J. R. Statist. Soc. A (2006)
169, *Part* 4, *pp.* 805–827

Multilevel modelling of complex survey data

Sophia Rabe-Hesketh

University of California, Berkeley, USA, and Institute of Education, London, UK

and Anders Skrondal

London School of Economics and Political Science, London, UK, and Norwegian Institute of Public Health, Oslo, Norway

[Received April 2005. Revised December 2005]

Summary. Multilevel modelling is sometimes used for data from complex surveys involving multistage sampling, unequal sampling probabilities and stratification. We consider generalized linear mixed models and particularly the case of dichotomous responses. A pseudolikelihood approach for accommodating inverse probability weights in multilevel models with an arbitrary number of levels is implemented by using adaptive quadrature. A sandwich estimator is used to obtain standard errors that account for stratification and clustering. When level 1 weights are used that vary between elementary units in clusters, the scaling of the weights becomes important. We point out that not only variance components but also regression coefficients can be severely biased when the response is dichotomous. The pseudolikelihood methodology is applied to complex survey data on reading proficiency from the American sample of the 'Program for international student assessment' 2000 study, using the Stata program gllamm which can estimate a wide range of multilevel and latent variable models. Performance of pseudo-maximum-likelihood with different methods for handling level 1 weights is investigated in a Monte Carlo experiment. Pseudo-maximum-likelihood estimators of (conditional) regression coefficients perform well for large cluster sizes but are biased for small cluster sizes. In contrast, estimators of marginal effects perform well in both situations. We conclude that caution must be exercised in pseudo-maximum-likelihood estimation for small cluster sizes when level 1 weights are used.

Keywords: Adaptive quadrature; Generalized linear latent and mixed model; Generalized linear mixed model; gllamm program; Multilevel model; Probability weighting; 'Program for international student assessment'; Pseudolikelihood; Sandwich estimator; Stratification

1. Introduction

Surveys often employ multistage sampling designs where clusters (or primary sampling units (PSUs)) are sampled in the first stage, subclusters in the second stage, etc., until elementary units are sampled in the final stage. This results in a multilevel data set, each stage corresponding to a level with elementary units at level 1 and PSUs at the top level L. At each stage, the units at the corresponding level are often selected with unequal probabilities, typically leading to biased parameter estimates if standard multilevel modelling is used. Longford (1995a, b, 1996), Graubard and Korn (1996), Korn and Graubard (2003), Pfeffermann *et al.* (1998) and others have discussed the use of sampling weights to rectify this problem in the context of two-level linear (or linear mixed) models, particularly random-intercept models. In this paper we consider generalized linear mixed models.

Address for correspondence: Sophia Rabe-Hesketh, 3659 Tolman Hall, University of California, Berkeley, CA 94720-1670, USA.
E-mail: sophiarh@berkeley.edu

 0964–1998/06/169805

When estimating models that are based on complex survey data, sampling weights are sometimes incorporated in the likelihood, producing a pseudolikelihood (e.g. Skinner (1989) and Chambers (2003)). For two-level linear models, Pfeffermann *et al.* (1998) implemented pseudo-maximum-likelihood estimation by using a probability-weighted iterative generalized least squares algorithm. For generalized linear mixed models, a weighted version of the iterative quasi-likelihood algorithm (e.g. Goldstein (1991)), which is analogous to probability-weighted iterative generalized least squares, is implemented in MLwiN (Rasbash *et al*., 2003). Unfortunately, this method is not expected to perform well since unweighted penalized quasi-likelihood often produces biased estimates, in particular when the responses are dichotomous (e.g. Rodríguez and Goldman (1995, 2001)). Furthermore, Renard and Molenberghs (2002) reported serious convergence problems and strange estimates when using MLwiN with probability weights for dichotomous responses.

A better approach for generalized linear mixed models is full pseudo-maximum-likelihood estimation, for instance via numerical integration. Grilli and Pratesi (2004) accomplished this by using SAS `NLMIXED` (Wolfinger, 1999) which implements maximum likelihood estimation for generalized linear mixed models by using adaptive quadrature. However, they had to resort to various tricks and the use of frequency weights at level 2 since probability weights are not accommodated. SAS `NLMIXED` is furthermore confined to models with no more than two levels. Another limitation is that it provides only model-based standard errors which are not valid for pseudo-maximum-likelihood estimation. Grilli and Pratesi (2004) therefore implemented an extremely computer-intensive nonparametric bootstrapping approach.

In this paper we describe full pseudo-maximum-likelihood estimation for generalized linear mixed models with any number of levels via adaptive quadrature (Rabe-Hesketh *et al*., 2005). Appropriate standard errors are obtained by using the sandwich estimator (Taylor linearization). Our approach is implemented in the Stata program `gllamm` (e.g. Rabe-Hesketh *et al*. (2002, 2004a) and Rabe-Hesketh and Skrondal (2005)), which allows specification of probability weights, as well as PSUs (if they are not included as the top level in the model) and strata. These methods are applied to the American sample of the 'Program for international student assessment' (PISA) 2000 study.

For linear mixed models Pfeffermann *et al.* (1998) pointed out that the scaling of the level 1 weights affects the estimates of the variance components, particularly the random-intercept variance, but may not have a large effect on the estimated regression coefficients (if the number of clusters is sufficiently large and the scaling constants do not depend on the responses). In contrast, for multilevel models for dichotomous responses we expect the estimated regression coefficients to be strongly affected by the scaling of the level 1 weights. This is because the regression coefficients are intrinsically related to the random-intercept variance. Specifically, for given marginal effects of the covariates on the response probabilities, the regression coefficients (which have conditional interpretations) are scaled by a multiplicative factor that increases as the random-intercept variance increases (see Section 3.2). Thus, the maximum likelihood estimates of the regression coefficients and the random-intercept variance are correlated in contrast with the linear case (e.g. Zeger *et al*. (1988)). As far as we are aware, this potential problem has not been investigated or pointed out before. Although Grilli and Pratesi (2004) considered pseudo-maximum-likelihood estimation for dichotomous responses, they focused mostly on the bias of the estimated random-intercept variance. Moreover, they simulated from models with small regression parameters (0 and 0.1), making it difficult to detect multiplicative bias unless it is extreme.

Using estimates from the multilevel model, approximate marginal effects can be obtained by rescaling the regression coefficients (conditional effects) according to the random-intercept

variance (e.g. Skrondal and Rabe-Hesketh (2004), page 125). We conjecture that these marginal effects will be less biased and less affected by the scaling of the level 1 weights than the original parameters. This would imply that marginal effects can be more reliably estimated in the presence of level 1 weights.

The plan of the paper is as follows. In Section 2 we briefly review descriptive and analytic inference for complex survey data with unequal selection probabilities. We then extend these ideas to multistage designs and introduce multilevel and generalized linear mixed models in Section 3. In Section 4 we suggest a pseudolikelihood approach to the estimation of multilevel and generalized linear mixed models incorporating sampling weights. We also describe various scaling methods for level 1 weights. In Section 5 we present a sandwich estimator for the standard errors of the pseudo-maximum-likelihood estimators, taking weighting into account. Having described the pseudolikelihood methodology, it is applied to a multilevel logistic model for complex survey data on reading proficiency among 15-year-old American students from the PISA 2000 study in Section 6. In Section 7 we carry out simulations to investigate the performance of pseudo-maximum-likelihood estimation using unscaled weights and different scaling methods. We also assess the coverage of confidence intervals based on the sandwich estimator and compare estimators by using different sampling designs at level 1. Finally, we close the paper with a discussion in Section 8.

2. Inverse probability weighting in surveys

In sample surveys, units are sometimes drawn with unequal selection probabilities. For example, lower selection probabilities may be assigned to units with higher data collection costs and higher selection probabilities to individuals from small subpopulations of particular interest. These *design probabilities* π_i for units i are a feature of the survey design and are assumed known before data analysis.

2.1. Descriptive inference

If the aim is to estimate finite population (census) quantities such as means, totals or proportions, a design-based approach is routinely used. Here the values of the variable of interest, y_i, are treated as fixed in a finite population and *design-based* inference considers the distribution of the estimator over repeated samples by using the same sampling design. The usual estimators such as the sample mean will be biased for the finite population quantity if the design probabilities are informative in the sense that they are related to the response y_i. A common solution is to use weighted estimators where the contribution of unit i is weighted by $w_i = 1/\pi_i$, the inverse probability of selection into the sample (e.g. Kish (1965) and Cochran (1977)). For instance, the Horvitz–Thompson estimator of the finite population mean is

$$\bar{y}^{\mathrm{HT}} = \frac{1}{\sum_i w_i} \sum_i w_i y_i.$$

In practice the construction of survey weights often also takes account of features other than design probabilities such as non-response adjustments and post-stratification. We shall return to this in Sections 6 and 8 but stick with design weights until then.

2.2. Analytic inference

Inverse probability weighting is also often used when the aim is *analytic inference*, such as estimation of the parameters of a data-generating mechanism or statistical *superpopulation* model.

Pfeffermann (1993, 1996) discussed this approach for estimating regression parameters $\boldsymbol{\beta}$ of a linear regression superpopulation model. Consider a hypothetical finite population for which the model holds. The properties of inference from survey data to model parameters can then be investigated by decomposing the problem into

(a) inference from the survey sample to the finite population and
(b) inference from the finite population to the model.

If data were available for the entire finite population, we could estimate $\boldsymbol{\beta}$ consistently by using ordinary least squares, giving the finite population parameters $\mathbf{b}_{\mathrm{p}}$. The estimator $\mathbf{b}_{\mathrm{p}}$ is thus *model consistent* for $\boldsymbol{\beta}$. However, in reality, we have only an estimator $\mathbf{b}_{\mathrm{s}}$ that is based on the sample and to make any claims about its consistency for $\boldsymbol{\beta}$ it must be demonstrated that $\mathbf{b}_{\mathrm{s}}$ is *design consistent* for $\mathbf{b}_{\mathrm{p}}$. Roughly speaking, this means that the estimator approaches the finite population parameter as both the finite population size and sample size tend to ∞ (e.g. Binder and Roberts (2003)); see Sen (1988) for a rigorous treatment.

Conventional estimators are not design consistent if the design probabilities are informative in the sense that they are related to the response even after conditioning on the covariates in the model (e.g. Pfeffermann (1996); see also Rubin (1976) and Little (1982)). In this case the model holding for the sample is different from the model holding for the finite population and superpopulation. Ignoring the sampling weights will therefore lead to biased estimates. For instance, if inclusion probabilities are positively correlated with the residual error in a linear regression model, the ordinary least squares estimator of the intercept will be positively biased.

To achieve design consistency the design variables determining the selection probabilities (or sometimes the weights themselves) could be included as covariates. This is an example of a disaggregated analysis because inference is conditional on the design variables. This approach is justifiable only if the extra conditioning does not alter the interpretation of the regression coefficients of interest in an undesirable manner (see also Pfeffermann (1996)). An alternative solution is to replace the usual estimators by their weighted counterparts, which is an example of an aggregated analysis. In the case of likelihood inference, this idea leads to a pseudolikelihood (e.g. Binder (1983), Skinner (1989) and Chambers and Skinner (2003)), where weights are incorporated as if they were frequency weights. The resulting estimator is design consistent and hence model consistent under suitable regularity conditions such as those discussed by Isaki and Fuller (1982) and Skinner (2005). However, this consistency typically comes at a price of reduced efficiency (e.g. Binder and Roberts (2003)).

3. Multistage sampling and multilevel models

3.1. Multistage sampling and probability weights

It is often not feasible to sample the elementary units i directly, for instance because the sampling frame is not known. Instead, *two-stage sampling* proceeds by sampling clusters or PSUs such as geographical regions or schools, in the first stage. Having obtained sampling frames for the sampled clusters, elementary units are subsequently sampled from the clusters in the second stage. If all units are included in the second stage, this is known as *cluster sampling*. Multistage sampling involves sampling (sub)clusters from clusters that were sampled in previous stages with elementary units sampled at the final stage. For notational simplicity, we consider two-stage sampling in this section.

At the initial stage, cluster j is sampled with probability π_j, $j = 1, \ldots, n^{(2)}$, and, at the subsequent stage, unit i is sampled with conditional probability $\pi_{i|j}$, $i = 1, \ldots, n_j^{(1)}$, given that cluster j was sampled in the first stage. Here and throughout the paper we use superscript (1) for 'level 1'

units i and (2) for 'level 2' units j. In a typical design, $n^{(2)}$ clusters are sampled with probabilities that are proportional to their sizes S_j (the number of units in the clusters),

$$\pi_j^{(2)} = n^{(2)} S_j \Big/ \sum_j S_j$$

and a constant number of units $n^{(1)}$ subsequently sampled for each cluster, corresponding to

$$\pi_{i|j}^{(1)} = n^{(1)}/S_j.$$

Such designs are self-weighting in the sense that all units have the same unconditional probability of selection,

$$\pi_{ij} = \pi_{i|j}^{(1)} \pi_j^{(2)} = n^{(2)} n^{(1)} \Big/ \sum_j S_j.$$

However, as we show in Section 4, the selection probabilities at each stage still need to be taken into account when taking a multilevel modelling approach.

3.2. Multilevel and generalized linear mixed models

Since there is usually unobserved heterogeneity between clusters even after conditioning on covariates, responses tend to be correlated within clusters. This dependence must be taken into account by using for instance multilevel modelling, which is an example of a disaggregated approach because the design variables defining the clusters are not aggregated over.

A two-level generalized linear mixed model (e.g. Breslow and Clayton (1993)) for response y_{ij} of unit i in cluster j can be specified as a generalized linear model with linear predictor

$$\nu_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}^{(2)\prime}\boldsymbol{\zeta}_j^{(2)}.$$

Here $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}^{(2)}$ are vectors of explanatory variables, $\boldsymbol{\beta}$ are fixed regression coefficients and $\boldsymbol{\zeta}_j^{(2)}$ are multivariate normal random effects varying over clusters with zero means and covariance matrix $\boldsymbol{\Psi}$. The conditional expectation μ_{ij} of y_{ij} (given $\boldsymbol{\zeta}_j^{(2)}$ and the covariates) is linked to the linear predictor ν_{ij} via a link function and the conditional distribution of y_{ij} is a member of the exponential family.

The regression parameters $\boldsymbol{\beta}$ represent conditional or cluster-specific effects of the covariates $\mathbf{x}_{ij}$ given the random effects $\boldsymbol{\zeta}_j^{(2)}$. For certain link functions such as the logit and probit the conditional effects will generally differ from the marginal or population-averaged effects (integrated over the random effects); see Ritz and Spiegelman (2004). Consider a two-level logistic random-intercept model for unit i in cluster j,

$$\log\left\{\frac{\Pr(y_{ij}=1|\mathbf{x}_{ij}, \zeta_j^{(2)})}{\Pr(y_{ij}=0|\mathbf{x}_{ij}, \zeta_j^{(2)})}\right\} = \nu_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \zeta_j^{(2)}, \qquad \zeta_j^{(2)} \sim N(0, \psi). \tag{1}$$

This model can alternatively be written in terms of a continuous latent response y_{ij}^* linked to the dichotomous observed response y_{ij} via a threshold model

$$\begin{aligned} y_{ij} &= I(y_{ij}^* > 0), \\ y_{ij}^* &= \mathbf{x}_{ij}'\boldsymbol{\beta} + \zeta_j^{(2)} + \varepsilon_{ij}, \end{aligned} \tag{2}$$

where $I(\cdot)$ is the indicator function and the ε_{ij} have independent logistic distributions with variance $\pi^2/3$. This latent response formulation is not only useful for the simulation that is described in Section 7 but also instrumental in investigating identification and equivalence in

latent variable models with categorical responses (e.g. Rabe-Hesketh and Skrondal (2001)). For the two-level logistic random-intercept model the total residual $\zeta_j^{(2)} + \varepsilon_{ij}$ has variance $\psi + \pi^2/3$ and intraclass correlation

$$\rho = \frac{\psi}{\psi + \pi^2/3}.$$

The marginal effects $\boldsymbol{\beta}^{\mathrm{M}}$ are therefore approximately related to the conditional effects $\boldsymbol{\beta}$ via

$$\begin{aligned}\boldsymbol{\beta}^{\mathrm{M}} &\approx \sqrt{\left(\frac{\pi^2/3}{\psi + \pi^2/3}\right)}\boldsymbol{\beta} \\ &= \sqrt{\left(\frac{1}{1 + 0.30\psi}\right)}\boldsymbol{\beta}, \end{aligned} \tag{3}$$

with $|\boldsymbol{\beta}^{\mathrm{M}}| < |\boldsymbol{\beta}|$ if $\psi > 0$. Aggregated approaches such as generalized estimating equations or ordinary logistic regression estimate the marginal coefficients $\boldsymbol{\beta}^{\mathrm{M}}$ directly. However, here we focus on conditional effects and variance components which are the parameters of primary interest in multilevel modelling.

All these ideas naturally generalize to models with more than two levels. For L levels the generalized linear mixed model has linear predictor

$$\nu = \mathbf{x}'\boldsymbol{\beta} + \sum_{l=2}^{L} \mathbf{z}^{(l)\prime}\boldsymbol{\zeta}^{(l)},$$

where subscripts have been omitted for notational simplicity. The random effects $\boldsymbol{\zeta}^{(l)}$ at each level l are multivariate normal with zero means and are uncorrelated with the random effects at the other levels.

4. Pseudo-maximum-likelihood estimation for multilevel models

4.1. Conventional likelihood

Let $\boldsymbol{\vartheta}$ be the vector of all parameters, including the fixed effects $\boldsymbol{\beta}$ and the unique elements of the covariance matrix of the random effects $\boldsymbol{\zeta}^{(l)}$ at levels $l = 2, \ldots, L$.

We let $\mathbf{y}_{jk(2)}$ denote the response vector for level 2 unit j in level 3 unit k (omitting subscripts for higher level units) and $\boldsymbol{\zeta}^{(l^+)} = (\boldsymbol{\zeta}^{(l)\prime}, \boldsymbol{\zeta}^{(l+1)\prime}, \ldots, \boldsymbol{\zeta}^{(L)\prime})'$. The log-likelihood contribution of a level 2 unit, conditional on the random effects at levels 3 and above, can be expressed as

$$\mathcal{L}_{jk}^{(2)}(\mathbf{y}_{jk(2)}|\boldsymbol{\zeta}_k^{(3^+)}) = \log\left[\int \exp\left\{\sum_{i=1}^{n_{jk}^{(1)}} \mathcal{L}_{ijk}^{(1)}(y_{ijk}|\boldsymbol{\zeta}_{jk}^{(2^+)})\right\} g^{(2)}(\boldsymbol{\zeta}_{jk}^{(2)})\,\mathrm{d}\boldsymbol{\zeta}_{jk}^{(2)}\right],$$

where $\mathcal{L}_{ijk}^{(1)}(y_{ijk}|\boldsymbol{\zeta}_{jk}^{(2+)})$ is the log-likelihood of a level 1 unit given all random effects and $g^{(2)}(\boldsymbol{\zeta}_{jk}^{(2)})$ is the multivariate normal density of the random effects at level 2. For notational simplicity we have omitted explicit reference to $\boldsymbol{\vartheta}$, $\mathbf{x}$ and $\mathbf{z}^{(l)}$ in the log-likelihoods and will continue to do so.

Using subscripts q for level $l-1$ units and r for level l units, the log-likelihood at level l conditional on $\boldsymbol{\zeta}^{([l+1]^+)}$ is

$$\mathcal{L}_r^{(l)}(\mathbf{y}_{r(l)}|\boldsymbol{\zeta}^{([l+1]^+)}) = \log\left[\int \exp\left\{\sum_{q=1}^{n_r^{(l-1)}} \mathcal{L}_{qr}^{(l-1)}(\mathbf{y}_{qr(l-1)}|\boldsymbol{\zeta}_r^{(l^+)})\right\} g^{(l)}(\boldsymbol{\zeta}_r^{(l)})\,\mathrm{d}\boldsymbol{\zeta}_r^{(l)}\right]. \tag{4}$$

Applying equation (4) recursively for $l = 2, \ldots, L$, we obtain the required log-likelihood as

$$\mathcal{L}(\mathbf{y}) = \sum_{t=1}^{n^{(L)}} \mathcal{L}_t^{(L)}(\mathbf{y}_t),$$

where t is the subscript for the highest level L and $\mathbf{y}$ is the vector of all responses.

4.2. Pseudolikelihood

Let $w_{q|r}^{(l-1)}$ denote the inverse probability that the qth level $l-1$ unit in the rth level l unit was selected conditionally on the rth level l unit having been selected. The log-pseudolikelihood is defined by replacing equation (4) with

$$\mathcal{L}_r^{(l)}(\mathbf{y}_{r(l)}|\boldsymbol{\zeta}^{([l+1]+)}) = \log\left[\int \exp\left\{\sum_{q=1}^{n_r^{(l-1)}} w_{q|r}^{(l-1)}\, \mathcal{L}_{qr}^{(l-1)}(\mathbf{y}_{qr(l-1)}|\boldsymbol{\zeta}_r^{(l+)})\right\} g^{(l)}(\boldsymbol{\zeta}_r^{(l)})\, \mathrm{d}\boldsymbol{\zeta}_r^{(l)}\right], \qquad l = 2, \ldots, L,$$

giving the log-pseudolikelihood as

$$\mathcal{L}(\mathbf{y}) = \sum_{t=1}^{n^{(L)}} w_t^{(L)}\, \mathcal{L}_t^{(L)}(\mathbf{y}_t). \tag{5}$$

Here the weights enter the log-pseudolikelihood as if they were frequency weights, representing the number of times that each unit should be replicated. Adaptive quadrature (e.g. Rabe-Hesketh *et al*. (2002, 2005)) provides good approximations to the integrals in the pseudolikelihood and this approach is implemented in `gllamm`.

It is clear from the form of the log-pseudolikelihood that we cannot simply use one set of weights based on the overall inclusion probabilities but must use separate weights at each level. Consequently, the self-weighting property of many multistage designs is lost.

4.3. Level 1 weights and bias in multilevel models

Unfortunately, pseudo-maximum-likelihood estimation is not as straightforward for multilevel models as it is for conventional single-level models. One issue that was discussed by Pfeffermann *et al.* (1998) for two-level linear mixed models is that, although consistency for the regression coefficients requires only that the number of clusters $n^{(2)}$ increases, both $n^{(2)}$ and the number of units $n_j^{(1)}$ per cluster must increase to ensure consistency for the variance components.

Another issue is that the scaling of weights, which is immaterial in single-level models, can now affect the estimates if the scaling is applied at level 1. Attempts have been made to devise methods for scaling the level 1 weights that reduce the bias in the variance components for small cluster sizes.

4.3.1. Bias of variance component estimators and effects of scaling

For simplicity, we shall consider a two-level linear variance components model

$$y_{ij} = \beta_0 + \zeta_j + \varepsilon_{ij}, \qquad \zeta_j \sim N(0, \psi), \quad \varepsilon_{ij} \sim N(0, \theta),$$

and assume that the data are balanced with $n_j^{(1)} = n^{(1)}$. To focus on the problems that are associated with level 1 weights, we let all level 2 units from the finite population be included in the sample.

Analytical expressions for the maximum likelihood estimators are given by (e.g. McCulloch and Searle (2001), page 39)

$$\hat{\beta}_0 = \bar{y}_{..}, \tag{6}$$

$$\hat{\theta} = \frac{\sum_{j=1}^{n^{(2)}} \sum_{i=1}^{n^{(1)}} (y_{ij} - \bar{y}_{.j})^2}{n^{(2)}(n^{(1)} - 1)} \tag{7}$$

and

$$\hat{\psi} = \frac{\sum_{j=1}^{n^{(2)}} (\bar{y}_{.j} - \bar{y}_{..})^2}{n^{(2)}} - \frac{\hat{\theta}}{n^{(1)}}. \tag{8}$$

(These estimators for θ and ψ only apply if $\hat{\psi}$ is positive.)

We can use level 1 weights $w_{i|j}^{(1)}$ by replacing all sums over i by weighted sums and $n^{(1)}$ by

$$w_{\cdot|j} = \sum_{i=1}^{n^{(1)}} w_{i|j}^{(1)},$$

which we shall refer to as the 'apparent' cluster size.

The problem with this approach is that the between-cluster variance ψ is overestimated. This can be seen by considering the weighted version of the first term in equation (8), letting the level 1 weights have constant sums $w_{\cdot|j}$ for simplicity. The expectation of the first term of the weighted estimator $\hat{\psi}$ is a sum of a contribution due to between-cluster variability ψ (not affected by the weighting) and a contribution due to within-cluster variability θ. The latter is given by

$$\frac{1}{n^{(2)}} E\left\{ \sum_{j=1}^{n^{(2)}} (\bar{\varepsilon}_{.j} - \bar{\varepsilon}_{..})^2 \right\} = \frac{n^{(2)} - 1}{n^{(2)}} H \frac{\theta}{w_{\cdot|j}} \approx H \frac{\theta}{w_{\cdot|j}}, \tag{9}$$

where the approximation is good when $n^{(2)}$ is large and

$$H = \frac{1}{n^{(2)}} \sum_{j=1}^{n^{(2)}} \frac{\sum_{i=1}^{n^{(1)}} w_{i|j}^2}{w_{\cdot|j}}.$$

Without weights ($w_{i|j} = 1$), $H = 1$ and subtraction of the weighted version of the second term in equation (8) makes $\hat{\psi}$ consistent. However, for large weights ($H > 1$), subtraction of the second term does not suffice (note that the weighted version of $\hat{\theta}$ has expectation less than θ). Intuitively, the scaling makes the clusters appear larger without reducing the between-cluster variability due to the ε_{ij}. This extra between-cluster variability is incorrectly attributed to ψ.

If sampling of level 1 units is within strata that are determined by the sign of ε_{ij} (as in Pfeffermann *et al.* (1998) and the simulations of this paper), this stratification will reduce the expectation in expression (9) (which is not compensated for by a corresponding change in $\hat{\theta}$), leading to smaller estimates $\hat{\psi}$ than in the unstratified case. We expect qualitatively similar behaviour for generalized linear mixed models where analytic investigation of estimators is not possible.

Scaling the weights at the top level L by multiplying by a scalar a simply results in the log-pseudolikelihood being rescaled and therefore does not affect the point estimates. In contrast, scaling the lower level weights does affect the parameter estimates even if a constant scaling

factor $a^{(l)}$ is used at level l. For the linear variance components model for balanced data, it is clear from equation (6) that rescaling the weights as $a^{(1)}w_{i|j}^{(1)}$ does not affect the estimator of β_0. However, for the variance components, we obtain

$$\hat{\theta}_a = \frac{w_{\cdot|j} - 1}{w_{\cdot|j} - 1/a^{(1)}}\hat{\theta}_w,$$

and

$$\hat{\psi}_a = \hat{\psi}_w + \frac{\hat{\theta}_w}{w_{\cdot|j}}\left(1 - \frac{w_{\cdot|j} - 1}{a^{(1)}w_{\cdot|j} - 1}\right),$$

where $\hat{\theta}_w$ and $\hat{\psi}_w$ denote the weighted estimators using weights $w_{i|j}^{(1)}$ and $\hat{\theta}_a$ and $\hat{\psi}_a$ denote the estimators using scaled weights $a^{(1)}w_{i|j}^{(1)}$. $\hat{\theta}_a$ decreases and $\hat{\psi}_a$ increases with the scaling factor $a^{(1)}$ for a given sample, but the effect of $a^{(1)}$ decreases when the apparent cluster size $w_{\cdot|j}$ based on the raw weights becomes large, for instance when the actual cluster size $n^{(1)}$ becomes large. The increase in $\hat{\psi}_a$ is again related to the increase in apparent cluster size $w_{\cdot|j}^a = a^{(1)}w_{\cdot|j}$.

4.3.2. Weighting schemes

The two most common scaling methods for the level 1 weights are as follows.

(a) *Method 1*: Longford (1995a, b, 1996) argued that the scaling factor $a_1^{(1)}$ should be determined so that the 'apparent' cluster size $w_{\cdot|j}^a$ equals the 'effective' sample size (e.g. Pothoff *et al.* (1992)),

$$\begin{aligned} w_{\cdot|j}^a &= \sum_{i=1}^{n_j^{(1)}} a_1^{(1)} w_{i|j}^{(1)} \\ &= \frac{w_{\cdot|j}^2}{\sum w_{i|j}^2} \leqslant n_j^{(1)}, \end{aligned}$$

so the scale factor, which was referred to as 'method 1' in Pfeffermann *et al.* (1998), becomes

$$a_1^{(1)} = \frac{w_{\cdot|j}}{\sum_{i=1}^{n_j^{(1)}} (w_{i|j}^{(1)})^2}.$$

This is motivated by unbiasedness of the resulting weighted moment estimator of θ which coincides with the maximum likelihood estimator in the balanced case (Longford (1996), page 336). Similarly, Pfeffermann *et al.* (1998) investigated the performance of the probability-weighted iterative generalized least squares estimators for the variance components in a two-level linear mixed model with a single random effect. Assuming that the level 1 weights (but not the level 2 weights) are approximately non-informative and that the weights are uncorrelated with the covariate multiplying the random effect, they showed that the estimators of both variance components are approximately unbiased if scaling method 1 is used.

(b) *Method 2*: another obvious choice of scaling factor is one that sets the apparent cluster size $w_{\cdot|j}^a$ equal to the actual cluster size $n_j^{(1)}$, which is referred to as scaling method 2 in Pfeffermann *et al.* (1998),

$$a_2^{(1)} = n_j^{(1)}/w_{\cdot|j}.$$

Simulations in Pfeffermann *et al.* (1998) suggest that this method works better than method 1 for informative weights. Such a scaling factor has also been used by Clogg and Eliason (1987) in a different context.

Instead of scaling the level 1 weights, Graubard and Korn (1996) suggested a 'method D' which does not use any weights at level 1.

(c) *Method D*: new level 2 weights w_j^* are constructed as

$$w_j^* = \sum_{i=1}^{n_j^{(1)}} w_{i|j} w_j,$$

and level 1 weights are $w_{i|j}^* = 1$.

Korn and Graubard (2003) pointed out that moment estimators of the variance components using these weights are approximately unbiased under non-informative sampling at level 1.

5. Sandwich estimator of the standard errors

From standard likelihood theory (e.g. Pawitan (2001), pages 372–374 and 407), the asymptotic covariance matrix of the maximum likelihood estimator is

$$\mathrm{cov}(\hat{\boldsymbol{\vartheta}}) = \mathcal{I}^{-1} \mathcal{J} \mathcal{I}^{-1}. \tag{10}$$

Here $\mathcal{I}$ is the expected Fisher information and

$$\mathcal{J} \equiv E\left\{ \frac{\partial \mathcal{L}(y;\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} \frac{\partial \mathcal{L}(y;\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}'} \right\}\Bigg|_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}_0},$$

where $\boldsymbol{\vartheta}_0$ is the true parameter vector and the expectations are over the (true) distribution of the responses given the covariates. For model-based standard errors, the sandwich form of the covariance matrix in equation (10) collapses to $\mathcal{I}^{-1}$ because $\mathcal{J} = \mathcal{I}$ if the likelihood represents the true distribution of the responses (given the covariates). The expected Fisher information $\mathcal{I}$ is typically estimated by the observed Fisher information I at the maximum likelihood estimates. Since the pseudolikelihood does not represent the distribution of the responses, the sandwich does not collapse. Instead, we estimate $\mathrm{cov}(\hat{\boldsymbol{\vartheta}})$ by $\widehat{\mathrm{cov}}(\hat{\boldsymbol{\vartheta}}) = I^{-1} J I^{-1}$, where I is the observed (pseudo-) Fisher information at the pseudo-maximum-likelihood estimates and the estimator J of $\mathcal{J}$ is obtained by exploiting the fact that the pseudolikelihood is a sum of independent cluster contributions so that

$$\begin{aligned} \frac{\partial \mathcal{L}(\mathbf{y};\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} &= \sum_{t=1}^{n^{(L)}} w_t^{(L)} \frac{\partial \mathcal{L}^{(L)}(\mathbf{y}_{t(L)};\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} \\ &\equiv \sum_{t=1}^{n^{(L)}} \mathbf{S}_t(\boldsymbol{\vartheta}). \end{aligned}$$

We then estimate $\mathcal{J}$ by

$$\begin{aligned} J &= \frac{n^{(L)}}{n^{(L)}-1} \sum_{t=1}^{n^{(L)}} \mathbf{S}_t(\hat{\boldsymbol{\vartheta}}) \mathbf{S}_t(\hat{\boldsymbol{\vartheta}})' \\ &\equiv \frac{n^{(L)}}{n^{(L)}-1} \sum_{t=1}^{n^{(L)}} \mathbf{s}_t \mathbf{s}_t', \end{aligned}$$

where $\mathbf{s}_t$ is the weighted score vector of the top level unit t.

We now consider a more complex design where the top level units of the multilevel model are clustered in even higher level clusters. We need to consider only the highest level clusters or PSUs which may have been sampled using stratified sampling. To accommodate this situation we shall use $\mathbf{s}_{hgt}$ for the weighted score vector of the top level unit t in stratum h, $h=1,\ldots,H$, and cluster g, $g=1,\ldots,G_h$, where t, $t=1,\ldots,N_{hg}$, is now an index within stratum h and cluster g. The gradient of the log-pseudolikelihood can then be expressed as

$$\left.\frac{\partial \mathcal{L}(\mathbf{y};\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}\right|_{\boldsymbol{\vartheta}=\hat{\boldsymbol{\vartheta}}} = \sum_{h=1}^{H}\sum_{g=1}^{G_h}\sum_{t=1}^{N_{hg}} \mathbf{s}_{hgt}.$$

The corresponding covariance matrix, taking stratification and additional clustering into account, becomes

$$J = \sum_{h=1}^{H} \frac{G_h}{G_h-1} \sum_{g=1}^{G_h} (\mathbf{s}_{hg\cdot} - \bar{\mathbf{s}}_{h\cdot\cdot})(\mathbf{s}_{hg\cdot} - \bar{\mathbf{s}}_{h\cdot\cdot})',$$

where

$$\mathbf{s}_{hg\cdot} = \sum_{t=1}^{N_{hg}} \mathbf{s}_{hgt},$$

$$\bar{\mathbf{s}}_{h\cdot\cdot} = \frac{1}{G_h}\sum_{g=1}^{G_h} \mathbf{s}_{hg\cdot}.$$

Pseudolikelihood inference for complex surveys is discussed in Skinner (1989). The sandwich estimator that is described in this section has been implemented in `gllamm`.

As we shall see in our application, the procedures that were described above allow us to adopt a hybrid aggregated–disaggregated approach where lower levels of substantive interest are explicitly included in the multilevel model, whereas PSUs are considered a nuisance and are used only to adjust the standard errors.

6. Application

We analyse data from the 2000 Organisation for Economic Co-operation and Development PISA study on reading proficiency among 15-year-old American students.

In a three-stage cluster sampling design, geographic areas (PSUs) were sampled at stage 1, schools at stage 2 and students at stage 3. Stage 1 yielded 52 PSUs. In stage 2, public schools with more than 15% minority students were twice as likely to be sampled as other schools. Within high minority and other schools, the probability of selection was proportional to an estimate of size. Of the 220 schools that were sampled, only 128 were both eligible and willing to participate. These schools were supplemented with 32 replacement schools each having similar characteristics to those of a non-participating school. In stage 3, up to 35 students were sampled from 160 schools. In public schools with more than 15% minority students, minority students were oversampled (Lemke *et al*. (2001), appendix 1), but otherwise all students aged 15 years had an equal chance of being selected. Many of the students sampled did not participate owing to ineligibility, withdrawal, exclusion or failure to take assessments. 145 schools with more than 50% student participation were classified as 'responding' and eight schools with between 25% and 50% responding as 'partially responding'. These 153 schools with a total of 3846 participating students are included in the PISA database.

The PISA data include weights at the school and student levels. According to the manual for the PISA 2000 database (Organisation for Economic Co-operation and Development, 2000), the provided school level weights v_j^{pr} (called WNRSCHBW) are design weights $w_j = 1/\pi_j$ adjusted for school non-response,

$$v_j^{\mathrm{pr}} = f_{1j} w_j.$$

Here f_{1j} compensates for non-participation by other schools that are similar to school j (in terms of variables including region, metropolitan or non-metropolitan status, percentage minority and percentage of students eligible for free lunch).

The provided student weights v_{ij}^{pr} (called W_FSTUWT) are given by

$$v_{ij}^{\mathrm{pr}} = f_{1j} f_{2j} f_{1j}^{\mathrm{A}} w_{i|j} w_j.$$

Here, $w_{i|j} = 1/\pi_{i|j}$ are design weights at level 1, f_{1j}^{A} adjusts for non-inclusion by some schools of 15-year-old students from grades other than the modal grade for 15-year-old students and f_{2j} adjusts for non-participation of students who are included in the sample. Note that all terms in v_{ij}^{pr} except $w_{i|j}$ are school specific and that these terms do not affect the rescaled version of v_{ij}^{pr} by using either method 1 or 2.

We consider the response variable [Proficiency], an indicator taking the value 1 for the two highest reading proficiency levels as defined in Organisation for Economic Co-operation and Development (2000). Specifically, the threshold 552.89 was applied to the weighted maximum likelihood estimates (Warm, 1989) of reading ability. Ability scoring was based on a partial credit model, estimated by maximum marginal likelihood on a subset of the international data; see Adams and Wu (2002) for details.

As student level explanatory variables we use gender and most of the family background variables that were considered in Lemke *et al*. (2001):

(a) [Female]—the student is female (dummy);
(b) [ISEI]—international socio-economic index (see Ganzeboom *et al*. (1992));
(c) [Highschool]—highest education level by either parent is high school (dummy);
(d) [College]—highest education level by either parent is college (dummy);
(e) [English]—the test language (English) is spoken at home (dummy);
(f) [Oneforeign]—one parent is foreign born (dummy);
(g) [Bothforeign]—both parents are foreign born (dummy).

We also consider *contextual* or *compositional* effects of socio-economic status, i.e. the difference between the between-school and within-school effects. This has attracted considerable interest in education (e.g. Willms (1986) and Raudenbush and Bryk (2002)). For instance, Willms (1986) argued that the benefits of comprehensive schooling depend to a large extent on whether the socio-economic mix of a school has an effect on students' outcomes above the effect of individual student characteristics. In addition to the student level socio-economic index [ISEI] we therefore also consider its school mean [MnISEI] as a school level covariate.

We use the two-level random-intercept logistic regression model (1) for student i in school j, where [Proficiency] is regressed on the covariates that are mentioned above. The PISA database does not include any identifier for the PSUs but this information was kindly provided by the National Council for Education Statistics. We do not include PSUs as a level in the model because the variance between PSUs (with undisclosed definition) does not appear to be of substantive interest and estimation would require knowledge of the PSU selection probabilities. PSUs were instead accounted for in the sandwich estimator of the standard errors. Because of missing data on some of the covariates (mostly for [Highschool], [College] and [ISEI]), estima-

Table 1. Maximum likelihood estimates (with model-based and robust standard errors) and pseudo-maximum-likelihood estimates by using scaling method 1 (with robust standard errors taking and not taking PSUs into account)

Parameter	*Unweighted maximum likelihood*				*Weighted pseudo-maximum-likelihood*		
	Estimate	*SE*	SE_R	SE_R^{PSU}	*Estimate*	SE_R	SE_R^{PSU}
β_0, [Constant]	−6.034	0.539	0.547	0.458	−5.878	0.955	0.738
β_1, [Female]	0.555	0.103	0.102	0.111	0.622	0.154	0.161
β_3, [ISEI]	0.014	0.003	0.003	0.003	0.018	0.005	0.004
β_4, [MnISEI]	0.069	0.009	0.009	0.009	0.068	0.016	0.018
β_5, [Highschool]	0.400	0.256	0.262	0.224	0.103	0.477	0.429
β_6, [College]	0.721	0.255	0.257	0.235	0.453	0.505	0.543
β_7, [English]	0.695	0.285	0.269	0.301	0.625	0.382	0.391
β_8, [Oneforeign]	−0.020	0.224	0.200	0.159	−0.109	0.274	0.225
β_9, [Bothforeign]	0.099	0.236	0.245	0.295	−0.280	0.326	0.292
ψ	0.271	0.086	0.082	0.088	0.296	0.124	0.115

tion was based on 2069 students from 148 schools in 46 PSUs. The rescaling of level 1 weights was based on the estimation sample.

For the 148 schools contributing to the analysis, the provided school level weights v_j^{pr} had mean 262, standard deviation 539, lower decile 32 and upper decile 499. For the 2069 students, the provided student level weights v_{ij}^{pr} had mean 843, standard deviation 410, lower decile 347 and upper decile 1353. Because of a very large intraclass correlation of 0.98, the rescaled student level weights using methods 1 or 2 were close to 1 and almost identical, both having standard deviations of 0.05 and lower and upper deciles of 0.94 and 1.07 respectively.

Estimates using (unweighted) maximum likelihood and pseudo-maximum-likelihood with scaling method 1 are shown in Table 1. We used 12-point and 20-point adaptive quadrature, giving the same results to the precision that is reported. Model-based standard errors SE are given (for maximum likelihood only), together with robust standard errors from the sandwich estimator not taking PSUs into account (SE_R) and taking PSUs into account (SE_R^{PSU}). Estimates using scaling method 2 (which are not shown) were almost identical to those using method 1 because the level 1 weights are close to 1.

The pseudo-maximum-likelihood estimates are in accordance with educational theory and previous research. For instance, controlling for other covariates, reading proficiency is better for females than for males, better for students with parents having higher levels of education and better for students having English as their home language. As expected, socio-economic status [ISEI] has a positive effect; for students from the same school, a one within-school standard deviation change in [ISEI] of 15.5 is associated with an increase in the log-odds of 0.22, controlling for the other covariates. For students from different schools, a 1 standard deviation change in school mean socio-economic status [MnISEI] of 8.9 is associated with an increase in the log-odds of 0.74 after controlling for student level [ISEI] and the other covariates. There is thus evidence of a contextual effect of socio-economic status. The intraclass correlation of the latent responses y_{ij}^* in equation (2), given the covariates, is estimated as about 0.08 (0.14 when the school level covariate [MnISEI] is excluded).

Many of the pseudo-maximum-likelihood estimates are very different from the corresponding unweighted maximum likelihood estimates. This illustrates the importance of weighting and

suggests that sampling probabilities are informative in the present application. However, the loss in efficiency due to weighting is also apparent with substantially larger standard errors for pseudo-maximum-likelihood estimators. Note that taking PSUs into account does not necessarily increase the standard errors.

7. Simulation

A Monte Carlo experiment was carried out to assess the performance of pseudo-maximum-likelihood estimation and the sandwich estimator.

First, dichotomous responses were simulated for a finite population from the two-level logistic random-intercept model (1) with linear predictor

$$\nu_{ij} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2ij} + \zeta_j^{(2)},$$

and $\beta_0 = \beta_1 = \beta_2 = 1$ and $\psi = 1$. Since performance of estimators may differ for coefficients of between-cluster and within-cluster covariates, we simulated both types of covariate. For both types of covariate we drew independent samples from a Bernoulli distribution with probability 0.5. For the between-cluster covariate x_{1j}, a single value was sampled for the entire cluster and, for the unit-specific covariate x_{2ij}, different values were first sampled for each unit and then the cluster mean was subtracted (so that x_{2ij} varied purely within clusters). The finite population had 500 level 2 units, each with the same number $N_j^{(1)}$ of level 1 units.

Second, we sampled from the finite population by using the following two-stage sampling design. A subset of the level 2 units were sampled by using stratified random sampling without replacement with approximate (due to rounding) sampling fractions

$$\pi_j^{(2)} \approx \begin{cases} 0.25 & \text{if } |\zeta_j^{(2)}| > 1, \\ 0.75 & \text{if } |\zeta_j^{(2)}| \leqslant 1. \end{cases}$$

The average overall sampling fraction is about 0.6, yielding about 300 level 2 units.

From each level 2 unit, level 1 units were sampled (again by using stratified random sampling without replacement) with approximate sampling fractions

$$\pi_{i|j}^{(1)} \approx \begin{cases} 0.25 & \text{if } \varepsilon_{ij} > 0 \\ 0.75 & \text{if } \varepsilon_{ij} \leqslant 0, \end{cases}$$

where ε_{ij} are the residuals in the latent response formulation (2). Sampling at level 1 was similar to the proportional allocation method that was used by Pfeffermann *et al.* (1998) and Grilli and Pratesi (2004) except that our sampling fraction is higher, nearly half the units ($n_j^{(1)} \approx N_j^{(1)}/2$) being sampled from each cluster.

By making the sampling probabilities at stages 1 and 2 dependent on the corresponding residuals $\zeta_j^{(2)}$ and ε_{ij}, we are ensuring that sampling is informative at both levels. In practice, sampling probabilities would depend on observed design variables and our simulation corresponds to the situation where these design variables are strongly associated with the residuals. In a school survey, sampling at stage 1 could be stratified by school size or type (e.g. private and public) where the oversampled schools are more homogeneous with school level residuals closer to 0 (with a Pearson correlation between stratification variable and absolute value of school level residual of 0.82). At stage 2, strata could be based on some student characteristic (e.g. minority status) which correlates with the student level residual (here a correlation of 0.76).

The weights $w_j^{(2)} = 1/\pi_j^{(2)}$ and $w_{i|j}^{(1)} = 1/\pi_{i|j}^{(1)}$ that were used in pseudo-maximum-likelihood estimation were based on the proportions of units that were actually sampled. (Because of the

small strata that are involved when sampling level 1 units from level 2 units, the proportion that was sampled could differ considerably from 0.25 and 0.75.)

We varied the cluster sizes of the finite population $N_j^{(1)} \in \{5, 10, 20, 50, 100\}$ and simulated 100 data sets for each condition. Although a stratified sampling design at level 1 may be unusual with cluster sizes as small as 5 or 10, these situations might correspond to longitudinal data (occasions nested in subjects) where missingness depends on a time-varying covariate that is correlated with the level 1 residual. Five estimation methods were used for each simulated data set:

(a) unweighted maximum likelihood,
(b) pseudo-maximum-likelihood using raw unscaled weights,
(c) pseudo-maximum-likelihood using scaling method 1,
(d) pseudo-maximum-likelihood using scaling method 2 and
(e) pseudo-maximum-likelihood using method D.

Estimation was performed by using `gllamm` with 12-point adaptive quadrature.

In Tables 2–5 we report means and standard deviations over the 100 replications for the estimates of the conditional regression parameters β_0 (the fixed intercept), β_1 (the regression coefficient for the cluster-specific covariate) and β_2 (the regression coefficient for the unit-specific covariate), and the random-intercept standard deviation $\sqrt{\psi}$. We also report the mean estimated marginal effects β_0^{M}, β_1^{M} and β_2^{M}, which were obtained by plugging the parameter estimates into approximation (3). We do not present the results for $N_j^{(1)} = 100$ as they are almost identical to those for $N_j^{(1)} = 50$.

When no weights are used, the deliberate undersampling of level 2 units with large absolute values $|\zeta_j^{(2)}|$ leads to a downward bias for the random-intercept standard deviation $\sqrt{\psi}$. The deliberate undersampling of level 1 units with positive values of ε_{ij} leads to a downward bias for the fixed intercept β_0. This downward bias for β_0 is also observed for the weighted estimates by using method D because this method uses the cluster averages of overall inclusion weights w_{ij} as level 2 weights whereas the level 1 sampling probabilities $\pi_{i|j}$ vary mostly within clusters.

Table 2. Cluster size $N_j^{(1)} = 5$: mean estimates and standard deviations

Parameter	*True value*	*Unweighted maximum likelihood estimate*	*Weighted pseudo-maximum-likelihood estimates*			
			Raw	*Method 1*	*Method 2*	*Method D*
Model parameters: conditional effects						
β_0	1	0.40	1.03	0.68	0.75	0.42
		(0.11)	(0.19)	(0.16)	(0.15)	(0.15)
β_1	1	1.08	1.19	0.96	0.98	1.05
		(0.18)	(0.32)	(0.26)	(0.26)	(0.25)
β_2	1	1.06	1.22	0.94	0.96	1.02
		(0.22)	(0.35)	(0.25)	(0.26)	(0.26)
$\sqrt{\psi}$	1	0.39	1.47	0.58	0.70	0.62
		(0.37)	(0.21)	(0.31)	(0.30)	(0.51)
Rescaled regression coefficients: approximate marginal effects						
β_0^{M}	0.88	0.39	0.80	0.64	0.70	0.38
β_1^{M}	0.88	1.04	0.92	0.91	0.90	0.96
β_2^{M}	0.88	1.02	0.94	0.89	0.89	0.93

Table 3. Cluster size $N_j^{(1)} = 10$: mean estimates and standard deviations

Parameter	*True value*	*Unweighted maximum likelihood estimate*	*Weighted pseudo-maximum-likelihood estimates*			
			Raw	*Method 1*	*Method 2*	*Method D*
Model parameters: conditional effects						
β_0	1	0.37	1.04	0.83	0.88	0.37
		(0.11)	(0.16)	(0.14)	(0.14)	(0.17)
β_1	1	1.13	1.06	0.91	0.94	1.13
		(0.14)	(0.23)	(0.20)	(0.20)	(0.22)
β_2	1	1.14	1.11	0.91	0.97	1.13
		(0.14)	(0.20)	(0.16)	(0.17)	(0.16)
$\sqrt{\psi}$	1	0.77	1.19	0.40	0.73	1.04
		(0.10)	(0.13)	(0.34)	(0.16)	(0.12)
Rescaled regression coefficients: approximate marginal effects						
β_0^{M}	0.88	0.34	0.87	0.79	0.82	0.32
β_1^{M}	0.88	1.04	0.89	0.87	0.87	0.98
β_2^{M}	0.88	1.05	0.93	0.88	0.90	0.98

Table 4. Cluster size $N_j^{(1)} = 20$: mean estimates and standard deviations

Parameter	*True value*	*Unweighted maximum likelihood estimate*	*Weighted pseudo-maximum-likelihood estimates*			
			Raw	*Method 1*	*Method 2*	*Method D*
Model parameters: conditional effects						
β_0	1	0.36	1.02	0.91	0.94	0.36
		(0.09)	(0.16)	(0.14)	(0.15)	(0.17)
β_1	1	1.16	1.05	0.94	0.97	1.16
		(0.14)	(0.22)	(0.20)	(0.21)	(0.23)
β_2	1	1.16	1.05	0.95	0.99	1.15
		(0.10)	(0.14)	(0.12)	(0.13)	(0.12)
$\sqrt{\psi}$	1	0.82	1.09	0.70	0.83	1.10
		(0.06)	(0.09)	(0.13)	(0.16)	(0.08)
Rescaled regression coefficients: approximate marginal effects						
β_0^{M}	0.88	0.32	0.87	0.84	0.85	0.31
β_1^{M}	0.88	1.06	0.89	0.87	0.88	0.99
β_2^{M}	0.88	1.05	0.90	0.89	0.89	0.98

As in linear random-intercept models, $\sqrt{\psi}$ is overestimated by using raw weights, less so as the cluster size increases, with very little bias for cluster sizes $N_j^{(1)}$ of 50 or more (corresponding to $n_j^{(1)} \gtrapprox 25$). One reason for the relatively good performance of the raw weights is the high sampling fractions at level 1, leading to moderate level 1 weights of about 4 and 1.3, whereas the sampling fractions in Pfeffermann *et al.* (1998) were smaller. In this paper the sampling fractions are the same regardless of cluster size $n_j^{(1)}$, making it possible to isolate the effect of

Table 5. Cluster size $N_j^{(1)} = 50$: mean estimates and standard deviations

Parameter	*True value*	*Unweighted maximum likelihood estimate*	*Weighted pseudo-maximum-likelihood estimates*			
			Raw	*Method 1*	*Method 2*	*Method D*
Model parameters: conditional effects						
β_0	1	0.35	1.01	0.96	0.98	0.35
		(0.08)	(0.13)	(0.12)	(0.12)	(0.14)
β_1	1	1.18	1.03	0.98	1.00	1.18
		(0.11)	(0.17)	(0.17)	(0.17)	(0.19)
β_2	1	1.18	1.02	0.98	0.99	1.17
		(0.06)	(0.08)	(0.07)	(0.07)	(0.07)
$\sqrt{\psi}$	1	0.87	1.05	0.87	0.94	1.14
		(0.04)	(0.08)	(0.08)	(0.07)	(0.08)
Rescaled regression coefficients: approximate marginal effects						
β_0^{M}	0.88	0.31	0.87	0.87	0.87	0.29
β_1^{M}	0.88	1.07	0.90	0.88	0.89	1.00
β_2^{M}	0.88	1.06	0.88	0.88	0.88	0.99

cluster size. This is in contrast with the results of Pfeffermann *et al.* (1998) where smaller cluster sizes were due to smaller sampling fractions, leading to confounding of these effects.

Scaling methods 1 and 2 both appear to overcorrect the positive bias for $\sqrt{\psi}$. This may be due to the within-cluster stratification based on the sign of ε_{ij} as discussed in Section 4.3.1. Scaling method 2 seems to perform better than method 1, giving results that are intermediate between those for raw weights and scaling method 1 as would be expected since the scaling constants tend to be closer to 1 than for method 1. The three methods employing both level 1 and level 2 weights (raw, method 1 and method 2) produce biased estimates for the regression coefficients whenever they are biased for the random-intercept standard deviation. Interestingly, these biases roughly cancel out in the expression (3) for the marginal effects.

Our simulation results appear to be consistent with the results in Grilli and Pratesi (2004) for informative sampling at both levels with small cluster sizes. For the level 2 variance, their unscaled fully weighted (our 'raw') estimators are upward biased whereas scaling method 2 overcorrects this bias. However, for the intercept and regression coefficients, the weighted estimators are less severely biased in Grilli and Pratesi (2004). As mentioned in Section 1, this could be due to the small true values for these parameters. In Grilli and Pratesi (2004), the sampling variability is considerably lower by using scaled weights than by using raw weights, whereas this difference is less pronounced in our simulations. The reason could be the lower sampling fractions that were used in their simulations.

To study the performance of the sandwich estimator, we simulated the model 1000 times for cluster size $N_j^{(1)} = 50$. We used raw level 1 weights, which produced only small biases for this cluster size. Table 6 shows the mean estimates and their standard deviations, as well as the mean standard errors and coverage for approximate 95% confidence intervals based on the normal distribution. The mean standard errors are almost identical to the standard deviations of the estimates. The coverage is close to the nominal level, even for the random-intercept standard deviation where the normality approximation may be dubious.

To investigate whether the relatively small bias for $\sqrt{\psi}$ using raw weights (and downward bias

Table 6. Coverage of 95% confidence intervals for cluster size $N_j^{(1)} = 50$ by using raw weights (1000 replications)

Parameter	*True value*	*Mean estimate*	*Standard deviation of estimate*	*Mean SE*	*95% confidence interval coverage*
β_0	1	1.01	0.13	0.13	94.1
β_1	1	1.02	0.18	0.18	94.7
β_2	1	1.03	0.08	0.08	94.1
$\sqrt{\psi}$	1	1.07	0.07	0.08	92.4

using scaled weights) is due to the within-cluster stratification as discussed in Section 4.3.1, we conducted further simulations for cluster size $N_j^{(1)} = 10$, using three stratification methods:

(a) stratification based on the sign of ε_{ij} as used in all simulations so far,
(b) the same design but with stratification determined by the sign of a standard normal random variable that was correlated 0.5 with ε_{ij} and
(c) stratification determined by the sign of a standard normal random variable that was uncorrelated with ε_{ij} and thus independent of the response.

Table 7 shows that the reasonable performance of the raw method that was seen earlier for case (a) deteriorates as the stratification becomes less related to the response. Scaling method 1 works very well for (c) where stratification is independent of the response. Qualitatively the same results (worse performance of the raw method and better performance of scaling method 1 as stratification becomes less related to the response) are obtained when the sampling fractions in both strata are 0.5. This suggests that the results are not due to varying the 'informativeness' of the weights (the correlation between the weights and the responses).

Table 7. Effect of stratification method: mean estimates and standard deviations for cluster size $N_j^{(1)} = 10$

Parameter	*True value*	*Results for raw weights and the following stratification methods:*			*Results for method 1 and the following stratification methods:*		
		(a)	*(b)*	*(c)*	*(a)*	*(b)*	*(c)*
Model parameters: conditional effects							
β_0	1	1.04	1.10	1.29	0.83	0.88	1.01
		(0.16)	(0.16)	(0.21)	(0.14)	(0.13)	(0.16)
β_1	1	1.06	1.11	1.26	0.91	0.92	0.99
		(0.23)	(0.26)	(0.30)	(0.20)	(0.23)	(0.25)
β_2	1	1.11	1.12	1.17	0.91	0.91	0.96
		(0.20)	(0.21)	(0.25)	(0.16)	(0.17)	(0.19)
$\sqrt{\psi}$	1	1.19	1.33	1.77	0.40	0.61	0.98
		(0.13)	(0.15)	(0.15)	(0.34)	(0.24)	(0.16)
Rescaled regression coefficients: approximate marginal effects							
β_0^{M}	0.88	0.87	0.88	0.92	0.79	0.83	0.89
β_1^{M}	0.88	0.89	0.89	0.90	0.87	0.86	0.86
β_2^{M}	0.88	0.93	0.90	0.83	0.88	0.86	0.84

8. Discussion

We have described a pseudolikelihood approach for generalized linear mixed modelling of data from complex sampling designs. Unlike previous contributions (e.g. Pfeffermann *et al.* (1998), Skinner and Holmes (2003) and Grilli and Pratesi (2004)), our approach can handle multilevel models with any number of levels, as well as allowing for stratification and PSUs that are not represented by a random effect in the model.

The pseudolikelihood methodology was applied to three-stage complex survey data on reading proficiency from the American PISA 2000 study, using the Stata program `gllamm` (e.g. Rabe-Hesketh *et al*. (2002, 2004a) and Rabe-Hesketh and Skrondal (2005)). The performance of pseudo-maximum-likelihood for two-level logistic regression using different methods for handling level 1 weights was investigated in a Monte Carlo experiment. This revealed that not only the estimated random-intercept variance but also the (conditional) regression coefficients were biased for small cluster sizes. Thus, considerable caution should be exercised in this case and sensitivity analyses should be conducted by comparing estimates from different scaling methods. It may also be useful to simulate a finite population from the estimated model, select a sample by mimicking the actual sampling design and investigate how well different methods recover the model parameters. The estimated marginal regression coefficients performed well even for small cluster sizes, suggesting that interpretation may best be confined to marginal effects in this case. We also conducted a small Monte Carlo experiment to investigate the performance of the sandwich estimator and found that the coverage was good.

The contribution of this paper is not confined to multilevel models but also applies to factor, item response, structural equation and latent class models. For these models we view the variables, indicators or items measuring the latent variables as level 1 units and the subjects as level 2 units, since the latent variables vary at the subject level (e.g. Skrondal and Rabe-Hesketh (2004)). Most previous pseudolikelihood approaches for latent variable models have been confined to weighting at level 2 (subjects) where the problem of appropriately scaling the weights does not arise. An exception is Skinner and Holmes (2003) who discussed structural equation models for longitudinal data by using weights both at level 2 (the subject) and level 1 (the occasion). Asparouhov (2005) considered pseudo-maximum-likelihood estimation for structural equation modelling with level 2 weights. Muthén and Satorra (1995), Stapleton (2002) and Skinner and Holmes (2003) considered a related approach where weighted mean and covariance matrices are used in the fitting functions of the weighted least squares estimator implemented in standard software for structural equation modelling. Wedel *et al*. (1998), Patterson *et al*. (2002) and Vermunt (2002) discussed pseudo-maximum-likelihood estimation of latent class models by using complex survey data with weighting at level 2 (subjects).

A major practical obstacle in using pseudo-maximum-likelihood estimation for multilevel modelling of complex survey data is that the necessary information is often not provided in publicly available data sets. For instance, many surveys include only a single overall weighting variable for the level 1 units, whereas the pseudolikelihood approach requires the weights corresponding to the levels of the hierarchical sampling design. Approaches to retrieving this information from the overall weights have been suggested by Kŏvacević and Rai (2003), pages 116–117, and Goldstein (2003), page 79, but little appears to be known regarding the performance of their approximations.

Non-response at any level can easily be addressed by adjusting the weights. However, post-stratification weights are typically constructed by considering sampling proportions for level 1 units by subpopulations such as males and females. These weights are not *conditional* weights dependent on the selected cluster j as required for the pseudolikelihood for multilevel models.

An exception may be cross-national surveys where the level 2 clusters are nations and post-stratification weights are constructed by nation. Another example where post-stratification weights can be used is panel surveys since the subject-specific weights are then at level 2.

Level 1 weights are not only used in standard multilevel models. In panel surveys, waves can be regarded as level 1 units and subjects as level 2 units. In this case π_j are the usual sample selection probabilities for the first panel wave whereas the level 1 weights are determined by non-response and attrition (Skinner and Holmes, 2003). Level 1 weighting to account for drop-out has also been used in generalized estimating equations (e.g. Robins *et al*. (1995)).

The pseudolikelihood methodology that is discussed here and implemented in `gllamm` accommodates the wide range of models subsumed in the generalized linear latent and mixed model framework of Rabe-Hesketh *et al*. (2004b) and Skrondal and Rabe-Hesketh (2004). In addition to conventional multilevel and latent variable models, this framework includes extensions such as multilevel structural equation models and models with nonparametric random effects or latent variable distributions (Rabe-Hesketh *et al*., 2003). Responses can be continuous, dichotomous, ordinal or nominal variables (Skrondal and Rabe-Hesketh, 2003), as well as counts and durations.

Acknowledgements

We are grateful to Mariann Lemke at the National Center for Education Statistics for providing us with an anonymized PSU identifier for the US PISA 2000 data and to Leonardo Grilli, Jouni Kuha and two reviewers for very helpful comments.

Appendix A: Stata commands for the application

The data without the PSU identifier can be downloaded from `http://www.blackwellpublishing.com/rss` and the `gllamm` program from `http://www.gllamm.org`.

Below are the Stata and `gllamm` commands for producing the estimates that are reported in Table 1.

```
insheet pisaUSA2000.txt, clear

*** Level 1 weights using scaling method 1
gen sqw = w_fstuwt^2
egen sumsqw = sum(sqw), by(id_school)
egen sumw = sum(w_fstuwt), by(id_school)
  gen pwt1s1 = w_fstuwt * sumw/sumsqw

*** Level 1 weights using scaling method 2
egen nj = count(female), by(id_school)
gen pwt1s2 = w_fstuwt * nj/sumw

egen mn_isei = mean(isei), by(id_school)

*** Maximum likelihood estimates (no weights)

gllamm pass_read female isei mn_isei high_school college test_lang ///
  one_for both_for, i(id_school) l(logit) f(binom) nip(12) adapt

* Robust standard errors
gllamm, robust

* Robust standard errors taking PSUs into account
gllamm, robust cluster(wvarstr)
```

```
*** Pseudo-maximum-likelihood estimates (scaling method 1)

matrix a = e(b)

gen pwt2 = wnrschbw
gen pwt1 = pwt1s1
gllamm pass_read female isei mn_isei high_school college test_lang ///
  one_for both_for, i(id_school) pweight(pwt) l(logit) f(binom) ///
  from(a) copy nip(12) adapt

* Robust standard errors taking PSUs into account
gllamm, cluster(wvarstr)
```

Here, the option `pweight(pwt)` means that the inverse probability weights for level l are in the variable `pwtl`, $l = 1, \ldots, L$. At least one of these variables must be defined. If there is no variable for a given level, the weights are assumed to equal 1. For pseudo-maximum-likelihood-estimation with scaling method 2, `pwt1` must be replaced by `pwt1s2`.

References

Adams, R. (2002) Scaling PISA cognitive data. In *PISA 2000 Technical Report* (eds R. Adams and M. Wu), pp. 99–108. Paris: Organisation for Economic Co-operation and Development.

Asparouhov, T. (2005) Sampling weights in latent variable modeling. *Struct. Equn Modlng*, **12**, 411–434.

Binder, D. A. (1983) On the variances of asymptotically normal estimators from complex surveys. *Int. Statist. Rev.*, **51**, 279–292.

Binder, D. A. and Roberts, G. R. (2003) Design-based and model-based methods for estimating model parameters. In *Analysis of Survey Data* (eds R. L. Chambers and C. J. Skinner), pp. 29–48. Chichester: Wiley.

Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *J. Am. Statist. Ass.*, **88**, 9–25.

Chambers, R. L. (2003) Introduction to Part A. In *Analysis of Survey Data* (eds R. L. Chambers and C. J. Skinner), pp. 13–27. Chichester: Wiley.

Chambers, R. L. and Skinner, C. J. (eds) (2003) *Analysis of Survey Data*. Chichester: Wiley.

Clogg, C. C. and Eliason, S. R. (1987) Some common problems in log-linear analysis. *Sociol. Meth. Res.*, **16**, 8–44.

Cochran, W. G. (1977) *Sampling Techniques*, 3rd edn. New York: Wiley.

Ganzeboom, H. G. B., De Graaf, P., Treiman, D. J. and de Leeuw, J. (1992) A standard international socio-economic index of occupational status. *Socl Sci. Res.*, **21**, 1–56.

Goldstein, H. (1991) Nonlinear multilevel models, with an application to discrete response data. *Biometrika*, **78**, 45–51.

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

Graubard, B. I. and Korn, E. L. (1996) Modeling the sampling design in the analysis of health surveys. *Statist. Meth. Med. Res.*, **5**, 263–281.

Grilli, L. and Pratesi, M. (2004) Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs. *Surv. Methodol.*, **30**, 93–103.

Isaki, C. T. and Fuller, W. A. (1982) Survey design under the regression super-population model. *J. Am. Statist. Ass.*, **77**, 89–96.

Kish, L. (1965) *Survey Sampling*. London: Wiley.

Korn, E. L. and Graubard, B. I. (2003) Estimating variance components by using survey data. *J. R. Statist. Soc.* B, **65**, 175–190.

Kŏvacević, M. S. and Rai, S. N. (2003) A pseudo maximum likelihood approach to multilevel modelling of survey data. *Communs Statist. Theory Meth.*, **32**, 103–121.

Lemke, M., Calsyn, C., Lippman, L., Jocelyn, L., Kastberg, D., Liu, Y., Roey, S., Williams, T., Kruger, T. and Bairu, G. (2001) *Outcomes of Learning: Results from the 2000 Program for International Student Assessment of 15-year-olds in Reading, Mathematics, and Science Literacy*. Washington DC: National Center for Education Statistics.

Little, R. J. A. (1982) Models for nonresponse in sample surveys. *J. Am. Statist. Ass.*, **77**, 237–250.

Longford, N. T. (1995a) Model-based methods for analysis of data from 1990 NAEP Trial State Assessment. *Research and Development Report NCES 95-696*. Washington DC: National Center for Education Statistics.

Longford, N. T. (1995b) *Models for Uncertainty in Educational Testing*. New York: Springer.

Longford, N. T. (1996) Model-based variance estimation in surveys with stratified clustered designs. *Aust. J. Statist.*, **38**, 333–352.

McCulloch, C. E. and Searle, S. R. (2001) *Generalized, Linear and Mixed Models*. New York: Wiley.

Muthén, B. O. and Satorra, A. (1995) Complex sample data in structural equation modeling. In *Sociological Methodology 1995* (ed. P. Marsden), pp. 267–316. Cambridge: Blackwell.
Organisation for Economic Co-operation and Development (2000) *Manual for the PISA 2000 Database*. Paris: Organisation for Economic Co-operation and Development (Available from `http://www.pisa.oecd.org/dataoecd/53/18/33688135.pdf`.)
Patterson, B. H., Dayton, C. M. and Graubard, B. I. (2002) Latent class analysis of complex sample survey data: application to dietary data (with discussion). *J. Am. Statist. Ass.*, **97**, 721–741.
Pawitan, Y. (2001) *In All Likelihood: Statistical Modelling and Inference using Likelihood*. Oxford: Oxford University Press.
Pfeffermann, D. (1993) The role of sampling weights when modeling survey data. *Int. Statist. Rev.*, **61**, 317–337.
Pfeffermann, D. (1996) The use of sampling weights for survey data analysis. *Statist. Meth. Med. Res.*, **5**, 239–261.
Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H. and Rasbash, J. (1998) Weighting for unequal selection probabilities in multilevel models. *J. R. Statist. Soc.* B, **60**, 23–40.
Pothoff, R. F., Woodbury, M. A. and Manton, K. G. (1992) 'Equivalent sample size' and 'equivalent degrees of freedom' refinements for inference using survey weights under superpopulation models. *J. Am. Statist. Ass.*, **87**, 383–396.
Rabe-Hesketh, S., Pickles, A. and Skrondal, A. (2003) Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statist. Modllng*, **3**, 215–232.
Rabe-Hesketh, S. and Skrondal, A. (2001) Parameterization of multivariate random effects models for categorical data. *Biometrics*, **57**, 1256–1264.
Rabe-Hesketh, S. and Skrondal, A. (2005) *Multilevel and Longitudinal Modeling using Stata*. College Station: Stata.
Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2002) Reliable estimation of generalized linear mixed models using adaptive quadrature. *Stata J.*, **2**, 1–21.
Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2004a) GLLAMM manual. *Technical Report 160*. Division of Biostatistics, University of California, Berkeley. (Available from `http://www.bepress.com/ucbbiostat/paper160/`.)
Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2004b) Generalized multilevel structural equation modeling. *Psychometrika*, **69**, 167–190.
Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2005) Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *J. Econometr.*, **128**, 301–323.
Rasbash, J., Browne, W. J. and Goldstein, H. (2003) *MLwiN 2.0 Command Manual, Version 2.0.01*. London: Institute of Education. (Available from `http://multilevel.ioe.ac.uk/download/comman20.pdf`.)
Raudenbush, S. W. and Bryk, A. S. (2002) *Hierarchical Linear Models*. Thousand Oaks: Sage.
Renard, D. and Molenberghs, G. (2002) Multilevel modeling of complex survey data. In *Topics in Modelling Clustered Data* (eds M. Aerts, H. Geys, G. Molenberghs and L. M. Ryan), pp. 263–272. Boca Raton: Chapman and Hall–CRC.
Ritz, J. and Spiegelman, D. (2004) A note about the equivalence of conditional and marginal regression models. *Statist. Meth. Med. Res.*, **13**, 309–323.
Robins, J. M., Rotnitzky, A. G. and Zhao, L. P. (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Statist. Ass.*, **90**, 106–121.
Rodríguez, G. and Goldman, N. (1995) An assessment of estimation procedures for multilevel models with binary responses. *J. R. Statist. Soc.* A, **158**, 73–89.
Rodríguez, G. and Goldman, N. (2001) Improved estimation procedures for multilevel models with binary response: a case-study. *J. R. Statist. Soc.* A, **164**, 339–355.
Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.
Sen, P. K. (1988) Asymptotics in finite populations. In *Handbook of Statistics*, vol. 6, *Sampling* (eds P. R. Krishnaiah and C. R. Rao), pp. 291–331. Amsterdam: North-Holland.
Skinner, C. J. (1989) Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys* (eds C. J. Skinner, D. Holt and T. M. F. Smith). Chichester: Wiley.
Skinner, C. J. (2005) On weight scaling for estimation in multilevel models using survey weights. Unpublished. Department of Social Statistics, University of Southampton, Southampton.
Skinner, C. J. and Holmes, D. J. (2003) Random effects models for longitudinal data. In *Analysis of Survey Data* (eds R. L. Chambers and C. J. Skinner). Chichester: Wiley.
Skrondal, A. and Rabe-Hesketh, S. (2003) Multilevel logistic regression for polytomous data and rankings. *Psychometrika*, **68**, 267–287.
Skrondal, A. and Rabe-Hesketh, S. (2004) *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton: Chapman and Hall–CRC.
Stapleton, L. (2002) The incorporation of sample weights into multilevel structural equation models. *Struct. Equn Modlng*, **9**, 475–502.
Vermunt, J. K. (2002) Discussion on 'Latent class analysis of complex sample survey data: application to dietary data'. *J. Am. Statist. Ass.*, **97**, 736–737.
Warm, T. A. (1989) Weighted likelihood estimation of ability in item response models. *Psychometrika*, **54**, 427–450.

Wedel, M., ter Hofstede, F. and Steenkamp, J.-B. E. M. (1998) Mixture model analysis of complex samples. *J. Classificn*, **15**, 225–244.
Willms, J. D. (1986) Social class segregation and its relationship to pupils' examination results in Scotland. *Am. Sociol. Rev.*, **51**, 224–241.
Wolfinger, R. D. (1999) Fitting non-linear mixed models with the new NLMIXED procedure. *Technical Report*. SAS Institute, Cary.
Zeger, S. L., Liang, K.-Y. and Albert, P. S. (1988) Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44**, 1049–1060.