other reasons the link function is often the inverse of some standard distribution function of a continuous random variable. Common choices in application are the inverses of the standard logistic, normal and extreme-value (Weibull) distributions which yield respectively probit, logit , and complementary log-log links.

A GLM can be extended by including random effects in the LP in the same way that linear fixed effects models are extended to linear mixed models. The resulting generalised linear mixed models (GLMMs) for ordered categories are only fairly recently receiving more than a minimum of attention. McCulloch (1999) reviews the general class of GLMMs and their estimation. These type of models are the focus of this paper. Often the random effects are hierarchical effects arising out of multilevel structured processes and data. In these cases the GLMMs for ordered categories can be readily seen as a particular generalised version of continuous response multilevel models, which have received much attention in the past decade or so.

For instance, in a three level educational structure of students within classes within educational institutions we may have a graded outcome represented by a set of S categories labeled for convenience s=1, 2, 3, …….., S, in order from low to high. For a logit model The cumulative probabilities $g_{ijk}^{(s)}$ of achieving at least grade s for student i in group j of institution k are modelled by

$$L(g_{ijk}^{(s)}) = \text{logit}(g_{ijk}^{(s)}) = \log_e(g_{ijk}^{(s)}/(1-g_{ijk}^{(s)}) = q^{(s)} - [(\mathbf{X}\hat{\mathbf{a}})_{ijk} + v_k + u_{jk}] \qquad \text{s=1, 2, 3,…..(S-1)} \qquad (1)$$

with $g_{ijk}^{(S)} = 1.$ by definition. The $q^{(s)}$ parameters are often referred to as category thresholds or cut points. The level 3 random effect v_{0k} and the level 2 random effect u_{0jk} are assumed independently normally distributed around zero with variances s_{v0}^2 and s_{u0}^2 respectively. **X** is a data matrix of covariates excluding an intercept term (not separately identifiable from the $q^{(s)}$). The vector of covariate effect parameters is $\hat{\mathbf{a}} = \{b_1, b_2 \ldots, b_L\}'$. For extended models some elements of this vector may also be random at any level A coefficient random at only level 3, for instance is then subscripted by k and we have. For such models the set of effects at a level are assumed jointly multivariate normal and covariances will usually be specified. Effects across levels are assumed independent. The negative sign on the right hand side of the model equation is for interpretational convenience in results. Substantively positive effects on the response itself shifts its distribution up and there is a negative effect on the cumulative probabilities and their logits. We note that there is no explicit level 1 disturbance term in these models since variation at this level is governed by the set of probabilities, although extra-multinomial variation can be introduced. In our later example we also use a probit model with $L(g_{ijk}^{(s)})$ as the inverse of the standard normal $\Phi^{-1}(g_{ijk}^{(s)})$. The logit model has the easy interpretation of a linear effects on cumulative log-odds constant across s. This is equivalent to a constant multiplier of cumulative odds yielding the attractive proportional odds property (McCullagh and Nelder, 1989).

A variety of estimation procedures have been suggested for these ordered category models. Harville and Mee (1984) with a probit model iterate between extensions of best linear unbiased prediction based estimates of fixed effects and a Taylor series expansion to approximate estimation of random parameters using restricted maximum likelihood. Jensen (1990) estimates with maximum likelihood in a two level probit model with a single random effect. The Gaussian quadrature methods of Anderson and Aitkin (1985) evaluate normal integrals in the likelihood. Ezzett and Whitehead (1991) have a crossover trial with a single random effect in a logit model. They propose Newton-Raphson numerical integration methods to evaluate integrals. Hedeker and Gibbons (1996) use maximum marginal likelihood. They describe a Fisher scoring solution for either a logit or probit response function and use multidimensional quadrature to numerically integrate over many random effects. These are implemented in the freely available MIXOR software (Hedeker and Gibbons,1998) which is the most flexible of the methods discussed so far. For our purposes we regard many of these methods to be limited by scope, number of levels possible, or

exclusion of random regression coefficients. All require complicated numerical integration which makes for computational restrictions. The analyses we present use the iterative generalised least squares procedures of the multilevel modelling software MlwiN (Rasbash, et. al. 1999) due to its wide availability, flexibility, adaptability and extensive macro facilities. Goldstein (1995) gives theoretical details of models can be approximated by a various linearisations and cast into forms of a standard linear model to which the MlwiN procedures can be applied. It also allows extra-multinomial variation at level 1 which may be appropriate if the conditional probabilities are mis-specified in various ways. We will discuss later how the inclusion of a multinomial term may be beneficial for estimation even if there is no a priori modelling reason for its inclusion. The method is available as a set of macros MULTICAT. The estimation options and details of how to set the model up are given in the manual by Yang et al (1998)). The results in this paper use a PQL2 option, so known since it uses second order terms in the linearisations leading to what are in effect penalised (predictive) quasi-likelihood estimators. The macros are under continuous review and a forthcoming official release (Yang, et. al. 2000a) will incorporate a probit link that is not currently available. It will also incorporate methods for cross-classified random effects models which we will now discuss.

2.2 Cross-Classified Models For Ordered Category Responses

In the introduction we motivated the desire to extend the hierarchical models to GLMMs with cross-classified random effects at certain levels. We consider only two level models with a crossing at level 2 . A simple logit model for ordered responses has the form

$$\log_e(g^{(s)}_{i(j_i j_2)}/(1-g^{(s)}_{i(j_i j_2)}) = q^{(s)} - [(\mathbf{X\hat{a}})_{i(j_1,j_2)} + u^1_{j_1} + u^2_{j_2}] \qquad \text{s=1, 2, 3,…..(S-1)} \qquad (2)$$

Level 1 units are now nested within cells (j_1, j_2) at level 2 of two cross-classification features indexed by j_1 and j_2 each with their separate additive random effects. It is also possible to introduce an interaction effects between these features. The structure of these random effects is no longer hierarchical . The effects are assumed independent at level 2 and as usual

normally distributed As with the hierarchical model the regression parameters can also be made random across either or both of the crossed level 2 units. In the application we discuss the model lodges Advanced level subject entries within a crossing of students and teaching groups

Goldstein (1997) discuses the iterative generalised least squares theory for cross-classified random effects for unbalanced data structures at any level in the hierarchy. Efficient algorithms for their implementation are evolved by Rasbash and Goldstein (1994) and a range of special commands are available in MlwiN to set these up. Raudenbush (1993) has conceptualised similar approaches in terms of exchangeable regressions. The units for one of the cross classified factors are declared as an hierarchical level in the usual way and effects are random variables exchangeable across these units. For the other cross-classification factor dummy variables indicative of unit membership are set up. The unit effects are the regression coefficients of these variables which are exchangeable within the hierarchical level units of the first cross-classified factor. In implementation this is achieved by constraining the variance of the dummy random effects to be equal. In effect the crossed effects model is reformulated to a strictly hierarchical one, which cam be analysed by multilevel methods. A variety of ways of improving computational efficiency through model set up are discussed in the cited references. Of particular importance are ways of taking advantage of any unique blockings of the certain combinations of crossed effects. In our modelling example, for instance, the crossing of students and groups is within colleges. Due to their small number college effects are treated through fixed effect dummy variables. However, although they are not treated as an additional level , they can be regarded as a factor by which the lower crossing is blocked

The theory and methods for crossed effects have so far been treated in terms of continuous response linear models. They can also be extended to multi-way crossings at a level and to be operative at several levels. In generalised models the type of crossing of effects discussed occur at levels which are represented only by terms in the LP. Thus procedures for reformulating the model in hierarchical terms carry over straightforwardly. Once this is done the resultant form may then be treated by means described in Section 2.1. For technical reasons only MULTICAT cannot at present deal with this synthesis of ordered category responses and crossed-effects, although the implementation is straightforward for nominal category responses. Thus a special stand alone macro, ORDCAT[3], has been written

[3] This together with user notes can be downloaded from

for these purposes and which will be incorporated in the future software developments (Yang, et. al.2000a).

2.3 Cross-Classified Weighted Random Effects

We desire to incorporate a further crossing factor above the level of A level entries but one where there are multiple effects on each observation. Several teachers are involved in a subject group and each teacher may teach several groups. The logit model we propose to handle this is of the following form:

$$\log_e(g^{(s)}_{i(j_i j_2)}/(1-g^{(s)}_{i(j_i j_2)}) = q^{(s)} - [(\mathbf{X}\hat{\mathbf{a}})_{i(j_1,j_2)} + u^1_{j_1} + u^2_{j_2} + \sum_{j_3}^{J_3} w_{i(j_1,j_2)j_3} u^3_{j_3}] \qquad (3)$$

The contribution of teachers to a group have their random effects $u^3_{j_3}$ weighted according to the proportion of teaching time provision of a group they taught. In the model the weights $w_{i(j_1,j_2)j_3}$ sum to unity across all teachers. For a particular teaching groups the majority of these weights will be zero. There may be substantive objections to the choice of weights. However, in implementation of sensitivity analyses, the teacher variance parameters have proved relatively robust to variations of this choice. Although slightly different in conception the model details are formally similar to those devised by Hill and Goldstein (1998) for multiple membership of units and missing level identifiers.

In principle the methods for strictly cross-classified structures can carry over to this sort of model. What is required is the replacement of the design matrix of dummy indicator variables for exchangeable crossed effects by the weight design matrix. Hill and Goldstein (1998) explain this. Again special MlwiN commands are available to do this. Once the matrix and other model details are set up the ordered category methods of ORDCAT can then handle it.

.3 TEACHER, STUDENT, AND GROUP EFFECTS IN AN ORDINAL MODEL FOR GCE ADVANCED LEVEL SUBJECT GRADES

3.1 Background

In studies of cost-effectiveness of GCE Advanced Level provision we have become aware that variations within institutions are sometimes for more relevance than institutional differences (Belfield, Fielding & Thomas, 1996) This is in concordance with a switching away in much educational effectiveness literature from interest in effective schools to effective classrooms. Here we will be concerned with a side of this problem, analysis of A level outcomes, suitably adjusted, at the teaching group level. The complexities of the structure of A level provision, and indeed the data we have, have motivated the implementation of the type of model discussed in the previous section. We will first mention this structure, associated data, and the methodological problems it poses. We hope that some of the methods we have proposed will have wider relevance since similar structures are quite common Although the concern is with educational achievement these issues are of general relevance..

3.2 Example Structure And Data

The conclusions in educational research about variations at classroom level within institutions also begs the question of what is going on at this level This problem of disentangling what is happening is clearly put by Coe and Fitz-Gibbon (1998) who say, 'The combination of subject taught, teacher and pupil group is, of course, unique for each class, and effects could be attributed to all three parts of this tripartite confound'. The way in which teachers and classes have sometimes been treated synonymously in single teacher courses is not entirely unequivocal. Also Hill and Rowe (1996) have commented on the sparsity of research on teacher effects and that this may be due to the fact of several teachers being involved in a classroom outcome. From the structure of our data, the information it contains and the methodological approaches we suggest, we have an opportunity to make some advances. Also the pupil group is unique for each class but it is often argued that satisfactory control cannot always be exercised by taken cognisance of all relevant pupil characteristics. Yet as students

cannot be treated as randomly allocated to teaching groups (classes) there may yet be important but unmeasured sources of student variation which are associated with selection into teaching groups. Any differences between groups may be partly reflecting these. A within group random student disturbance caters for unmeasured variation but may not adequately cater for these systematic effects which may be confounded with those of groups. We hope in hierarchical situations that sufficient control has been exercised. The disentangling of many of these possible confounding factors is an important set of problems, which, in a particular context we try to go some way towards addressing here. The additional feature of our data which is both a problem in itself but also presents an opportunity to resolve these other problems is the fact of multiple representation of students in groups. In strictly hierarchical situation we hope that with sufficient control unmeasured student influences are not systematically related to group effects. In the present situation since entries are correlated across groups due to common individual unmeasured influences school effects may be confounded with those of students. Considering a cross classification of students and groups resolves some of these issues

The subset of the A level data set used has 3717 entries nested within 317 teaching groups (classes), which in turn are nested within 6 colleges. The subset comes from entries in these colleges from full course students, i.e. from those taking at least three subjects. On the surface this is a normal 3-level structure, although due to their small number colleges have been treated as a fixed effect blocking factor in later analyses. However, between three and five entries are made by each of 1522 students and there is thus an additional crossed nesting of entries within student at level 2. Ignoring this crossing may result in teaching group differences being confounded with students and disentangling class such confounding effects is the concern raised above. We compare possible effects in later analyses, Modelling of unbalanced cross-classified designs is a sound methodological development to handle this

There are 145 teachers involved in the teaching groups and no teaching group has less than two teachers. Teacher effects are of obvious interest in their own right. However, without separate consideration they may confound with other features of group heterogeneity.

The response variable we use is the A level grade with six ordered categories A, B, C, D, E, Fail. The main explanatory control variable used is a measure of prior ability , a standardised score on the General Certificate of Secondary Education (STGC) taken by students before they embarked on A levels. Also used are gender, seven broad subject groups and the college attended. Dummy variables have been created for subject groups relative to a Social Science base group. For the six colleges with a range of sizes we use the acronyms

FEC, SFC , and TC respectively for three college types, Further, Sixth Form and Tertiary. Dummies are formed relative to the medium sized FEC base.

Table 1. Parameter Estimates For Hierarchical , Cross-classified And Weighted Random Effects Models For Performance In Subjects At General Certificate Of Education At Advanced Level In Six Colleges For Post Compulsory School Aged Students

.

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

-

3.3 The Example Results

The first two columns of Table 1 present estimates for a base and elaborated hierarchal model for entries within teaching groups and ignores the complexities discussed. This was the type of model originally entertained before the possible extra specifications were considered. It forma a comparative base against which to judge the specification of complex random effects. The second two columns introduce the comparable cross-classified student effects. Disentangling the teacher and group effects is attempted in the model results in the final two columns.

The same set of elaborating explanatory fixed effect covariates was used in each case. Also, in the results, a single random effect is illustrated for each factor in each case and there are no differential random regression parameters. The results presented were a culmination of much deeper exploration of potential fixed effects for which data were available and random effects which might have proved promising. A range of statistical evaluation procedures ere used including the Wald test procedures available in MlwiN. Effects not presented proved unfruitful on both substantive and inferential grounds. Included among trial fixed effects were teaching group context effects such as size and aggregate process variables such as attrition from the course. Differential teaching group variation in the STGCSE coefficients with nothing interesting emerging for this situation.

A main object of the analysis has been to specify teaching group variation and effects for input into further work on cost-effectiveness. However, some comment may be offered on the fixed estimates. The broad pattern of fixed effect coefficients give a similar impression across the differently specified models. In any case some differences may emerge in generalised models due to the necessary implicit scale changes between model developments. We might expect better standard error estimates as we refine the specification and if these specifications are more appropriate. This may have implications for detailed inferences on the fixed parameters were such to be performed . A quadratic term is required for this in all models . There is a marked ceiling to these STGC scores and they are skewed to this ceiling. These factors may explain the quadratic effect. However, higher order polynomials are not required for the performance function as happens with some linear points scores models of the same response (Fielding, 1998). The negative gender coefficients indicate that girls make less progress than males. By contrast, although not illustrated here, positive female effects emerge if STGC is not controlled and unadjusted performance is the issue. This phenomenon is also encountered in an analysis of a national

1997 cohort by Yang et. al. (2000). There is also a negative interaction indicating that girls have a smaller STGC 'slope' effect. This will mean that lower ability girls will make more progress than similar boys but vice-versa at higher ability levels. There are some important subject effects. There is a vigorous debate in the literature about whether results such as this mean that Mathematics, Sciences and Languages can be perceived as more difficult. (Fitz-Gibbon and Vincent, 1997; Goldstein and Cresswell, 1996; Newton, 1997). This will not be pursued here. The six colleges represent a range of sizes and types found in British post school education (Belfield, et. al. 1996). College dummies are relative to a medium sized Further Education college. It is known that college size and type do make a difference. They have been introduced here in fixed effects as relevant block adjustment controls. There are too few in this data to draw generalisations apart from differences between specific colleges in the data.. In the tables both sets of dummies characterise the teaching group and teacher levels.

In generalised linear multilevel models there is no explicit level 1 variance term estimated since variation at that level is governed by conditional expectations in the form of the conditional probabilities. Thus it is not easy to assess the relative sizes of residual variance components within models and changes in these variances as we move from a model to a more elaborate one. These features are one of the preoccupations of interpretation of linear multilevel models. However, one interpretation of a logit model is that it characterises the conditional probabilities of responses falling in certain groups of values on a latent variable (*lv*) underlying the ordered categories. If this *lv* follows a linear multilevel model but with a standard logistic distribution for the level 1 variance then the logit model ensues. The estimable parameters $q^{(s)}$ may be viewed as cut-points for the categories on this *lv* scale (McCullagh, 1980).. For identification reasons the *lv* is standardised and the implicit latent variable whatever the model specification always has the logistic variance $p^2/3$ at level 1. A similar relationship exists between the probit model and a normal *lv* with unit variance at level 1. For a particular model it is fairly straightforward then to assess the higher level variance components relative to the latent level variance. However, differently specified models will have different scales. consequent on the re-standardisation at each stage. Thus, for instance, we might expect introducing an important explanatory variable to reduce the variability in the underlying response at level 1. However there is always an implicit scale change to standardise this variance to be always equal $p^2/3$. The scale change will also result in rescaling of all fixed parameters, random effects, and their variances. Thus without

taking these scale changes into account it is difficult to compare estimates and in particular reductions or otherwise in higher level variance components. Fielding and Yang (1999) discuss this problem and devise a very approximate method of assessing these scale changes. It then becomes possible to rescale and compare parameter estimates across models. The method is very rough but it involves finding a factor which rescales each model to the same scale as base models similar to that of column 1 of Table 1 A set of logistic conditional mean scores for the ordered groups are found from the grouped distribution implied by the base model. These constant set of scores are then applied to the grouped level distribution in a later model development and a variance approximated. This compared to the variance for the base distribution gives a scale factor, The roughness of the approximation may be seen from the fact of grouping of an unobservable continuous variable but it does serve to indicate the order of magnitude of the changes. We present no detailed results here on the application of this method but it is readily applicable to model results in situations where we wished to have detailed comparisons of variance changes from the base model to the explanatory fixed effects model under each scenario. We will content ourselves with examining the relative sizes of the variance components within each model and how these compare across models. This is sufficient for present purposes.

Since the cut-point parameters on the *lv* interpretation should be invariant to anything but the scale changes, we might expect their estimates to reflect this. With these facts in mind we might examine the differences in detail in the fixed parameter estimates in the models of Table 1. These must be evaluated in the light of the relationships to extra controls that the introduction of further random effects implies. Effects on log odds mirrored in the coefficient estimates are net of random effects.. Thus we might expect some changes when student heterogeneity is introduced into the teaching group models, since they are then net of unmeasured student attributes. On the *lv* linear model interpretation there will also be consequent scale changes. The reduction in student variable coefficients is proportionately in line with changes in the cut point estimates indicative of rescaling. However, there are uneven changes in the subject and college dummies. They are not consonant with scale changes and those of the associated net log-odds. Part of the reason for this may be the clustering of student entries into certain subject groups and the attraction of some colleges for certain types of student. The mathematics and science effects are much more sharply defined. On cursory investigation, the weighted teacher model would appear to have similar implicit conditional *lv* variability to the teaching group model. Further the cut-points and student variable coefficients have similar values. Mathematics and language effects relative

to Social Sciences are no longer significant. It might be conjectured that subject effects observed in earlier models might be inextricably bound up to some extents to the type of teachers that deliver them. On introducing a teacher effect the net effects of subjects will thus change. Similar comments may be made about the changing pattern of college effects. There is quite a lot of complexity in these patterns, which might be unraveled by deeper investigation and more extensive data. The results do, however, pose some intriguing questions in the study of educational progress. They cannot be fully investigated here. As a final detailed point about the fixed estimates we may note the minor changes in their estimated standard errors as variance specifications are refined. However, it may be pointed out that in most statistical investigations the accuracy of these estimates is sensitive to what is assumed about the specifications of variance. In general more appropriate specifications lead to better inferences.

The variance component estimates and their relative sizes across the models raise many interesting issues of both methodological and substantive nature. In Table 1 we have presented the estimates and also in each case expressed them relative to the total variance assuming standard logistic at level 1. In the teaching group model the covariates reduce the teaching group and entry variation proportionately. This is seen in the similar percentages (17.7 and 18.2) attributable to groups relative to standardised entry *lv* variance ($\pi^2/3$). The approximate scale calculation yields a variance reduction of the order of 30%. Introducing a student cross-classified random effect into the base model reveals two interesting features. Firstly part of the teaching group variation is now explicable by the differences of students selected into them. Students do not make an independent contribution within groups since their effects are common to certain groups. Secondly , variation amongst students is fairly high at 30% of total variation. However it is relatively much less than the 60.5% of residual variation at the entry level that remains when student and group differences have been accounted for. On this evidence there is much variation between the A level grades of subjects taken by each student which cannot be accounted for by a teaching group effect or subject preferences. This point is conventionally recognised by some university admissions officers who specify sets of particular grade achievements for specific subjects rather than rely on aggregate points scores. For many purposes the latter hides the diversity in addition to being a dubious scaling device. In the model with student effects the greatest relative impact of the control covariates is on the student variance, which may be expected. Although we do not present results here a comparison of teaching group residuals from the two models

demonstrates considerable differences in our assessment of group effects. This has had practical significance in comparing groups within colleges and stresses the importance of isolating the common influence of unmeasured attributes of students selected into the groups..

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

effects have been provided. What is further required is more attention to study designs in relevant research and the collection of detailed data, particularly on teachers and teaching scenarios. Perhaps the recent interest in randomised controlled trials educational and teaching interventions is the way forward.

4 THE EXTRA MULTINOMIAL PARAMETER

In all the results presented an extra-multinomial parameter has been estimated. This is the parameter that multiplies the multinomial variance covariance matrix of the multivariate indicator variables representing the category of response at level 1. It may arise in situations where the model under discussion mis-specifies the conditional category probabilities. If the parameter is unity we have multinomial variation. Discussion of has been left to the end since there are some general issues of interpretation. In one sense its introduction is justified on certain accumulating evidence that it improves estimation of model parameters Allowing this parameter to be free seems to take up any features of the data that make the imposition of multinomial variation over rigid. Further, simulations of multinomial variation for certain structures indicate it may be better to unconstrain it even where it is known that multinomial variation holds in the sampled structure. (Yang, 1997**;** Fielding and Yang, 1999) These references and Wright (1997) also note that with certain sorts of sparse data and even with known multinomial variation, analyses result in the parameter being estimated quite often much different from unity. Our example data is quite a sparse structure for reasons indicated above. However, and importantly, its inclusion seems to improve the estimation of other parameters in models. The differences in emphasis on the role of this parameter stem from its provenance in controlled experimental designs and models. In such situations and for marginal single level analyses the multiplying parameter follows fairly readily from clustering in the data (another level) or important uncontrolled variables not specified in the probability model (e.g. Williams, 1982). For complex survey data and multilevel designs the issue is not so clear-cut. Models can only ever be approximate reflections of complex mechanisms going on in the data and ranges of possibilities and specifications will always be open to improvement. An important source of extra-multinomial variation is often an important level in the structure being missing from the model. Hierarchical models of children within schools often ignore a possible clustering effect due to classrooms (Fielding, 2000b). The balance in the data structure itself will also mean that model estimation is

affected. The introduction of a multinomial parameter, then, may be regarded as a rough provision for these inherent possibilities. If as seems to be the case the estimation is improved, then so much the better. It is on these grounds that it is used in the results here. Of course, unless more is known a priori about potential sources or , for instance, that sparse data dictates a preference for its inclusion, then the actual interpretation of the value may be difficult. It is often a matter of speculation that requires further evidence and investigation.. As has been seen the sparseness and balance of the crossing may have an effect. In Table 1 the estimate of 0. 696 indicating under-dispersion may be a result of the few observations per student or may be due to the lack of specification of the teacher effects. To some extent either or both of these may be washed out when the effects are marginalised jointly in the teaching group only model. Much more would need to be know about allocation of teachers and students to groups to evaluate these possibilities. However, they give an idea of the range of speculation that is possible. The important lesson is that the parameter should usually be included to reflect a range of possible but unforeseen complexities, data sparseness, and to improve estimation. This is recommended irrespective of any attempted interpretation of the value of its estimate.

5 CONCLUDING REMARKS ABOUT ESTIMATION

The methods presented in this paper may be seen as a part of a general development of analytical tools to analyse complex structures using survey and other data that correspond to classical experimental designs but are more flexible (Raudenbush, 1993). Alongside the development of methodology it is recognised that attention to study design and collection of more appropriate data is necessary if questions about these complex structures are to be fully handled

The estimation procedures suggested have been based on penalised quasi-likelihood. using the flexible macro facilities available within the software MlwiN. Alternative estimation procedures which currently have implementable software have been suggested. These are mostly based on a maximum likelihood approaches. At present the applicability of these are limited except for basic structures. The root of this limitation seems to be a computational one involving numerical maximisation of complicated integrals. A good review of alternative theoretical methods has been given by McCulloch (1999).

It is known that the quasi-likelihood are not unbiased. However, accumulating evidence has indicated that the PQL2 estimation for the type of response used in the examples can provide reasonable estimates (Yang, 1997). McCulloch (1999) is a little more sceptical but focuses in the main on binary responses. The methods of Kuk (1995) using bootstrapping can be applied to bias-adjust the PQL2 estimates. These are also implementable through MlwiN and have been tried in the hierarchical teaching group model of Table 1. Bias adjustment proved minimal.. The bootstrap methods are computer intensive and cannot so far be routinely applied quickly for the more complex structures.

.We have conducted some limited simulations using our data structure and a model using the fixed and random parameter estimates in model with student effects (column 4 of Table 1). We simulated ten data sets with multinomial variation at the entry level. From these data sets we fitted the models using PQL2 with and without a multinomial parameter. Each fit takes about an hour on a Pentium II 450 Mz so the practical limitations of extensive exercises of this sort are limited The fits showed reasonable estimation of fixed parameters. Across the data sets the model variances were underestimated by factors of order 15-25% when the extra multinomial parameter was constrained to have a multinomial value of unity, as it should be for the simulated sets. The effect on student and group variances were proportionate. When this parameter was unconstrained it was estimated around unity but the variances were proportionally underestimated but now by factors of between 2-9% only.. In the sparse structures in Table 2 of Wright (1997), although he does not explicitly mention it in the text, higher level variances are overestimated, but no information is offered for constrained multinomial fits. No general conclusion about directions of impact can appear to be made. Our limited evidence does, however, seem to support the argument that the extra-multinomial parameter be left unconstrained.

Clearly there is a need for further development of tractable general estimation procedures for the type of flexible models discussed. The quasi-likelihood procedures are one such an approach and the biases appear not to be too serious on the limited evidence available. Clearly though future research may extend further the knowledge of the statistical and computational properties of these methods. At the moment they have the virtue of being flexibly applicable and now implementable in widely available software. As research develops these methods may be improved and adapted or completely new methods may arise.. A promising approach which adapts PQL2 for binary responses using data augmentation for a cross-classified structure is given by Clayton and Rasbash (1999). By considering parallel models for crossed factors this improves considerably on computational

efficiency. This approach also incorporates the Bayesian approaches of Monte-Carlo Markov Chain methods (MCMC). The latter approaches may have wider applicability for the sort of situation that has been considered here, as further advances are made. They are currently being developed as an alternative class of estimation approaches within MlwiN. One idea that is currently being researched for ordered category responses utilises the latent variable concept. At various stages of the iterative process ordered responses are replaced by simulated responses of the underlying latent variable. This means that the complex structure models can be handled within the framework of linear continuous response models. There is no doubt that this is a lively area of current interest and a range of methods are being investigated, developed and evaluated. The future holds promise for the ultimate aim of very general and statistically and computationally efficient procedures which are readily accessible..

5 REFERENCES

Anderson, D. A., & Aitkin, M. (1985). Variance component models with binary responses. *Journal of the Royal Statistical Society, Series B, 47*, 203-210.

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

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

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

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

Fielding, A. (1998). Methodological innovations in the use of teaching groups for evaluating cost- effectiveness. *International Congress on School Effectiveness*

and Improvement, 1998 Proceedings. CD Rom, ISBN 0-902252-62-3; School of Education, University of Manchester.

Fielding, A. (1999). Why use arbitrary points scores? Ordered categories in models of educational progress. *Journal of the Royal Statistical Society , Series A, 162, 3,* 303-328.

Fielding, A. (2000a). *Scores and categories for ordinal responses in multilevel and weighted random effects models: Applications in educational research*. Departnent of Economics Discussion Paper 00-02: University of Birmingham.

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

Fielding, A., & 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.

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

Goldstein, H. (1987). Multilevel variance component models. *Biometrika, 74,* 430-431.

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

Goldstein, H. (1995). *Multilevel statistical models*. London: Edward Arnold.

Goldstein, H. & Sammons, P. (1997).The influence of secondary and junior schools on sixteen year examination performance: A cross-classified multilevel analysis. *School Effectiveness and School Improvement, 8, 2,* 219-230.

Harville, D. A., & Mee, R. W. (1984). A mixed model procedure for analysing ordered categorical data. *Biometrics, 40,* 393-408.

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

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

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

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

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

Kuk, A. Y. C. (1995). Asymptotically unbiased estimation in generalised linear models with random effects. *Journal of the Royal Statistical Society, Series B, 57,* 395-407

McCulloch, C. E. (1999). An introduction to generalised linear models. In H. Friedl, A. Berghold, & G. Kauermann (eds), *Statistical modelling: Proceedings of the fourteenth workshop on statistical modelling.* Graz, Austria: Technical University of Graz

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B, 42,* 109-127.

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

Newton, P. E. (1997). Measuring comparability of standard between subjects: Why our statistical techniques do not make the grade. *British Educational Research Journal, 23, 4,* 433-449.

Rasbash, J., & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random effects 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., & 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.

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

Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics, 31*, 144-148

Wright, D. (1997). Extra-binomial variation in multilevel logistic models with sparse structures. *British Journal of Mathematical and Statistical Psychology, 50*, 21-29

Yang, M (1997). Multilevel models for multiple category responses by MLn. *Multilevel Models Newsletter, 8,1,* 9-15, Institute of Education, University of London

Yang, M., Rasbash, H. , & Goldstein, H. (1998). *MlwiN macros for advanced multilevel modelling, Version 1.0*. Multilevel Models Project: Institute of Education, University of London

Yang, M., Rasbash, H. , Goldstein, H., & Fielding, A. (2000a). *MlwiN macros for advanced multilevel modelling, Version 2.0*. Multilevel Models Project: Institute of Education, University of London. (forthcoming)

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

.