A comparison of Bayesian and likelihood-based methods for fitting multilevel models

William J. Browne
University of Nottingham, UK

and David Draper†
University of California, Santa Cruz, USA

Summary. We use simulation studies, whose design is realistic for educational and medical research (as well as other fields of inquiry), to compare Bayesian and likelihood-based methods for fitting variance-components (VC) and random-effects logistic regression (RELR) models. The likelihood (and approximate likelihood) approaches we examine are based on the methods most widely used in current applied multilevel (hierarchical) analyses: maximum likelihood (ML) and restricted ML (REML) for Gaussian outcomes, and marginal and penalized quasi-likelihood (MQL and PQL) for Bernoulli outcomes. Our Bayesian methods use Markov chain Monte Carlo (MCMC) estimation, with adaptive hybrid Metropolis-Gibbs sampling for RELR models, and several diffuse prior distributions ($\Gamma^{-1}(\epsilon, \epsilon)$ and $U(0, \frac{1}{\epsilon})$ priors for variance components). For evaluation criteria we consider bias of point estimates and nominal versus actual coverage of interval estimates in repeated sampling. In two-level VC models we find that (a) both likelihood-based and Bayesian approaches can be made to produce approximately unbiased estimates, although the automatic manner in which REML achieves this is an advantage, but (b) both approaches had difficulty achieving nominal coverage in small samples and with small values of the intraclass correlation. With the three-level RELR models we examine we find that (c) quasi-likelihood methods for estimating random-effects variances perform badly with respect to bias and coverage in the example we simulated, and (d) Bayesian diffuse-prior methods lead to well-calibrated point and interval RELR estimates. Given that the likelihood-based methods we study are considerably faster computationally than MCMC and that a number of models are typically fit during the model exploration phase of a multilevel study, one possible analytic strategy suggested by our results is a hybrid of likelihood-based and Bayesian methods, with (i) REML and quasi-likelihood estimation (for their computational speed) during model exploration (but with awareness of the possible understatement of random-effects variances and their uncertainty bands) and (ii) diffuse-prior Bayesian estimation using MCMC to produce final inferential results. Other analytic strategies based on less approximate likelihood methods are also possible but would benefit from further study of the type summarized here.

Keywords: Adaptive MCMC, bias, calibration, diffuse priors, hierarchical modeling, hybrid Metropolis-Gibbs sampling, intraclass correlation, IGLS, interval coverage, MQL, mixed models, PQL, RIGLS, random-effects logistic regression, REML, variance-components models

1. Introduction

Multilevel models, for data possessing a nested hierarchy and—more generally—for the expression of uncertainty at several levels of aggregation, have gained dramatically in scope of application in the past 15 years, in fields as diverse as education and health policy (e.g., Goldstein et al. 1993, Draper 1995a, Goldstein and Spiegelhalter 1996). Statisticians and substantive researchers who use such models now have a variety of options in approaches to inference, with a corresponding variety of computer programs: to mention four, the maximum-likelihood (ML) Fisher-scoring approach in `VARCL` (Longford 1987); ML via iterative generalized least squares (IGLS) and restricted IGLS (RIGLS, or REML) for Gaussian outcomes, and quasi-likelihood methods (MQL and PQL) for dichotomous

†*Address for correspondence*: Department of Applied Mathematics and Statistics, Baskin School of Engineering, University of California, 1156 High Street, Santa Cruz CA 95064 USA (email `draper@ams.ucsc.edu`, web `www.ams.ucsc.edu/~draper`).

outcomes, in MLwiN (Goldstein 1986, 1989; Rasbash et al. 2000); empirical-Bayes estimation using the EM algorithm in HLM (Bryk et al. 1988); and fully-Bayesian inference in BUGS (Spiegelhalter et al. 1997) and MLwiN. This variety of fitting methods can lead to confusion, however: ML and Bayesian analyses of the same data can produce rather different point and interval estimates, and the applied multilevel modeler may well be left wondering what to report.

1.1. Example 1: The Junior School Project

The Junior School Project (JSP; Mortimore et al. 1988; Woodhouse et al. 1995) was a longitudinal study of about 2,000 pupils from 50 primary schools chosen randomly from the 636 Inner London Education Authority (ILEA) schools in 1980. Here we will examine a random subsample of $N = 887$ students taken from $J = 48$ schools. A variety of measurements were made on the students during the four years of the study, including background variables (such as gender, age at entry, ethnicity, and social class) and measures of educational outcomes such as mathematics test scores (on a scale from 0 to 40) at year 3 (math3) and year 5 (math5). Both mathematics scores had distributions with negative skew due to a ceiling effect, with some students piling up at the maximum score, but transformations to normality produced results almost identical to those using the raw data (we report the latter). A principal goal of the study was to establish whether some schools were more effective than others in promoting pupils' learning and development, after adjusting for background differences.

Two simple baseline analyses that might be undertaken early on, before more complicated modeling, are as follows.

- Thinking (incorrectly) of the data as a simple random sample (SRS) from the population of ILEA pupils in the early 1980s, the mean mathematics score at year 5 would be estimated as 30.6 with a repeated-sampling standard error (SE) of 0.22, but this ignores the large estimated intracluster (within-school) correlation of $\hat{\rho} = +0.12$ for this variable. The correct SE, from standard survey-sampling results (e.g., Cochran 1977) or the Huber-White sandwich estimator (Huber 1967, White 1980, as implemented in the package Stata; StataCorp 2004), is 0.43, almost double the SRS value. There is clearly scope for multilevel modeling here to account correctly for the nested structure of the data.

- Consider next a variance-components (VC) model,

$$\begin{array}{l} y_{ij} = \beta_0 + u_j + e_{ij}, \quad i = 1, \ldots, n_j, \quad j = 1, \ldots, J, \\ \sum_{j=1}^{J} n_j = N, \quad u_j \overset{\text{IID}}{\sim} N(0, \sigma_u^2), \quad e_{ij} \overset{\text{IID}}{\sim} N(0, \sigma_e^2), \end{array} \tag{1}$$

  where y_{ij} is the math5 score for pupil i in school j; this model would generally be fit before a random-slopes regression model relating math5 to math3 is examined. (In our terminology i indexes level 1 of the model and j level 2. (1) is sometimes referred to as a *mixed linear model* for its combination of *fixed effects* (β_0) and *random effects* (the u_j and e_{ij}).) As noted above, the parameters in this model may be estimated in at least two ways: likelihood-based and Bayesian approaches. Maximum likelihood (ML) in turn may be based on iterative generalized least squares (IGLS, or some other equivalent method), or approximately unbiased estimation with restricted maximum likelihood (REML, based for instance on RIGLS) (Goldstein 1986, 1989) may be preferred. Table 1 presents the results of ML, REML, and Bayesian fitting of model (1), in the latter case using a diffuse prior to be discussed in Section 2.3.1 ($U\left(0, \frac{1}{\epsilon}\right)$ on the variance scale).

  While there is little difference in the three methods on point estimates for β_0 and σ_e^2 and on SEs/posterior standard deviations (SDs) for the latter quantity, (a) the posterior SD for β_0 is about 5% larger than the SE from ML and REML (note that the Bayesian uncertainty assessment essentially coincides with the cluster-sampling SE 0.43 mentioned earlier), (b) the Bayesian estimate of σ_u^2 is 14–17% larger than the likelihood estimates, and (c) the posterior SD for σ_u^2 is 18–21% larger than the ML/REML SEs. Moreover, the default likelihood results (point

estimates and estimated asymptotic SEs) in the ML computer programs in most widespread current use do not include interval estimates, encouraging investigators either to report no intervals at all (a practice to be frowned upon) or to use large-sample 95% Gaussian intervals of the form (estimate $\pm 1.96\,\widehat{SE}$). The bottom part of Table 1 compares Gaussian intervals based on REML estimates with Bayesian 95% posterior probability intervals, and it may be seen that in particular the two methods give quite different answers for σ_u^2. What should someone trying to arrive at substantive conclusions based on the JSP data report?

Table 1. *A comparison of ML, REML, and Bayesian fitting (with a diffuse prior) in model (1) applied to the JSP data. Figures in parentheses in the upper table are SEs (for the ML methods) or posterior SDs (for the Bayesian method). Bayesian point estimates are posterior means, and 95% central posterior intervals are reported.*

Point Estimates	Parameter		
Method	β_0	σ_u^2	σ_e^2
ML	30.6	5.16	39.3
	(0.400)	(1.55)	(1.92)
REML	30.6	5.32	39.3
	(0.404)	(1.59)	(1.92)
Bayesian with	30.6	6.09	39.5
diffuse priors	(0.427)	(1.91)	(1.94)

95% Interval Estimates	Parameter		
Method	β_0	σ_u^2	σ_e^2
REML (Gaussian)	$(29.8, 31.4)$	$(2.22, 8.43)$	$(35.5, 43.0)$
Bayesian	$(29.8, 31.5)$	$(3.18, 10.6)$	$(35.9, 43.5)$

1.2. Example 2: The Guatemalan Child Health Study

The 1987 Guatemalan National Survey of Maternal and Child Health (Pebley and Goldman 1992) was based on a multistage cluster sample of 5,160 women aged 15–44 years living in 240 communities, with the goal of increased understanding of the determinants of health for mothers and children in the period during and after pregnancy. The data have a three-level structure—births within mothers within communities—and one analysis of particular interest estimated the probability of receiving modern (physician or trained nurse) prenatal care as a function of covariates at all three levels. Rodríguez and Goldman (1995) studied a subsample of 2,449 births by 1,558 women who (a) lived in the 161 communities with accurate cluster-level information and (b) had some form of prenatal care during pregnancy. The random-effects logistic regression (RELR) model they examined is

$$\begin{gathered}(y_{ijk}\,|\,p_{ijk}) \overset{\text{indep}}{\sim} \text{Bernoulli}(p_{ijk}) \qquad \text{with} \\ \text{logit}(p_{ijk}) = \beta_0 + \beta_1 x_{1ijk} + \beta_2 x_{2jk} + \beta_3 x_{3k} + u_{jk} + v_k,\end{gathered} \tag{2}$$

where y_{ijk} is a binary indicator of modern prenatal care or not and where $u_{jk} \sim N(0, \sigma_u^2)$ and $v_k \sim N(0, \sigma_v^2)$. In this formulation $i = 1, \ldots, I_{jk},\ j = 1, \ldots, J_k$, and $k = 1, \ldots K$ index the level 1, 2, and 3 units, respectively, corresponding to births, mothers, and communities, and the variables x_1, x_2, and x_3 are composite scales, because the original Pebley-Goldman model contained many covariates at each level. The original Rodríguez-Goldman data set is not publicly available; however, these

authors simulated 25 data sets with the same structure but with known parameter values, and they have kindly made these simulated data sets available to us.

As in Example 1, several likelihood-based and Bayesian fitting methods for model (2) are available: the main (approximate) likelihood alternatives (e.g., Goldstein 1995) currently employed with greatest frequency by multilevel modelers in substantive fields of inquiry (based upon empirical usage in the recent literature) are marginal quasi-likelihood (MQL) and penalized (or predictive) quasi-likelihood (PQL), in both of which the investigator has to specify the order of the Taylor-series approximation, and a variety of prior distributions may be considered in the Bayesian approach. Table 2 summarizes a comparison between first-order MQL, second-order PQL, and Bayesian fitting—again with a particular diffuse prior to be discussed in Section 2.3.1 ($U\left(0, \frac{1}{\epsilon}\right)$ on the variance scale with $\epsilon \to 0$)—on the Rodríguez-Goldman simulated data set number 1 (the true values of the parameters are given in the first row of this table). Here the differences are much more striking than those in Table 1: many MQL estimates are badly biased, and—although PQL does achieve some improvements—its estimates of β_2 and the variance components are still substantially too low, leading to dramatically different intervals for the variances. Because we have the luxury of knowing the right answer in this simulation context, it is easy to see which fitting method has produced better results on this one data set (and Section 4.2 will demonstrate that this table accurately reflects the superiority of Bayesian methods in models like (2) when compared with quasi-likelihood approaches, at least in settings similar to the Guatemalan Health study), but—if these data arose as the result of an actual sample survey—a researcher trying to draw substantive conclusions about variability within and between mothers and communities would certainly wonder which figures to publish.

Table 2. *A comparison of first-order MQL, second-order PQL and Bayesian fitting (with a diffuse prior) in model (2) applied to the Rodríguez-Goldman simulated Guatemalan child health data set number 1. Figures in square brackets in the upper table are true parameter values; figures in parentheses in the upper table are SEs (for the ML methods) or posterior SDs (for the Bayesian method). Bayesian point estimates are posterior means, and 95% central posterior intervals are reported.*

Point Estimates	Parameter					
Method	β_0 [0.65]	β_1 [1.0]	β_2 [1.0]	β_3 [1.0]	σ_v^2 [1.0]	σ_u^2 [1.0]
MQL_1	0.491 (0.149)	0.791 (0.172)	0.631 (0.081)	0.806 (0.189)	0.546 (0.102)	0.000 —
PQL_2	0.641 (0.186)	0.993 (0.201)	0.795 (0.099)	1.06 (0.237)	0.883 (0.159)	0.486 (0.145)
Bayesian with diffuse priors	0.675 (0.209)	1.050 (0.225)	0.843 (0.115)	1.124 (0.268)	1.043 (0.217)	0.921 (0.331)

95% Interval Estimates	Parameter					
Method	β_0	β_1	β_2	β_3	σ_v^2	σ_u^2
PQL_2 (Gaussian)	(0.276, 1.01)	(0.599, 1.39)	(0.601, 0.989)	(0.593, 1.52)	(0.571, 1.19)	(0.202, 0.770)
Bayesian	(0.251, 1.07)	(0.611, 1.50)	(0.626, 1.078)	(0.586, 1.62)	(0.677, 1.52)	(0.334, 1.63)

1.3. Outline of the paper

Our interest is in comparing likelihood-based and Bayesian methods for fitting variance-components and random-effects logistic regression models, using bias and interval coverage behavior in repeated

sampling as evaluation criteria. Following a brief literature review below, Section 2 describes the fitting methods we compare; Sections 3 and 4 cover simulation study details and results for VC and RELR models, respectively; and Section 5 offers some conclusions and discussion. Browne and Draper (2000) and Browne et al. (2002) contain results that parallel those presented here for random-slopes regression models and multilevel models with heteroscedasticity at level 1, respectively.

We focus in this paper on the likelihood-based (and approximate likelihood) methods most readily available (given current usage patterns of existing software) to statisticians and substantive researchers making frequent use of multilevel models: ML and REML in VC models, and MQL and PQL in RELR models. Other promising likelihood-based approaches—including (a) methods based on Gaussian quadrature (e.g., Pinheiro and Bates 1995; see Section 5 for a software discussion); (b) the nonparametric maximum likelihood methods of Aitkin (1999a); (c) the Laplace-approximation approach of Raudenbush et al. (2000); (d) the work on hierarchical generalized linear models of Lee and Nelder (2000); and (e) interval estimation based on ranges of values of the parameters for which the log likelihood is within a certain distance of its maximum, for instance using profile likelihood (e.g., Longford 2000)—are not addressed here. It is evident from the recent applied literature that, from the point of view of multilevel analyses currently being conducted to inform educational and health policy choices and other substantive decisions, the use of methods (a–e) is not (yet) as widespread as REML and quasi-likelihood approaches. In particular, methods such as Gaussian quadrature may produce poor results in RELR models if not used carefully (see Lesaffre and Spiessens 2001 for a striking example); we intend to report elsewhere on a thorough comparison of quadrature with the methods examined here.

Statisticians are well aware that the highly skewed repeated-sampling distributions of ML estimators of random-effects variances in multilevel models with small sample sizes are not likely to lead to good coverage properties for large-sample Gaussian approximate interval estimates of the form $\hat{\sigma}^2 \pm 1.96\,\widehat{SE}\left(\hat{\sigma}^2\right)$, but with few exceptions the profession has not (yet) responded to this by making software for improved likelihood interval estimates widely available to multilevel modelers. In Sections 3 and 4 we document the extent of the poor coverage behavior of the Gaussian approach, and we offer several simple approximation methods, based only on information routinely output in multilevel software, which exhibit improved (although still not in some cases satisfactory) performance. Note that we are not advocating interval estimates for random-effects variances based on normal approximations in small samples; we are merely documenting how badly these intervals—which are all that will be readily available to many users of popular likelihood-based software packages—may behave, even with a variety of improvements to them.

1.4. Previous literature on comparisons between multilevel fitting methods

The literature on Bayesian and likelihood-based methods for fitting VC and RELR models is vast, e.g., Aitkin (1996, 1999b); Besag et al. (1995); Bryk and Raudenbush (1992); Corbeil and Searle (1976); Daniels and Gatsonis (1999); Draper (2004); Gelfand and Smith (1990); Goldstein (1995); Harville and Zimmerman (1996); Kahn and Raftery (1996); Kass and Steffey (1989); Lee and Nelder (1996); Longford (1987, 1997); and Searle, Casella, and McCulloch (1992) (for a competing approach based on best linear unbiased prediction, see, e.g., Henderson 1950 and Robinson 1991). Comparisons between multilevel fitting methods are less abundant, but some theoretical work has been done to demonstrate the equivalence of several of the leading approaches to fitting multilevel models: for instance, Raudenbush (1994) showed that Fisher scoring is equivalent to ML, and empirical-Bayes estimates based on the EM algorithm may be seen to coincide with maximum likelihood results in many Gaussian models (e.g., Goldstein 1995). Less work is available comparing the performance of the approaches in terms of bias of point estimates and calibration of interval estimates.

In the VC model (1), Box and Tiao (1973) reviewed results of Klotz et al. (1969) and Portnoy (1971) which contrast the mean squared error (MSE) behavior of the following estimators of σ_u^2: the classical unbiased estimator based on mean squares (e.g., Scheffé 1959), the ML estimator, and the mean and mode of the marginal posterior distribution for σ_u^2 with several choices of relatively diffuse priors. They found, over all values of the intraclass (intracluster) correlation $\rho = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2}$ they examined, that (a) the MSEs of the ML and posterior-mode estimators are comparable and much smaller than that of the unbiased estimator, and (b) the posterior mean is, by a substantial margin, the worst estimator on MSE grounds. Box and Tiao criticized MSE as an arbitrary criterion for performance assessment, and resisted the distillation of an entire posterior distribution down to a single point estimate. We are sympathetic with their position—from the Bayesian viewpoint the choice of posterior summaries should ideally be based on decision criteria arising from possible actions when using models like (1) and (2) to solve real-world problems—but we nevertheless find it relevant, particularly in the context of general-purpose multilevel modeling software (where the eventual use of the output is far from clear), to examine operating characteristics such as bias and interval coverage. See Rubin (1984) for a good discussion of the relevance of frequency properties in Bayesian inference, and Carlin and Louis (2001, Chapter 4) for an evaluation in the spirit of the one presented here for some simpler non-hierarchical Gaussian and binary-outcome models.

Hulting and Harville (1991) compared frequentist and Bayesian methods of fitting the mixed-effects linear model

$$y = X\beta + Zs + e, \tag{3}$$

where y is an $n \times 1$ vector of quantitative outcomes, β is a $p \times 1$ vector of fixed effects, X and Z are known matrices, $s_i \sim N(0, \sigma_s^2)$, and $e_i \sim N(0, \sigma_e^2)$; the VC model (1) is a special case of (3). These authors obtained results broadly in agreement with those in Section 3.2 below, although they focused on predictive inferences about quantities of the form $W = \lambda'\beta + \delta' s$ and examined different frequentist estimators than the ones we consider. Chaloner (1987) carried out a similar frequentist/Bayesian comparison in model (1); however she used different diffuse prior distributions, focused on the variance ratio $\tau = \frac{\sigma_u^2}{\sigma_e^2} = \frac{\rho}{1-\rho}$ in her results on interval estimation, and conducted a less extensive simulation study than that reported here. See Swallow and Monahan (1984), Brown and Burgess (1984) and Huber et al. (1994) for additional simulation results comparing various non-Bayesian estimation methods in VC models, and Singh et al. (1998) for Bayesian and non-Bayesian comparisons in small-area estimation.

In model (2), Rodríguez and Goldman (1995) used the structure of the Guatemalan child health data to examine how well quasi-likelihood methods compare with fitting a standard logistic regression model and ignoring the multilevel structure. As noted in Section 1.2, their approach involved creating simulated data sets based on the original structure but with known true values for the fixed effects (the β_l in model (2)) and variance parameters. They considered the MQL method and showed that estimates of the fixed effects produced by MQL were even worse, in terms of bias, than estimates produced by standard logistic regression disregarding the hierarchical nature of the data. Goldstein and Rasbash (1996) considered the same problem but used the PQL method, and showed that the results produced by second-order PQL estimation were far better than for MQL, but still biased, in the Rodríguez-Goldman problem. Breslow and Clayton (1993) presented some brief comparisons between quasi-likelihood methods and a version of rejection Gibbs sampling in RELR models proposed by Zeger and Karim (1991); also see Natarajan and Kass (2000) for simulation results in a RELR model fit by the Zeger-Karim approach. Rodríguez and Goldman (2001) obtained results that parallel ours (in Section 4.2) with respect to bias of PQL random-effects variance estimates in REML models (and showed that a parametric bootstrap approach yields considerable improvement), but they have no

corresponding findings on interval estimates.

2. Methods for fitting multilevel models

2.1. Iterative generalized least squares (IGLS/ML) and restricted ML (RIGLS/REML)

Iterative generalized least squares (IGLS/ML; Goldstein 1986) is a sequential refinement procedure based on GLS estimation. The method can fit all Gaussian multilevel models, and has been described in detail elsewhere (e.g., Goldstein 1995). Briefly, equations such as (1) are expressed in the usual general linear model form $Y = X\beta + e^*$ (for example, in (1) X is a vector of 1s, $\beta = \beta_0$, and $e^*_{ij} = u_j + e_{ij}$), in which the vector e^* has mean 0 and covariance matrix V; and then the observation is made that (i) if V were known, β could be estimated by GLS, yielding $\hat{\beta}$, and (ii) if β were known, one could form the residuals $\tilde{Y} = Y - X\beta$, calculate $Y^* = \tilde{Y}\tilde{Y}^T$, stack the columns of Y^* into one long column vector Y^{**}, and define a linear model $Y^{**} = Z^*\theta + \epsilon$, where Z^* is the design matrix for the random-effects parameters θ (in (1) $\theta = \left(\sigma_u^2, \sigma_e^2\right)^T$). Another application of GLS then yields $\hat{\theta}$. Starting with an initial estimate of the fixed effect(s) β from ordinary least squares, IGLS iterates between steps (i) and (ii) to convergence, which is judged to occur when two successive sets of estimates differ by no more than a given tolerance (on a component-by-component basis). As with many ML procedures, IGLS produces biased estimates in small samples, often in particular underestimating random-effects variances because the sampling variation of $\hat{\beta}$ is not accounted for in the algorithm above. Defining the residuals instead as $\tilde{Y}^* = Y - X\hat{\beta}$ and $\hat{Y}^* = \tilde{Y}^* \left(\tilde{Y}^*\right)^T$, Goldstein (1989) showed that

$$E(\hat{Y}^*) = V - X\left(X^T V^{-1} X\right)^{-1} X^T, \tag{4}$$

so that the ML estimates can be bias-adjusted by adding the second term on the right-hand side of (4) to $\hat{Y}^*$ at each iteration. This is restricted IGLS (RIGLS), which coincides with restricted maximum likelihood (REML) in Gaussian models such as (1). Estimated asymptotic standard errors of ML and REML estimates are based on the final values at convergence of the covariance matrices for $\hat{\beta}$ and $\hat{\theta}$, expressions for which are given by Goldstein (1995).

2.2. Marginal and penalized quasi-likelihood (MQL and PQL)

ML and REML are relevant to linear multilevel models with Gaussian outcomes; different likelihood-based methods are needed with models for dichotomous outcomes, such as (2). Following Goldstein (1995), in the simpler case of a two-level structure a reasonably general multilevel model for the binary outcome y_{ij} has the form

$$\begin{gathered}(y_{ij} \,|\, p_{ij}) \sim \text{Bernoulli}(p_{ij}) \quad \text{with} \\ p_{ij} = f\Big(X_{ij}\beta + Z_{ij}^{(1)} e_{ij} + Z_{ij}^{(2)} u_j\Big),\end{gathered} \tag{5}$$

where $f(l)$ has a nonlinear character such as $\text{logit}^{-1}(l) = \left(1 + e^{-l}\right)^{-1}$. One approach to the fitting of (5) is through quasi-likelihood methods, which proceed (e.g., Breslow and Clayton 1993) by linearizing the model via Taylor series expansion; for instance, with H_t as a suitably chosen value around which to expand, the $f(\cdot)$ expression in (5) for the ijth unit at iteration $(t+1)$ may be approximated by

$$\begin{gathered}f(H_t) + X_{ij}\left(\beta_{t+1} - \beta_t\right) f'(H_t) + \\ \left(Z_{ij}^{(1)} e_{ij} + Z_{ij}^{(2)} u_j\right) f'(H_t) + \tfrac{1}{2}\left(Z_{ij}^{(1)} e_{ij} + Z_{ij}^{(2)} u_j\right)^2 f''(H_t)\end{gathered} \tag{6}$$

in terms of parameter values estimated at iteration t. The simplest choice, $H_t = X_{ij}\beta_t$, the fixed-part predicted value of the argument of f in (5), yields the marginal quasi-likelihood (MQL) algorithm.

This can be improved upon by expanding around the entire current predicted value for the ijth unit, $H_t = X_{ij}\beta_t + Z_{ij}^{(1)}\hat{e}_{ij} + Z_{ij}^{(2)}\hat{u}_j$, where $\hat{e}_{ij}$ and $\hat{u}_j$ are the current estimated random effects; when this is combined with an improved approximation obtained by replacing the second line in (6) with

$$\begin{aligned} &\left[Z_{ij}^{(1)}\left(e_{ij}-\hat{e}_{ij}\right)+Z_{ij}^{(2)}\left(u_j-\hat{u}_j\right)\right] f'(H_t)+ \\ &\tfrac{1}{2}\left[Z_{ij}^{(1)}\left(e_{ij}-\hat{e}_{ij}\right)+Z_{ij}^{(2)}\left(u_j-\hat{u}_j\right)\right]^2 f''(H_t)\,, \end{aligned} \tag{7}$$

the result is the penalized or predictive quasi-likelihood (PQL) algorithm. The order of an MQL or PQL algorithm refers to how many terms are used in the Taylor expansion underlying the linearization; for example, (6) is based on expansion up to second order and leads to MQL_2 and PQL_2 estimates. Estimated asymptotic standard errors for MQL/PQL estimates typically derive from a version of observed Fisher information based on the quasi-likelihood function underlying the estimation process; see Breslow and Clayton (1993) for details.

2.3. Markov chain Monte Carlo

The Bayesian fitting of both VC and RELR models involves, as usual in the Bayesian approach, the updating from prior to posterior distributions for the parameters via appropriate likelihood functions; but in both of these model classes closed-form exact expressions for most or all of the relevant joint and marginal posterior distributions are not available (see Chapter 5 of Box and Tiao 1973 for some limited analytical results in the VC model (1)). Instead we rely here on sampling-based approximations to the distributions of interest via Markov chain Monte Carlo (MCMC) methods (e.g., Gilks et al. 1996): we use a Gibbs sampling approach in the VC model (cf. Seltzer 1993) and an adaptive hybrid Metropolis-Gibbs method for random-effects logistic regression.

2.3.1. Diffuse priors for multilevel models

As with the Bayesian analysis of all statistical models, broadly speaking two classes of prior distributions are available for multilevel models: (a) diffuse and (b) informative, corresponding to situations in which (a) little is known about the quantities of interest *a priori* or (b) substantial prior information is available, for instance from previous studies judged relevant to the current data set. In situation (a), on which we focus in this paper, it seems natural to seek prior specifications that lead to well-calibrated inferences (e.g., Dawid 1985), which we take here to mean point estimates with little bias and interval estimates whose actual coverage is close to the nominal level (in both cases in repeated sampling).

There is an extensive literature on the specification of diffuse priors (e.g., Bernardo and Smith 1994; Kass and Wasserman 1996; Spiegelhalter et al. 1997; Gelman, Carlin, et al. 2003), leading in some models to more than one intuitively reasonable approach. It is sometimes asserted in this literature that the performance of the resulting Bayesian estimates is broadly insensitive, with moderate to large sample sizes, to how the diffuse prior is specified. In preliminary studies we found this to be the case for fixed effects in both the VC and RELR model classes, and in what follows we use (improper) priors that are uniform on the real line $\mathbb{R}$ for such parameters (these are functionally equivalent to proper Gaussian priors with huge variances). As others (e.g., DuMouchel 1990) have elsewhere noted, however, we found large differences in performance across plausible attempts to construct diffuse priors for random-effects variances in both model classes. Intuitively this is because the effective sample size for the level–2 variance in a two-level analysis with J level–2 units and N total level–1 units (typically $J << N$) is often much closer to J than to N; in other words, in the language of Example 1, even with

data on hundreds of pupils the likelihood information about the between-school variance can be fairly weak when the number of schools is modest, so that prior specification can make a real difference in such cases.

The off-the-shelf (improper) choice for a diffuse prior on a variance in many Bayesian analyses is $p(\sigma^2) \propto \frac{1}{\sigma^2}$, which is equivalent to assuming that $\log\left(\sigma^2\right)$ is uniform on $\mathbb{R}$. This is typically justified by noting that the posterior for σ^2 will be proper even for very small sample sizes; but (e.g., DuMouchel and Waternaux 1992) this choice can lead to improper posteriors in random-effects models. We avoid this problem by using two alternative diffuse (but proper) priors, both of which produce proper posteriors:

- A locally uniform prior for σ^2 on $(0, \frac{1}{\epsilon})$ for small positive ϵ (Gelman and Rubin 1992, Carlin 1992), which is equivalent to a Pareto$(1, \epsilon)$ prior for the precision $\lambda = \frac{1}{\sigma^2}$ (Spiegelhalter et al. 1997); and
- A $\Gamma^{-1}(\epsilon, \epsilon)$ prior for σ^2 (Spiegelhalter et al. 1997), for small positive ϵ.

Both of these priors are members of the scaled inverse chi-squared $\chi^{-2}(\nu, s^2)$ family (e.g., Gelman, Carlin, et al. 2003); this is equivalent to an inverse gamma $\Gamma^{-1}\left(\frac{\nu}{2}, \frac{\nu}{2}s^2\right)$ distribution, where ν is the prior effective sample size and s^2 is a prior estimate of σ^2. The $U(0, \frac{1}{\epsilon})$ and $\Gamma^{-1}(\epsilon, \epsilon)$ priors above are formally specified by the choices $(\nu, s^2) = (-2, 0)$ and $(2\epsilon, 1)$, respectively (in the former case in the limit as $\epsilon \to 0$). We have found that results are generally insensitive to the specific choice of ϵ in the region of 0.001 (the default setting in Spiegelhalter et al. 1997); we report findings with this value. (We also studied the effects of a gently data-determined prior for σ^2—$\chi^{-2}(\epsilon, \hat{\sigma}^2)$ for small ϵ, with REML or PQL estimates used for $\hat{\sigma}^2$—but found that its results were indistinguishable from those of the $\Gamma^{-1}(\epsilon, \epsilon)$ prior.) See, e.g., Daniels (1999) and Gelman (2004) for alternatives to the diffuse priors for variance parameters in hierarchical models which we examine here.

2.3.2. Gibbs sampling in the VC model

The unknown quantities in the VC model can be split into four groups: the fixed effect β_0, the level–2 residuals u_j, the level–2 variance σ_u^2, and the level–1 variance σ_e^2. Typically the parameters $(\beta_0, \sigma_u^2, \sigma_e^2)$ are of principal interest, but Gibbs sampling in this model proceeds most smoothly by treating the level–2 residuals as latent variables and sampling in turn from the full conditional distributions $p(\beta_0|y, \sigma_u^2, \sigma_e^2, u)$, $p(u_j|y, \sigma_u^2, \sigma_e^2, \beta_0)$, $p(\sigma_u^2|y, \beta_0, u, \sigma_e^2)$, and $p(\sigma_e^2|y, \beta_0, u, \sigma_u^2)$ (here y and u are the N– and J–vectors of responses and residuals, respectively).

With $\chi^{-2}(\nu_u, s_u^2)$ and $\chi^{-2}(\nu_e, s_e^2)$ priors for σ_u^2 and σ_e^2, respectively, the full conditionals for model (1) have simple and intuitively reasonable Gaussian and inverse gamma forms (cf. Seltzer et al. 1996):

$$\begin{aligned}
\left(\beta_0|y, \sigma_e^2, u\right) &\sim N\left[\tfrac{1}{N}\textstyle\sum_{ij}\left(y_{ij} - u_j\right), \tfrac{\sigma_e^2}{N}\right], \\
\left(u_j|y, \sigma_u^2, \sigma_e^2, \beta_0\right) &\sim N\left[\tfrac{\hat{D}_j}{\sigma_e^2}\textstyle\sum_{i=1}^{n_j}\left(y_{ij} - \beta_0\right), \hat{D}_j\right], \\
\left(\sigma_u^2|u\right) &\sim \Gamma^{-1}\left[\tfrac{J+\nu_u}{2}, \tfrac{1}{2}\left(\nu_u s_u^2 + \textstyle\sum_{j=1}^{J} u_j^2\right)\right], \quad \text{and} \\
\left(\sigma_e^2|y, \beta_0, u\right) &\sim \Gamma^{-1}\left[\tfrac{N+\nu_e}{2}, \tfrac{1}{2}\left(\nu_e s_e^2 + \textstyle\sum_{ij} e_{ij}^2\right)\right],
\end{aligned} \tag{8}$$

where $\hat{D}_j = \left(\frac{n_j}{\sigma_e^2} + \frac{1}{\sigma_u^2}\right)^{-1}$ and $e_{ij} = y_{ij} - \beta_0 - u_j$.

It is possible to improve upon the Monte Carlo efficiency of the simple Gibbs sampler (8) in VC models with re-parameterization (e.g., Roberts and Sahu 1997), and Metropolis-Gibbs hybrids based

on block updating of the residuals (as in Browne and Draper 2000 for RELR models) may also lead to Monte Carlo acceleration; we do not pursue these possibilities here. It is worth noting in this context that the hierarchical centering parameterization introduced by Gelfand et al. (1995) only leads to better mixing in VC models if $\sigma_e^2 < \sigma_u^2$, which almost never occurs with educational data.

2.3.3. Adaptive hybrid Metropolis-Gibbs sampling in RELR models

Gibbs sampling in RELR models is not straightforward. For example, in the simple model

$$\begin{aligned} (y_{ij} \,|\, p_{ij}) &\sim \text{Bernoulli}(p_{ij}), \quad \text{where} \\ \text{logit}(p_{ij}) &= \beta + u_j, \quad u_j \sim N(0, \sigma_u^2), \end{aligned} \tag{9}$$

and assuming uniform priors, the full conditional for β is

$$p(\beta|y, u, \sigma_u^2) \propto \prod_{ij} \left(1 + e^{-\beta - u_j}\right)^{-y_{ij}} \left(1 + e^{\beta + u_j}\right)^{y_{ij}-1}. \tag{10}$$

This distribution does not lend itself readily to direct sampling. Rejection sampling (Zeger and Karim 1991) is possible, and the software package `BUGS` (Spiegelhalter et al. 1997) employs adaptive rejection sampling (ARS; Gilks and Wild 1992). In this paper we use a hybrid Metropolis-Gibbs approach in which (a) Gibbs sampling is employed for variances and (b) univariate-update random-walk Metropolis sampling with Gaussian proposal distributions is used for fixed effects and residuals; see Browne (1998) for details. As with VC models we take uniform priors on $\mathbb{R}$ for fixed effects and $\chi^{-2}(\nu, s^2)$ priors for the variances of random effects. The fixed effects and residuals may also be block-updated using multivariate normal proposal distributions; Browne and Draper (2000) describes comparisons between these two Metropolis alternatives and documents the pronounced Monte Carlo efficiency advantage of the hybrid Metropolis-Gibbs approach over alternatives such as ARS in RELR models, where the former (with block updating) was 1.7 to 9.0 times faster than the latter in achieving the same accuracy of posterior summaries in the examples studied.

Metropolis sampling with univariate normal proposals requires specification of the variances of the proposal distributions. We use scaled versions of the estimated covariance matrices of REML or PQL estimates to set the initial values of the proposal distribution variances, but optimal scaling factors for many multilevel models are not known (Gelman, Roberts, et al. 1995 contains useful results in simple non-hierarchical settings). Our preferred method for specifying the proposal distribution variances is adaptive (see, e.g., Müller 1993 and Gilks et al. 1998 for other approaches to adaptive Metropolis sampling). From starting values based on the estimated covariance matrices, we first employ a sampling period of random length (but with an upper bound) during which the proposal distribution variances are adaptively tuned and eventually fixed for the remainder of the run; this is followed by the usual burn-in period (see Section 2.3.4); and then the main monitoring run from which posterior summaries are calculated occurs. The tuning of the proposal distribution variances is based on achieving an acceptance rate r for each parameter that lies within a specified tolerance interval $(r - \delta, r + \delta)$.

The algorithm examines empirical acceptance rates in batches of 100 iterations, comparing them for each parameter with the tolerance interval and modifying the proposal distribution appropriately before going on to the next batch of 100. With r^* as the acceptance rate in the most recent batch and σ_p as the proposal distribution SD for a given parameter, the modification performed at the end of each batch is as follows:

$$\text{If} \quad r^* \geq r, \quad \sigma_p \to \sigma_p \left[2 - \left(\frac{1 - r^*}{1 - r}\right)\right], \quad \text{else} \quad \sigma_p \to \frac{\sigma_p}{\left(2 - \frac{r^*}{r}\right)}. \tag{11}$$

This modifies the proposal standard deviation by a greater amount the farther the empirical acceptance rate is from the target r. If r^* is too low, the proposed moves are too big, so σ_p is decreased; if r^* is too high, the parameter space is being explored with moves that are too small, and σ_p is increased. If the r^* values are within the tolerance interval during three successive batches of 100 iterations, the parameter is marked as satisfying its tolerance condition, and once all parameters have been marked the overall tolerance condition is satisfied and adapting stops. After a parameter has been marked it is still modified as before until all parameters are marked, but each parameter only needs to be marked once for the algorithm to end. To limit the time spent in the adapting procedure an upper limit is set (we typically use 5,000 iterations) and after this time the adapting period ends regardless of whether the tolerance conditions are met (in practice this occurs rarely). Values of $(r, \delta) = (0.5, 0.1)$ appear to give near-optimal univariate-update Metropolis performance for a wide variety of multilevel models (Browne and Draper 2000).

2.3.4. Starting values and burn-in strategy

In MCMC sampling with multilevel models it is natural to use as starting values the likelihood and quasi-likelihood results from ML/REML in VC models and from MQL/PQL in REML models. We have found that marginal posteriors in multilevel models of data sets with all but the tiniest sample sizes, even with diffuse priors, are almost invariably unimodal (but see Liu and Hodges 1999 for a cautionary note); this encourages a relatively short burn-in period without fear of missing significant posterior mass in all but the most unusual of situations. We have found burn-ins of 500 iterations to be more than adequate in both the VC and RELR model classes when likelihood-based starting values are used.

3. Variance-components models

3.1. Simulation study design

We have conducted a large simulation study of the properties of Bayesian and likelihood-based estimation methods in the VC model (1). The design of this study was based on the JSP data set introduced in Section 1.1. The numbers n_j of pupils per school in the subsample of $N = 887$ students described in that section averaged 18.5, with an SD of 10.3 and a range from 5 to 61 (i.e., the sampling design across the $J = 48$ schools was quite unbalanced). To examine the effects of J and the distribution of the n_j in the simulations, we removed one pupil at random from each of the 23 largest schools to yield $N = 864$ students, an average of 18 per school. We then varied the number of schools included in the study, with schools chosen so that the average number of pupils per school was maintained at 18 and the sizes of the individual schools were well spread out. We considered four sizes of sampling experiment—6, 12, 24 and 48 schools—with a total of 108, 216, 432 and 864 pupils respectively, and examined one balanced and one unbalanced design in each case. The resulting 8 study designs are given in Table 3. The school-level sample sizes in the cases with unequal n_j were chosen to resemble the actual (highly positively skewed) distribution of class size in the JSP data.

The other factors that varied in our simulations were the true values given to the parameters of model (1): β_0, σ_u^2, and σ_e^2. The fixed effect, β_0, is typically of lesser importance in VC models; we fixed it at 30 throughout all runs. The two variances are more interesting; we chose three possible values for each of these parameters. The between-schools variance, σ_u^2, took the values 1, 10 and 40, and we set the between-pupils variation, σ_e^2, to 10, 40 and 80. For realism in the educational context of the JSP data we only examined cases in which $\sigma_e^2 \geq \sigma_u^2$.

A full-factorial experiment varying both size/balance of the classroom samples and true parameter values was both computationally prohibitive and unnecessary (preliminary investigation revealed little

or no interaction between these two factors), so we (a) made one set of runs varying the sample sizes as in Table 3, while holding the parameters fixed at values similar to those in the JSP data ($\beta_0 = 30, \sigma_u^2 = 10$, and $\sigma_e^2 = 40$), and (b) held the sample sizes constant at the values specified by design 7 in Table 3 (the layout most similar to the JSP data), and varied the parameters across seven settings—$(\sigma_u^2, \sigma_e^2) = (1,10), (1,40), (1,80), (10,10), (10,40), (10,80), (40,80)$, giving rise to intraclass correlation values from 0.012 to 0.5—in all cases with $\beta_0 = 30$. We created 1,000 simulated data sets in each cell of the experimental grid; see the Appendix for additional simulation details.

Table 3. *Summary of study designs for the VC model (1) simulations.*

Design (J)	Number of pupils per school (n_j)												Total number of pupils (N)
1 (6)	5	10	13	18	24	38							108
2 (6)	18	18	18	18	18	18							108
3 (12)	5	8	10	11	11	12	13	15	20	24	26	61	216
4 (12)	18	18	18	18	18	18	18	18	18	18	18	18	216
5 (24)	5	7	8	10	10	11	11	12	12	13	13	14	432
	15	16	18	19	20	21	23	24	26	29	34	61	
6 (24)	(18 for all schools)												432
7 (48)	5	6	7	8	8	10	10	10	11	11	11	11	864
	12	12	12	12	13	13	13	13	14	14	15	15	
	16	16	17	18	18	19	19	20	20	21	21	21	
	23	24	24	24	25	26	27	29	34	37	38	61	
8 (48)	(18 for all schools)												864

3.2. VC results

3.2.1. Estimator bias

All methods of estimating β_0 we examined yielded negligible bias values; for brevity we omit details. Tables 4 and 5 present Monte Carlo estimates of the relative bias of eight methods of estimating σ_u^2 and σ_e^2 in the VC model (1). Two of these methods are likelihood-based (ML and REML), the other six Bayesian: two priors for the variances ($\Gamma^{-1}(\epsilon, \epsilon)$ and $U(0, \frac{1}{\epsilon})$) crossed with three methods of summarizing the posterior distribution for the purpose of point estimation (mean, median, mode). In Table 4 σ_u^2 and σ_e^2 were held constant at 10 and 40, respectively, with the results varying across the eight study designs in Table 3; in Table 5 study design 7 was maintained while (σ_u^2, σ_e^2) varied across seven settings. All methods were close to unbiased for the pupil-level variance σ_e^2, because—even in the smallest study designs—data on 108 or more pupils were available (in particular all relative bias estimates for σ_e^2 in the simulations summarized in Table 5 were less than 1%, and we omit these values for brevity). A number of clear conclusions emerge from these tables; we describe the results in the language of schools and pupils, with obvious extension to other settings.

- Bias for all methods drops steadily with increasing N, and tends to be somewhat smaller with balanced designs than when substantial imbalance is present. In Table 5 the magnitude of the bias of estimates of σ_u^2 generally decreases as the intraclass correlation ρ increases from near 0 to 0.5.

- ML estimates of σ_u^2 are biased low with the smallest designs; this is effectively remedied by the REML bias correction.

Table 4. *Estimates of relative bias for the variance parameters in VC model (1) with a variety of methods and study designs. The true values of σ_u^2 and σ_e^2 were 10 and 40, respectively. Figures in parentheses are Monte Carlo SEs.*

σ_u^2 Relative Bias (%)		Study Number							
Estimation Method		1	2	3	4	5	6	7	8
ML		–22.6	–20.1	–11.9	–9.8	–2.4	–4.1	–2.1	–2.0
		(2.1)	(2.0)	(1.6)	(1.4)	(1.1)	(1.1)	(0.9)	(0.8)
REML		–1.0	0.0	–1.0	0.4	3.1	1.0	0.5	0.5
		(2.5)	(2.4)	(1.7)	(1.5)	(1.2)	(1.2)	(0.9)	(0.8)
$\Gamma^{-1}(\epsilon,\epsilon)$ Prior	Mean	49.1	51.4	18.4	20.3	12.0	9.7	4.7	4.8
		(4.1)	(4.0)	(2.2)	(2.1)	(1.3)	(1.3)	(0.9)	(0.9)
	Median	–6.7	–0.6	–1.9	1.1	3.5	1.7	0.9	1.0
		(2.9)	(2.8)	(1.9)	(1.8)	(1.2)	(1.2)	(0.9)	(0.8)
	Mode	–33.6	–31.6	–27.3	–24.1	–12.8	–13.4	–7.7	–7.1
		(1.9)	(1.9)	(1.5)	(1.4)	(1.1)	(1.1)	(0.8)	(0.8)
Uniform$(0,\frac{1}{\epsilon})$ Prior	Mean	481	450	74.9	70.9	30.8	26.7	12.5	12.0
		(10.2)	(9.7)	(2.7)	(2.6)	(1.4)	(1.4)	(1.0)	(0.9)
	Median	140	133	40.6	39.0	20.1	16.9	8.3	8.0
		(5.1)	(4.9)	(2.3)	(2.2)	(1.5)	(1.3)	(0.9)	(0.9)
	Mode	107	94.3	1.2	0.4	0.8	–1.1	–1.0	–0.8
		(3.8)	(3.6)	(1.7)	(1.6)	(1.2)	(1.2)	(0.9)	(0.8)

σ_e^2 Relative Bias (%)		Study Number							
Estimation Method		1	2	3	4	5	6	7	8
ML		–0.42	–0.45	–0.02	–0.16	–0.31	–0.15	–0.04	–0.09
REML		–0.42	–0.41	–0.03	–0.16	–0.31	–0.15	–0.04	–0.09
$\Gamma^{-1}(\epsilon,\epsilon)$	Mean	2.8	2.8	1.6	1.4	0.3	0.4	0.3	0.2
	Median	1.1	1.4	0.9	0.7	–0.0	0.1	0.1	0.1
	Mode	–1.2	–1.2	–0.4	–0.6	–0.7	–0.6	–0.2	–0.3
Uniform$(0,\frac{1}{\epsilon})$	Mean	3.5	3.6	2.0	1.9	0.7	0.8	0.4	0.4
	Median	1.8	2.3	1.4	1.3	0.3	0.5	0.3	0.3
	Mode	–0.6	–0.5	0.0	–0.1	–0.3	–0.2	–0.1	–0.1

Note: The Monte Carlo SEs for all rows in the σ_e^2 portion of this table were 0.5 (designs 1 and 2), 0.3 (designs 3 and 4) and 0.2 (designs 5–8).

- Posterior means with the $\Gamma^{-1}(\epsilon,\epsilon)$ prior for the school-level variance σ_u^2 are sharply biased high with small sample sizes; this largely disappears when posterior medians are used with this prior. The exception to this pattern occurs when σ_e^2 is 40–80 times larger than σ_u^2, a situation which gave all of the methods trouble but which arguably casts doubt on the need for random effects at level 2 in the first place. Posterior modes with the $\Gamma^{-1}(\epsilon,\epsilon)$ prior are uniformly biased on the low side, sometimes substantially.

- The $U(0,\frac{1}{\epsilon})$ prior can produce huge positive biases when attention focuses on the posterior mean, but has good bias properties with all but the smallest sample sizes when the mode is used as a point estimate. There is clearly a trade-off between choice of prior distribution and choice of posterior summary; the need for these choices gives REML the advantage on bias grounds in small samples.

- The behavior of the two priors is understandable given their shape on the σ^2 scale: $\Gamma^{-1}(\epsilon, \epsilon)$ priors have a sharp spike near 0, which has no effect when the likelihood is concentrated away from 0 but which can create appreciable negative bias when the data evidence for positive σ^2 is weak. By contrast $U(0, \frac{1}{\epsilon})$ priors do not have this defect, but claiming in the prior that σ^2 is as likely to be 500 (say) as it is to be 10 creates substantial positive bias when the true value is near 10 but sample sizes are small, leading to a relatively diffuse likelihood.

Table 5. *Estimates of relative bias for the variance parameter σ_u^2 in VC model (1) with a variety of methods and true parameter values. All runs use study design 7. Figures in parentheses are Monte Carlo SEs. Column headings record the true values of σ_u^2, σ_e^2, and the intraclass correlation ρ.*

$\boxed{\sigma_u^2}$ Relative Bias (%)		$\sigma_u^2; \sigma_e^2/\rho$						
Estimation Method		1; 80/ 0.012	1; 40/ 0.024	1; 10/ 0.091	10; 80/ 0.111	10; 40/ 0.200	40; 80/ 0.333	10; 10/ 0.500
ML		–3.4 (3.0)	–6.5 (2.1)	–3.1 (1.1)	–2.8 (1.0)	–2.1 (0.9)	–1.9 (0.8)	–1.7 (0.7)
REML		7.2 (3.2)	0.3 (2.1)	0.4 (1.1)	0.4 (1.0)	0.5 (0.9)	0.5 (0.8)	0.5 (0.7)
$\Gamma^{-1}(\epsilon, \epsilon)$ Prior	Mean	–22.8 (2.5)	–18.5 (2.1)	3.2 (1.2)	3.7 (1.1)	4.7 (0.9)	4.9 (0.8)	4.9 (0.8)
	Median	–47.9 (2.5)	–31.7 (2.2)	–1.7 (1.2)	–0.8 (1.1)	0.9 (0.9)	1.4 (0.8)	1.6 (0.7)
	Mode	–60.0 (1.7)	–50.1 (1.7)	–15.1 (1.1)	–12.7 (1.0)	–7.7 (0.8)	–5.5 (0.7)	–4.5 (0.7)
Uniform$(0, \frac{1}{\epsilon})$ Prior	Mean	84.5 (3.1)	39.6 (2.2)	15.9 (1.2)	14.8 (1.1)	12.5 (1.0)	11.2 (0.9)	10.6 (0.8)
	Median	61.0 (3.1)	27.1 (2.1)	10.5 (1.2)	9.8 (1.1)	8.3 (0.9)	7.4 (0.8)	6.9 (0.8)
	Mode	15.7 (2.4)	–4.2 (1.8)	–3.9 (1.1)	–2.9 (1.0)	–1.0 (0.9)	–0.1 (0.8)	0.4 (0.7)

3.2.2. Interval performance

We also monitored the coverage and length of interval estimates for the parameters in the VC model (1). To construct Bayesian $100(1-\gamma)\%$ intervals we simply used the $100\frac{\gamma}{2}\%$ and $100\left(1-\frac{\gamma}{2}\right)\%$ quantiles of the relevant posterior distributions (as estimated by MCMC). With the likelihood methods we examined six approaches: the first was intended (as in Examples 1 and 2) to reflect the behavior of many practitioners of multilevel modeling who are presented in the output of the standard computer programs with nothing more than an estimate and a standard error; the second through fifth are simple computationally inexpensive small-sample adjustments to the first for variance components; and the sixth is an idealized version of likelihood interval estimation for variances, assuming knowledge of the sampling distribution which would not be available with a single sample. For brevity we present ML results only for the first method.

- Method 1 used intervals of the form $\left[\hat{\theta} \pm \Phi^{-1}\left(1-\frac{\gamma}{2}\right)\widehat{SE}\left(\hat{\theta}\right)\right]$ based on asymptotic normality of the MLE.

- In the case of variance parameters, method 2 approximates the sampling distribution of the likelihood estimate by a $\Gamma(\alpha, \beta)$ distribution (preliminary work suggested that this approximation was reasonable for moderate to large sample sizes). In this approach we equated the mean $\frac{\alpha}{\beta}$ of the gamma distribution to $\hat{\sigma}^2$ and the variance $\frac{\alpha}{\beta^2}$ to $\hat{V}\left(\hat{\sigma}^2\right)$, obtaining $[\hat{\alpha}, \hat{\beta}] = \left[\hat{\sigma}^4/\hat{V}\left(\hat{\sigma}^2\right), \hat{\sigma}^2/\hat{V}\left(\hat{\sigma}^2\right)\right]$, and then used quantiles of the corresponding gamma distribution to generate the interval endpoints. (In the smaller study designs the distribution of the REML estimate is a mixture of a point mass at 0 and an approximate gamma distribution conditional on being positive. Any attempt to achieve further improvement in a small-sample likelihood-based approximation would have to cope with the spike at 0.)

- Methods 3 and 4 use Taylor series and transformations to normality. Suppose that the sampling distribution of $g\left(\hat{\sigma}^2\right)$ is approximately Gaussian for some invertible function g, and $\hat{\sigma}^2$ is approximately unbiased. Then by the Δ–method $g\left(\hat{\sigma}^2\right)$ has approximate mean $g\left(\sigma^2\right)$ and variance $\left[g'\left(\sigma^2\right)\right]^2 V\left(\hat{\sigma}^2\right)$, and an approximate $100(1-\gamma)\%$ confidence interval for σ^2 is therefore of the form
$$g^{-1}\left[g\left(\hat{\sigma}^2\right) \pm \Phi^{-1}\left(1-\frac{\gamma}{2}\right)\left|g'\left(\hat{\sigma}^2\right)\right| \widehat{SE}\left(\hat{\sigma}^2\right)\right]. \tag{12}$$
Method 3 takes the sampling distribution of $g\left(\hat{\sigma}^2\right)$ to be approximately lognormal and uses $g(\cdot) = \ln(\cdot)$, and method 4 employs the Wilson-Hilferty (1931) optimal transformation to normality for gamma random variables, $g(\cdot) = (\cdot)^{\frac{1}{3}}$. Both of these methods fail when $\hat{\sigma}^2 = 0$ because of division by zero in the derivative calculation in (12).

- Method 5, which uses a variance-stabilizing (VS) transformation, is based on the observation by Longford (2000) that (a) the ML estimate of the variance ratio $\tau = \frac{\sigma_u^2}{\sigma_e^2}$ is highly correlated with its estimated asymptotic standard error $\widehat{SE}\left(\hat{\tau}\right)$, and (b) this dependence is removed asymptotically by working instead with $\eta = \ln\left(\bar{n}^{-1} + \tau\right)$, where $\bar{n}$ is a suitably chosen mean of the numbers n_j of level–1 units per level–2 unit (we found that harmonic means work best). This suggests building a Gaussian interval estimate on the η scale, relying on the large-sample result $V(\hat{\eta}_{\mathrm{ML}}) = \frac{2}{J}$, and back-transforming to obtain an interval for τ. We then convert this into an interval for σ_u^2 by using the REML estimate of σ_e^2 in place of σ_e^2, which should yield good performance in our context because the total number N of level–1 units in our simulations never drops below 108. The resulting intervals for σ_u^2 have the form
$$\hat{\sigma}_e^2\left\{\exp\left[\ln\left(\bar{n}^{-1} + \frac{\hat{\sigma}_u^2}{\hat{\sigma}_e^2}\right) \pm \Phi^{-1}\left(1-\frac{\gamma}{2}\right)\sqrt{\frac{2}{J}}\right] - \bar{n}^{-1}\right\}. \tag{13}$$
These intervals may have a negative left endpoint when σ_u^2 is small in relation to σ_e^2; in many uses of model (1) this is undesirable, but (as Longford points out) reformulations of the model exist in which negative values of τ are sensible subject to a positive-definite constraint on the implied covariance matrix of y.

- To estimate "best possible" (idealized) performance of the likelihood intervals for variances (method 6), we reasoned as follows. As in method 2, the sampling distribution for a likelihood estimate such as $\hat{\sigma}_u^2$ should be approximately gamma, with parameters $(\hat{\alpha}, \hat{\beta})$ which depend on the study design and underlying model parameters, and if these $(\hat{\alpha}, \hat{\beta})$ values were known an interval estimate for σ_u^2 could be formed by analogy with the usual result with an IID Gaussian sample of size n: $\frac{(n-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$, i.e., $\hat{\sigma}^2 \sim \Gamma\left(\frac{n-1}{2}, \frac{n-1}{2\sigma^2}\right)$. In each of the cells of our simulation

grid we therefore used maximum likelihood (e.g., Johnson et al. 1994) to estimate $(\hat{\alpha}, \hat{\beta})$ from the 1,000 simulation replications, set $\hat{n} = 2\hat{\alpha} + 1$, and constructed 1,000 idealized interval estimates of the form

$$\left[\frac{(\hat{n}-1)}{\chi^2_{\hat{n}-1,1-\frac{\gamma}{2}}} \hat{\sigma}^2, \frac{(\hat{n}-1)}{\chi^2_{\hat{n}-1,\frac{\gamma}{2}}} \hat{\sigma}^2 \right], \tag{14}$$

where $\chi^2_{k,\gamma}$ is the γ quantile of the χ^2_k distribution. This method also fails when $\hat{\sigma}^2 = 0$ because the MLEs of the parameters of a gamma distribution are undefined if any of the data values are zero.

Tables 6 and 7 present the actual coverage and mean length of nominal 95% interval estimates for σ^2_u. The following conclusions are evident from these tables, and from other simulation results not presented here (for more details see Browne 1998, which is available on the web at `www.ams.ucsc.edu/~draper`).

- Intervals for σ^2_e (not shown) had close to nominal coverage with all methods, and will not be discussed further. The coverage of ML/REML intervals for the fixed effect β_0 (also not shown) was below nominal with 6–12 schools and 108–216 pupils (study designs 1–4) but approached nominal levels with larger sample sizes. Bayesian interval coverage for β_0 with both $\Gamma^{-1}(\epsilon, \epsilon)$ and $U(0, \frac{1}{\epsilon})$ priors for the variance components was close to nominal in all designs examined (β_0 and σ^2_u are correlated in the posterior, so the prior for σ^2_u affects inferences about β_0).

- The effects of imbalance in the design were small but nonzero, and intuitively reasonable: holding the total number of pupils constant, balance yielded narrower intervals and generally better coverage.

- As was the case with bias in the estimation of σ^2_u (Table 5), interval performance generally improved as the intraclass correlation ρ increased away from 0 (Table 7). Even with data on 48 schools and 864 pupils, both likelihood-based and Bayesian methods can have difficulty apportioning variation within and between schools when σ^2_e is much larger than σ^2_u.

- ML produced Gaussian intervals for σ^2_u that were consistently too narrow to achieve good coverage. REML improved on this but still fell below nominal coverage in all situations examined, using both the Gaussian and gamma intervals. In the two smallest designs the lognormal and cube root REML intervals failed to exist 4–5% of the time, and over-covered when they did not fail (and the lognormal intervals continued to over-cover in designs 3 and 4), but with 24 or more level–2 units (schools) both transformation-based methods improved on the Gaussian and gamma intervals and achieved coverages close to nominal. The lognormal and cube root REML intervals failed to exist 7–19% of the time when $\rho \leq 0.024$, but—as mentioned earlier—in such situations the need for VC modeling is unclear. The VS intervals sharply over-covered when τ was small and had a negative left endpoint 4–43% of the time in the smallest designs, but performed well otherwise.

- Bayesian intervals for σ^2_u with the $U(0, \frac{1}{\epsilon})$ prior had actual coverages at or close to nominal levels in all study designs and parameter settings examined. The $\Gamma^{-1}(\epsilon, \epsilon)$ intervals undercovered to some extent (actual levels near 90% at nominal 95%) when the number of level–2 units or the variance ratio τ were small, but performed well in all other situations. Note, however, that the $U(0, \frac{1}{\epsilon})$ intervals were extremely wide with small samples (Table 6); further work is needed to see if other prior specifications might yield narrower but still well-calibrated intervals in such situations.

Table 6. *Performance of interval estimates of σ_u^2 in VC model (1) with a variety of methods and study designs. The top table gives actual coverages of nominal 95% intervals; the bottom table records mean interval lengths. The true values of σ_u^2 and σ_e^2 were 10 and 40, respectively. Figures in parentheses in the bottom table are Monte Carlo SEs, and in the top table (i) values in square brackets report the percentage of time REML yielded variance estimates of zero and (ii) values in curly brackets record the percentage of time the VS intervals had a negative left endpoint (LE).*

$\boxed{\sigma_u^2}$ Coverage (%)		Study Number							
Estimation Method		1	2	3	4	5	6	7	8
ML	Gaussian	71.9	73.3	80.9	83.0	89.5	88.5	91.4	90.7
REML	Gaussian	78.5	80.4	86.2	87.1	91.2	90.2	92.4	91.1
	Gamma	84.1	85.9	90.0	91.0	93.1	92.3	93.8	92.4
	Lognormal*	99.1	98.7	98.4	98.2	95.0	94.8	94.5	93.9
	Cube Root*	99.3	98.3	93.1	94.5	93.9	93.5	94.0	93.5
	VS	90.7	89.1	92.9	93.3	94.0	93.4	94.8	93.1
	Idealized*	94.5	93.6	95.5	95.7	93.7	94.5	94.7	94.8
REML	% zero $\hat{\sigma}_u^2$	[4.8%]	[3.6%]	[0.4%]	[0%]	[0%]	[0%]	[0%]	[0%]
VS	% LE < 0	{43%}	{27%}	{12%}	{3.9%}	{0.2%}	{0.1%}	{0%}	{0%}
$\Gamma^{-1}(\epsilon, \epsilon)$ Prior		88.9	88.5	92.2	93.7	94.0	93.9	93.8	93.4
Uniform$(0, \frac{1}{\epsilon})$ Prior		91.5	90.7	94.2	93.0	93.8	93.5	93.0	93.2

$\boxed{\sigma_u^2}$ Interval Length		Study Number							
Estimation Method		1	2	3	4	5	6	7	8
ML	Gaussian	23.2	22.9	18.6	18.0	14.1	13.4	9.9	9.6
		(0.5)	(0.5)	(0.3)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)
REML	Gaussian	28.3	27.4	20.4	19.5	14.7	13.9	10.2	9.8
		(0.6)	(0.6)	(0.3)	(0.3)	(0.1)	(0.1)	(0.1)	(0.1)
	Gamma	27.0	26.3	19.9	19.1	14.5	13.8	10.1	9.8
		(0.6)	(0.5)	(0.3)	(0.3)	(0.1)	(0.1)	(0.1)	(0.1)
	Lognormal*	—	—	—	—	16.0	15.1	10.6	10.2
		—	—	—	—	(0.1)	(0.1)	(0.1)	(0.1)
	Cube Root*	36.9	32.9	21.5	20.5	15.0	14.2	10.2	9.9
		(3.8)	(1.2)	(0.3)	(0.3)	(0.1)	(0.1)	(0.1)	(0.1)
	VS	36.7	33.9	23.4	21.8	15.8	14.7	10.6	10.1
		(0.7)	(0.7)	(0.3)	(0.3)	(0.1)	(0.1)	(0.1)	(0.1)
	Idealized*	131	112	43.6	39.7	19.7	19.4	12.4	11.8
		(3.2)	(2.6)	(0.7)	(0.7)	(0.2)	(0.2)	(0.1)	(0.1)
$\Gamma^{-1}(\epsilon, \epsilon)$ Prior		59.5	58.1	29.5	28.4	17.5	16.6	11.0	10.6
		(1.4)	(1.3)	(0.4)	(0.4)	(0.2)	(0.2)	(0.1)	(0.1)
Uniform$(0, \frac{1}{\epsilon})$ Prior		299	273	46.7	44.0	21.0	19.7	11.8	11.5
		(5.0)	(4.7)	(0.6)	(0.6)	(0.2)	(0.2)	(0.1)	(0.1)

Notes: Monte Carlo SEs for coverage rates in the top table ranged from 0.3% (for estimates near 99%) to 1.4% (for estimates near 70%). The dashes in the lognormal interval lengths replace enormous numbers arising from division by near-zero values. *In the lognormal, cube-root, and idealized cases, interval coverages and lengths were based only on the replications in which the estimate was nonzero.

Table 7. *Performance of interval estimates of σ_u^2 in VC model (1) with a variety of methods and true parameter values. The top table gives actual coverages of nominal 95% intervals; the bottom table records mean interval lengths. All runs use study design 7. In the top table column headings record the true values of σ_u^2, σ_e^2, and the intraclass correlation ρ; in the bottom table just σ_u^2 and σ_e^2 are noted, and the columns are sorted not by ρ but by σ_u^2. Figures in parentheses in the bottom table are Monte Carlo SEs, and in the top table (i) values in square brackets report the percentage of time REML yielded variance estimates of zero and (ii) values in curly brackets record the percentage of time the VS intervals had a negative left endpoint (LE).*

$\boxed{\sigma_u^2}$ Coverage (%)		$\sigma_u^2; \sigma_e^2/\rho$						
Estimation Method		1; 80/ 0.012	1; 40/ 0.024	1; 10/ 0.091	10; 80/ 0.111	10; 40/ 0.200	40; 80/ 0.333	10; 10/ 0.500
ML	Gaussian	78.5	88.0	90.7	90.1	91.4	92.1	91.4
REML	Gaussian	80.4	89.4	91.8	91.7	92.4	92.9	92.7
	Gamma	75.7	88.7	93.7	93.5	93.8	93.2	93.9
	Lognormal*	92.1	94.6	95.5	95.0	94.5	94.3	94.6
	Cube Root*	95.4	96.9	94.1	94.3	94.0	94.3	94.1
	VS	99.2	98.9	94.5	94.6	94.8	94.4	94.5
	Idealized*	90.7	94.6	94.6	94.9	94.7	95.5	95.8
REML	% 0 Estimate	[19%]	[7.0%]	[0.1%]	[0%]	[0%]	[0%]	[0%]
VS	% LE < 0	{92%}	{74%}	{1.2%}	{0.2%}	{0%}	{0%}	{0%}
$\Gamma^{-1}(\epsilon, \epsilon)$ Prior		89.5	88.6	92.6	93.5	93.8	93.8	94.3
Uniform$(0, \frac{1}{\epsilon})$ Prior		95.9	95.5	93.0	93.1	93.0	92.9	92.8

$\boxed{\sigma_u^2}$ Interval Length		$\sigma_u^2; \sigma_e^2$						
Estimation Method		1; 10	1; 40	1; 80	10; 10	10; 40	10; 80	40; 80
ML	Gaussian	1.27 (0.01)	2.39 (0.03)	3.49 (0.1)	8.41 (0.1)	9.93 (0.1)	11.8 (0.1)	35.7 (0.2)
REML	Gaussian	1.30 (0.01)	2.47 (0.03)	3.66 (0.1)	8.59 (0.1)	10.2 (0.1)	12.1 (0.1)	36.5 (0.2)
	Gamma	1.29 (0.01)	2.34 (0.03)	3.31 (0.1)	8.56 (0.1)	10.1 (0.1)	12.0 (0.1)	36.2 (0.2)
	Lognormal*	1.40 (0.01)	— —	— —	8.86 (0.1)	10.6 (0.1)	12.8 (0.1)	37.8 (0.3)
	Cube Root*	1.32 (0.01)	40.3 (25.5)	14.1 (4.0)	8.65 (0.1)	10.2 (0.1)	12.2 (0.1)	36.8 (0.2)
	VS	1.41 (0.01)	3.15 (0.02)	5.53 (0.03)	8.83 (0.06)	10.6 (0.1)	12.9 (0.1)	37.7 (0.3)
	Idealized*	1.81 (0.02)	7.46 (0.15)	13.6 (0.3)	10.0 (0.1)	12.4 (0.1)	16.0 (0.2)	42.9 (0.3)
$\Gamma^{-1}(\epsilon, \epsilon)$ Prior		1.40 (0.01)	2.28 (0.03)	2.95 (0.1)	9.34 (0.1)	11.0 (0.1)	13.0 (0.1)	39.6 (0.3)
Uniform$(0, \frac{1}{\epsilon})$ Prior		1.53 (0.01)	2.99 (0.03)	4.71 (0.05)	9.97 (0.1)	11.8 (0.1)	14.2 (0.1)	42.5 (0.3)

Notes: See Table 6.

- In some cases REML estimated asymptotic standard errors underestimated the actual sampling variabilities they were meant to estimate. This may be seen from the substantially improved performance of the idealized intervals over the REML gamma intervals in small samples, and is

also clear from a comparison of the mean value of the REML squared standard errors for $\hat{\sigma}_u^2$ with the sample variance of the 1,000 simulated $\hat{\sigma}_u^2$ values: across studies 1–8 in Table 6, ratios of the form $\left\{\text{mean}\left[\widehat{SE}^2\left(\hat{\sigma}^2\right)\right] / \hat{V}\left(\hat{\sigma}^2\right)\right\}$ came out $(.854, .837, .920, .910, 1.02, .933, .899, .925)$, respectively, i.e., the REML squared SEs underestimated the sampling variance on average by 15–16% in studies 1–2 (see Longford 2000 for a theoretical explanation of this phenomenon).

4. Random-effects logistic regression models

4.1. Simulation study design

We have also conducted a large simulation study of the properties of quasi-likelihood and Bayesian estimation methods in the RELR model (2). The design of this study was based on the Rodríguez-Goldman data set introduced in Section 1.2. Conditioning on both the covariates $(x_{1ijk}, x_{2jk}, x_{3k})$ and the true parameter values $(\beta_0 = 0.65, \beta_1 = \beta_2 = \beta_3 = \sigma_u^2 = \sigma_v^2 = 1.0)$ used by Rodríguez and Goldman (1995) in their likelihood-based simulation study, we used model (2) to create 500 simulation replications of the Rodríguez-Goldman data structure, each with 161 communities, 1,558 mothers, and 2,449 births.

For each simulated data set we estimated the six parameters using two quasi-likelihood methods—MQL_1 and PQL_2—and Bayesian fitting with two priors. In the quasi-likelihood estimation we used a convergence tolerance (maximum relative change in parameter values from one iteration to the next) of 0.01. For Bayesian estimation we used MCMC with (improper) uniform priors on $\mathbb{R}$ on the β_l and two prior distributions on the variance components: $\Gamma^{-1}(\epsilon, \epsilon)$ and (improper) uniform on $(0, \infty)$, functionally equivalent to a proper $U(0, \frac{1}{\epsilon})$ prior for small ϵ. We used the adaptive hybrid Metropolis-Gibbs method described in Section 2.3.3, with a maximum adaptation period of 5,000 iterations, a target acceptance rate of 44%, a burn-in from PQL_2 starting values of 500 iterations, and a monitoring run of 25,000 iterations (based on Raftery-Lewis default accuracy recommendations).

4.2. RELR results

Table 8 and Figures 1 and 2 present our simulation findings. For each of the six parameters Table 8 contrasts the mean estimate, and coverage and length of nominal 95% intervals, for the various estimation methods, using posterior means as Bayesian point estimates (medians and modes gave essentially the same results); the table also summarizes large-sample Gaussian intervals—and gamma, lognormal, and idealized intervals as in Section 3.2.2 for the variance parameters—based on the quasi-likelihood methods (the cube root results were inferior to those from the lognormal approximation, and the VS method is not readily adaptable to this setting since there is no direct estimate of the level–1 variance). Figures 1 and 2 give calibration plots for the six parameters (three in each figure), in which nominal and actual coverage of $100(1-\gamma)\%$ intervals for $\gamma = .01, .02, \ldots, .99$ are contrasted for the various estimation methods, using Gaussian intervals for MQL_1 and PQL_2 for the fixed effects and adding the PQL_2 lognormal intervals for the variance parameters. The following conclusions may be drawn from these summaries.

- MQL_1 yielded sharply biased estimates and very poor coverage properties, especially for the random-effects variances (e.g., the MQL_1 point estimate of the level–2 variance σ_u^2 was 0 in 58% of the simulated data sets). PQL_2 produced a considerable improvement, but bias and undercoverage with the Gaussian intervals were still noticeable, especially for σ_u^2. The lognormal intervals offered some improvement but still exhibited substantial undercoverage.

Table 8. *Mean estimates (top table), and coverage (middle table) and length (bottom table) of nominal 95% intervals, for four estimation methods in RELR model (2) with the Rodríguez-Goldman data structure. True values of the parameters are given in square brackets in the top table. 95% central posterior Bayesian intervals are reported, and figures in parentheses are Monte Carlo SEs.*

Mean Estimate		Parameter					
Estimation Method		β_0 [0.65]	β_1 [1.0]	β_2 [1.0]	β_3 [1.0]	σ_v^2 [1.0]	σ_u^2 [1.0]
MQL_1		0.474	0.741	0.753	0.727	0.550	0.026
		(0.007)	(0.007)	(0.004)	(0.009)	(0.004)	(0.002)
PQL_2		0.612	0.945	0.958	0.942	0.888	0.568
		(0.009)	(0.009)	(0.005)	(0.011)	(0.009)	(0.010)
Bayesian	$\Gamma^{-1}(\epsilon,\epsilon)$ Priors	0.638	0.991	1.006	0.982	1.023	0.964
		(0.010)	(0.010)	(0.006)	(0.012)	(0.011)	(0.018)
	$U(0,\infty)$ Priors	0.655	1.015	1.031	1.007	1.108	1.130
		(0.010)	(0.010)	(0.005)	(0.013)	(0.011)	(0.016)

Actual Coverage (%)		Parameter					
Estimation Method		β_0	β_1	β_2	β_3	σ_v^2	σ_u^2
MQL_1	Gaussian	76.8	68.6	17.6	69.6	2.4	0.0
		(1.9)	(2.1)	(1.7)	(2.1)	(0.7)	(—)
PQL_2	Gaussian	92.0	96.2	90.8	89.8	77.6	26.8
		(1.2)	(0.9)	(1.3)	(1.4)	(1.9)	(2.0)
	Gamma	—	—	—	—	81.0	31.4
		—	—	—	—	(1.8)	(2.1)
	Lognormal	—	—	—	—	84.2	37.4
		—	—	—	—	(1.6)	(2.1)
	Idealized	—	—	—	—	93.6	83.4
		—	—	—	—	(1.1)	(1.7)
Bayesian	$\Gamma^{-1}(\epsilon,\epsilon)$ Priors	93.2	96.4	92.6	92.2	94.4	88.6
		(1.1)	(0.8)	(1.2)	(1.2)	(1.0)	(1.4)
	$U(0,\infty)$ Priors	93.6	96.4	92.8	93.6	92.2	93.0
		(1.1)	(0.8)	(1.2)	(1.1)	(1.2)	(1.1)

Mean Interval Length		Parameter					
Estimation Method		β_0	β_1	β_2	β_3	σ_v^2	σ_u^2
MQL_1	Gaussian	0.589	0.681	0.327	0.746	0.404	0.177
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
PQL_2	Gaussian	0.735	0.796	0.400	0.930	0.638	0.591
		(0.003)	(0.002)	(0.001)	(0.003)	(0.005)	(0.002)
	Gamma	—	—	—	—	0.636	0.586
		—	—	—	—	(0.005)	(0.003)
	Lognormal	—	—	—	—	0.641	0.635
		—	—	—	—	(0.005)	(0.004)
	Idealized	—	—	—	—	0.851	1.25
		—	—	—	—	(0.009)	(0.022)
Bayesian	$\Gamma^{-1}(\epsilon,\epsilon)$ Priors	0.798	0.875	0.463	1.01	0.878	1.25
		(0.004)	(0.003)	(0.002)	(0.004)	(0.009)	(0.015)
	$U(0,\infty)$ Priors	0.828	0.895	0.476	1.05	0.948	1.32
		(0.003)	(0.002)	(0.002)	(0.004)	(0.008)	(0.011)

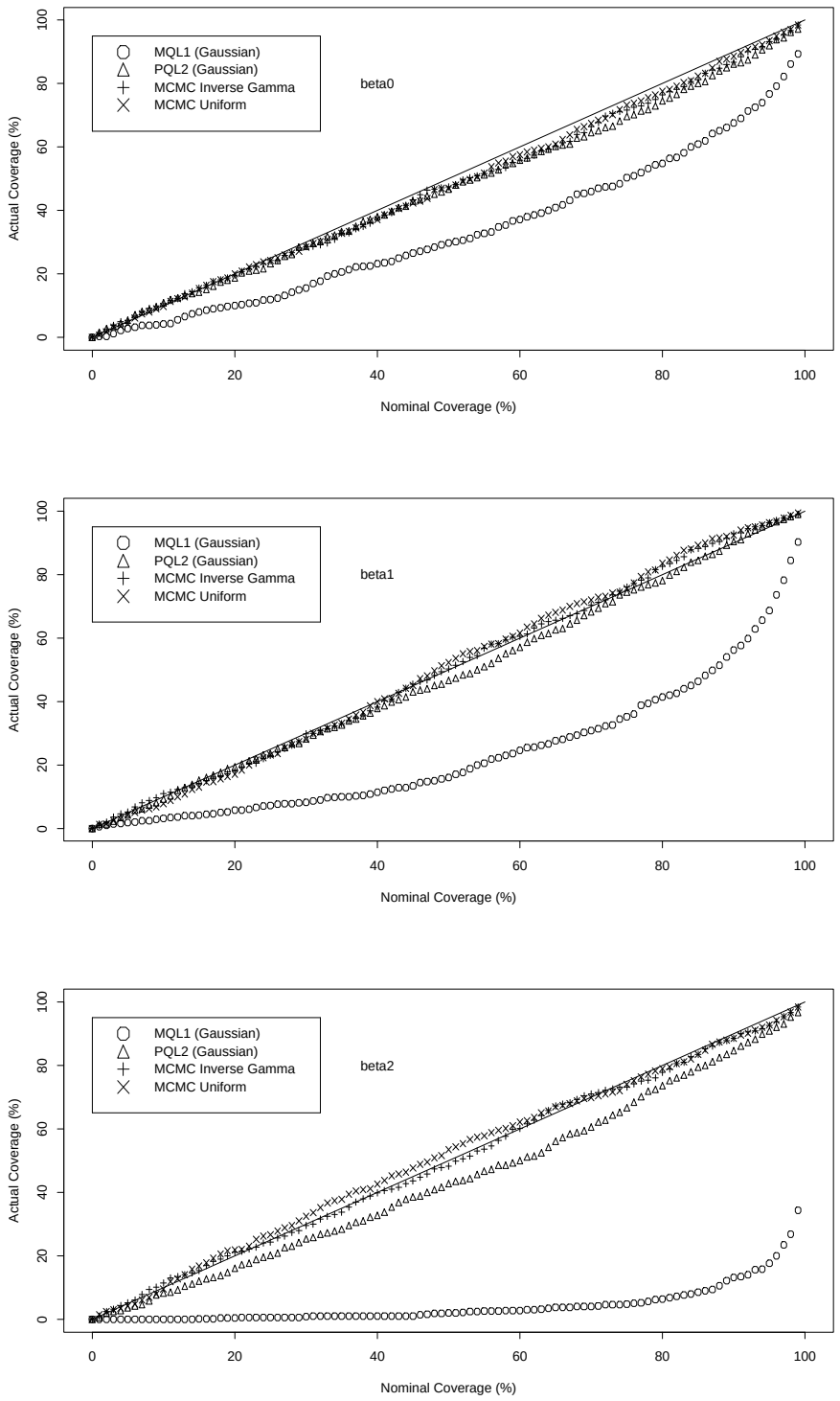

Figure 1. *Actual versus nominal coverage of four estimation methods for the parameters β_0, β_1 and β_2 in the RELR model (2).*

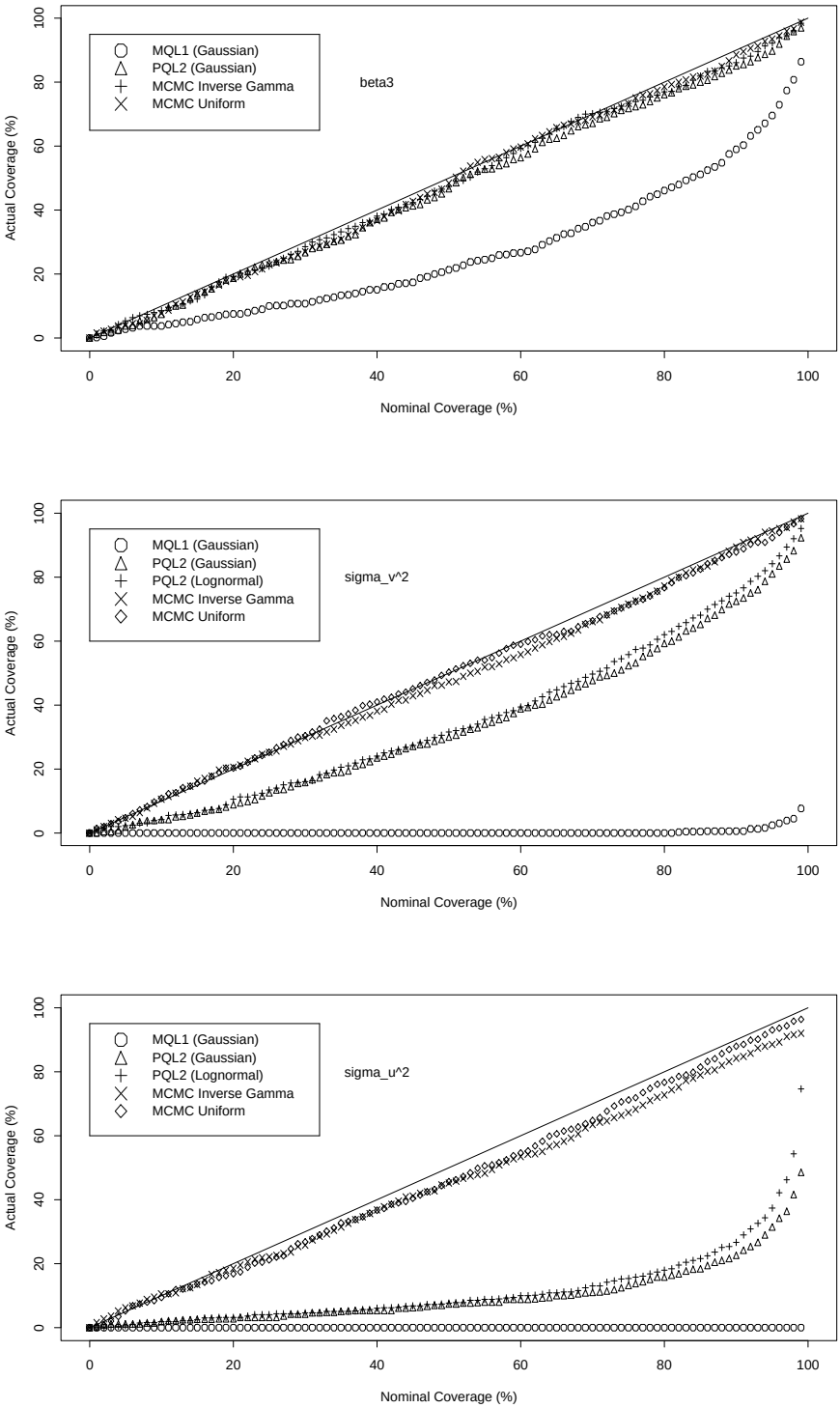

Figure 2. *Actual versus nominal coverage of four estimation methods for the parameters* β_3, σ_v^2 *and* σ_u^2 *in the RELR model (2).*

- PQL_2 underperformed for the variance estimates both because the PQL estimates are biased low and because the PQL standard errors are too small (see Engel 1998 and Lee and Nelder 2000 for theoretical results that support this conclusion). As was the case with the VC model, this may be seen in two ways: (a) by the improved performance of the idealized interval estimates and (b) through the ratios $\left\{\text{mean}\left[\widehat{SE}^2\left(\hat{\sigma}^2\right)\right] / \hat{V}\left(\hat{\sigma}^2\right)\right\}$, which were 0.447 and 0.672 for σ_u^2 and σ_v^2, respectively, i.e., the typical estimated variance of $\hat{\sigma}_u^2$ in any given simulated data set was only about 45% of the actual sampling variance across the 500 data sets. This seems to be largely a small-sample problem for PQL—even though each simulated data set had 2,449 births, the average number of women per community (the most important determinant of the accuracy of $\hat{\sigma}_u^2$) was only $\frac{1{,}558}{161} \doteq 9.7$—but note that even with 161 communities the PQL performance for σ_v^2 was also unsatisfactory.

- Bayesian estimates with both priors were close to unbiased and well calibrated for all parameters, with actual coverage values close to nominal at all levels in Figures 1 and 2.

5. Discussion and conclusions

In two large simulations studies whose design is realistic for educational and medical research, we have examined the performance of likelihood-based and Bayesian methods of fitting variance-components (VC) and random-effects logistic regression (RELR) models, focusing on the likelihood-based approaches in most frequent current use in the applied multilevel-modeling literature. Our main findings are as follows.

- In two-level VC models with a wide variety of sample sizes and true parameter values,
  - both likelihood-based (ML and REML) and Bayesian (diffuse-prior) methods can be made to yield approximately unbiased point estimates, in the likelihood case by using REML rather than ML estimates, and in the Bayesian case by choosing one of several combinations of diffuse priors and posterior point summaries. The automatic nature of this choice for REML represents an advantage for the likelihood-based approach as far as bias is concerned with small samples; however
  - both approaches experienced difficulty in attaining nominal coverage of interval estimates in two situations: when (i) the number J of level–2 units and/or (ii) the variance ratio $\tau = \frac{\sigma_u^2}{\sigma_e^2}$ between levels 2 and 1 (or equivalently the intraclass correlation $\rho = \frac{\tau}{\tau+1}$) are small.

- In the three-level RELR model we studied (which had 161 units at level 3, an average of 9.7 level–2 units per level–3 unit, and a total of 2,449 level–1 units),
  - quasi-likelihood methods performed badly in terms of bias of point estimates and coverage of interval estimates for random-effects variances; and
  - Bayesian methods with diffuse priors were well-calibrated in both point and interval estimation for all parameters of the model.

Our RELR results, narrowly construed, apply only to the 3–level model (2) with sample sizes like those in Example 2, but our quasi-likelihood conclusions are consistent with broad theoretical predictions

made by Engel (1998) and Lee and Nelder (2000), and our Bayesian calibration findings are in line with those in other RELR settings we have simulated.

These results bear comment both methodologically and in their practical implications for applied multilevel modeling in health care, education, and other fields. On the methodological side,

- Further study is needed to see if alternative diffuse priors can remedy the undercoverage of Bayesian intervals (and achieve approximate unbiasedness without the need to select a method of posterior summary depending on the problem) with small numbers of level–2 units in 2–level VC models; we intend to report on this elsewhere. Likelihood-based intervals of the kind we have studied here underperform in that situation for a reason that may be harder to remedy: the insistence on maximization (rather than integration) over the other parameters of a highly-skewed likelihood surface with its marginal maximum at $\sigma_u^2 = 0$, leading to zero point estimates in small samples with some frequency when the true value is well away from 0;

- The usual quasi-likelihood computer output in RELR models may not be trustworthy either for point estimation or uncertainty assessment, in the latter case because the estimated asymptotic standard errors can be systematically too small when the mean numbers of level–k units per level–$(k+1)$ unit (and/or the number of level–M units in an M–level model) are small for $k \geq 1$; and

- There is an expectation, expressed formally in the Bernstein-von Mises theorem (e.g., Freedman 1999; also see Samaniego and Reneau 1994 and Severini 1994), that likelihood and diffuse-prior Bayesian results will be close in large samples; and this will typically occur when parametric models with a modest number of parameters are fit to data not possessing a hierarchical structure. However,

  - what looks like a large sample in multilevel modeling may not be so large in reality, because the effective sample sizes for variances of random effects at levels greater than 1 in the hierarchy are mainly governed not by the total number of level–1 units (which will often be large) but by the numbers of units at the other levels, which are often much smaller; and
  - exact-likelihood methods for non-Gaussian multilevel models have until fairly recently been difficult to implement (because evaluation of the likelihood function involves integrating over the random effects), with the result that approximate methods such as quasi-likelihood techniques in RELR models have gained widespread use, and the Bernstein-von Mises theorem says nothing about agreement between Bayesian and *approximate* likelihood approaches unless the approximation is good.

On the practical side, as mentioned in Section 1.3, likelihood methods that may prove superior to quasi-likelihood have recently been under development, based on (a) Gaussian quadrature (e.g., Pinheiro and Bates 1995; see the `SAS` procedure `MIXED` for VC model fitting and the packages `EGRET`, `MIXOR`, and `LIMDEP`, the `SAS` procedure `NLMIXED`, and the `SAS` macro `NLINMIX` for examples of quadrature implementations in RELR models. Note however that, since these programs are only applicable to 2–level designs, they could not be used on the RELR models in this paper), (b) nonparametric maximum likelihood (Aitkin 1999a, supported by `GLIM4` macros written by the author), (c) Laplace approximations (Raudenbush et al. 2000, available in `HLM`), (d) hierarchical generalized linear models (Lee and Nelder 2000, as implemented in `GENSTAT` macros), and (e) profile likelihood (e.g., Longford 2000); and parametric bootstrapping of PQL estimates (Rodríguez and Goldman 2000, for instance

using `MLwiN`) may well lead to significant improvement in RELR models as well. Large-scale simulation results on the calibration of these approaches in small samples (particularly the quality of interval estimates) do not yet appear to be abundant.

One important likelihood-Bayesian comparison we have not addressed is computational speed, where ML/REML and MQL/PQL approaches have a distinct advantage (for example, PQL_2 fitting of model (2) to the Rodríguez-Goldman data set in Table 2 takes 4 seconds on a 2GHz PC versus 2.7 minutes using MCMC with 25,000 monitoring iterations). It is common practice in statistical modeling to examine a variety of models on the same data set before choosing a small number of models for reporting purposes (although this practice by itself encourages underpropagation of model uncertainty, e.g., Draper 1995b). The results of this paper suggest (but do not insist upon) a hybrid modeling strategy, in which likelihood-based methods like those studied here are used in the model exploration phase (but with awareness of the possible understatement of random-effects variances and their uncertainty bands) and Bayesian diffuse-prior methods are used for the reporting of final inferential results. Other analytic strategies based on less approximate likelihood methods are also possible but would benefit from further study of the type summarized here.

Acknowledgements

We are grateful to the UK EPSRC and ESRC for financial support for William J. Browne, and to Murray Aitkin, David Clayton, Constantine Gatsonis, Andrew Gelman, Wally Gilks, Harvey Goldstein, Sander Greenland, Jim Hodges, Youngjo Lee, Dennis Lindley, Nick Longford, John Nelder, Jon Rasbash, Steve Raudenbush, Tony Robinson, Michael Seltzer, and David Spiegelhalter for references and comments on this and related papers and presentations. Membership on this list does not imply agreement with the ideas expressed here, nor are any of these people responsible for any errors that may be present.

Appendix: Computing details

In the VC simulations, to decide how long to monitor the Gibbs-sampling output we estimated time per iteration and calculated Raftery-Lewis (1992) diagnostics as a function of the total number of pupils N. This revealed that the smaller designs in Table 3 needed longer monitoring runs to satisfy Raftery-Lewis default accuracy constraints but took less time per iteration, leading to the following monitoring run lengths M: 50,000 in studies 1 and 2, 30,000 in 3 and 4, 20,000 in 5 and 6, and 10,000 in studies 7 and 8. The full set of VC simulations took 1.8 GHz-months of CPU time on 3 Sun SPARCstations and a Pentium-based PC.

The data sets in Examples 1 and 2, and `WinBUGS` and `MLwiN` programs to fit models (1) and (2) to those examples, are available on the web at `www.ams.ucsc.edu/~draper`.

References

Aitkin M (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, **6**, 251–262.

Aitkin M (1999a). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, **55**, 117–128.

Aitkin M (1999b). Meta-analysis by random-effects modelling in generalized linear models. *Statistics in Medicine*, **18**, 2343–2351.

Bernardo JM, Smith AFM (1994). *Bayesian Theory*. New York: Wiley.

Besag J, Green P, Higdon D, Mengersen K (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science*, **10**, 3–41.

Box GEP, Tiao GC (1973). *Bayesian Inference in Statistical Analysis*. New York: Wiley.

Breslow NE, Clayton DG (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Brown KG, Burgess MA (1984). On maximum likelihood and restricted maximum likelihood approaches to estimation of variance components. *Journal of Statistical Computation and Simulation*, **19**, 59–77.

Browne WJ (1998). *Applying MCMC Methods to Multilevel Models*. PhD dissertation, Department of Mathematical Sciences, University of Bath, UK.

Browne WJ, Draper D (2000). Implementation and performance issues in the Bayesian fitting of multilevel models. *Computational Statistics*, **15**, 391–420.

Browne WJ, Draper D, Goldstein H, Rasbash J (2002). Bayesian and likelihood methods for fitting multilevel models with complex level–1 variation. *Computational Statistics and Data Analysis*, forthcoming.

Bryk AS, Raudenbush SW (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. London: Sage.

Bryk AS, Raudenbush SW, Seltzer M, Congdon R (1988). *An Introduction to HLM: Computer Program and User's Guide (Second Edition)*. Chicago: University of Chicago Department of Education.

Carlin B (1992). Discussion of "Hierarchical models for combining information and for meta-analysis," by Morris CN, Normand SL. In *Bayesian Statistics 4*, Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds.), 336–338. Oxford: Clarendon Press.

Carlin B, Louis TA (2001). *Bayes and Empirical Bayes Methods for Data Analysis*, second edition. London: Chapman & Hall.

Chaloner K (1987). A Bayesian approach to the estimation of variance components in the unbalanced one-way random-effects model. *Technometrics*, **29**, 323–337.

Cochran WG (1977). *Sampling Techniques (Third Edition)*. New York: Wiley.

Corbeil RR, Searle SR (1976). Restricted maximum likelihood (REML) estimation of variance components in mixed models. *Technometrics*, **18**, 31–38.

Daniels MJ, Gatsonis C (1999). Hierarchical generalized linear models in the analysis of variations in health care utilization. *Journal of the American Statistical Association*, **94**, 29–42.

Dawid AP (1985). Calibration-based empirical probability. *Annals of Statistics*, **13**, 1251–1274.

Draper D (1995a). Inference and hierarchical modeling in the social sciences (with discussion). *Journal of Educational and Behavioral Statistics*, **20**, 115–147, 233–239.

Draper D (1995b). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society Series B*, **57**, 45–97.

Draper D (2004). *Bayesian Inference and Prediction*. New York: Springer-Verlag, forthcoming.

DuMouchel W (1990). Bayesian meta-analysis. In *Statistical methodology in the pharmaceutical sciences*, Berry D (ed.), pp. 509–529. New York: Marcel Dekker.

DuMouchel W, Waternaux C (1992). Discussion of "Hierarchical models for combining information and for meta-analysis," by Morris CN, Normand SL. In *Bayesian Statistics 4*, Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds.), 338–341. Oxford: Clarendon Press.

Engel B (1998). A simple illustration of the failure of PQL, IRREML and APHL as approximate ML methods for mixed models for binary data. *Biometrical Journal*, **40**, 141–154.

Freedman D (1999). On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Annals of Statistics*, **27**, 1119–1140.

Gelfand AE, Sahu SK, Carlin BP (1995). Efficient parameterizations for generalized linear mixed models (with discussion). In *Bayesian Statistics 5*, Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds.), 165–180. Oxford: Clarendon Press.

Gelfand A, Smith AFM (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Gelman A (2004). Prior distributions for variance parameters in hierarchical models. Technical report, Department of Statistics, Columbia University.

Gelman A, Carlin JB, Stern HS, Rubin DB (2003). *Bayesian Data Analysis*, second edition. London: Chapman & Hall.

Gelman A, Roberts GO, Gilks WR (1995). Efficient Metropolis jumping rules. In *Bayesian Statistics 5*, Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds.), 599–607. Oxford: Clarendon Press.

Gelman A, Rubin DB (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**, 457–511.

Gilks WR, Richardson S, Spiegelhalter DJ (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

Gilks WR, Roberts GO, Sahu SK (1998). Adaptive Markov chain Monte Carlo sampling through regeneration. *Journal of the American Statistical Association*, **93**, 1045–1054.

Gilks WR, Wild P (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**, 337–348.

Goldstein H (1986). Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika*, **73**, 43–56.

Goldstein H (1989). Restricted unbiased iterative generalised least squares estimation. *Biometrika*, **76**, 622–623.

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

Goldstein H, Rasbash J (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, **159**, 505–513.

Goldstein H, Rasbash J, Yang M, Woodhouse G, Pan H, Nutall D, Thomas S (1993). A multilevel analysis of school examination results. *Oxford Review of Education*, **19**, 425–433.

Goldstein H, Spiegelhalter DJ (1996). League tables and their limitations: statistical issues in comparisons of institutional performance (with discussion). *Journal of the Royal Statistical Society, Series A*, **159**, 385–444.

Harville DA, Zimmermann AG (1996). The posterior distribution of the fixed and random effects in a mixed-effects linear model. *Journal of Statistical Computation and Simulation*, **54**, 211–229.

Henderson CR (1950). Estimation of genetic parameters (abstract). *Annals of Mathematical Statistics*, **21**, 309–310.

Huber DA, White TL, Hodge GR (1994). Variance-component estimation techniques compared for two mating designs with forest genetic architecture through computer simulation. *Theoretical and Applied Genetics*, **88**, 236–242.

Huber PJ (1967). The behavior of maximum likelihood estimates under non-standard conditions. In *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability*. Berkeley CA: University of California Press, **1**, 221–233.

Hulting FL, Harville DA (1991). Some Bayesian and non-Bayesian procedures for the analysis of comparative experiments and for small-area estimation: computational aspects, frequentist properties, and relationships. *Journal of the American Statistical Association*, **86**, 557–568.

Johnson NL, Kotz S, Balakrishnan N (1994). *Continuous Univariate Distributions*, Volume 1 (second edition). New York: Wiley.

Kahn MJ, Raftery AE (1996). Discharge rates of Medicare stroke patients to skilled nursing facilities: Bayesian logistic regression with unobserved heterogeneity. *Journal of the American Statistical Association*, **91**, 29–41.

Kass RE, Steffey D (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association*, **84**, 717–726.

Kass RE, Wasserman L (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**, 1343–1370.

Klotz JH, Milton RC, Zacks S (1969). Mean square efficiency of estimators of variance components. *Journal of the American Statistical Association*, **64**, 1383–1394.

Lee Y, Nelder JA (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B*, **58**, 619–678.

Lee Y, Nelder JA (2001). Hierarchical generalized linear models: a synthesis of generalized linear models, random-effect models, and structured dispersion. *Biometrika*, **88**, 987–1006.

Lesaffre E, Spiessens B (2001). On the effect of the number of quadrature points in a logistic random-effects model: an example. *Journal of the Royal Statistical Society, Series C*, **50**, 325–335.

Liu JN, Hodges JS (2003). Posterior bimodality in the balanced one-way random-effects model. *Journal of the Royal Statistical Society, Series B*, **65**, 247–255.

Longford NT (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, **74**, 817–827

Longford NT (1997). Comment on "Improved approximations for multilevel models with binary responses," by Goldstein H, Rasbash J. *Journal of the Royal Statistical Society, Series A*, **160**, 593.

Longford NT (2000). On estimating standard errors in multilevel analysis. *The Statistician*, **49**, 389–398.

Mortimore P, Sammons P, Stoll L, Lewis D, Ecob R (1988). *School Matters*. Wells: Open Books.

Müller P (1993). A generic approach to posterior integration and Gibbs sampling. Technical Report, Institute of Statistics and Decision Sciences, Duke University.

Natarajan R, Kass RE (2000). Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, **95**, 227–237.

Pebley AR, Goldman N (1992). Family, community, ethnic identity, and the use of formal health care services in Guatemala. *Working Paper 92–12*, Princeton NJ: Office of Population Research.

Pinheiro JC, Bates DM (1995). Approximations to the log-likelihood function in the non-linear mixed-effects model. *Journal of Computational and Graphical Statistics, 4*, 12–35.

Portnoy S (1971). Formal Bayes estimation with application to a random-effects model. *Annals of Mathematical Statistics*, **42**, 1379–1388.

Raftery AL, Lewis S (1992). How many iterations in the Gibbs sampler? In *Bayesian Statistics 4*, Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds.), 763–774. Oxford: Clarendon Press.

Rasbash J, Browne WJ, Goldstein H, Yang M, Plewis I, Healy M, Woodhouse G, Draper D, Langford I, Lewis T (2000). *A User's Guide to* `MLwiN`, Version 2.1, London: Institute of Education, University of London.

Raudenbush SW (1994). Equivalence of Fisher scoring to iterative generalized least squares in the normal case, with application to hierarchical linear models. Technical Report, College of Education, Michigan State University.

Raudenbush SW, Yang M-L, Yosef M (2000). Maximum likelihood for hierarchical models via high-order multivariate Laplace approximations. *Journal of Computational and Graphical Statistics*, **9**, 141–157.

Roberts GO, Sahu SK (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society, Series B*, **59**, 291–318.

Robinson GK (1991). That BLUP is a good thing: the estimation of random effects (with discussion). *Statistical Science*, **6**, 15–51.

Rodríguez G, Goldman N (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, **158**, 73–89.

Rodríguez G, Goldman N (2001). Improved estimation procedures for multilevel models with binary responses: a case study. *Journal of the Royal Statistical Society, Series A*, **164**, 339–355.

Rubin DB (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, **12**, 1151–1172.

Samaniego FJ, Reneau DM (1994). Toward a reconciliation of the Bayesian and frequentist approaches to point estimation. *Journal of the American Statistical Association*, **89**, 947–957.

Scheffé H (1959). *The Analysis of Variance*. New York: Wiley.

Searle SR, Casella G, McCulloch CE (1992). *Variance Components*. New York: Wiley.

Seltzer MH (1993). Sensitivity analysis for fixed effects in the hierarchical model: a Gibbs sampling approach. *Journal of Educational Statistics*, **18**, 207–235.

Seltzer MH, Wong WH, Bryk AS (1996). Bayesian analysis in applications of hierarchical models: issues and methods. *Journal of Educational and Behavioral Statistics*, **21**, 131–167.

Severini TA (1991). On the relationship between Bayesian and non-Bayesian interval estimates. *Journal of the Royal Statistical Society, Series B*, **53**, 611–618.

Singh AC, Stukel DM, Pfefferman D (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society, Series B*, **60**, 377-396.

Spiegelhalter DJ, Thomas A, Best NG, Gilks WR (1997). *BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.60*. Cambridge: Medical Research Council Biostatistics Unit.

StataCorp (2004). *Stata Statistical Software: Release 8.0*. College Station TX: Stata Corporation.

Swallow WH, Monahan JF (1984). Monte Carlo comparison of ANOVA, MIVQUE, REML, and ML estimators of variance components. *Technometrics*, **26**, 47–57.

White H (1980). A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica*, **48**, 817–830.

Wilson EB, Hilferty MM (1931). The distribution of chi-square. *Proceedings of the US National Academy of Sciences*, **17**, 684.

Woodhouse G, Rasbash J, Goldstein H, Yang M, Howarth J, Plewis I (1995). *A Guide to* `MLn` *for New Users*. London: Institute of Education, University of London.

Zeger SL, Karim MR (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79–86.