Statistical Modelling 2003; **3**: 127–153

Multilevel ordinal models for examination grades

Antony Fielding[1], Min Yang[2] and Harvey Goldstein[2]
[1]Department of Economics, University of Birmingham, UK
[2]Institute of Education, University of London, UK

Abstract: In multilevel situations graded category responses are often converted to points scores and linear models for continuous normal responses fitted. This is particularly prevalent in educational research. Generalized multilevel ordinal models for response categories are developed and contrasted in some respects with these normal models. Attention is given to the analysis of a large database of the General Certificate of Education Advanced Level examinations in England and Wales. Ordinal models appear to have advantages in facilitating the study of institutional differences in more detail. Of particular importance is the flexibility offered by logit models with nonproportionally changing odds. Examples are given of the richer contrasts of institutional and subgroup differences that may be evaluated. Appropriate widely available software for this approach is also discussed.

Key words: educational grades; GCE Advanced Level; logit; MLwiN; MULTICAT; multilevel models; nonproportional odds; ordinal responses

Data and software link available from: http://stat.uibk.ac.at/SMIJ
Received: March 2002; **revised:** October 2002, March 2003; **accepted:** March 2003

1 Introduction

In the England and Wales public examination systems the reporting of pass results is by grades: A*–E for General Certificate of Secondary Education (GCSE) and grades A–E for the Advanced (A) Level General Certificate of Education. Principally for purposes of selection to higher education, the A-level grades are converted to a University Central Admissions Service (UCAS) tariff of points scores. These are scored for each subject examination taken (A = 10, B = 8, C = 6, D = 4, E = 2, with F indicating unclassified or fail at 0). They are then often summed to provide a total points score for each candidate in assessing student achievement. Typical diets are three or four of these subjects but some students take more or fewer. Those taking fewer offer on occasion other qualifications at this level.

Most research on A-level examinations to date, particularly in studies of school effectiveness, has used either the point score by subject or summative scores and they are used in this form by the Government in the production of 'performance indicators' or 'league tables' (O'Donoghue *et al.*, 1996). One of the drawbacks to the use of scores is that information is lost about the actual distribution of grades in particular subjects

Address for correspondence: Antony Fielding, Department of Economics, University of Birmingham, Birmingham B15 2TT, UK. E-mail: a.fielding@bham.ac.uk

10.1191/1471082X03st052oa

when inferences are made at the level of the school. A school mean score in a particular subject could be the result of different individual student grade distributions. Thus an average score could be produced by most students performing very close to the median grade or by some performing very well and some very badly; the distinction between two such schools is potentially important.

The present paper develops explanatory models for the actual grades and compares these with the standard point scoring system. The aim is to gain additional insights from using the former as opposed to the latter models. This is done using multilevel models that recognize the essentially hierarchical nature of examination data with students nested within schools. For the point scoring system standard normal theory models are applied, while for the grades less well-known ordered categorical response models are used. A technical advantage of the latter models lies in the fact that they do not require strong scaling assumptions, but merely the existence of an ordering. They are also not subject to estimation problems arising from grouped observations of an assumed continuous response scale. Further they do not require a basic normality assumption over the scale, although this is not problematic in the point scores used in our examples. It is also possible that the use of the ordered categorical models will result in models with fewer higher order fixed or random effect parameters to fit the data. Many of these comparative criteria are reviewed in work by Fielding (1999, 2002).

In Section 2 we discuss the source of our database, the variables available and the educational context of the application. Section 3 reviews normal theory continuous response multilevel models and we stress the importance in our context of allowing random regression coefficients. Multilevel models for ordered categories are introduced in section 4. Existing work on such models using a variety of estimation procedures (for example, Ezzett and Whitehead, 1991; Jansen, 1990; Harville and Mee, 1984; Hedeker and Gibbons, 1994, 1996; Saei and McGilchrist, 1996; Saei *et al.*, 1996; Chan *et al.*, 1998) is extended because methods in our software provide more flexibility in allowing random coefficient specifications. We also consider estimation of specific random residual effects. Section 5 compares results on the normal points score and ordered models for our application and highlights methodological and substantive points of interest. Sections 6, 7, and 8 detail some important consequences about prior ability GCSE effects, the prediction of A-level grades, and the use of residual estimates for institution value added. In Section 9 we extend ordinal logit models by considering nonproportional changing odds for fixed-effect variables, such as gender, in similar ways to Hedeker and Mermelstein (1998). Saei and McGilchrist (1998) also allow nonproportional fixed time effects in panel data. However, we entertain the possibility of more complexity and also consider developments by considering nonproportional multilevel random effects for our institutions. The latter prove very informative in the context of our application. We conclude in the discussion by focusing on the practical significance of the results, show that the more complex models may improve fit, and consider future directions.

While a central thrust of this paper is methodological, some important substantive results emerge, especially in terms of gender differences and institutional variation. These are highlighted by the use of the variants of the multilevel ordered category response models considered.

2 Data and source

In this paper we utilize information provided by the UK Department for Education and Employment (DfEE) from its database of linked A/AS-level and GCSE examination results (O'Donoghue *et al.*, 1996). The AS (Advanced Supplementary) qualifications are intermediate ones usually taken after one year of study. Grades are usually scored at one-half the corresponding A-level grade. Each individual's outcomes for each subject and qualification are recorded. Additionally a limited range of information is available on certain background factors. We have student's gender, date of birth, and previous educational achievements, as well as the type of their educational establishment, local education authority, region, and examination board.

In this study, we concentrate on A-level outcomes in 1997 for two subjects: Chemistry and Geography. There are a number of reasons for these particular choices of subject. The principal reason is that these two subjects are popular, giving a reasonable number of entries that could be matched to previous General Certificate of Secondary (GCSE) results. These will be used as prior achievement variables. They are the normal secondary school qualifications taken by most students at the end of compulsory education prior to any further advanced study. Total A-level entries are 30 910 in 2409 institutions for Chemistry and 33 276 in 2317 institutions for Geography. AS-level entries are only 3.5 and 3.8% respectively of combined totals. Given this small incidence and also noting that for present purposes modelling essentially different outcomes simultaneously would add to model complexity, the AS entries are not included in analyses. In both subjects only 1.8% of students had several A-level entries and in these cases all entries except the final one scored zero. Thus the single best entry, indicative of achievement in that subject, was entered into our analyses without loss of substantive meaning. Another reason for the choice of these two subjects is that they have distinct distributions of grades. As Table 1 shows, the grade

Table 1 Frequency distributions on A-level Chemistry and Geography (1997) on cases that had matching students' GCSE results

	Chemistry (2409 institutions)[a]				Geography (2317 institutions)[b]			
Grades	Number	Overall (%)	Male (%)	Female (%)	Entry	Overall (%)	Male (%)	Female (%)
A	6680	21.6	21.8	21.4	4170	12.5	10.8	14.7
B	6666	21.6	21.2	22.1	7407	22.3	20.4	24.5
C	5732	18.5	18.0	19.2	7885	23.7	23.7	23.7
D	4611	14.9	14.8	15.0	6297	18.9	20.3	17.2
E	3606	11.7	11.7	11.6	4271	12.8	14.2	11.1
F	3615	11.7	12.5	10.7	3246	9.8	10.6	8.7
Total	30 910	100.0	100.0	100.0	33 276	100.0	100.0	99.9

[a]Average A-level point score: 5.83 (Males = 5.78; Females = 5.89) Average GCSE point score: 6.30 (Males = 6.16; Females = 6.47).
[b]Average A-level point score: 5.47 (Males = 5.23; Females = 5.76) Average GCSE point score: 5.85 (Males = 5.70; Females = 6.04).

distribution for Chemistry was skewed towards grades A and B, while that for Geography was more nearly symmetrical. They therefore provide good examples in comparing model sensitivities to distributional shape.

It was also decided to omit extreme small outlying groups of 0.36% of Chemistry and 0.37% of Geography students who had very low average GCSE scores well separated from the main distribution (3 or less using scores discussed below).

In many analyses of aggregate educational performance scores, transformation by normalizing has been a practical way of overcoming problems in modelling due to the presence of marked ceilings and floors in the score range. This also helps with model assumptions of normal errors (Goldstein, 1995). In this paper, with single subject grade score responses experimentation with normalizing transformations did not noticeably improve the error distribution of the data compared to using the raw point score. With a limited number of discrete values the effect of grouping is a likely caveat, but this is present even under transformation. A further point is that effects are more easily interpretable on the raw points score scale and also make comparisons between models for scores and the grades more direct.

A mean centred average GCSE score, GA, is derived from all GCSE subjects of the student with scores A* = 7, A = 6, B = 5, C = 4, D = 3, E = 2. This is used as a prior attainment covariate in modelling. Also used are available student level covariates, gender of the student (females = 1; males = 0) and centred age. The cohort is aged between 18 and 19 years with a mean of 18.5 years. We also introduce the mean of GA (Sch-GA) and standard deviation of GA (Sch-SD) at the level of the institution as possible institution-level effects. Institutions were also formed into 11 categories according to their admission policy and type of funding. Most are publicly funded at the local level as Local Education Authority (LEA) Maintained Schools. Of these, schools in LEAs having no selection by ability at entry (most LEAs) are Comprehensive (M/C). In selective areas schools are usually Selective (M/S) with the rest designated Modern (M/M). There are also Grant Maintained Institutions funded directly from central Government with a similar selection typology according to their local area (GM/C, GM/S, GM/M). Independent schools are privately funded and usually fee-paying and are either Selective or NonSelective (IND/S, IND/NS). Sixth Form Colleges (SFC) and Further Education Colleges (FE) are institutions catering specifically for students beyond the compulsory education age of 16 years and are funded directly by central Government through a funding council. In the main there is a heavier concentration of A-level work in the SFCs. There is a small miscellaneous range of other types (Other). In models dummy explanatory variables were formed with M/C as the base category. The examination boards involved in the study were Associated Examining Board (AEB), Cambridge (Camb), London, Oxford, Joint Matriculation Board (JMB), Oxford–Cambridge joint delegacy (OXCAM), and the Welsh board, WJEC. The latter has only a few entries and did not show obvious difference from AEB in data exploration. Thus these two boards were combined to form the base of dummies for other boards. Fuller details of these variables and their educational context in the UK are given by Yang and Woodhouse (2001).

3 Statistical models for point scores

As a base for evaluating further developments we can formulate a standard variance components model for points with institutions at level 2 and students at level 1:

$$y_{ij} = \beta_0 + u_{0j} + e_{ij} \tag{3.1}$$

Here y_{ij} denotes the UCAS scored response for the grade of an A-level subject offered by the ith student from the jth school. The term u_{0j} is the jth institution random effect and assumed $\sim N(0, \sigma^2_{u0})$. The within-institution student level disturbance is $e_{ij} \sim N(0, \sigma^2_e)$. We note an implicit normality assumption for the response which further means it is assumed continuous. For our grade-scored data this is strictly untenable but may sometimes be assumed to hold approximately for the arbitrary scale on which the points are located.

We can now add to the model covariates such as are introduced in the previous section. We can allow covariates that are polynomial terms in continuous variables, interactions between main factors, and so on. We write

$$y_{ij} = \beta_0 + \mathbf{X}_{ij}\boldsymbol{\beta} + u_{0j} + e_{ij} \tag{3.2}$$

Here $\boldsymbol{\beta}$ is a vector of fixed-effect coefficients associated with such factors and covariates in the data vector $\mathbf{X}_{ij}$. Goldstein (1995) gives terminology and details of fitting such types of model.

As other researchers have shown, we need in general to fit random coefficients models to adequately describe institution-level variation (O'Donoghue *et al.*, 1996; Goldstein and Spiegelhalter, 1996; Yang and Woodhouse, 2001). Extending by these means we have a model of the form

$$y_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{u}_j + e_{ij}$$

with

$$\mathbf{X}_{ij} = \{1, x_{1ij}, \dots\}, \quad \mathbf{Z}_{ij} = \{1, z_{1ij}, \dots\} \tag{3.3}$$

$\boldsymbol{\beta}^T$ is now $\{\beta_0, \beta_1, \dots\}$, with $\mathbf{u}_j^T = \{u_{0j}, u_{1j}, \dots\}$.

Usually, but not always, most of the Z variables are a subset of the X variables. The elements of $\mathbf{u}_j^T$ are random variables at the institution level and are assumed dependent multivariate normal with expectation zero.

In the following analyses variants of models of types (3.1), (3.2), (3.3) are developed for A-level Chemistry and Geography point scores separately.

4 Multilevel models for ordered categories

We now exposit parallel formulations modelling ordered grade responses directly without reference to explicit scoring scales. The six categories of response A–F are

denoted by integer labels $s = 1, 2, 3, 4, 5, 6$. Following the single level model methods of McCullagh and Nelder (1989) we use generalized linear models with cumulative probabilities associated with responses as dependent. For the ith student from the jth institution the probability of a grade higher than that represented by s is denoted by $\gamma_{ij}^{(s)}$. We have $0 < \gamma_{ij}^{(1)} < \gamma_{ij}^{(2)} < \cdots < \gamma_{ij}^{(5)} < \gamma_{ij}^{(6)} = 1$. We note that although probabilities for s are cumulated upwards, those of the ordered grades are cumulated downwards. This proves convenient for interpretation. It is the changing nature of this whole probability distribution for individuals in response to fixed and random explanatory effects that we now wish to model. In continuous (normal) response models by contrast we model conditional expectations given the set of these effects.

However, we usually desire models in which effects operate in a linear and additive fashion. A monotonic 'link' transformation of a set of cumulative probabilities on the $[0, 1]$ scale to the real line usually facilitates this in generalized linear models. In general this link transformation can be any inverse distribution function of a continuous variable. In particular the logit (inverse logistic), complementary log–log (inverse Weibull) and probit (inverse normal) are frequently used. Conceptually a set of thresholds or cut-points on this link scale for each individual are determined by the individual's probability distribution over the grades and vice versa (Bock, 1975). Thus in our case a link transformation of $\gamma_{ij}^{(s)}$ $(s = 1, 2, \ldots, 6)$ corresponds to sequential positions on the whole real line $(\alpha_{ij}^{(1)}, \alpha_{ij}^{(2)}, \alpha_{ij}^{(3)}, \alpha_{ij}^{(4)}, \alpha_{ij}^{(5)}, +\infty)$ with $\alpha_{ij}^{(s)}$ constituting thresholds for the grades. Fixed and random effects operate linearly on these thresholds and hence indirectly on the probabilities over the ordered grades. A related conception used in ordinal models by many researchers (for example, McCullagh, 1980; Hedeker and Gibbons, 1994; Fielding, 1999) is through the notion of an unmeasured and arbitrarily scaled latent variable. This is assumed to underlie the ordered grades and varies continuously along the real line. The ordered categories represent contiguous intervals on this variable with unknown but fixed thresholds. The latent response is assumed governed by a linear model, and in our case a multilevel linear model. Different distributional assumptions about the latent variable may be shown to generate particular generalized linear models for ordered categories of the type under consideration. There are some advantages in these ideas since interpretation of results can be made directly on the scale of the latent variable. However, here we shall not be directly concerned with such an interpretation since our principal aim is to compare the different kinds of inferences arising from the normal points and ordinal models. However, Fielding and Yang (1999) further discuss this idea. Also, when we allow more complexity in randomly varying thresholds as we do later in the paper, it is not clear that latent variable interpretations can be easily adapted.

Goldstein (1995) discusses the formulation of these models in a multilevel context. In the main we deal with logit models. Comparable to the base variance components model (3.1) is

$$\text{logit}\{\gamma_{ij}^{(s)}\} = \log\left(\frac{\gamma_{ij}^{(s)}}{1 - \gamma_{ij}^{(s)}}\right) = \alpha_{ij}^{(s)} = \alpha^{(s)} + u_{0j} \qquad (s = 1, 2, 3, 4, 5) \qquad (4.1)$$

A fit to model (4.1) estimates a series of marginal location cut-points conceptually similar to the intercept of model (3.1). Again for the jth educational establishment there

is a single random effect u_{0j}, which is assumed to be $N(0, \sigma^2_{u_0})$ distributed. Individual responses follow a multinomial distribution determined by their set of grade probabilities, although estimation procedures can allow for extra-multinomial variation (Goldstein, 1995). Even when multinomial variation may be assumed there may be advantages in estimator quality by allowing an extra-multinomial parameter to operate (see Yang, 1997; Fielding and Yang, 1999). We have allowed it in our present analyses though its estimate usually suggests multinomial variation is appropriate.

Analogously to model (3.2) we extend the basic model by adding to the model appropriate fixed-effect covariates. We now have

$$\text{logit}(\gamma_{ij}^{(s)}) = \log\left(\frac{\gamma_{ij}^{(s)}}{1-\gamma_{ij}^{(s)}}\right) = \alpha_{ij}^{(s)} = \alpha^{(s)} + \mathbf{X}_{ij}\boldsymbol{\beta} + u_{0j} \tag{4.2}$$

This model possesses the proportional odds property (McCullagh, 1980). For all s the fixed or random effects operate on cumulative odds by constant multiplicative factors. More detailed explanation of this and an illustration of parameter interpretation is given by Yang (2001). A referee of this paper has suggested that as written this model might imply that the sign of a β_k is the reverse of the direction of the effect of a variable on the underlying response. This is usually a consequence of an upward shift in probabilities cumulated upwards on the ordered scale implying a downward shift in the response. This often causes confusion in interpreting results. Attempts to remove this by inserting negative signs before the β_k and random effects have often been suggested, but this may cause further confusion (Fielding, 1999). In our formulation and as defined our cumulation is downward on the grade scale and both these interpretational difficulties are removed. The signs of the coefficients will be the same as the direction of effects on the underlying response. We feel that this may possibly be adopted in standard practice to good effect.

Further, by analogy with the random coefficients model (3.3) we have

$$\text{logit}(\gamma_{ij}^{(s)}) = \log\left(\frac{\gamma_{ij}^{(s)}}{1-\gamma_{ij}^{(s)}}\right) = \alpha_{ij}^{(s)} = \alpha^{(s)} + \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{u}_j \tag{4.3}$$

This has similar normality assumptions about the vector of random components $\mathbf{u}_j$.

5 Comparison of results between the Normal point score and the ordinal models

To fit the normal models we use the iterative generalized least squares (IGLS) procedures of MLwiN (Rasbash *et al.*, 1999). Ordinal model results all use penalized quasi-likelihood in the MLwiN macros MULTICAT (Yang *et al.*, 1998), incorporating the improved second-order procedures (PQL2) of Goldstein and Rasbash (1996). Yang (1997) discusses the validity and statistical properties of these estimators.

In this section we compare parameter estimates and individual institution residual estimates for the normal and ordinal models. Table 2 provides the comparative results for base models (3.1) and (4.1). Tables 3 and 4 give results for two variants of models (3.2) and (4.2). First we adjust for a range of background characteristics of student, institution, and examination board, but exclude the main intake ability characteristic, student GCSE average. Table 4 then adds variables derived from the latter in various ways. In particular polynomial terms in GA are introduced (GA^2, GA^3, GA^4), interactions of these with gender (GA-F, GA^2-F, GA^3-F), and the aggregated institutional level intake score. This is a useful extension since it draws a distinction between control for extraneous factors affecting raw performance and assessing progress using initial intake achievement covariates. This is standard in educational performance research where it is desired to highlight types of control on institutional effects (Willms, 1992). We do not attempt model selection here and include many relevant parameter estimates that on diagnosis are not statistically significant. Our purpose is a broad comparison of the models within frameworks familiar in educational research.

Because the parameters associated with the same variable relate to different scales under the model type comparisons we report as precision measures the standardized *t*-ratios of parameter estimates to estimated standard errors. The broad pattern of covariate effects are the same under normal and ordinal model assumptions. Generally, Ezzett and Whitehead (1991) have commented that major effects will emerge and are relatively insensitive to model formulation, though we may expect size of estimates to differ somewhat. In our case precision measures of regression parameter estimates are very similar between model types. Formal tests on these yield the same inferential conclusions. Impressions from either model type closely agree. No real difference in impact on broad substantive interpretation emerges. Precision of the school-level variance is slightly higher in all cases for ordinal models but the improvement is barely discernable.

We can comment on the broadly similar patterns of covariate effects for the models of Table 3, which do not adjust for intake achievement. In general, better performances come from females and younger students. Compared to the base M/C schools, selective schools of all types have significantly higher achievement. Modern schools have lower performance but this is not statistically discernable for M/M in Chemistry. Sixth form colleges and others (mainly sixth form centres in schools) also perform better but in neither case are results statistically significant for Geography. FE colleges have much lower achievement generally. GM/C and IND/NS schools are not significantly different from their maintained comprehensive (MC) counterparts. The estimates of dummy parameters for boards relative to AEB/WJEC reveal significantly higher average grade performances for OXCAM, JMB, and CAMB Chemistry examinations. Oxford has significantly worse performance in Geography. Other board effects are not significantly different from the base.

The results in Table 4 have the additional prior achievement covariates at both institutional and student levels. Effect estimates in this table thus relate to 'adjusted performance' and relate to progress over the course of A-level study. As before, conclusions from models (3.2b) and (4.2b) are comparable. In both subjects younger students make more progress in addition to having higher general achievement. However, boys now make more progress than girls despite girls being higher achievers,

Table 2 Model estimates for variance component model (3.1) and basic ordinal model (4.1)

Model (3.1)					Model (4.1)				
	Chemistry		Geography			Chemistry		Geography	
Parameter	Estimate	Precision[a]	Estimate	Precision	Parameter	Estimate	Precision	Estimate	Precision
β_0	5.349		5.250		$\alpha^{(1)}$	−1.881		−2.405	
					$\alpha^{(2)}$	−0.668		−0.913	
					$\alpha^{(3)}$	0.248		0.230	
					$\alpha^{(4)}$	1.106		1.274	
					$\alpha^{(5)}$	2.089		2.406	
σ^2_u	2.829	24.90	2.017	24.41	σ^2_u	1.190	25.50	0.995	25.38
σ^2_e	8.507		7.228		Extra-multinomial variation	0.945		0.959	

[a]Precision = estimate/standard error. This measure was not calculated for the intercept in the Normal point score model nor for the thresholds in the ordinal model, or for level 1 parameters, as they relate to noncomparable quantities across the two approaches.

as also noted by O'Donoghue *et al.* (1996). M/S and GM/S selective schools now lose their significant comparative advantage when progress rather than raw performance is the criterion. However, the effect of IND/S is statistically significant in both subjects. M/M, GM/C, and IND/NS schools make no significantly different progress from M/C in either subject. GM/M seem to do worse but the effect is significant only for Geography. Sixth form colleges achieve higher progress in Chemistry than the base M/C type but no longer have the advantage in Geography. FE colleges show significantly lower progress in Geography but this does not carry over for Chemistry. The pattern of board effects for adjusted performance in Chemistry is similar to those on raw performance exhibited in Table 3. However, CAMB now joins Oxford in having significantly worse adjusted performance in Geography. Other board effects are again not distinguishable. All these general substantive findings are similar to those of Yang and Woodhouse (2001) based on aggregate A/AS points scores for the whole database.

6 The nature of GCSE effects

Prior achievement as measured by GCSE results have formed an input into the second group of models. It is this fact that often enables researchers to treat institutional effects as 'adjusted' and form a basis for a 'value-added' criterion. The way it operates in combination with other factors has been likened to the economic concept of an educational production function. This paper and others (Goldstein and Thomas, 1996; O'Donoghue *et al.*, 1996; Yang and Woodhouse, 2001) show that this production function should include many nonlinear terms in the covariates. Thus, in Table 4 polynomial terms of order up to four in prior achievement have been included to allow a possibly necessary fine nonlinear graduation of the response to this variable, particularly at the extremes. Institution context factors such as Sch-GA and Sch-SD

Table 3 Model estimates and precision of estimates for models (3.2) and (4.2) without adjusting for GCSE average score[a]

Model (3.2a)					Model (4.2a)				
	Chemistry		Geography			Chemistry		Geography	
Parameter	Estimate	Precision	Estimate	Precision	Parameter	Estimate	Precision	Estimate	Precision
β_0	4.318		4.571		$\alpha^{(1)}$	−2.511		−2.883	
					$\alpha^{(2)}$	−1.297		−1.382	
					$\alpha^{(3)}$	−0.383		−0.233	
					$\alpha^{(4)}$	0.470		0.814	
					$\alpha^{(5)}$	1.446		1.946	
Female	0.169	4.40	0.553	16.7	Female	0.099	4.30	0.383	17.6
Age	−0.004	−0.82	−0.003	−0.64	Age	−0.003	−1.07	−0.002	−0.63
M/S	1.239	6.49	1.291	7.51	M/S	0.755	6.22	0.869	7.27
M/M	−0.481	−0.86	−0.948	−2.78	M/M	−0.300	−0.87	−0.624	−2.69
GM/C	0.011	0.10	0.086	0.90	GM/C	0.001	0.02	0.073	1.11
GM/S	1.513	9.70	1.288	9.20	GM/S	0.933	9.39	0.885	9.05
GM/M	−2.119	−2.84	−2.321	−5.28	GM/M	−1.296	−2.78	−1.728	−5.59
IND/S	2.187	23.3	1.694	18.7	IND/S	1.407	23.5	1.185	18.9
IND/NS	0.076	0.27	0.214	0.81	IND/NS	0.041	0.24	0.167	0.93
SFC	0.569	3.95	0.109	0.86	SFC	0.343	3.73	0.069	0.78
FE	−0.993	−7.22	−1.012	−8.25	FE	−0.653	−7.55	−0.704	−8.32
Other	1.083	3.30	0.195	0.59	Other	0.603	2.93	0.150	0.66
Camb	0.700	5.11	−0.193	−1.87	Camb	0.441	5.19	−0.122	−1.71
London	0.169	1.31	0.230	2.71	London	0.097	1.20	0.149	2.55
Oxford	−0.136	−0.59	−1.267	−6.54	Oxford	−0.092	−0.64	−0.835	−6.28
JMB	0.725	5.50	0.231	2.17	JMB	0.458	5.55	0.149	2.04
OXCAM	1.085	7.48	0.228	1.61	OXCAM	0.654	7.27	0.123	1.26
σ_u^2	1.662	21.71	1.236	21.50	σ_u^2	0.698	22.83	0.623	22.57
σ_e^2	8.521		7.173		Extra-multinomial variation	0.951		0.959	

[a]See note on precision results in Table 2.

Table 4 Model estimates and precision of estimates for models (3.2) and (4.2) adjusting for GCSE average[a]

Model (3.2b)					Model (4.2b)				
	Chemistry		Geography			Chemistry		Geography	
Parameter	Estimate	Precision	Estimate	Precision	Parameter	Estimate	Precision	Estimate	Precision
β_0	5.272		5.224		$\alpha^{(1)}$	−2.950		−3.614	
					$\alpha^{(2)}$	−1.104		−1.404	
					$\alpha^{(3)}$	0.247		0.276	
					$\alpha^{(4)}$	1.436		1.718	
					$\alpha^{(5)}$	2.704		3.147	
Female	−0.900	−24.7	−0.348	−11.3	Female	−0.720	−23.4	−0.304	−11.1
Age	−0.039	−10.4	−0.026	−7.94	Age	−0.035	−11.7	−0.025	−8.61
M/S	−0.033	−0.23	0.060	0.44	M/S	−0.065	−0.50	0.069	0.53
M/M	0.521	1.25	−0.127	−0.49	M/M	0.395	1.10	−0.057	−0.23
GM/C	−0.019	−0.22	0.085	1.17	GM/C	−0.017	−0.23	0.099	1.43
GM/S	0.114	0.95	0.023	0.21	GM/S	0.069	0.65	0.038	0.36
GM/M	−1.366	−2.45	−1.088	−3.23	GM/M	−1.251	−2.57	−1.174	−3.58
IND/S	0.266	3.24	0.239	3.05	IND/S	0.242	3.08	0.247	3.34
IND/NS	−0.161	−0.77	−0.040	−0.20	IND/NS	−0.147	−0.82	−0.051	−0.27
SFC	0.477	4.44	−0.010	−0.10	SFC	0.402	4.38	−0.025	−0.27
FE	−0.159	−1.53	−0.445	−4.69	FE	−0.179	−1.96	−0.414	−4.60
Other	0.699	2.86	−0.329	−1.31	Other	0.577	2.70	−0.361	−1.51
Camb.	0.628	6.16	−0.400	−5.06	Camb.	0.579	6.51	−0.358	−4.77
London	−0.005	−0.05	0.091	1.40	London	0.005	0.06	0.084	1.35
Oxford	−0.303	−1.76	−1.397	−9.44	Oxford	−0.212	−1.41	−1.281	−9.22
JMB	0.476	4.83	−0.036	−0.44	JMB	0.460	5.34	−0.046	−0.60
OxCam	1.207	11.2	−0.014	−0.13	OxCam	1.054	11.2	−0.044	−0.42
GA	3.309	79.2	2.808	93.9	GA	2.600	65.2	2.437	77.6
GA^2	0.255	7.31	0.236	7.61	GA^2	0.484	15.1	0.377	12.6
GA^3	−0.404	−17.0	−0.219	−17.5	GA^3	−0.035	−1.32	−0.047	−3.36
GA^4	−0.099	−9.61	−0.042	−5.45	GA^4	−0.013	−1.30	−0.007	−0.92
GA-F	−0.048	−0.86	0.088	2.87	GA-F	−0.201	−4.24	0.049	1.69
GA^2-F	0.275	7.45	0.066	2.48	GA^2 − F	0.114	2.93	0.028	1.08
GA^3-F	0.087	3.38	N/A	N/A	GA^3 − F	0.076	2.97	N/A	N/A
Sch-GA	0.175	2.82	0.108	1.86	Sch-GA	0.151	2.76	0.094	1.71
Sch-SD	0.210	1.71	0.329	2.79	Sch-SD	0.146	1.38	0.333	3.03
σ^2_u	0.910	21.52	0.733	21.77	σ^2_u	0.754	22.85	0.705	23.11
σ^2_e	4.820		4.112		Extra-multinomial variance	0.928		0.945	

[a]See note on precision results in Table 2.

will also often have a discernible influence. Sch-D may, for instance, be useful in examining what, if any, is the impact of homogeneity of school intake on progress. The gender differentiation of GCSE effects is represented by terms for the interaction with the female gender dummy variable.

As indicated by the *t*-ratio precision measures in Table 4 the normal model evidenced significant polynomial terms in GCSE average up to the fourth order. The ordinal logit model required only a quadratic function. As a result of polynomial effects, the nature of the interaction of gender with prior ability cannot be simply expressed as a single additive element. However, the normal model exhibited interactions for both subjects but the ordinal model only for Chemistry. This may be due to the skewed nature of its response distribution over grades or could be related to differential impacts of ceilings and floors. In the context of primary school progress, Fielding (1999) notes that ordered category models seem to be more parsimonious in many circumstances and require fewer complex fixed-effect terms. It is further noted there that this requirement may be conditioned by the response distributions. The results here seem to confirm these impressions. The normal model seems much more sensitive to the actual values of GCSE scores than the logit model.

In both types of model the effects of school intake contexts appear relatively small, but with Sch-GA having a marginally significant effect in Chemistry and within-school heterogeneity (Sch-SD) having a positive effect in Geography.

7 The use of ordinal models in predicting grade distributions

In A-level work teachers are expected to predict A-level performances for university entrance purposes. Normal models give score predictions that may be difficult to relate to grades. A more useful approach might be to evaluate the 'chances' that a certain student will achieve certain grade thresholds. Ordinal models have a very useful role in this area by predicting directly the probability that students with given background characteristics and initial ability will achieve certain grades. Normal models can only do this indirectly from the conditional means and variances and assumptions that grade boundaries are placed appropriately along the points scale (5.0–7.0, for instance, for grade C). Using estimates from Model (4.2b) in Table 4, we illustrate in Figure 1 the 'chances' of achievement estimated for two female students having the same set of background covariate values but different GCSE average scores. Student 1, with a high GCSE score of 7.5, has a very high chance of achieving a Chemistry grade no less than B, while Student 2, with a GCSE score of 5, is most likely to achieve a grade no higher than D. Estimates from Model (3.2b) in Table 4 for these two students give A-level point score predictions of 9.8 and 1.9, respectively. Although they are roughly equivalent to grades A and E, they represent conditional expectations only. Using individual level variance estimates a predicted grade distribution could also be calculated from the assumed normal (for example, probability of grade C would be found from the area between 5 and 7). However, assumed normality on the underlying arbitrary raw points scale is crucial when applied in this context. Model estimates may be reasonably robust to departures from the normality assumption. Interval predictions

may not be quite so robust. Even relatively small departures from normality in the true distribution over the raw points scale might yield quite different predictions. This is a further aspect of sensitivity to the arbitrary score scale and normal assumptions over it. No such fixing of a scale or constraining distributional assumptions are required for ordinal models.

8 Value added estimates using school residuals

We have also incorporated random coefficients in the models as in (3.3) and (4.3). Detailed results are not illustrated here. However, significant random coefficients at the institutional level for both model types and subjects were the female gender and the first order GA term. Thus we now have three random effects for each institution, which could be estimated by MLwiN residual procedures. Models (3.3) and (4.3) in our discussion and diagram below relate to fits of models with these two extra random effects added to the models of Table 4. However, we focus on the intercept residual estimates only. Since GA was centred these represent institution average adjusted effects or 'value added,' as they are often termed in the educational literature. They relate to males with an average GCSE score, having allowed in the model for possible differential institutional effects on students of different gender and prior achievement. This approach gave us a set of homogenous institution residual estimates that enable us to investigate mild changes of assessment of institutional 'added value' between the model approaches. We also diagnostically checked certain model assumptions for ensuring the comparability between models. The distributions of the standardized institution inter-

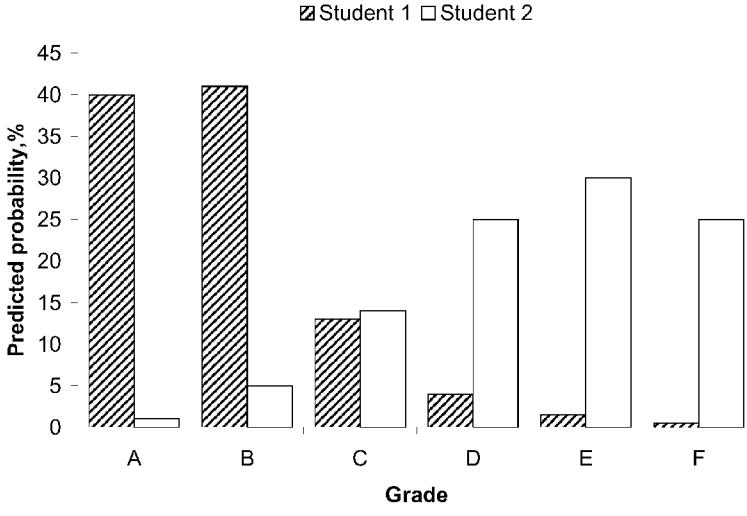

Figure 1 Predicted distribution of A-level grades for Chemistry for Ordinal model (4.2b) for two students (Student 1: female, 18.5 years old, from an independent selective school, with examination board Oxford–Cambridge, and overall GCSE score 7.5. Student 2: same as Student 1, but a lower GCSE average score of 5)

cept residuals for the two models on both A-level subjects are very close to normality as seen by the normal plots in Figures 2 and 3. Residuals from the model approaches are closely related. The correlation coefficients and rank correlations between the institution residual estimates from each pair of models are given in Table 5. Note that these correlations are somewhat inflated because of 'shrinkage' factors. Nevertheless, they and inspection of the residual scatter plots, not illustrated here, suggest a strong agreement between model types for institution 'value added' estimates. Given the fairly complex and full modelling of covariates and effects we would usually not expect otherwise. However, they are not perfect and there is scope for some movement in the positioning of individual institutions. Correlations can be relatively insensitive to these. Even mild sensitivity to model formulation is potentially of substantive interest. In particular, we might investigate any dependency on models of the identification of institutions at the extremes of the range of effects.

Thus we now examine in detail some selected institutions. We choose four for each subject. Two of these are in the middle of the distribution of the institution residual estimates for Chemistry normal models. These are also examined for Geography. Further, for each subject separately, two extreme institutions are selected. Some details on the selected institutes are listed in Table 6. Table 7 shows the ranks of the residuals of the selected institutions in each model, the residual estimates, and their standard errors. Also shown are 95% overlap intervals (Goldstein and Healy, 1995) converted into equivalent intervals of ranks.

The results show that extreme schools are detected with either model. In the middle of the distribution, however, there are often considerable differences in rankings. There

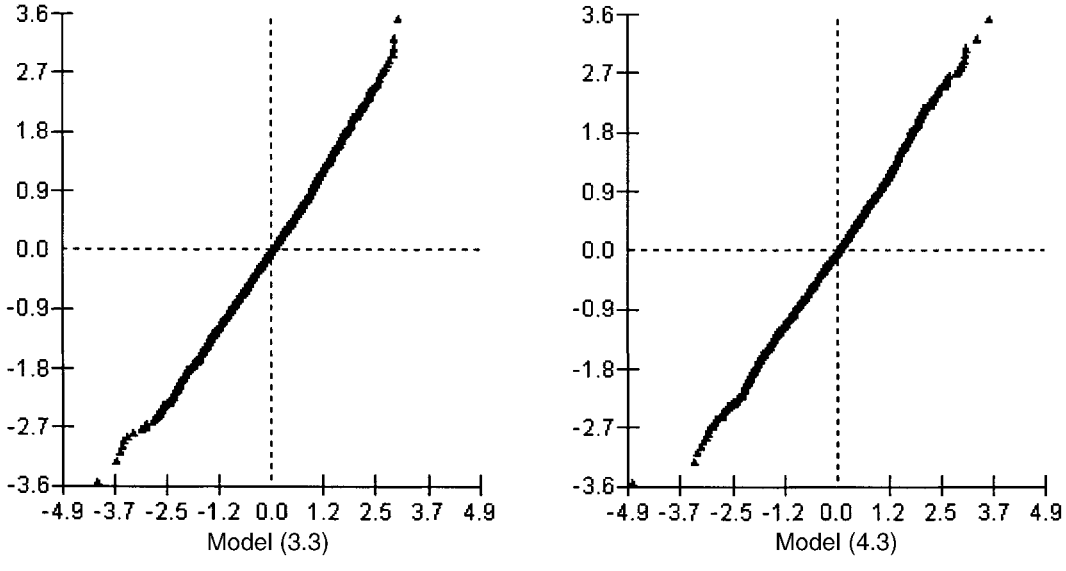

Figure 2 Normal score (*y*-axis) by standardized residual (*x*-axis) for the Normal model (3.3) and ordinal model (4.3) for Chemistry

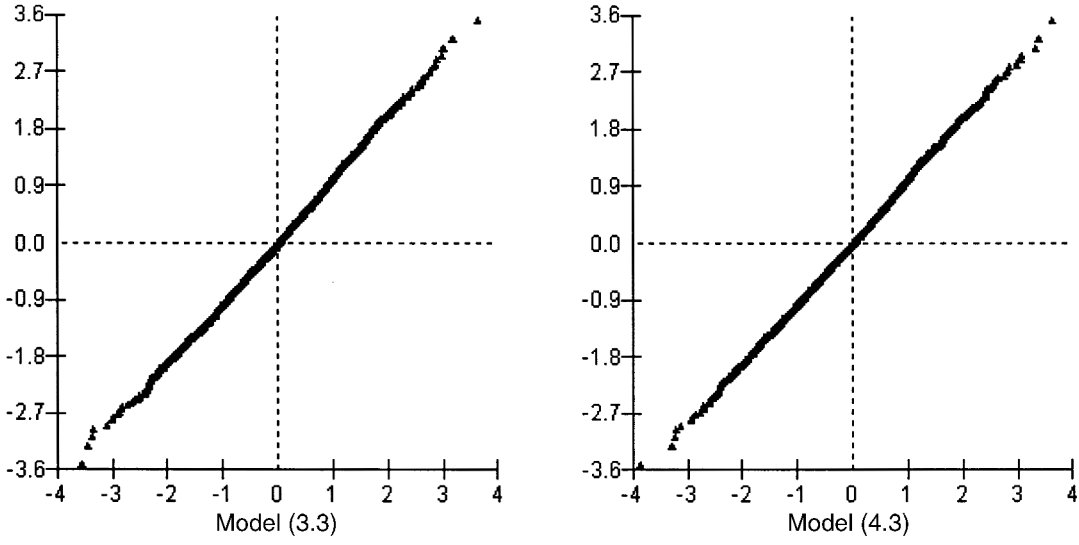

Figure 3 Normal score (*y*-axis) by standardized residual (*x*-axis) for the Normal model (3.3) and ordinal model (4.3) for Geography

is important sensitivity of 'league table' position to the chosen adjustment model even when both include the same covariates. The intervals for institutions 1 and 2 on Chemistry, whose ranks are only 66 apart (1317 – 1251), overlap considerably for the normal model. However, even in the ordinal model for this case where rankings are about 1000 apart (1706 – 702), and much more clearly separated, there is still considerable overlap of the intervals. For Geography, institutions 1 and 2, being less concentrated in the middle of the rank range, have separable intervals under both models, but less clearly so for the ordinal model. However, it requires rankings differing by 1502 and 1472, respectively, to achieve these separations. Although the extreme institutions have a much more consistent ranking there are one or two other features worthy of note. For Geography, institution 5 (highest ranked on the ordinal model) and institution 2 have relatively short rank differences of 111 and 300 for normal and ordinal models, respectively. However, at the top end of the range their intervals do not overlap for the ordinal model and only just for the normal. Rank intervals for extreme cases are very short. There appears to be a clearer separation between pairs as we move away from the middle of the distribution. The same phenomenon occurs at the lower end.

Table 5 Correlation coefficients and rank correlations between institution residual estimates

	Chemistry	Geography
Establishment residuals	0.982	0.968
Ranks of establishment residuals	0.983	0.970

Table 6 Selected institutes with institutional characteristics

Instn.	Number of A level entries by gender (male : female)		Exam board	Instn. type	Mean GSCE point score of individual students		Mean A-level point score of individual students	
	Chemistry	Geography			Chemistry	Geography	Chemistry	Geography
1	19 : 28	29 : 29	CAMB	6th Form	5.84	5.74	5.19	4.93
2	0 : 14	0 : 30	OXCAM	IND/S	7.46	6.78	9.71	7.93
3	6 : 0		AMB	IND/S	6.79		3.00	
4	14 : 12		CAMB	G/C	5.28		6.54	
5		45 : 27	CAMB	6th Form		5.70		6.75
6		17 : 21	London	6th Form		5.33		1.42

9 Extensions of the ordinal model

So far the ordinal logit models have the proportional odds feature implied by fixed cut-point thresholds not varying across observations. However, more flexibility can be introduced by allowing interactions of thresholds with covariates or allowing random thresholds effects across level 2. For example, as it stands, the fitted ordinal model in Table 4 suggests that the additive main effect of gender on the cumulative log-odds is the same across grades. Covariate changes shift only the location of the entire grade distribution keeping the relative odds proportional. We can relax this by interacting covariates with cut-points. For instance, interacting with gender means the cut-points for males and females are no longer separated by a single additive gender effect, and gender can affect cumulative odds nonproportionally across grades. Indeed, there is some preliminary evidence that a proportional odds gender effect may not be tenable. In Table 1 in Geography, for instance, more female students achieved the top two grades than males and vice versa for the bottom two grades. The distributional differences between genders may be more complex than a constant shift in cumulative log-odds. The suggested nonproportional extensions can be fitted fairly readily by adaptation of the quasi-likelihood procedures in the MLwiN MULTICAT macros that we use.

9.1 Model with nonproportional changing odds

Model (9.1) below extends the fixed part of Model (4.2b) in these directions. We focus on nonproportional gender effects because these have evoked our interest. For ease of exposition we revert to the model with a single variance at level 2. We have investigated extensions to Model (4.3) with little difference of substance to the arguments we present. This type of model has been called a multilevel thresholds of change model (MTCM) by Hedeker and Mermelstein (1998).

Table 7 School value added estimates for the Normal and ordinal models

	Rank of residuals		Residual estimate (S.E.)		Ranks corresponding to 95% overlap intervals of residuals	
School	Normal model (3.3)	Ordinal model (4.3)	Normal model (3.3)	Ordinal model (4.3)	Normal model (3.3)	Ordinal model (4.3)
Chemistry						
1	1251	1706	−0.02 (0.38)	−0.35 (0.26)	1209 ~ 2020	1209 ~ 2068
2	1317	702	−0.07 (0.58)	0.37 (0.69)	71 ~ 1940	50 ~ 1971
3	2406	2405	−2.11 (0.65)	−2.08 (0.56)	2324 ~ 2408	2347 ~ 2409
4	1	2	2.75 (0.43)	2.31 (0.46)	1 ~ 23	1 ~ 11
Geography						
1	1615	1773	−0.30 (0.30)	−0.46 (0.30)	985 ~ 2045	1226 ~ 2133
2	113	301	1.04 (0.50)	0.69 (0.51)	6 ~ 663	46 ~ 1198
5	2	1	2.09 (0.27)	2.69 (0.31)	1 ~ 10	1 ~ 4
6	2316	2317	−2.28 (0.36)	−2.94 (0.39)	2308 ~ 2317	2316 ~ 2317

Letting t_{ij} be 1 if the ith person from institute j is female, we write

$$\text{logit}(\gamma_{ij}^{(s)}) = \alpha_{ij}^{(s)} = \alpha^{(s)} + \omega^{(s)} t_{ij} + \mathbf{X}_{ij}\boldsymbol{\beta} + u_{0j} \tag{9.1}$$

Estimates of the cut-points $\alpha^{(s)}$ determine the cumulative grade distribution of males conditional on other explanatory variables, and estimates of $(\alpha^{(s)} + \omega^{(s)})$ those of females. Similar terms could be introduced for other explanatory covariates and higher order interactions are also possible. McCullagh and Nelder (1989, p. 155) and also Hedeker and Mermelstein (1998) comment that for continuous covariates this may unfortunately lead to negative fitted values for some values of covariates. In our case we have checked that this would not occur inside the observed range if we were to entertain nonproportional odds for covariates in our data such as GA.

The results of fitting model (9.1) are displayed in Table 8. They suggest definite interactions of cut-points with gender and hence a nonproportional effect of gender on cumulative odds.

The ratios of cumulative odds of males to females at each threshold are displayed in Figure 4 and contrasted with the constant proportional odds ratios of Model (4.2b). It is clearly seen that the overall negative effect of females estimated by Model (4.2b) on Chemistry was mainly because female students achieve relatively few high grades, having adjusted for their GCSE average score. For Geography there is a different pattern with relatively more females in middle grades and slightly more failures than a proportionality assumption would warrant. Differences in skewness of the distributions of grades in the two subjects may play a role and the importance of these is underlined by this type of ordinal model.

9.2 Random institution effects on cut-points for the distribution over grades

In the same way that we consider odds changing nonproportionally for different values of covariates we can allow nonproportionality of the random institution effect. This is achieved by allowing the cut-points to vary randomly across institutions. Thus we now generalize model (9.1), which had a single random effect, by incorporating a set of grade specific cut-points $(\alpha^{(s)} + u_{0j}^{(s)}, s = 1, 2, 3, 4, 5)$ for each institution. The model is

$$\text{logit}(\gamma_{ij}^{(s)}) = \alpha_{ij}^{(s)} = \alpha^{(s)} + \omega^{(s)} t_{ij} + \mathbf{X}_{ij}\boldsymbol{\beta} + u_{0j}^{(s)} \tag{9.2}$$

Here, $\boldsymbol{u}_{0j} = \{u_{0j}^{(1)}, u_{0j}^{(2)}, \ldots, u_{0j}^{(5)}\}' \sim MVN(\mathbf{0}, \boldsymbol{\Omega}_{u_0})$. The $\boldsymbol{\Omega}_{u_0}$ is a (5×5) covariance matrix of the separate random effects. For simplicity we also assume that the interacting gender coefficients are fixed, that is, there is no differential institutional effect by gender.

The estimates for Model (9.2) are also given in Table 8. The institution random effect parameters are shown separately in the lower part of Table 9. The fixed part results show main effects similar to those estimated by Model (9.1) for both Chemistry and Geography but there are some changes to the base (male) and female cut-points. From the random parameter estimates we see that there is relatively more institutional variation at grade A and F thresholds in Geography. For Chemistry the F threshold

Table 8 Parameter estimates for Models (9.1) and (9.2) by subject[a]

	Model (9.1)				Model (9.2)			
	Chemistry		Geography		Chemistry		Geography	
Variable	Estimate	Precision	Estimate	Precision	Estimate	Precision	Estimate	Precision
$\alpha^{(1)}$	−2.899	−26.1	−3.551	−34.8	−2.775	−32.3	−3.600	−35.0
$\alpha^{(2)}$	−1.084	−9.85	−1.401	−14.4	−0.955	−11.4	−1.344	−13.9
$\alpha^{(3)}$	0.228	2.07	0.261	2.69	0.334	3.98	0.318	3.28
$\alpha^{(4)}$	1.384	12.6	1.710	17.4	1.490	17.5	1.789	18.3
$\alpha^{(5)}$	2.598	23.2	3.158	31.3	2.764	31.4	3.340	32.7
$\omega^{(1)}$	−0.856	−17.8	−0.434	−8.04	−0.841	−17.9	−0.408	−7.56
$\omega^{(2)}$	−0.786	−21.8	−0.312	−8.91	−0.775	−22.8	−0.304	−8.00
$\omega^{(3)}$	−0.697	−19.4	−0.271	−8.47	−0.695	−20.4	−0.269	−8.68
$\omega^{(4)}$	−0.621	−15.2	−0.289	−7.61	−0.623	−16.0	−0.278	−7.51
$\omega^{(5)}$	−0.488	−9.38	−0.333	−6.53	−0.476	−9.33	−0.310	−6.08
Age	−0.036	−12.0	−0.025	−8.33	−0.036	−12.0	−0.026	−8.67
M/S	−0.064	−0.50	0.069	0.53	−0.078	−0.62	0.089	0.68
M/M	0.395	1.10	−0.057	−0.23	0.350	0.98	−0.077	−0.31
GM/C	−0.017	−0.23	0.099	1.43	−0.012	−0.17	0.099	1.43
GM/S	0.070	0.65	0.039	0.36	0.078	0.75	0.022	0.21
GM/M	−1.258	−2.57	−1.176	−3.57	−1.319	−2.63	−1.125	−3.27
IND/S	0.241	3.35	0.248	3.35	0.244	3.44	0.251	3.39
IND/NS	−0.149	−0.83	−0.051	−0.27	−0.149	−0.84	−0.044	−0.24
SFC	0.403	4.20	−0.025	−0.27	0.434	4.67	−0.012	−0.13
FE	−0.182	−1.98	−0.413	−4.59	−0.115	−1.25	−0.368	−4.04
Other	0.583	2.71	−0.361	−1.50	0.631	2.99	−0.345	−1.43
Camb.	0.581	6.53	−0.359	−4.79	0.547	6.22	−0.444	−5.92
London	0.006	0.07	0.084	1.35	−0.019	−0.23	0.033	0.53
Oxford	−0.215	−1.42	−1.283	−9.23	−0.223	−1.49	−1.429	−10.1
JMB	0.460	5.35	−0.047	−0.61	0.454	5.28	−0.121	−1.57
OXCAM	1.054	11.2	−0.045	−0.43	1.026	11.0	−0.077	−0.74
GA	2.545	62.1	2.417	73.2	2.497	89.2	2.410	83.1
GA^2	0.487	15.2	0.367	12.2	0.459	27.0	0.371	18.6
GA^3	−0.029	−1.12	−0.047	−3.35	n/a	n/s	−0.023	−2.30
GA-F	−0.070	−1.32	0.089	2.41	−0.052	−1.08	0.090	2.57
GA^2-F	0.111	2.78	0.055	1.96	0.104	3.25	0.043	1.59
GA^3-F	0.064	2.46	n/a	n/a	0.055	2.89	n/a	n/a
Sch-GA	0.150	2.73	0.094	1.71	0.119	2.16	0.105	1.88
Sch-SD	0.144	1.36	0.333	3.03	n/a	n/a	0.323	2.94
σ_u^2	0.759	23.0	0.706	22.8	School level variance–covariance, see Table 9			
Extra-multinomial variation	0.928	St. err = 0.003	0.946	St. err = 0.004	0.871	St. err = 0.003	0.853	St. err = 0.003

[a]See note in Table 2 on definition of precision measure.

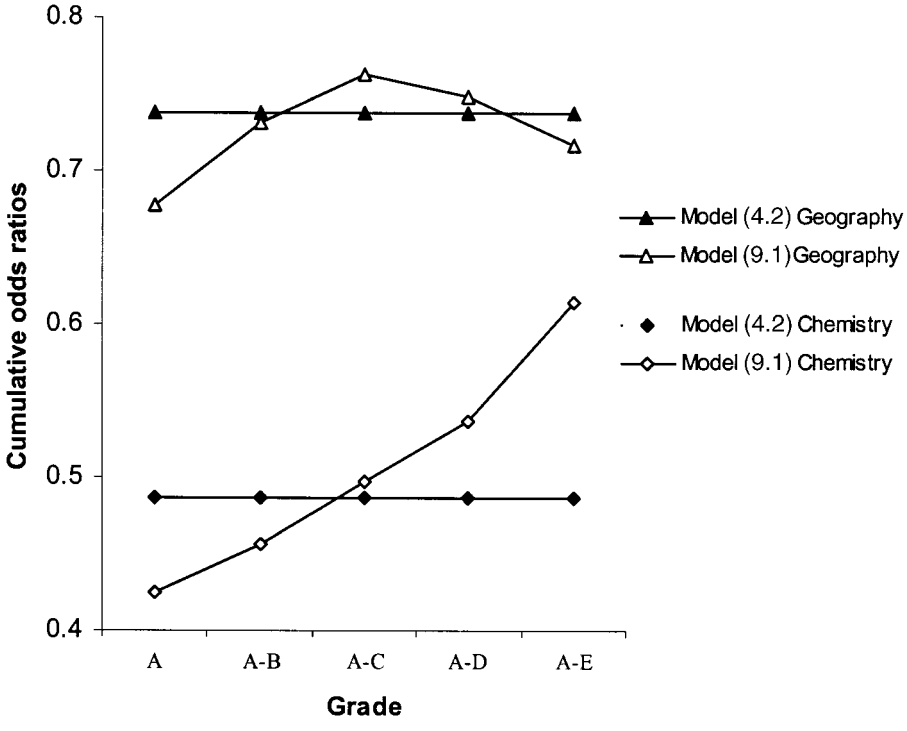

Figure 4 Gender effects on grade threshold probabilities: ratios of cumulative odds between females and males estimated by Models (4.2) and (9.1)

parameter exhibits most variation. Such sources of institutional variation at crucial thresholds may be potentially of more substantive interest then overall average levels of adjusted performance.

Detailed diagnostic normal plots of all estimated standardized residuals from Model (9.2), though not illustrated, showed good agreement with the normality assumption for both subjects.

For the same institutions previously investigated in Section 8, estimates of the full set of cut-point residuals in Model (9.2) are illustrated in Figures 5(a) and (b). These show a relatively constant pattern of effects over the grades on Chemistry for Institutions 2, 3, and 4 and are not much different from results observed in Table 7. These institutions could be interpreted as having relatively uniform effect on students across all levels of ability. Institution 1 has a below average conditional expectation of achieving at least either of the top two thresholds, about the same as average for above grade C, and above average for the proportion not failing or achieving above grade D. It would be interesting to examine the practice at this institution, which seems to have a better than expected overall pass rate but a relatively lower than expected achievement at the top end of the grading. This pattern is further displayed in detail for males in Figure 6(a), which contrasts the predicted A-level grades in Chemistry for typical males in Institution 1 with similar males in the base group of students. It will be noted that compared to

the similar base group there appear relatively fewer in the bottom three categories but much higher proportions in Grades B and C.

For Geography the effect of Institutions 1 and 2 are approximately proportional across grade thresholds. Institutions 5 and 6, occupying the highest and lowest positions in Table 7, have a profile of threshold effects that are parallel and consequently with a similar relative pattern but at different absolute levels. Given their general levels, the size of their effects declines relative to all institutions as we move through the grade thresholds. They are relatively rather better at targeting top grades than they are at getting students above low thresholds and passing. Institution 6, for instance, is not too far from average in its effect on top grade chances (and better than Institution 1) but its low overall position is a result of 'deficiencies' at lower thresholds. Figure 6(b) presents the predicted distributions of A-level Geography grades for males of mean age with mean GCSE in Institution 5 at the top of the scale and Institution 6 at the bottom. Although as expected Institution 5 has a much higher proportion of Grade As, there is little difference between Institution 6 and that of the base group in this respect. The impact of failures on the overall position of Institution 6 is obvious from this diagram. There is a further important general caveat for predictions for particular institutions. These use residual estimates which have uncertainty and require some caution as stressed by Goldstein and Spiegelhalter (1996). Often the residuals are based on relatively small numbers of students so that the standard errors of estimates can be quite large. Thus, for example, in the present comparisons in Geography, Institution 1 has a standard error for the grade B cut-point of 0.299. Conditional on the fixed-point estimates a 95% confidence interval for the logit can be constructed. Converting to the cumulative probabilities gives an approximate interval of 28.1 to 56.1% for above grade B. In principle, for more detailed analysis, intervals can be constructed for overall predictions of full grade distributions for each institution.

10 Discussion

In this paper we have demonstrated a flexible range of models for educational grades treated as ordered responses. The operational definition of the outcome variable is at no

Table 9 Variance–covariance estimates of the cut-points for Chemistry (first line) and Geography (second line) at school level, SE in parentheses, correlation coefficients in the upper triangle of the table

	A	Above B	Above C	Above D	Above E
A	0.806 (0.047) 1.157 (0.068)	0.92 0.88	0.83 0.70	0.78 0.56	0.48 0.61
Above B	0.699 (0.037) 0.803 (0.043)	0.714 (0.037) 0.725 (0.037)	0.94 0.92	0.88 0.80	0.63 0.81
Above C	0.659 (0.036) 0.658 (0.040)	0.701 (0.035) 0.683 (0.033)	0.785 (0.039) 0.755 (0.036)	0.97 0.95	0.80 0.96
Above D	0.658 (0.038) 0.549 (0.043)	0.700 (0.035) 0.630 (0.033)	0.807 (0.039) 0.770 (0.036)	0.883 (0.045) 0.863 (0.043)	0.92 0.99
Above E	0.479 (0.055) 0.716 (0.045)	0.594 (0.041) 0.758 (0.041)	0.791 (0.043) 0.908 (0.044)	0.965 (0.050) 1.002 (0.050)	1.236 (0.070) 1.194 (0.066)

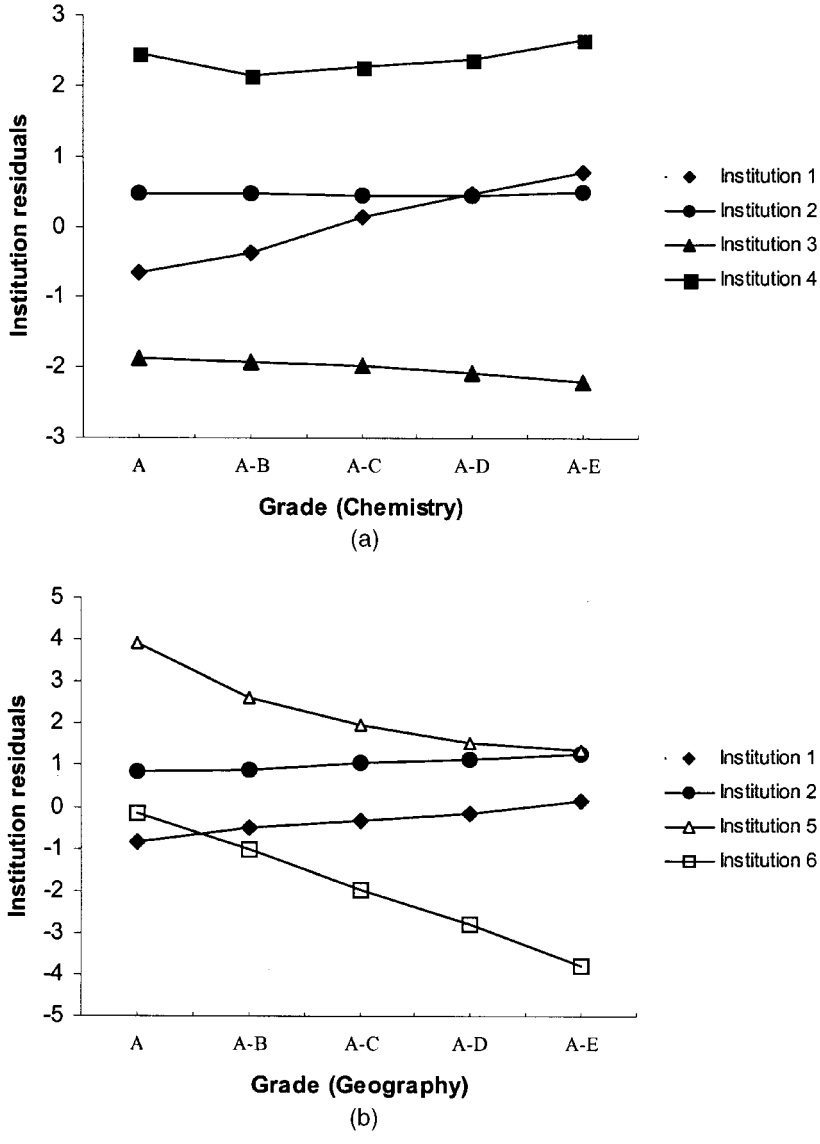

Figure 5 Plots of residuals of four institutions for (a) A-level Chemistry, logit scale and (b) A-level Geography, logit scale

higher a level of measurement than this. Assumptions of continuous response multilevel models with scores may thus be inappropriate, particularly since there are few scored grades. Statistical objections have ranged from those about the scaling implied by arbitrariness in scoring through to continuous distribution properties applied to discrete measurements and to bias in estimation due to grouping. A review of some literature on

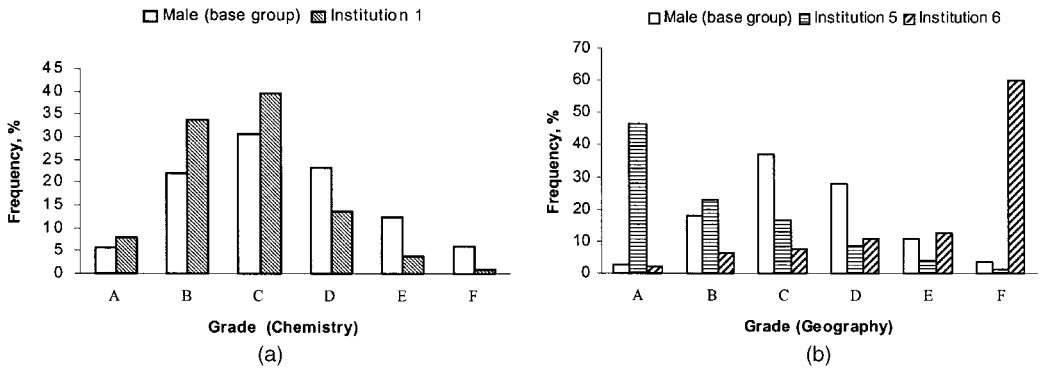

Figure 6 Predicted distributions of (a) A-level Chemistry for males of mean age with mean GCSE score: overall base group of maintained comprehensive schools with AEB board compared with Institution 1 (a sixth form college and Cambridge Board) and (b) A-level Geography for males of mean age with mean GCSE score: overall base group of maintained comprehensive schools with AEB board compared with Institution 5 (a sixth form college and Cambridge Board) and Institution 6 (a sixth form college with London Board)

this is given by Fielding (1999) and a recent general contribution by Kampen and Swyngedouw (2000).

From a practical and substantive view Normal linear models are certainly useful and have the virtue of familiarity with very accessible methodology and software. Certainly, this paper shows that the general scale of fixed effects is relatively insensitive to model formulation. Estimated precision of the fixed-effect estimates are also comparable, although little is known about how the estimated precision is affected by the discrete nature of the observed data. However, for institution-specific details there are considerable differences between the models. In this respect it may be argued that ordinal models make fewer restrictive assumptions about the response distribution and provide conclusions that are less open to substantive query. Although the intercept residuals from models (3.3) and (4.3) are highly correlated, the location of specific institutions in their range can vary greatly. For instance, Institution 2 for Chemistry is at the 29th percentile of ranked residuals on the ordinal model and at the 55th percentile for the Normal model (Table 7). It is true that these positions are both subject to uncertainty but the question arises as to the appropriateness of the modelling if we want to draw substantive 'value added' conclusions.

From a practical point of view in educational research the ordinal models also offer as much information as do normal models, and it could be argued more. The 'newness' of ordinal models and (until recently) lack of suitable software may have acted as a practical deterrent, but this is being remedied. The ability to convey predictive information through probability distributions, which cannot easily be done using standard models, is a particular advantage. Because grades and levels are standard modes of reporting, it may obviously be useful to relate the interpretations of results to these. A predicted point score for an individual, even when contextualized in terms of the conditional mean of a continuous distribution, has less ready an interpretation when grades and levels are the medium of converse. The implications of the use of ordinal models in such practical areas as target setting within schools may be clear.

Ordinal models also seem to be capable of extension in substantively useful ways. Their characterization in terms of grade probability responses permits flexible parameterization for a variety of conditions. Our analyses have mainly been concerned to comment on the practical significance of this. However, these additional complexities also considerably improve the fit of models. A simultaneous Wald test (available in MLwiN) on parameters $\omega^{(s)}$, the interactions between the gender variable and the cut-points in the fixed part of (9.1) yielded significant $\chi^2_5 = 132.8$ for Geography and $\chi^2_5 = 580.8$ for Chemistry. Secondly, a restriction that all 15 variance and covariance parameters of the separate random effects at the institution level in (9.2) are equal to a common parameter value reduces it to the single effect model of (9.1). An approximate Wald test on this yields very significant $\chi^2_{14} = 164.9$ for Geography and $\chi^2_{14} = 261.7$ for Chemistry. From a practical view we have seen, for instance, that by allowing cut-points to interact with gender we can study gender differences in distributions in greater detail. As discussed these differences go further than simply differences in average performance. By allowing random cut-points institutional differences can also be exhibited in more meaningful ways. They can be compared at important thresholds rather than through simple mean levels of adjusted achievements. Thus, Institution 1 in Chemistry has lower grade A achievements than expected but it also has lower failures. Institution 6 in Geography has a considerable failure rate but does quite well in achieving high grades compared to other typical institutions. Differences between institutions in such respects might well engage the interest of effectiveness researchers and policy makers as of much if not more relevance than differences in 'average' achievement or progress.

An aspect of the ordinal model that we have not discussed in any detail is the nature of the link function. We have focused on the familiar logit. However, we have also carried out some investigations using a probit link, which will be available in the latest issue of MULTICAT. A probit link is often interpreted in terms of normally distributed latent variable. Thus it might seem to fit more easily into comparisons with Normal linear models. There is a conventional wisdom in the generalized linear modelling literature (for example, McCullagh and Nelder, 1989; Greene, 2000) that important results are relatively insensitive to this choice of link. This is often attributed to the similarity of the logistic and normal distributions except at the tails. Preliminary results show some differences but none are startling. However, methodological work comparing logit, probit, and other links in the multilevel context is under way and needs further advancing.

If, as we claim, ordinal models are worthy of more extensive application they need also to be developed further in a number of important directions. Ordinal models with cross-classified random effects at higher levels have been considered by Fielding and Yang (1999). Multivariate response multilevel models for continuous variables are developed and quite widely used in education (Goldstein and Sammons, 1997; Yang *et al.*, 2001). In our investigations with Model (9.1) we compared the two sets of cut-point residuals for institutions that had Geography and Chemistry in common. A general impression conveyed was that there were two major groups of institutions. One group was those institutions whose effects for the two subjects were similar relative to all schools. However, another major group had relatively high 'adjusted' performances in one subject together with a low achieve-

ment in the other. There are some interesting practical questions here about the differential 'effectiveness' of schools in different A-level subjects and the relationships between subject grades at both student and institutional level. Multivariate ordinal response multilevel model developments are required for this. Their characterization is not so easy as the analogous continuous variable specifying normal correlation structures. Nor would their estimation be as easily adaptable from standard available procedures. Promising lines of inquiry, which we have started to investigate for multilevel structures, are log-linear characterizations of the multivariate distributions and multivariate logit and probit (Joe, 1997; Lesaffre and Molenberghs, 1991; Molenberghs and Lesaffre, 1994). Other developments we envisage are multivariate models for mixed continuous and ordered category responses, and to parallel the longitudinal binary response models of Yang *et al.* (2000), variants for ordered categorizations. The latter situation has received some attention in the generalized estimating equation (GEE) literature (Lipsitz *et al.*, 1994) but these are population averaging methods. As such they concentrate mainly on ways of obtaining efficient fixed-effects estimates and cannot at present be used to investigate the detailed structure of multilevel effects.

Acknowledgements

This study is part of the project Application of Advanced Multilevel Modelling Methods for the Analysis of Examination Data, supported by the Economic and Social Research Council (ESRC) of the UK under Grant award R000237394. The Department of Education and Employment (DfEE) kindly provided the raw examination data. The ESRC has also given support to Antony Fielding's Visiting Research Fellowship at the Multilevel Models Project under award H51944500497 of the Analysis of Large and Complex Datasets programme. Valuable comments were received from James Carpenter and reviewers of an earlier version of the paper.

References

Bock RD (1975) *Multivariate Statistical Methods in Behavioral Research.* New York: McGraw-Hill.

Chan JSK, Kuk AYC, Bell J, McGilchrist CE (1998) The analysis of methadone clinic data using marginal and conditional logistic models with mixture or random effects. *Australian and New Zealand Journal of Statistics*, **40**(1), 1–10.

Ezzett F, Whitehead J (1991) A random effects model for ordinal responses from a crossover trial. *Statistics in Medicine*, **10**, 901–07.

Fielding A (1999) Why use arbitrary points scores? ordered categories in models of educational progress. *Journal of the Royal Statistical Society Series A*, **A162**, 303–30.

Fielding A (2002) Ordered category responses and random effects in multilevel and other complex structures: scored and generalised linear models. In Reise, S, Duan N eds. *Multilevel Modelling: Methodological Advances, Issues and Applications.* New Jersey: Erlbaum.

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

Goldstein H (1995) *Multilevel Statistical Models*, 2nd edn. London: Edward Arnold.

Goldstein H, Healy M Jr (1995) The graphical presentation of a collection of means. *Journal of the Royal Statistical Society Series A*, **158**, 505–13.

Goldstein H, Rasbash J (1996) Improved estimation in multilevel models with binary responses. *Journal of the Royal Statistical Society Series A*, **159**, 505–13.

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**, 219–30.

Goldstein H, Spiegelhalter DJ (1996) League tables and their limitations: statistical issues in comparisons of institutional performance (with discussion). *Journal of the Royal Statistical Society Series A*, **159**, 385–443.

Goldstein H, Thomas S (1996) Using examination results as indicators of school and college performance. *Journal of the Royal Statistical Society Series A*, **159**, 149–63.

Greene WH (2000) *Econometric analysis*, 4th edn. Upper Saddle Valley, New Jersey: Prentice-Hall.

Harville DA, Mee RW (1984) A mixed-model procedure for analysing ordered categorical data. *Biometrics*, **40**, 393–408.

Hedeker D, Gibbons RD (1994) A random-effects ordinal regression model for multilevel analysis. *Biometrics*, **50**, 933–44.

Hedeker D, Gibbons RD (1996) MIXOR: a computer program for mixed effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine*, **49**, 157–76.

Hedeker D, Mermelstein J (1998) A multilevel thresholds of change model for analysis of stages of change data. *Multivariate Behavioral Research*, **33**(4), 427–55.

Jansen J (1990) On the statistical analysis of ordinal data when extra-variation is present. *Applied Statistics*, **39**(1), 75–84.

Joe H (1997) *Multivariate Models and Dependency Concepts.* Boca Raton, Florida: CHC.

Kampen J, Swyngedouw M (2000) The ordinal controversy revisited. *Quality & Quantity*, **34**, 87–102.

Lesaffre E, Molenberghs G (1991) Multivariate probit analysis: a neglected procedure in medical statistics. *Statistics in Medicine*, **10**, 1391–1403.

Lipsitz SR, Kim K, Zhao L (1994) Analysis of repeated categorical data using generalised estimating equations. *Statistics in Medicine*, **13**, 1149–63.

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

McCullagh P, Nelder JA (1989) *Generalised Linear Models*, 2nd edn. London: Chapman and Hall.

Molenberghs G, Lesaffre E (1994) Marginal modelling of correlated ordinal data using a multivariate Plackett distribution. *Journal of the American Statistical Association*, **89**, 633–44.

O'Donoghue C, Thomas S, Goldstein H, Knight, T (1996) *1996 DfEE study of value added for 16–18 year olds in England. DfEE Research Series*, March, London: Department for Education and Employment.

Rasbash J, Browne W, Goldstein H, Yang M *et al.* (1999) *A User's Guide to MLwiN.* London: Multilevel Models Project, Institute of Education, University of London.

Saei A, McGilchrist CA (1996) Random component threshold models. *Journal of Agricultural, Biological and Environmental Statistics*, **1**, 288–96.

Saei A, McGilchrist CA (1998) Longitudinal threshold models with random components. *Journal of the Royal Statistical Society Series D (The Statistician)*, **47**, 365–75.

Saei A, Ward J, McGilchrist CA (1996) Threshold models in a methadone programme evaluation. *Statistics in Medicine*, **15**(20), 2253–60.

Willms JD (1992) *Monitoring School Performance: A Guide for Educators.* Lewes: Falmer Press.

Yang M (1997) Multilevel models for multiple category responses by MLn simulation. *Multilevel Modelling Newsletter*, **9**(1), 9–15, London: Institute of Education, University of London.

Yang M (2001) Multinomial regression. In Leyland A and Goldstein H eds. *Multilevel Modelling of Health Statistics.* Chichester: Wiley.

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

Yang M, Rasbash J, Goldstein H (1998) *MLwiN Macros for Advanced Multilevel Modelling.* London: Multilevel Models Project, Institute of Education, University of London.

Yang M, Woodhouse G (2001) Progress from GCSE to A and AS level: institutional and gender differences, and trends over time. *British Journal of Education Research*, **27**, 2.

Yang M, Goldstein H, Browne W, Woodhouse G (2001) Multivariate multilevel analysis of examination results. *Journal of the Royal Statistical Analysis Series A*, **165**, 137–53.