Handbook of Statistics, Vol. 26
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(06)26012-7

12

Multilevel Models in Psychometrics

Fiona Steele and Harvey Goldstein

1. Introduction

Before describing the basic multilevel model, it is useful to reflect on why such models are useful. For many years, social researchers, especially in education, discussed the 'units of analysis' problem, one version of which has also been called the 'ecological fallacy' (see Robinson, 1951). At one extreme, it is possible to study relationships among variables ignoring group structures. At the other extreme we can work solely with group, say school, averages in exploring relationships. Aitkin and Longford (1986) set out the statistical issues associated with various procedures. In an earlier analysis Aitkin et al. (1981) reworked a well-known study on teaching styles which used student level data but ignored school membership (Bennett, 1976). They showed that formerly 'significant' results became non-significant when a multilevel model was used. Woodhouse and Goldstein (1989) showed how the use solely of aggregate level data based upon school means could lead to unstable and misleading conclusions. In addition to the problem of making misleading inferences, failure to model both students and schools simultaneously makes it impossible to study the extent to which school and student characteristics interact to influence the response measurement or measurements. A useful compilation of some early work in the area of psychometrics is the volume edited by Bock (1989).

Multilevel models overcome these problems wherever we have any kind of hierarchical structure, such as individuals grouped within households, themselves grouped within areas (a three-level model). Repeated measures data are an example of a two-level hierarchy with measurement occasions (level 1 units) grouped within individuals (level 2 units). We can readily incorporate multivariate data within this framework. In the simple case of a set of responses on a sample of individuals we think of the set of response variables as forming the lowest level of the data hierarchy; that is, measurements are nested within individuals. Further levels, such as school, can then be added above that of individual. Models for repeated measures and multivariate data are discussed in Sections 3 and 4 and structural equation models are introduced in Section 6.

While the ability to fit models to hierarchically structured data is important, there are many applications where the data structures are more complex. Suppose a student is classified as belonging sequentially to a particular combination of primary (elementary)

school and secondary (high) school and we have followed a sample of such students through each school and wish to relate measurements made at the end of secondary school to those made earlier in the primary schools. The students will be identified by a *cross classification* of primary schools and secondary schools. Note that even if we did not have prior measurements, but *did* have identification of the primary and secondary schools we could still carry out a cross classified analysis. Another example is where students are simultaneously classified by the school they attend and the area where they live, both classifications affecting the value of a response variable. Models for such cross classifications will be discussed in Section 5.

A further complication occurs when we cannot assign a lower level unit to a single higher level unit. Suppose that, during secondary schooling many pupils move between schools. If our response is, say, a test score at the end of secondary school, then for such a student we will need to share out the school 'effect' among all the schools attended, using a suitable weighting function. Another example occurs with spatial data where the affect of area will be shared among the area where someone resides and surrounding areas, with weights a function of geographical or other distance. Such models find considerable use in epidemiological studies. These multiple membership models will also be discussed in Section 5.

2. Basic Models for Two-level Hierarchical Data Structures

We begin with a description of simple multilevel models for data from populations with a two-level hierarchical structure. More detailed accounts of these and more general models can be found in Bryk and Raudenbush (2002), Goldstein (2003), Longford (1993) and Snijders and Bosker (1999).

2.1. *Random intercept model*

For simplicity consider a simple data structure where a response y_{ij} is measured on individual i in school j ($i = 1, \ldots, n_j$; $j = 1, \ldots, J$), together with a single explanatory variable x_{ij}. For example, the response might be an examination score measured on students at age 16 years and the explanatory variable a test score measured on the same students five years earlier at age 11. Instead of schools we could think of any grouping of individuals, such as households or areas. We wish to model a relationship between the individual response and the explanatory variable, taking into account the effect of school on the mean response. We shall assume in what follows that we are dealing with a continuously distributed response, and for simplicity that this has a Normal distribution. Our data have a simple two-level structure with the schools as higher level units and students as lower level units. The simplest multilevel model that we can fit to such a structure can be written as follows

$$\begin{aligned} y_{ij} &= \beta_{0j} + \beta_1 x_{ij} + e_{ij}, \\ \beta_{0j} &= \beta_0 + u_{0j}, \end{aligned} \tag{2.1}$$

where $\text{var}(e_{ij}) = \sigma_e^2$ and $\text{var}(u_{0j}) = \sigma_{u0}^2$.

The model may also be written as a single equation

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{0j} + e_{ij},$$

where $\beta_0 + \beta_1 x_{ij}$ is referred to as the *fixed part* of the model, and $u_{0j} + e_{ij}$ is the *random part*. The random variable u_{0j} represents the departure of the jth school's intercept from the overall population intercept term β_0. The slope coefficient β_1 is for the present assumed to be the same for all the schools. As mentioned we shall develop the model initially assuming that the random variables have a Normal distribution: $e_{ij} \sim N(0, \sigma_e^2)$, $u_{0j} \sim N(0, \sigma_{u0}^2)$. This model is sometimes called a variance components model, owing to the fact that the residual variance is partitioned into components corresponding to each level in the hierarchy. The variance between schools is σ_{u0}^2, and the variance between students within a given school is σ_e^2.

The similarity between individuals in the same school is measured by the *intra-class correlation* (where, here, 'classes' are schools):

$$\frac{\sigma_{u0}^2}{\sigma_{u0}^2 + \sigma_e^2}.$$

The intra-class correlation (ICC) measures the extent to which the y-values of students in the same school resemble each other as compared to those from individuals in different schools. The ICC may also be interpreted as the proportion of the total residual variation that is due to differences between schools, and is referred to as the *variance partition coefficient* (VPC) as this is the more usual interpretation (see Goldstein, 2003, pp. 16–17).

Having fitted (2.1) we can obtain estimates for the residuals (u_{0j}, e_{ij}) by estimating their expected values, given the data and estimates of the parameters $(\beta_0, \beta_1, \sigma_e^2, \sigma_{u0}^2)$. Of particular interest in our example are estimates of the level 2 residuals u_{0j}, which represent school effects on attainment at age 16, adjusted for prior attainment x_{ij}. In ordinary least squares (OLS) regression residual estimates are obtained simply by subtracting the predicted values of y_{ij} from their observed values, i.e., $r_{ij} = y_{ij} - \hat{y}_{ij}$. In multilevel models, with residuals at more than one level, a more complex procedure is needed. Estimates of u_{0j} are obtained by taking the average of the raw residuals r_{ij} for each school j and multiplying the result by a *shrinkage factor*. This shrinkage factor pulls the estimate of u_{0j} towards zero when the between-school variance σ_{u0}^2 is small relative to the within-school variance σ_e^2, or when the number of students sampled in a school n_j is small.

2.2. *Fixed versus random effects*

An alternative way of allowing for school effects would be to use an analysis of variance (ANOVA) or fixed effects model, which would involve including as explanatory variables a set of dummy variables that indicate the school to which a student belongs. While ANOVA can also be used to compare any number of schools, the random effects approach has a number of advantages over fixed effects models. First, if there are J schools to be compared, then $J-1$ parameters are required to capture school effects and, therefore, if J is large, a large number of parameters need to be estimated. In contrast, in

a random effects model only one additional parameter, the between-school variance σ^2_{u0}, is estimated regardless of the number of schools.

Second, the origins of ANOVA lie in experimental design where there are typically a small number of groups under comparison and all groups of interest are sampled. Often we have only a sample of groups (e.g., a sample of schools) and it is the population of groups from which our sample was drawn which is of interest. The ANOVA model does not allow inferences to be made beyond the groups in the sample. The key point about the random variable u_{0j} in the random effects model is that it allows us to treat the samples of units as coming from a universe or population of such units. Thus, the schools (and students) chosen are not typically the principal focus of interest; they are regarded as a random sample from a population of schools and we are concerned with making statements about that population, for example in terms of its mean and variance.

Finally, in a fixed effects model the effects of level 2 explanatory variables cannot be separately estimated. Such variables are confounded with the level 2 effects because any level 2 variable can be expressed as a linear combination of the dummy variables for higher level units. This is a serious drawback of the ANOVA approach as one is often interested in exploring the extent to which the level 2 variation can be explained by observed level 2 characteristics. In the random effects model level 2 variables are not confounded with the level 2 effects u_{0j}, and therefore their effects may be estimated while simultaneously accounting for level 2 variance due to unobserved factors.

2.3. *Random slope model*

We can elaborate (2.1) by allowing the coefficient β_1 to vary across schools:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij}, \quad \beta_{0j} = \beta_0 + u_{0j}, \quad \beta_{1j} = \beta_1 + u_{1j}, \tag{2.2}$$

where

$$e_{ij} \sim N(0, \sigma^2_e), \qquad \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N(0, \Omega_u), \qquad \Omega_u = \begin{pmatrix} \sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1} \end{pmatrix}.$$

Model (2.2) is often referred to as a *random coefficient* model by virtue of the fact that the coefficients β_{0j} and β_{1j} in the first equation of (2.2) are random quantities, each having a variance with a covariance between them. Now the slope for school j, β_{1j}, is given by β_1, the average slope across all schools, plus a random departure u_{1j}. The terms u_{0j} and u_{1j} are random departures from β_0 and β_1, or residuals at the school level. As more explanatory variables are introduced into the model, so we can choose to make their coefficients random at the school level thereby introducing further variances and covariances, and this will lead to models with complex covariance structures. One of the aims of multilevel modelling is to explore such potential structures and also to attempt to explain them in terms of further variables.

When coefficients of explanatory variables are permitted to vary randomly across level 2 units, the level 2 variance is no longer constant but depends on those variables with random coefficients. For example, in model (2.2) the school-level variance is

$$\text{var}(u_{0j} + u_{1j}x_{ij}) = \sigma^2_{u0} + 2\sigma_{u01}x_{ij} + \sigma^2_{u1}x^2_{ij}, \tag{2.3}$$

i.e., a quadratic function of x_{ij}.

2.4. *Complex level 1 variation*

We have seen how in a random coefficient model, the level 2 variance is a function of explanatory variables (Eq. (2.3)). We can also allow the level 1 variance to depend on explanatory variables. This leads to a model with *complex level 1 variation* or *heteroskedasticity* at level 1.

Suppose, for example, that we wish to explore whether the between-student variation within schools differs for boys and girls. We introduce a second explanatory variable x_{2ij} which indicates a student's gender (coded 1 for girls and 0 for boys). We then specify a model in which both the mean and variance of y_{ij} depend on gender by including x_{2ij} in the fixed part of the model and allowing separate level 1 residuals for girls and boys. The random intercept version of this model can be written

$$\begin{aligned} y_{ij} &= \beta_{0j} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + e_{0ij}(1 - x_{2ij}) + e_{1ij} x_{2ij}, \\ \beta_{0j} &= \beta_0 + u_{0j}. \end{aligned} \tag{2.4}$$

From (2.4) the student-level variance is $\sigma^2_{e0}(1 - x_{2ij}) + \sigma^2_{e1} x_{2ij}$ which reduces to σ^2_{e0} for boys and σ^2_{e1} for girls.

2.5. *Example*

We now give examples of applications of the models described above using educational data on 4059 London school children from 65 schools (see Rasbash et al., 2005, for further details). The dependent variable is a normalised exam score at age 16. We consider two explanatory variables: a standardised reading test score at age 11 and gender (coded 1 for girls and 0 for boys). Table 1 shows the results from fitting three two-level models to these data. Also presented are the -2 log-likelihood values for the fitted models, which may be used to carry out likelihood ratio tests to compare the fit of nested models.

We begin by considering a simple random intercept model. From this model the ICC (or VPC) is estimated as $0.088/(0.088 + 0.562) = 0.135$. Thus, after accounting for the effects of prior attainment and gender, 13.5% of the remaining variance in age 16 scores is due to differences between schools.

The next model fitted is an elaboration of the random intercept model in which the effect of the age 11 reading test score on attainment at age 16 is allowed to vary across schools, a random slope (or coefficient) model. The difference in -2 log-likelihood value (i.e., deviance change) between this and the random intercept model is 42.6, on 2 degrees of freedom. We therefore conclude that the random slope model is a significantly better fit to the data. In this model, the individual school slopes vary about a mean of 0.553 with estimated variance 0.019. There is a positive covariance between the intercepts and slopes, estimated as 0.015, suggesting that schools with higher intercepts tend to have steeper slopes; on average the effect of prior attainment is stronger in schools with a high mean age 16 score. We note that this conclusion is not invariant under shifts in the values of predictors with random coefficients. In this example, the intercept for a given school is that school's mean attainment at age 16 for boys with a

Table 1
Results from fitting two-level models to educational attainment at age 16

	Random intercept		Random slope for age 11 score		Complex level 1 variance	
Parameter	Est.	(SE)	Est.	(SE)	Est.	(SE)
Fixed						
β_0 (Intercept)	−0.095	(0.043)	−0.112	(0.043)	−0.112	(0.043)
β_1 (Age 11 score)	0.560	(0.012)	0.553	(0.020)	0.553	(0.020)
β_2 (Girl)	0.171	(0.033)	0.176	(0.032)	0.175	(0.032)
Random: Between-school						
σ^2_{u0} (Intercept)	0.088	(0.017)	0.086	(0.017)	0.086	(0.017)
σ^2_{u1} (Age 11 score)	–	–	0.015	(0.004)	0.015	(0.004)
σ_{u01} (Intercept/slope covariance)	–	–	0.019	(0.007)	0.019	(0.007)
Random: Within-school						
σ^2_e (Total)	0.562	(0.013)	0.550	(0.012)	–	–
σ^2_{e0} (Boys)	–	–	–	–	0.587	(0.021)
σ^2_{e1} (Girls)	–	–	–	–	0.525	(0.015)
−2 log-likelihood	9330.0		9287.4		9281.4	

mean age 11 score. A different estimate would be obtained for the intercept-slope covariance if age 11 scores were not centred. The positive covariance also implies that the between-school variance increases with prior attainment (see Eq. (2.3)).

The final model allows the within-school between-student residual variance to depend on gender. A likelihood ratio test statistic for a comparison of this model with the constant level 1 variance model is 6 on 1 d.f., so there is strong evidence of heteroskedasticity at the student level. The estimated within-school variance is 0.587 for boys and 0.525 for girls. Thus while on average boys have a lower age 16 score than girls, their scores have greater dispersion than the girls'.

3. Models for repeated measures

The models described in the previous section can also be applied in the analysis of repeated measures where observations over time are the level 1 units and individuals are at level 2. We illustrate multilevel modelling of longitudinal data using a dataset consisting of nine measurements made on 26 boys between the ages of 11 and 13.5 years, approximately three months apart (Goldstein, 1989). Figure 1 shows the mean heights by the mean age at each measurement occasion.

We assume that growth can be represented by a polynomial function with coefficients varying randomly across individuals. Other functions are possible, including fractional

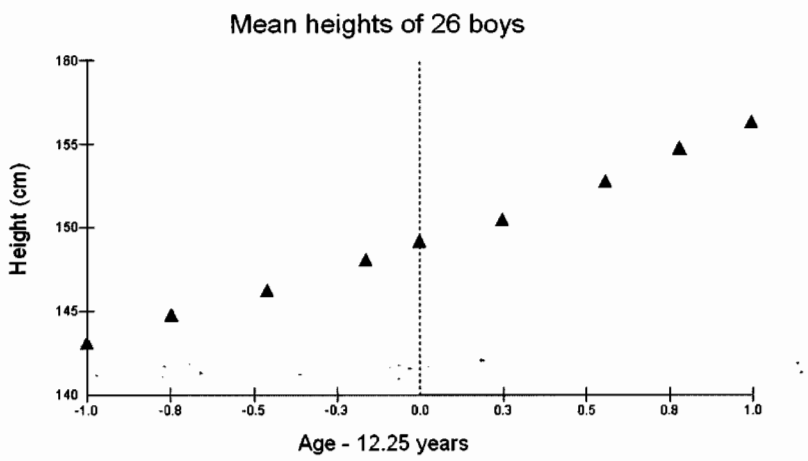

Fig. 1. Mean height by mean age at nine measurement occasions.

polynomials or nonlinear functions, but for simplicity we confine ourselves to examining a fourth order polynomial centred at an origin of 12.25 years. The model we fit can be written as follows:

$$
\begin{aligned}
&y_{ij} = \sum_{h=0}^{4} \beta_{hj} t_{ij}^{h} + e_{ij}, \\
&\beta_{0j} = \beta_0 + u_{0j}, \qquad \beta_{1j} = \beta_1 + u_{1j}, \qquad \beta_{2j} = \beta_2 + u_{2j}, \\
&\beta_{3j} = \beta_3, \qquad \beta_{4j} = \beta_4
\end{aligned}
\tag{3.1}
$$

where

$$
\begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{pmatrix} \sim N(0, \Omega_u), \qquad \Omega_u = \begin{pmatrix} \sigma_{u0}^2 & & \\ \sigma_{u01} & \sigma_{u1}^2 & \\ \sigma_{u02} & \sigma_{u12} & \sigma_{u2}^2 \end{pmatrix},
$$

$$
e_{ij} \sim N\left(0, \sigma_e^2\right).
$$

Eq. (3.1) defines a two-level model with level 1 being 'measurement occasion' and level 2 'individual boy'. Note that we allow only the coefficients up to the second order to vary across individuals; in the present case this provides an acceptable fit. The level 1 residual term e_{ij} represents the unexplained variation within individuals about each individual's growth trajectory. Table 2 shows the maximum likelihood parameter estimates for this model.

For each boy we can also estimate their random effects or 'residuals' (u_{0j}, u_{1j}, u_{2j}), and use these to predict their growth curve at each age. Figure 2 shows these predicted curves. Growth over this period exhibits a seasonal pattern with growth in the summer being about 0.5 cm greater than growth in the winter. Since the period of the growth cycle is a year this is modelled by including a simple cosine term, which could also have a random coefficient.

Table 2
Height modelled as a fourth degree polynomial on age

Fixed effects	Estimate	Standard error
β_0 (Intercept)	149.0	1.54
β_1 (t)	6.17	0.35
β_2 (t^2)	1.13	0.35
β_3 (t^3)	0.45	0.16
β_4 (t^4)	−0.38	0.30

Random effects
Level 2 (individual) correlation matrix: variances on diagonal

	u_{0j} (Intercept)	u_{1j} (t)	u_{2j} (t^2)
u_{0j} (Intercept)	61.6		
u_{1j} (t)	0.61	2.8	
u_{2j} (t^2)	0.22	0.66	0.64

Level 1 (occasion) variance = 0.22

−2 log-likelihood = 625.4

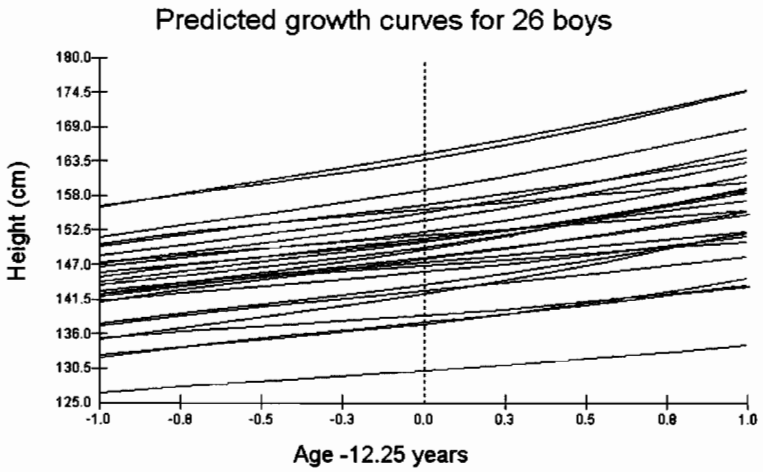

Fig. 2. Predicted growth curves from model with random coefficients for linear and quadratic terms.

In our example we have a set of individuals all of whom have nine measurements. This restriction, however, is not necessary and (3.1) does not require either the same number of occasions per individual nor that measurements are made at equal intervals, since time is modelled as a continuous function. In other words we can combine data from individuals with very different measurement patterns, some of whom may only have been measured once and some who have been measured several times at irregular intervals. This flexibility, first noted by Laird and Ware (1982), means that the multi-

level approach to fitting repeated measures data is to be preferred to previous methods based upon a traditional multivariate formulation assuming a common set of fixed occasions.

In these models it is assumed that the level 1 residual terms (e_{ij}) are independently distributed. We may relax this assumption, however, and in the case of repeated measures data this may be necessary, for example where measurements are taken very close together in time. Goldstein et al. (1994) show how to model quite general nonlinear covariance functions and in particular of the form $\text{cov}(e_t, e_{t-s}) = \sigma_e^2 \exp(-g(\alpha, s))$, where s is the time difference between occasions. This allows the correlation between occasions to vary smoothly as a function of their (continuous) time difference. A simple example is where $g = \alpha s$, which in discrete time produces a first-order autoregression, AR(1), model.

For further discussion of multilevel modeling of longitudinal data see Singer and Willett (2003) and Muthén (1997).

4. Models for multivariate response data

Repeated measures data consist of multiple observations on individuals over time. Suppose again that we have more than one observation per individual, but that the observations are for different variables, e.g., responses to a set of items in a questionnaire. In the examples given below we have multivariate responses on students within schools, where y_{rij} denotes the rth response ($r = 1, \ldots, R$) on individual i in school j. In this section we consider two types of multilevel model for multivariate response data. We begin with a multivariate model, which is appropriate when the focus is on estimating the effects of explanatory variables on each response, while accounting for the correlation between responses at the student and school level. We then consider a multilevel factor model which assumes that the pairwise correlations between responses are explained by their mutual dependency on one or more latent variables (factors).

4.1. Multivariate models

A two-level random intercept model for multivariate response data may be written

$$y_{rij} = \beta_{0r} + \beta_{1r} x_{ij} + u_{0rj} + e_{rij}, \tag{4.1}$$

where $u_{0j} = \{u_{0rj}\} \sim N_R(0, \Omega_u)$ and $e_{ij} = \{e_{rij}\} \sim N_R(0, \Omega_e)$.

Eq. (4.1) can be viewed as a 3-level model where the level 1 units are the within-student measurements. The explanatory variables are a set of R dummy variables that indicate the responses (with coefficients β_{0r}), and their interactions with the covariate x_{ij} (with coefficients β_{1r}). The coefficients of the dummy variables are assumed to vary randomly across students (at level 2) and schools (level 3) to obtain the student and school residuals, e_{rij} and u_{0rj}.

There are two main advantages to fitting a multivariate model rather than carrying out a separate analysis for each response. First, we can obtain estimates of the pairwise correlations between responses at each level adjusting for the effects of x_{ij}.

Table 3
Results from fitting a bivariate multilevel model to written paper and coursework exam scores

	Written		Coursework	
	Estimate	(SE)	Estimate	(SE)
Intercept	49.4		69.7	
Female	−2.5	(0.6)	6.8	(0.7)
School-level				
Variance	46.8	(9.2)	75.2	(14.6)
Covariance	24.9	(8.9)	–	–
(Correlation)	*(0.4)*	–		
Student-level	124.6	(4.4)	180.1	(6.2)
Variance				
Covariance	73.0	(4.2)	–	–
(Correlation)	*(0.5)*	–		

Second, individuals with missing data on one or more response can be retained in the analysis; under a 'missing at random' assumption efficient estimates of coefficients and covariance structures at each level are obtained. This relaxation of the requirement for balanced multivariate data is particularly useful for the analysis of data from rotation designs where each respondent answers a random subset from a pool of questions.

We illustrate the use of multivariate models in an analysis of students' scores on two components of a science examination taken by 1905 students in 73 English schools (see Rasbash et al., 2005, for details). The first component is a traditional written paper, and the second is a piece of coursework. A total of 382 responses were missing, but students with only one missing score remain in the analysis sample. The scores on both components have been rescaled so that their maximum is 100, thus enabling comparison of covariate effects across responses. We consider one covariate, the student's gender (coded 1 for female and 0 for male). The results from fitting model (4.1) to these data are presented in Table 3. From the fixed part estimates we conclude that while girls perform worse than boys on the written paper, they do better on the coursework. Turning to the random part of the model, we see that there is greater variation in coursework scores at both the student and school level. The correlations between the coursework and written scores at the student and school level are, respectively, 0.4 and 0.5.

4.2. *Multilevel factor models*

In a multilevel factor model, the student and school level correlations between pairs of responses are assumed to be explained by one or more factors at each level. For simplicity, we assume that there is a single factor at each level. The factor model is an

extension of (4.1) and can be written

$$y_{rij} = \beta_{0r} + \beta_{1r}x_{ij} + \lambda_r^{(2)}\eta_j^{(2)} + \lambda_r^{(1)}\eta_{ij}^{(1)} + u_{0rj} + e_{rij}, \tag{4.2}$$

where $\eta_{ij}^{(1)}$ and $\eta_j^{(2)}$ are the factors at the student and school level respectively, and $\lambda_r^{(1)}$ and $\lambda_r^{(2)}$ are the factor loadings. The factors are assumed to be normally distributed. Conditional on the factors and x_{ij}, the responses are assumed to be independent. Thus the residuals u_{0rj} and e_{rij} (often referred to as 'uniquenesses') are now assumed to be uncorrelated across responses. See Goldstein and Browne (2002), Muthén (1994), Skrondal and Rabe-Hesketh (2004) and Steele (2005) for further discussion of multilevel factor analysis.

We illustrate the application of multilevel factor analysis using a dataset of science scores for 2439 students in Hungarian schools (Goldstein, 2003, Chapter 6). The data consist of scores on four test booklets: a core booklet with components in earth science, physics and biology, two biology booklets and one in physics. There are therefore six possible test scores (one earth science, three biology, and two physics). Each student responds to a maximum of five tests, the three tests in the core booklet plus a randomly selected pair of tests from the other booklets. The analysis presented below is based on standardised test scores.

The results from a two-level factor model, with a single factor at each level, are presented in Table 4. The factor variances at the student and school level are estimated as 0.127 (SE = 0.016) and 0.057 (SE = 0.024), respectively. For ease of interpretation, standardised leadings are calculated for each factor as (omitting subscripts) $\lambda_r^{(k)*} = \lambda_r^{(k)}\sqrt{\text{var}(\eta^{(k)})}$, $k = 1, 2$. Because, at each level, the standardised loadings have the same sign across responses, we interpret the factors as student- and school-level measures of overall attainment in science. Biology R3 has the lowest loading. The poor fit for this test is reflected in a relatively high residual variance estimate at both levels. Thus only a small amount of the variance in the scores for this biology test is explained by the student and school level factors, i.e., the test has a low communality.

Table 4
Estimates from two-level factor model with one factor at each level

	Student-level			School-level		
	$\lambda_r^{(1)}$ (SE)	$\lambda_r^{(1)*}$	σ_{er}^2 (SE)	$\lambda_r^{(2)}$ (SE)	$\lambda_r^{(2)*}$	σ_{ur}^2 (SE)
E. Sc. core	1†	0.36	0.712 (0.023)	1†	0.24	0.098 (0.021)
Biol. core	1.546 (0.113)	0.55	0.484 (0.022)	2.093 (0.593)	0.50	0.015 (0.011)
Biol. R3	0.583 (0.103)	0.21	0.892 (0.039)	0.886 (0.304)	0.21	0.033 (0.017)
Biol. R4	1.110 (0.115)	0.40	0.615 (0.030)	1.498 (0.466)	0.36	0.129 (0.029)
Phys. core	1.665 (0.128)	0.59	0.422 (0.022)	2.054 (0.600)	0.49	0.036 (0.013)
Phys. R2	1.558 (0.133)	0.56	0.526 (0.030)	1.508 (0.453)	0.36	0.057 (0.018)

†Parameter constrained for model identification.

5. Models for non-hierarchical structures

5.1. Cross classified models

In the example of children moving from primary to secondary school we have a cross classified structure which can be modelled as follows:

$$y_{i(j_1 j_2)} = (X\beta)_{i(j_1 j_2)} + u_{j_1} + u_{j_2} + e_{i(j_1 j_2)}, \tag{5.1}$$

($j_1 = 1, \ldots, J_1$; $j_2 = 1, \ldots, J_2$; $i = 1, \ldots, n$), in which the score of student i, belonging to the combination of primary school j_1 and secondary school j_2, is predicted by a linear 'regression' function denoted by $(X\beta)_{i(j_1,j_2)}$. The random part of the model is given by two level 2 residual terms, one for the primary school attended by the student (u_{j_1}) and one for the secondary school attended (u_{j_2}), together with the usual level 1 residual term for each student. We note that the latter may be further modelled to produce complex level 1 variation (see Section 2.4), allowing for example for separate variances for males and females.

As an example consider the analysis carried out by Goldstein (2003, Chapter 11) who fitted cross classified models to 3435 students who attended 148 primary schools and 19 secondary schools in Fife, Scotland. The dependent variable is the overall exam attainment at age 16 and a verbal reasoning score measured on entry to secondary school is included as an explanatory variable. The principal aim was to separate the effect of primary school attended from that of secondary school. Table 5 shows results from two alternative model specifications: a two-level hierarchical model, which ignores the information on primary schools, and the cross classified model.

From the two-level model, the type of model typically used in school effectiveness studies, we would estimate that secondary schools explain $0.28/(0.28 + 4.26) \times 100 = 6.2\%$ of the residual variance in age 16 scores (after accounting for age 12 verbal reasoning). The cross classified model takes into account both secondary and primary school

Table 5
Estimates from analysis of examination scores using hierarchical and cross classified models

	2-level model		Cross classified model	
	Estimate	(SE)	Estimate	(SE)
Fixed				
Intercept	5.99		5.98	
Verbal reasoning	0.16	(0.003)	0.16	(0.003)
Random				
σ^2_{u1} (primary)	–	–	0.27	(0.06)
σ^2_{u2} (secondary)	0.28	(0.06)	0.011	(0.021)
σ^2_e (student)	4.26	(0.10)	4.25	(0.10)
−2 log-likelihood	17172.0		14845.6	

effects on age 16 attainment, and is clearly a much better fit to the data than the two-level hierarchical model (LR statistic = 2326.4, 1 d.f.). From the cross classified model we would conclude that it is primary schools rather than secondary schools that have the strongest effect on attainment; the proportion of variance due to secondary schools is now only 0.24%, compared to 6.0% due to primary schools. The substantive importance of this finding for studies of schooling is that it becomes necessary to take account of achievement during periods of schooling prior to the one immediately being considered (secondary here). However, Goldstein (2003) notes that one reason for the substantially larger primary school variance may be that secondary schools are generally larger than primary schools, so that the sampling variance is smaller.

5.2. *Multiple membership models*

Turning to multiple membership models we consider just the secondary schools from the above example and suppose that we know, for each individual, the weight w_{ij_2} associated with the j_2th secondary school attended by student i with $\sum_{j_2=1}^{J_2} w_{ij_2} = 1$. These weights, for example, may be proportional to the length of time a student is in a particular school during the course of the longitudinal study. Note that we allow the possibility that for some (perhaps most) students only one school is involved, so that one of these weights is one and the remainder are zero. Note that when all level 1 units have a single non-zero weight of 1 we obtain the usual purely hierarchical model. We can write the following model for the case of membership of just two schools (1,2):

$$y_{i(1,2)} = (X\beta)_{i(1,2)} + w_{i1}u_1 + w_{i2}u_2 + e_{i(1,2)}, \tag{5.2}$$

where $w_{i1} + w_{i2} = 1$.

A more general model is

$$y_{i\{j\}} = (X\beta)_{i\{j\}} + \sum_{h\in\{j\}} w_{ih}u_h + e_{i\{j\}}, \tag{5.3}$$

where $\sum_{h\in\{j\}} w_{ih} = 1$, $\text{var}(u_h) = \sigma_u^2$, and $\text{var}(\sum_{h\in\{j\}} w_{ih}u_h) = \sigma_u^2 \sum_{h\in\{j\}} w_{ih}^2$. (The notation $h \in \{j\}$ means for any school h that belong to the set of schools $\{j\}$.) In the particular case of membership of just two schools with equal weights we have $w_{i1} = w_{i2} = 0.5$, $\text{var}(\sum_h w_{ih}u_h) = \sigma_u^2/2$. In other words the contribution to the level 2 variation is just half that for a student who remains in one school, since in the former case the level 2 contribution is averaged over two (random) schools. Note that if we ignore the multiple membership of schools and simply assign students, say, to the final school that they attend, we will underestimate the true extent of between-school variation. This is because, for those students who do attend more than one school, the true level 2 variation is less than that for students who attend a single school. In the model, however, we assume that the level 2 variation for these students is the same as that for those attending a single school, with the result that the overall level 2 variation is underestimated.

A slightly different notation to describe membership relationships is used by Browne et al. (2001). This is particularly useful when we have very complex structures involving mixtures of hierarchical, crossed and multiple membership classifications. Essentially it

works by requiring just a single unique identification for each lowest level observation, in the present case a student. Each student then has a relationship with every other type of unit, here primary and secondary schools. The model specifies which classifications are involved and the data structure specifies precisely which schools are involved for each student. Thus our cross classified model would be written as follows

$$y_i = (X\beta)_i + u^{(2)}_{secondary(i)} + u^{(3)}_{primary(i)} + e_i \quad (i = 1, \ldots, n), \tag{5.4}$$

where *primary*(i) and *secondary*(i) refer respectively to the primary and secondary schools attended by student i. The superscripts for the random variables identify the classification; where this is absent, and if there is no ambiguity, it is assumed to be the lowest level, classification (1). Using this notation the multiple membership model would be written as

$$y_i = (X\beta)_i + \sum_{h \in school(i)} w_{i,h} u^{(2)}_h + e_i, \tag{5.5}$$

where $\sum_{h \in school(i)} w_{i,h} = 1$, and $\text{var}(u^{(2)}_h) = \sigma^2_u$.

We can have mixtures of cross classifications and multiple memberships. Thus, for example, pupils may move between schools and also between areas for the cross classification of schools by areas. Such a model can be written in the form

$$y_{i\{j_1\}\{j_2\}} = (X\beta)_{i\{j_1\}\{j_2\}} + \sum_{h \in \{j_1\}} w_{1ih} u_{1h} + \sum_{h \in \{j_2\}} w_{2ih} u_{2h} + e_{i\{j_1\}\{j_2\}}, \tag{5.6}$$

where $\sum_h w_{1ih} = W_1$, $\sum_h w_{2ih} = W_2$, $\text{var}(u_{1h}) = \sigma^2_{u1}$, and $\text{var}(u_{2h}) = \sigma^2_{u2}$.

There are now two sets of higher-level units (schools and areas) which influence the response, each set having a multiple membership structure. Another application of such a model is for household data where individuals move among households and among addresses.

5.3. *Representing complex data structures*

When we have complex mixtures of hierarchies, cross classifications and multiple memberships, a straightforward way of representing these becomes important. Browne et al. (2001) use simple diagrams for representing such complex structures. Figure 3 represents the cross classified structure described in the example of Section 5.1. The single directional lines indicate a membership relation, and here students are members of just one secondary school and one primary school. Where multiple membership is involved, two parallel lines are used (see Figure 4).

To illustrate the flexibility of these models and their representation using classification diagrams, consider the example of modelling learning groups where the response is modelled at the group level and we have data where each student moves among groups. We can formulate this as a multiple membership model where groups (level 1) 'belong' to individuals (level 2). Suppose, in addition to measuring outcomes at the group level (y_{1j}) we also have a measure of achievement or attitude at the student level (y_{2i}). Recalling that the groups are defined as level 1 units, the group response will have an individual component and this will generally be correlated with the response at the student

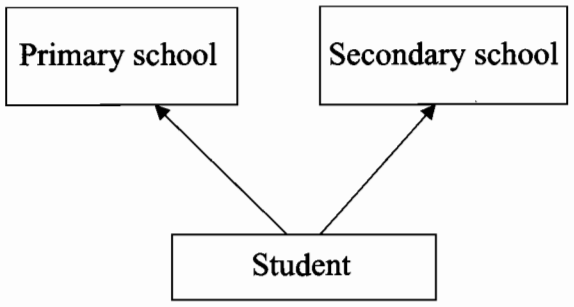

Fig. 3. Classification diagram showing students nested within a cross-classification or primary and secondary schools.

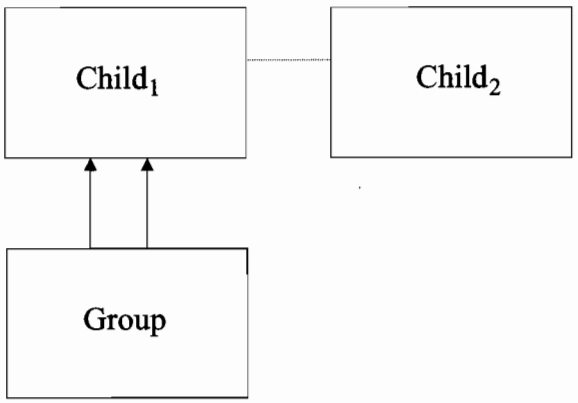

Fig. 4. Classification diagram for a bivariate response with multiple membership structure.

level. We could therefore write such a model as

$$
\begin{aligned}
&y_{1i} = (X_1\beta_1)_i + \sum_{j \in group(i)} w_{i,j} u_{1j}^{(2)} + e_i^{(1)}, \qquad \sum_{j \in group(i)} w_{i,j} = 1, \\
&y_{2j} = (X_2\beta_2)_j + u_{2j}^{(2)}, \\
&\operatorname{cov}\left(u_{1j}^{(2)}, u_{2j}^{(2)}\right) \neq 0.
\end{aligned}
\tag{5.7}
$$

Eq. (5.7) defines a bivariate response model with one response at each level. The first equation refers to a group response and, given suitable data with individuals belonging to different groups, can be used to estimate individual and group effects. The second equation models an individual student response, and from the complete model we can directly estimate the correlation between a student's contribution to the group response and their individual response. Figure 4 shows the relationships using a double arrow for the multiple membership of groups within children and a dotted line joining the two child 'effects' to indicate a bivariate response model.

We can also identify those individuals who may be discrepant, say with low contributions to the group response but high individual effects, and this might be an important diagnostic for learning potential. An alternative formulation for some purposes would be to incorporate the individual level measure as a covariate in the model for the group response. If, however, we had sequential measures on individuals then we might wish to fit trend terms with random coefficients and then the full bivariate response formulation becomes necessary (see Goldstein, 2003, Chapter 5, for such a model in the purely hierarchical case). Further elaborations can be introduced, for example by modelling classes of students (containing learning groups) within teachers and/or schools and so forth.

6. Further extensions

6.1. Further levels

In Sections 2–4 above, models for two-level hierarchical structures were described. All models can be extended to handle further levels of clustering. For example, sampling classes within schools would lead to a three-level structure. If we denote by y_{ijk} the response for student i in class j in school k, the random intercept model in (2.1) generalises to

$$y_{ijk} = \beta_{0jk} + \beta_1 x_{ijk} + e_{ijk}, \quad \beta_{0jk} = \beta_0 + u_{0jk} + v_{0k}, \tag{6.1}$$

where u_{0jk} is the random effect for class j in school k, and v_{0k} is the random effect for school k, both of which are assumed to follow normal distributions. The variance of u_{0jk} represents the between-class within-school variation. Eq. (6.1) may be extended to allow the coefficient of x_{ijk} to vary randomly across classes and/or schools.

6.2. Categorical responses

In the above sections we have described models for continuous responses. These models may be generalised to handle different types of response, including binary, ordered and unordered categorical, count and duration data. For example, a two-level logit model for binary responses may be written

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_{0j} + \beta_1 x_{ij}, \quad \beta_{0j} = \beta_0 + u_{0j} \tag{6.2}$$

where $\pi_{ij} = \Pr(y_{ij} = 1)$. Models for non-normal responses are described in Goldstein (2003, Chapter 4), and further applications are given in Rasbash et al. (2005).

It is also possible to handle mixtures of different response types in a multivariate model. For example, Goldstein (2003, pp. 105–107) describes an application where a binary response indicating whether an individual smokes, is modelled jointly with a continuous response (defined only for smokers) for the number of cigarettes smoked per day. More recently, multilevel factor models have been developed for mixtures of binary, ordinal and continuous items (Goldstein et al., 2006; Skrondal and Rabe-Hesketh, 2004; Steele and Goldstein, 2006).

6.3. Structural equation and item response models

Multilevel factor models may be generalised to structural equation models (SEM). Eq. (4.2) defines the measurement component of a SEM, which describes the relationship between the observed responses and the latent variables, possibly conditional on covariates. In a SEM, a structural component is added in which the latent variables at each level may depend on covariates and other latent variables. Detailed accounts of multilevel SEM include Muthén (1994) and Skrondal and Rabe-Hesketh (2004). A summary is given by Steele (2005).

The traditional (two parameter) item response model (Lord, 1980) is essentially a one-factor model with binary responses which can be written as

$$\text{logit}(\pi_{ri}) = \beta_r + \lambda_r \eta_i + e_{ri}, \quad \pi_{ri} = \Pr(y_{ri} = 1) \tag{6.3}$$

for a set of binary responses (r) for each individual (i). This is a special case of (4.2) but with a logit rather than an identity link function; we can also use a probit link function. In its original formulation each individual's score (or 'latent trait' value), η_i, was treated as a fixed effect but now would generally be formulated as a random effect in (6.3). Model (6.3) is readily extended to the multilevel case, as in (4.2), and to cross classified and multiple membership structures. It can also be extended to handle ordered categorical responses as in 'partial credit' models and a discussion of such models and estimation procedures is given by Fox (2005).

7. Estimation procedures and software

Estimation procedures can conveniently be divided into those based upon maximum likelihood, or approximations such as quasi-likelihood, and those based upon Bayesian Markov Chain Monte Carlo (MCMC) methods. We first look at likelihood-based procedures.

In the case of normally-distributed responses, the two most common procedures are the EM algorithm and the iterative generalized least squares (IGLS) or the related Fisher scoring algorithm. Goldstein (2003, Chapter 2 appendices) gives details of these. These methods are iterative and implemented in major statistics packages including SAS (SAS Institute, 1999), SPlus (Insightful, 2001), SPSS (SPSS, 2003), and Stata (StataCorp, 2005) as well as specialist multilevel modelling packages such as HLM (Raudenbush et al., 2001) and MLwiN (Rasbash et al., 2005). For non-hierarchically structured Normal responses, cross-classified models can be fitted using SAS, SPSS and MLwiN, and the Stata program GLLAMM (Rabe-Hesketh and Skrondal, 2005) while multiple membership models are implemented in MLwiN.

Where responses are discrete, for example binary or count data, we have generalized linear models and estimation tends to become more complicated. Maximum likelihood estimation is commonly carried out using a search procedure whereby the parameter space is explored in a search for the maximum of the likelihood. The computation of the likelihood for a given set of parameter values is commonly carried out using 'quadrature' but this can be very time consuming when there are many random parameters. Thus, for example, Ng et al. (2006) use a simulation-based procedure (see Goldstein,

2003, Appendix 4.3) to compute the likelihood and show that for a large number of parameters this has important timing advantages. Because of this, several approximate procedures are in use. One set of these uses quasi-likelihood estimation based on a Taylor series linearization. Known as marginal or penalized (predictive) quasi-likelihood (MQL, PQL) these will often provide satisfactory estimates but are known to be biased in cases where data are sparse or variances are large. Another procedure that has good properties is a Laplace integral transformation (Raudenbush et al., 2000). Any of these procedures can be used to provide starting values for either full maximum likelihood or MCMC estimation (see below). Another approach is to use an iterated bootstrap which will converge to unbiased estimates but has the disadvantages that it is time-consuming and does not directly provide standard error estimates. We shall not go into details of MCMC estimation procedures here but refer the reader to Browne et al. (2001) and Browne (2004). Multilevel models for discrete response data can be estimated using SAS, Stata or GLLAMM (all of which use quadrature), HLM (PQL and Laplace) and MLwiN (MQL/PQL and MCMC). Very general multilevel models, including those considered in this review, can be fitted using MCMC methods in WinBUGS (Spiegelhalter et al., 2000).

Multilevel factor analysis is implemented in Mplus (Muthén and Muthén, 2004; using two-stage weighted least squares), GLLAMM (adaptive quadrature), and MLwiN (using MCMC). Mplus and GLLAMM can also be used to fit more general structural equation models to any mixture of normal and discrete responses.

Recently published reviews of some of the packages mentioned above are those of De Leeuw and Kreft (2001), Zhou et al. (1999) and Fein and Lissitz (2000). Full reviews of the multilevel modelling capabilities of most mainstream statistical and specialist packages are maintained by the Centre for Multilevel Modelling (http://www.mlwin.com/softrev).

8. Resources

The methodological literature on multilevel modelling is growing rapidly as is the literature on applications. The Centre for Multilevel Modelling endeavours to maintain a selection of the methodological literature and links to other resources such as web sites and training materials. A collection of data sets together with training materials and a version of the MLwiN package that will work with these data sets, is freely available at http://tramss.data-archive.ac.uk. Another useful resource for multilevel modelling is http://www.ats.ucla.edu/stat/mlm/default.htm.

There is a very active email discussion group that can be accessed and joined at http://www.jiscmail.ac.uk/lists/multilevel.html. The group serves as a means of exchanging information and suggestions about data analysis.

References

Aitkin, M., Bonnet, S.N., Hesketh, J. (1981). Teaching styles and pupil progress: a reanalysis. *British Journal of Educational Psychology* **51**, 170–186.

Aitkin, M., Longford, N. (1986). Statistical modelling in school effectiveness studies. *Journal of the Royal Statistical Society A* **149**, 1–43.

Bennett, S.N. (1976). *Teaching Styles and Pupil Progress*. Open Books, London.

Bock, R.D. (Ed.) (1989). *Multilevel Analysis of Educational Data*. Academic Press, San Diego, CA.

Browne, W.J. (2004). *MCMC Estimation in MLwiN*. Institute of Education, London.

Browne, W., Goldstein, H., Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling* **1**, 103–124.

Bryk, A.S., Raudenbush, S.W. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd ed. Sage Publications, Newbury Park, CA.

De Leeuw, J., Kreft, I. (2001). Software for multilevel analysis. In: Leyland, A., Goldstein, H. (Eds.), *Multilevel Modelling of Health Statistics*. Wiley, Chichester.

Fein, M., Lissitz, R.W. (2000). Comparison of HLM and MLwiN multilevel analysis software packages: a Monte Carlo investigation into the quality of the estimates. Working Paper. University of Maryland.

Fox, J.P. (2005). Multilevel IRT using dichotomous and polytomous response data. *British Journal of Mathematical and Statistical Psychology* **58**, 145–172.

Goldstein, H. (1989). Flexible models for the analysis of growth data with an application to height prediction. *Rev. Epidem. et Sante Public* **37**, 477–484.

Goldstein, H. (2003). *Multilevel Statistical Models*, 3rd ed. Arnold, London.

Goldstein, H., Browne, W.J. (2002). Multilevel factor analysis modelling using Markov Chain Monte Carlo (MCMC) estimation. In: Marcoulides, G., Moustaki, I. (Eds.), *Latent Variable and Latent Structure Models*. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 225–243.

Goldstein, H., Healy, M.J.R., Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine* **13**, 1643–1655.

Goldstein, H., Bonnet, G., Rocher, T. (2006). Multilevel multidimensional structural equation models for the analysis of comparative data on educational performance. *Journal of Educational and Behavioral Statistics*. In press.

Insightful (2001). *SPlus 6 for Windows Guide to Statistics*. Insightful Corporation, Seattle, WA.

Laird, N.M., Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

Longford, N.T. (1993). *Random Coefficient Models*. Oxford University Press, New York.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Muthén, B.O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research* **22**, 376–398.

Muthén, B.O. (1997). Latent variable growth modelling with multilevel data. In: Berkane, M. (Ed.), *Latent Variable Modeling with Applications to Causality*. Springer-Verlag, New York, pp. 3149–3161.

Muthén, L.K., Muthén, B.O. (2004). *Mplus User's Guide, Version 3.0*. Muthén & Muthén, Los Angeles, CA.

Ng, E.S.W., Carpenter, J.R., Goldstein, H., Rasbash, J. (2006). Estimation of generalised linear mixed models with binary outcomes by simulated maximum likelihood. *Statistical Modelling* **6**, 23–42.

Rabe-Hesketh, S., Skrondal, A. (2005). *Multilevel and Longitudinal Modeling using Stata*. Stata Press, College Station, TX.

Rasbash, J., Steele, F., Browne, W., Prosser, B. (2005). *A User's Guide to MLwiN Version 2.0*. University of Bristol.

Raudenbush, S.W., Bryk, A., Cheong, Y.F., Congdon, R. (2001). *HLM 5: Hierarchical Linear and Nonlinear Modeling*. SSI, Lincolnwood.

Raudenbush, S.W., Yang, M., Yosef, M. (2000). Maximum likelihood for generalised linear models with nested random effects via high-order multivariate Laplace approximation. *Journal of Computational and Graphical Statistics* **9**, 141–157.

Robinson, W.S. (1951). Ecological correlations and the behaviour of individuals. *American Sociological Review* **15**, 351–357.

SAS Institute (1999). *SAS/STAT User's Guide, Version 7.1*. SAS Institute, Cary, NC.

Singer, J.D., Willett, J.B. (2003). *Applied Longitudinal Data Analysis*. Oxford University Press, New York.

Skrondal, A., Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Chapman & Hall/CRC, Boca Raton, FL.

Snijders, T.A.B., Bosker, R.J. (1999). *Multilevel Analysis*. Sage Publications, London.

Spiegelhalter, D.J., Thomas, A., Best, N.G. (2000). *WinBUGS Version 1.3 User Manual*. Medical Research Council Biostatistics Unit, Cambridge.

SPSS (2003). *SPSS. Advanced Models 12.0*. SPSS, Chicago.

StataCorp (2005). *Stata 9.0 Base Reference Manual*. Stata Press, College Station, TX.

Steele, F. (2005). Structural equation modeling: multilevel. In: Everitt, B., Howell, D. (Eds.), In: *Encyclopedia of Statistics in Behavioral Science*, vol. 4. John Wiley and Sons, Chichester, pp. 1927–1931.

Steele, F., Goldstein, H. (2006). A multilevel factor model for mixed binary and ordinal indicators of women's status. *Sociological Methods and Research*. In press.

Woodhouse, G., Goldstein, H. (1989). Educational performance indicators and LEA league tables. *Oxford Review of Education* **14**, 301–319.

Zhou, X., Perkins, A.J., Hui, S.L. (1999). Comparisons of software packages for generalized linear multilevel models. *American Statistician* **53**, 282–290.