Maximum Likelihood for Generalized Linear Models with Nested Random Effects via High-Order, Multivariate Laplace Approximation
Author(s): Stephen W. Raudenbush, Meng-Li Yang and Matheos Yosef
Source: *Journal of Computational and Graphical Statistics*, Vol. 9, No. 1 (Mar., 2000), pp. 141-157

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of America
Stable URL: http://www.jstor.org/stable/1390617
Accessed: 15-12-2016 16:55 UTC

Maximum Likelihood for Generalized Linear Models With Nested Random Effects via High-Order, Multivariate Laplace Approximation

Stephen W. RAUDENBUSH, Meng-Li YANG, and Matheos YOSEF

Nested random effects models are often used to represent similar processes occurring in each of many clusters. Suppose that, given cluster-specific random effects b, the data y are distributed according to $f(y|b, \theta)$, while b follows a density $p(b|\theta)$. Likelihood inference requires maximization of $\int f(y|b, \theta)p(b|\theta)db$ with respect to θ. Evaluation of this integral often proves difficult, making likelihood inference difficult to obtain. We propose a multivariate Taylor series approximation of the log of the integrand that can be made as accurate as desired if the integrand and all its partial derivatives with respect to b are continuous in the neighborhood of the posterior mode of $b|\theta, y$. We then apply a Laplace approximation to the integral and maximize the approximate integrated likelihood via Fisher scoring. We develop computational formulas that implement this approach for two-level generalized linear models with canonical link and multivariate normal random effects. A comparison with approximations based on penalized quasi-likelihood, Gauss–Hermite quadrature, and adaptive Gauss–Hermite quadrature reveals that, for the hierarchical logistic regression model under the simulated conditions, the sixth-order Laplace approach is remarkably accurate and computationally fast.

Key Words: Hierarchical models; Mixed models; Numerical integration.

1. INTRODUCTION

Nested random effects models are often useful to represent an estimation process that is replicated over many clusters. Researchers have used these models, for example, to synthesize results from experiments replicated in each of many independent studies (Berkey, Hoaglin, Mosteller, and Colditz 1995; DerSimonian and Laird 1986; Raudenbush and Bryk 1985; Rubin 1981); to combine repeated measures data from each of many subjects (Laird and Ware 1982; Strenio, Weisberg, and Bryk 1983); and to compare regression coefficients defined on each of many schools (Aitkin and Longford 1986; deLeeuw and Kreft 1986; Goldstein 1986; Raudenbush and Bryk 1986). Many applications and statistical issues are reviewed in recent books (Bock 1989; Bryk and Raudenbush 1992;

Stephen W. Raudenbush is Professor, School of Education, Univeristy of Michigan, 610 East University Avenue, Ann Arbor, MI 48109 (E-mail: RAUDEN@UMICH.EDU). Meng-Li Yang is Assistant Professor, Nan-Tai Institute of Technology, Taiwan. Matheos Yosef is Doctoral Student, Michigan State University, 125 West Hoover, #1B, Ann Arbor, MI 48103.

Journal of Computational and Graphical Statistics, Volume 9, Number 1, Pages 141–157

Diggle, Liang, and Zeger 1994; Goldstein 1995; and Longford 1993).

Following Lindley and Smith's (1972) classic exposition, we view the nested random effects model as a special case of a hierarchical model: the data y depend on a parameter vector b through $f(y|b,\theta)$, while b, in turn, depends on the parameter vector θ through a prior or mixing distribution $p(b|\theta)$. ML estimation requires maximization of the integral

$$L(\theta;y) = \int f(y|b,\theta)p(b|\theta)db \tag{1.1}$$

with respect to θ. In some important cases, the integral (1.1) can be evaluated analytically and maximization proceeds by standard methods such as the EM algorithm (Dempster, Laird, and Rubin 1977; Dempster, Rubin, and Tsutakawa 1981), Fisher scoring (Goldstein 1986; Longford 1987), or Newton–Raphson (Lindstrom and Bates 1988). In general, however, integration is intractable. The purpose of this article is to propose and illustrate a solution to this problem that applies to generalized linear models with nested random effects. Our strategy is to approximate the log of the integrand by its fully multivariate Taylor expansion of sufficiently high order to ensure accuracy, and then to integrate using Laplace's method. We then maximize the (approximate) integrated likelihood using Fisher scoring.

Section 2 reviews past work on maximum likelihood for the generalized linear model with nested random effects. Section 3 develops the high-order Laplace approximation, and Section 4 derives computational formulas. Section 5 illustrates application in the case of a logistic regression with random effects. Section 6 presents simulations that compare the approach to alternative approximations based on penalized quasi-likelihood, Gaussian quadrature, and adaptive Gaussian quadrature. We conclude by sketching directions for future research.

2. HIERARCHICAL GENERALIZED LINEAR MODELS WITH NORMAL RANDOM EFFECTS

2.1 The Model

Consider the vector of responses y_j from cluster $j, j = 1, \ldots, J$. Conditional on cluster-specific random effects b_j, these data are distributed according to $f(y_j|b_j,\beta)$, a member of the exponential family (see Cox and Hinkley 1974, p. 27) with conditional mean $E(y_j|b_j,\beta) = \mu_j$, and canonical link function $g(\mu_j) = \eta_j = X_j\beta + Z_jb_j$. The random effect b_j has a density $p(b_j|\Psi)$. We thus have a generalized linear model with random effects (McCullagh and Nelder 1989). This is a special case of (1.1) with $y = (y_1, \ldots, y_J)^T$, $b = (b_1, \ldots, b_J)^T$, and $\theta = (\beta, \Psi)$. If $p(b|\Psi)$ is conjugate to $f(y|b,\beta)$, the integral (1.1) can, in some important cases, be evaluated analytically (Lee and Nelder 1996). In particular, when both f and p are normal densities and η is the identity link, we have the Lindley–Smith (1972) model.

However, the utility of the conjugate prior is limited in cases other than the normal. For example, the conjugate prior for the binomial likelihood is the beta. The beta will be tractable only if the elements of b are independent. Yet many applications require flexibility in specifying b, allowing for the possibility that the elements of b covary (Lee and Nelder 1996). Only the normal and t priors (see Seltzer 1993 and Thum 1997 for applications of the t prior) appear practical as parametric representations for multiple dependent random effects in hierarchical models.

2.2 Approximations to Maximum Likelihood

If the random effects are assumed multivariate normal, integral (1.1) will be tractable only when the likelihood is also normal and the canonical link is the identity link $\eta = \mu$. (See Pinheiro and Bates (1995) for a normal case with nonlinear link.) For non-normal likelihoods or nonidentity links, much of the research on estimation theory has involved strategies for coping with intractable integrals of this type. Three strategies are prominent: (1) quasi-likelihood inference; (2) Gauss-Hermite approximations; and (3) Monte Carlo integration.

Stiratelli, Laird, and Ware (1984) estimated the parameters of a logistic regression model with nested, normally distributed random effects by approximating the joint posterior density of $b, \beta, \Psi|y$ with a multivariate normal density having the same mode and curvature at the mode as the true posterior. They posed a uniform prior for $\theta = (\beta, \Psi)$, in which case the joint posterior of $b, \beta, \Psi|y$ is proportional to the integrand of (1.1). Wong and Mason (1985) used essentially the same approach. Direct maximization of this approximate joint posterior avoids the difficult integration of (1.1). Lee and Nelder (1996) referred the integrand of (1.1) as the "h-likelihood" and discussed properties of estimates of b, β based on its direct maximization.

Several authors have extended the approach in different ways. For example, see Belin et al. (1993); Breslow and Clayton (1993); Gilks (1987); McGilchrist (1994); Schall (1991); Wolfinger (1993); Goldstein (1991); Longford (1993). Following Breslow and Clayton (1993), we term this approach penalized quasi-likelihood (PQL). However, Breslow and Lin (1995) showed that for logistic regression models with nested random effects, PQL estimates of the normal covariance matrix and, hence, of β, are biased and inconsistent (see also Goldstein and Rasbash 1996; and Rodriguez and Goldman 1995). Bias is most serious when the random effects have large variance and the binomial denominator is small.

An alternative approach that will produce consistent and asymptotically unbiased estimates is to approximate ML estimation by approximating integral (1.1) as closely as desired, and then to maximize the approximate integral. Anderson and Aitkin (1985) applied Gauss–Hermite quadrature to evaluate the likelihood and maximized that likelihood in the case of a logistic regression model with one random effect per cluster of observations, y_j. Hedeker and Gibbons (1994, 1996) applied Gauss–Hermite quadrature to evaluate the required integral for ordinal probit and logistic models with multivariate normal priors. See also Tsutakawa (1985) for the case of a Poisson-normal mixture.

Pinheiro and Bates (1995) used adaptive Gauss–Hermite quadrature to approximate ML estimation for a nested random effects model with normal data and nonlinear link. Using the adaptive approach, the variable of integration (the random effect) is centered around its approximate posterior mode rather than around its mean of zero. In principle, this approach will produce more accurate results than nonadaptive quadrature, especially when the dispersion of the random effects is large. This approach has been implemented in an experimental version of SAS (Wolfinger 1999). Wei and Tanner (1990), Karim (1991), and Chan (1994) used Monte Carlo integration to evaluate integrals such as (1.1). Numerical integration via Gaussian quadrature becomes progressively difficult as the number of correlated random effects per cluster increases, while Monte Carlo integration is computationally intensive and provides stochastic rather than numerical convergence. Stochastic convergence can be difficult to assess.

An alternative approach to the approximation of integrals such as (1.1) uses Laplace's method. Breslow and Lin (1995) used a fourth-order Laplace approximation to correct the bias associated with PQL in the case of nested random effects models with a single random effect per cluster. Lin and Breslow (1996) extended this bias-correction strategy to the case of multiple independent random effects per cluster. Our aim is to extend this logic to higher order approximations and to multiple dependent random effects per cluster. Rather than using the method to correct bias, we view the approximated integral as the likelihood and maximize it to make inferences about θ. The approach has the following advantages: (1) integration per cluster is fully multivariate with arbitrary dimension; (2) the approximation is accurate to any degree required; (3) convergence is numerical rather than stochastic; and (4) computations are remarkably fast.

3. HIGH-ORDER MULTIVARIATE LAPLACE INTEGRATION

The approximation of integrals via Laplace's method has been widely used to find posterior distributions (Lindley, 1980; Kass, Tierney, and Kadane 1990) and to approximate likelihoods (Solomon and Cox 1992; Liu and Pierce 1993; Breslow and Lin 1995; Shun and McCullagh 1995; Lin and Breslow 1996; Shun 1997). In the standard application, the natural log of the integrand is expanded in a second-order Taylor series; higher order terms decrease with the sample size, making the approximation accurate in large samples. Shun and McCullagh (1995) and Shun (1997) noted, however, that in many interesting cases, the dimension of the integral increases as a function of the sample size. In these cases, the standard Laplace approximation is not valid because the error of the second-order approximation does not diminish with the sample size. They illustrate this point in the case of a cross-classified random effects model with a scalar random effect for each row and column of a two-dimensional array. In this case, the random effects must be integrated out of the joint likelihood of the data and the random effects to obtain the marginal likelihood of the data alone. The dimension of the integral has an order of the square root of the sample size.

The nested random effects model of interest in this article is simpler than the crossed case in that the required integral is the product of independent integrals, one for each cluster. However, we seek an approach that allows the random effects from each cluster to be correlated and to have arbirtrary dimension, q. Across all clusters, the integral thus has a dimension of Jq, and the ratio of sample size to the overall dimension of the integral is thus n/q, where n is the average sample size per cluster. The omitted terms in the sixth-order approximation of the likelihood are $O(n^{-2})$ for fixed q. However, the general framework we propose allows addition of yet higher order terms.

We begin with a compact and general representation of Taylor's theorem using the matrix notation and algebra of Magnus and Neudecker (1988).

Definition 1. *Infinite multivariate Taylor series. Let $h(b)$, a scalar function of a vector b, and all its partial derivatives be continuous in a neighborhood* **N** *of $\tilde{b}$. Then for b in* **N**,

$$\begin{aligned} h(b) &= h(\tilde{b}) + h^{(1)}(\tilde{b})(b-\tilde{b}) + \frac{1}{2}(b-\tilde{b})^T h^{(2)}(\tilde{b})(b-\tilde{b}) \\ &\quad + \frac{1}{3!}\left[(b-\tilde{b})^T \otimes (b-\tilde{b})^T\right] h^{(3)}(\tilde{b})(b-\tilde{b}) \\ &\quad + \cdots + \frac{1}{K!}\left[\overset{K-1}{\otimes} (b-\tilde{b})^T\right] h^{(K)}(\tilde{b})(b-\tilde{b}) + \cdots \\ &= h(\tilde{b}) + h^{(1)}(\tilde{b})(b-\tilde{b}) + \sum_{k=2}^{\infty} \frac{1}{k!}\left[\overset{k-1}{\otimes} (b-\tilde{b})^T\right] h^{(k)}(\tilde{b})(b-\tilde{b}), \end{aligned} \tag{3.1}$$

where

$$h^{(k)}(\tilde{b}) = \left.\frac{\partial\, vech^{(k-1)}(b)}{\partial b^T}\right|_{b=\tilde{b}},$$

and $\overset{k}{\otimes} u = u \otimes u \otimes \cdots \otimes u$, there being k u's in the Kronecker product. For example, $\overset{3}{\otimes} u = u \otimes u \otimes u$.

Next, we apply Laplace's method to integrate $\exp\{h(b)\}$:

Theorem 1. *Suppose $h(b)$ has a maximum at $b = \tilde{b}$. Then, for $b \in R^q$,*

$$\begin{aligned} \int_{R^q} \exp(h(b))db &= (2\pi)^{q/2}|V|^{1/2} \exp\left[h(\tilde{b})\right] \\ &\quad \times E\left\{\exp\left[\sum_{k=3}^{\infty} \frac{1}{k!}\left(\left[\overset{k-1}{\otimes} (b-\tilde{b})^T\right] h^{(k)}(\tilde{b})(b-\tilde{b})\right)\right]\right\}. \end{aligned} \tag{3.2}$$

Here $E(\cdot)$ is an expectation operator, taken over a multivariate normal distribution with mean vector 0, and covariance matrix $V = [-h^{(2)}(\tilde{b})]^{-1}$.

Proof: Because $\tilde{b}$ maximizes $h(b)$, $h^{(1)}(\tilde{b})(b-\tilde{b})$ disappears. We then view this integrand as the product of the constant $\exp h(\tilde{b})$, the normal kernel

$$\exp\left\{-\frac{1}{2}(b-\tilde{b})^T V^{-1}(b-\tilde{b})\right\},$$

and the exponential of the sum of the remaining terms in the series. The integral thus has the form of an expectation taken over a multivariate normal density $N(0, V)$. This completes the proof. □

The next problem is to evaluate the expectation in Equation (3.2); that is, $E\exp(R)$, where $R = \sum_{k=3}^{\infty} T_k$, with $T_k = \frac{1}{k!}[\overset{k-1}{\otimes} (b-\tilde{b})^T]h^{(k)}(\tilde{b})(b-\tilde{b})$. Setting $\exp(R) = 1 + R + R^2/2 + \cdots + R^k/k! + \cdots$, we need to evaluate $E(R), E(R^2) \ldots$. In evaluating $E(R)$ and $E(R^2)$, we find the following theorem useful.

Theorem 2.

$$
\begin{aligned}
E\left[T_k\right] &= 0 \;\; \textit{for odd} \;\;\; k, k > 2; \\
E\left[T_k\right] &= \frac{(k-1)(k-3)\ldots 3}{k!} vec^T \left(\overset{k/2}{\otimes} V \right) vec[h^{(k)}(\tilde{b})] \\
& \quad \textit{for even} \;\;\; k, k > 2. \qquad (3.3)
\end{aligned}
$$

$$
\begin{aligned}
E\left[T_k T_l\right] &= 0 \;\; \textit{for odd} \;\;\; (k+l), k \;\;\; \textit{and} \;\;\; l > 2; \\
E\left[T_k T_l\right] &= \frac{(k+l-1)(k+l-3)\ldots 3}{k!l!} vec^T \left(\overset{(k+l)/2}{\otimes} V \right) vec[h^{(k)}(\tilde{b}) \otimes h^{(l)}(\tilde{b})] \\
& \quad \textit{for even} \;\;\; (k+l), k, \;\;\; \textit{and} \;\;\; l > 2. \qquad (3.4)
\end{aligned}
$$

This theorem is proved in the Appendix.

4. COMPUTATIONAL FORMULAS FOR TWO-LEVEL MODELS

The response vector y_j is composed of items y_{ij} for observation i in cluster j, $i = 1, \ldots, n_j$. Similarly, $E(y_{ij}|b_j) = \mu_{ij}$ with link function $g(\mu_{ij}) = \eta_{ij} = X_{ij}^T \beta + Z_{ij}^T b_j$, where X_{ij} and Z_{ij} are known vectors of explanatory variables, β is a $p \times 1$ vector of fixed effects, b_j is a $q \times 1$ vector of random effects distributed as $N(0, D)$, and $D = D(\Psi)$ is a function of $q(q+1)/2$ unique covariance parameters in Ψ.

Suppose now that $f(y_j|b_j, \beta)$ belongs to the exponential family and that η_{ij} is the canonical link (McCullagh and Nelder 1989). Then we have the conditional likelihood

$$f(y_j|b_j, \beta) = \exp\{l_j\},$$

with

$$l_j = \sum_{i=1}^{n_j} \left\{ \left[y_{ij}\eta_{ij} - \delta(\mu_{ij}) \right] / \alpha(\phi) + \gamma(y_{ij}, \phi) \right\},$$

where α, δ, and γ are arbitrary functions of their arguments.

To get the marginal likelihood, we wish to integrate out b_j from the joint density of y_j and b_j, $j = 1, 2, \ldots, J$:

$$
\begin{aligned}
L &= \int \prod_j f(y_j|b_j, \beta) p(b_j|D) db_j \\
&= \prod_j \frac{1}{(2\pi)^{q/2}} |D|^{-1/2} \int \exp\left(l_j - \frac{1}{2} b_j^T D^{-1} b_j \right) db_j. \qquad (4.1)
\end{aligned}
$$

To apply Laplace's method, we regard $l_j - \frac{1}{2} b_j^T D^{-1} b_j$ as $h(b)$ in Theorem 1 for $b = (b_1, \ldots, b_J)^T$, and, given a maximizer $\tilde{b}(\beta, D)$ of $h(b)$, find derivatives up to the sixth order:

1. $h_j(\tilde{b}_j) = \tilde{l}_j - \frac{1}{2} \tilde{b}_j^T D^{-1} \tilde{b}_j$, where $\tilde{l}_j$ is l_j evaluated at $\tilde{b}_j$.

2. $h_j^{(1)}(\tilde{b}_j) = \tilde{l}_j^{(1)} - \tilde{b}_j^T D^{-1}$, where

$$\tilde{l}_j^{(1)} = \left.\frac{\partial l_j}{\partial b^T}\right|_{b=\tilde{b}} = (y_j - \tilde{\mu}_j)^T Z_j/\alpha(\phi) = (y_j^* - \tilde{\eta}_j)^T \tilde{W}_j Z_j/\alpha(\phi).$$

Here $y_j^* = \tilde{W}_j^{-1}(y_j - \tilde{\mu}_j) + \tilde{\eta}_j$, the linearized dependent variable (McCullagh and Nelder 1989); $\tilde{W}_j$ is diag$[\tilde{w}_{ij}]$, with $\tilde{w}_{ij} = d\tilde{\mu}_{ij}/d\tilde{\eta}_{ij}$, the derivative of μ_{ij} with respect to η_{ij}, evaluated at $\tilde{b}_{ij}$.

3. $h_j^{(2)}(\tilde{b}_j) = \tilde{l}_j^{(2)} - D^{-1}$, with $\tilde{l}_j^{(2)} = -Z_j^T \tilde{W}_j Z_j/\alpha(\phi_j)$, the second derivative of l_j evaluated at $\tilde{b}_j$.

4. For $k \geq 3$, $h_j^{(k)}(\tilde{b}_j) = \tilde{l}_j^{(k)} = -\sum_{i=1}^{n_j} m_{ij}^{(k)}(\overset{k}{\otimes} Z_{ij}^T)/\alpha(\phi_j)$, where $\tilde{m}_{ij}^{(k)}$ is the $(k-1)$th derivative of μ_{ij} with respect to η_{ij}, evaluated at $\tilde{b}_j$. In the case of binary y_{ij} with logit link, $w_{ij} = \mu_{ij}(1 - \mu_{ij})$ and the second to the fifth derivatives are

$$\begin{array}{rclrcl} \tilde{m}_{ij}^{(3)} & = & \tilde{w}_{ij}(1 - 2\tilde{\mu}_{ij}) & \tilde{m}_{ij}^{(4)} & = & \tilde{w}_{ij}(1 - 6\tilde{w}_{ij}), \\ \tilde{m}_{ij}^{(5)} & = & \tilde{m}_{ij}^{(3)}(1 - 12\tilde{w}_{ij}) & \tilde{m}_{ij}^{(6)} & = & \tilde{m}_{ij}^{(4)}(1 - 12\tilde{w}_{ij}) - 12\tilde{m}_{ij}^{(3)2}. \end{array} \quad (4.2)$$

In the case of count data ($y_{ij} \in \{0, 1, \ldots\}$) drawn from a conditional Poisson distribution with log link, $w_{ij} = m_{ij}^{(k)} = \mu_{ij}$ for all k. When y_{ij} is conditionally gamma distributed with reciprocal link, $w_{ij} = \mu_{ij}^2$, $m_{ij}^{(k)} = (k-1)!\mu_{ij}^k$. In the normal case, $w_{ij} = 1$ and $m_{ij}^{(k)} = 0$ for $k > 2$. The constants $\alpha(\phi)$, as is well known, are unity for the binomial and Poisson, var$(y_{ij}|b_j) = \sigma^2$ for the normal, and $-1/v$ for the gamma, where var$(y_{ij}|b_j) = \mu_{ij}^2/v$. All derivatives of l_j are evaluated at $\tilde{b}_j$. To avoid cumbersome notation, however, we henceforth drop the "$\sim$" except for the l_j's and $\tilde{b}_j$ itself.

Applying Laplace's method to Equation (4.1), we have

$$L = \prod_{j=1}^{J} \left\{ (2\pi)^{(-q/2)} |D|^{(-1/2)} \exp\left(\tilde{l}_j - \frac{1}{2}\tilde{b}_j^T D^{-1} \tilde{b}_j\right) \right.$$
$$\left. \times \int \exp\left\{-\frac{1}{2}\left(b_j - \tilde{b}_j\right)^T V_j^{-1}\left(b_j - \tilde{b}_j\right)\right\} \exp(R_j) db_j \right\}, \quad (4.3)$$

where the correction term $R_j = \sum_{k=3}^{\infty} T_{kj}$, with $T_{kj} = \frac{1}{k!}[\overset{k-1}{\otimes}(b_j - \tilde{b}_j)^T] h_j^{(k)}(\tilde{b}_j)(b_j - \tilde{b}_j)$. Note that $h_j^{(1)}(\tilde{b}_j)$ vanishes from Equation (4.2) because we choose $\tilde{b}_j$ to be the maximizer of $h_j(b_j)$; that is, $\tilde{b}_j = D\tilde{l}_j^{(1)} = (Z_j^T W_j Z_j + D^{-1})^{-1} Z_j^T W_j (y_j^* - X_j\beta)$. We obtain $\tilde{b}_j$ through iteratively solving this equation and substituting the new $\tilde{b}_j$ into y_j^* and W_j.

As mentioned earlier, $\exp(R_j) = 1 + R_j + (1/2)R_j^2 + \cdots$. In the following illustrative examples and data simulations we use the approximation with $E\exp(R_j) \approx 1 + E(T_{4j}) + E(T_{6j}) + (1/2)E(T_{3j}^2)$ and find it highly accurate, although the method allows us to go as far as we wish. We also note that Lindley (1980) has approximated posterior derivatives up to the sixth order, using the same terms as here, and that Liu and Pierce (1993) had approximations up to the fourth order for univariate integrals. The full expansion of the correction term involves $T_4, T_6, T_3^2/2, T_8, T_3T_5/2, T_4^2/2, \ldots$. The magnitude of the higher order terms diminishes for two main reasons. First, in the binomial and Poisson cases, the factorial denominator rapidly increases. Second, the terms diminish as a function of the cluster size, n. Terms T_4 and $T_3^2/2$ are $O(n^{-1})$ while $T_6, T_3T_5/2, T_4^2/2$ are all $O(n^{-2})$ and T_8 is $O(n^{-3})$. Higher order terms are $O(n^{-3})$ or smaller and have very large factorial denominators. The implication is that, for many applications, adding T_4 without adding $T_3^2/2$ would do little to improve the approximation. For some applications, it may be useful to add $T_3T_5/2, T_4^2/2$, though in our experience with the logistic case and cluster sizes of interest, these terms have been negligible. Using the sixth-order approximation and applying Equations (3.1) and (3.2), we approximate (4.1), up to a constant, as

$$L \approx \prod_{j=1}^{J} |D|^{-1/2} |Z_j^T W_j Z_j + D^{-1}|^{-1/2} \exp\left\{\tilde{l}_j - (1/2)(\tilde{b}_j^T D^{-1} \tilde{b}_j)\right\} \times \left[1 + E(T_{4j}) + E(T_{6j}) + (1/2)E(T_{3j}^2)\right]. \quad (4.4)$$

Following Shun and McCullagh (1995) and Shun (1997), an alternative approximation to (4.4) is

$$L \approx \prod_{j=1}^{J} |D|^{-1/2} |Z_j^T W_j Z_j + D^{-1}|^{-1/2} \exp\left\{\tilde{l}_j - (1/2)(\tilde{b}_j^T D^{-1} \tilde{b}_j)\right\} \times \exp\left\{E(T_{4j}) + E(T_{6j}) + \frac{1}{2}E(T_{3j}^2)\right\}. \quad (4.5)$$

In the following illustrative analyses and in the simulations, approximations (4.4) and (4.5) led to essentially identical results. Taking the log of (4.4) and applying algebraic simplifications (see Yang 1998), the log marginal likelihood becomes

$$\log(L) \approx \frac{-J}{2} \log|D| + \frac{1}{2}\sum_{j=1}^{J} \log|V_j| + \sum_{j=1}^{J}\left(\tilde{l}_j - \frac{1}{2}\tilde{b}_j^T D^{-1}\tilde{b}_j\right) + \sum_{j=1}^{J} \log A_j, \quad (4.6)$$

where

$$\begin{aligned} A_j &= 1 + E(T_{4j}) + E(T_{6j}) + (1/2)E(T_{3j}^2) \\ &= 1 - \frac{1}{8}\sum_{i}^{n_j} m_{ij}^{(4)} B_{ij}^2 - \frac{1}{48}\sum_{i}^{n_j} m_{ij}^{(6)} B_{ij}^3 + \frac{15}{72} k_j^T V_j k_j, \end{aligned}$$

with $V_j^{-1} = -h_j^{(2)}(\tilde{b}_j) = Z_j^T W_j Z_j + D^{-1}$, $B_{ij} = Z_{ij}^T V_j Z_{ij}$, and $k_j = \sum_i^{n_j} m_{ij}^{(3)} Z_{ij} B_{ij}$.

We use approximate Fisher scoring (Green 1984) to maximize (4.6), yielding highly accurate approximation to the ML estimates for β and Ψ, the vector of the distinct elements in D. Approximate Fisher scoring (Green 1984) requires only first derivatives, enabling us to avoid the complex second derivatives. Fisher scoring iteratively solves for $\theta = (\beta, \Psi)$ by using the equation $\theta^{\text{new}} - \theta^{\text{old}} = \sum_{j=1}^{J}(S_j S_j^T)^{-1} S_j$, where the $S_j = (S_{\beta j}^T, S_{\Psi j}^T)^T$ is the score vector of the jth cluster, S_β and S_Ψ being the derivatives of the log-marginal likelihood (4.6) with respect to β and Ψ, respectively.

In getting these derivatives, however, we need to take into consideration that $h_j(\tilde{b}_j)$ is evaluated at $\tilde{b} = \tilde{b}(\beta, \Psi) = V_j Z_j^T W_j (y_j^* - X_j \beta)$. We solve this interdependence with implicit differentiation of $\tilde{b}_j$ with respect to β and Ψ, respectively:

$$\frac{\partial \tilde{b}_j}{\partial \beta^T} = -V_j Z_j^T W_j X_j, \tag{4.7}$$

$$\frac{\partial \tilde{b}_j}{\partial \Psi^T} = [(y_j^* - \eta_j)^T W_j Z_j \otimes V_j D^{-1}]. \tag{4.8}$$

After repeated use of matrix algebra in Magnus and Neudecker (1988) we have the score vectors (Yang 1998)

$$S_{\beta j} = X_j^T W_j (y_j^* - X_j \beta - Z_j \tilde{b}_j) + \sum_i^{n_j} f_{ij} v_{ij} + \frac{1}{A_j} \sum_i^{n_j} c_{ij} v_{ij}, \tag{4.9}$$

and

$$\begin{aligned} S_{\Psi j} &= \frac{1}{2} E^T \mathrm{vec}[D^{-1}(\hat{D}_j - D) D^{-1}] - \frac{1}{2} \sum_i^{n_j} m_{ij}^{(3)} B_{ij} E^T \mathrm{vec}[Q_{ij}] \\ &\quad + \frac{1}{A_j} \left\{ \sum_i^{n_j} c_{ij} E^T \mathrm{vec}[Q_{ij}] + E^T \left[\sum_i^{n_i} \left(-\frac{1}{4} m_{ij}^{(4)} B_{ij} \right.\right.\right. \\ &\quad \left.\left.\left. - \frac{1}{16} m_{ij}^{(6)} B_{ij}^2 + \frac{15}{36} m_{ij}^{(3)} a_{ij} \right) \mathrm{vec}(F_{ij}) + \frac{15}{72} \mathrm{vec} \left\{ D^{-1} V_j k_j k_j^T V_j D^{-1} \right\} \right] \right\}, \end{aligned} \tag{4.10}$$

where $f_{ij} = \frac{-1}{2} m_{ij}^{(3)} B_{ij}$, $G_j = X_j^T W_j Z_j V_j$, $v_{ij} = X_{ij} - G_j Z_{ij}$, and $c_{ij} = -\frac{1}{8} m_{ij}^{(5)} B_{ij}^2 + \frac{1}{4} m_{ij}^{(3)} Z_{ij}^T H_j Z_{ij} - \frac{1}{48} m_{ij}^{(7)} \beta_{ij}^3 + \frac{1}{16} m_{ij}^{(3)} Z_{ij}^T p_j Z_{ij} + \frac{15}{36} m_{ij}^{(4)} B_{ij} a_{ij} - \frac{15}{72} m_{ij}^{(3)} a_{ij}^2 - \frac{15}{36} m_{ij}^{(3)} Z_{ij}^T h_j Z_{ij}$, with

$$\begin{array}{rcl rcl} H_j &=& \sum_i^{n_j} B_{ij} m_{ij}^{(4)} V_j Z_{ij} Z_{ij}^T V_j, & m_{ij}^{(7)} &=& m_{ij}^{(5)}(1 - 12 w_{ij}) - 36 m_{ij}^{(3)} m_{ij}^{(4)}, \\ p_j &=& \sum_i^{n_j} m_{ij}^{(6)} B_{ij}^2 V_j Z_{ij} Z_{ij}^T V_j, & a_{ij} &=& Z_{ij}^T V_j k_j, \\ h_j &=& \sum_i^{n_j} m_{ij}^{(3)} a_{ij} V_j Z_{ij} Z_{ij}^T V_j, & \hat{D}_j &=& \tilde{b}_j \tilde{b}_j^T + V_j, \\ E &=& \frac{d\mathrm{vec}D}{d\phi^T}, & Q_{ij} &=& D^{-1} V_j Z_{ij} (y_j^* - \eta_j)^T W_j Z_j, \quad \text{and} \\ F_{ij} &=& D^{-1} V_j Z_{ij} Z_{ij}^T V_j D^{-1}. & & & \end{array}$$

Computational formulas for the Poisson-normal and gamma-normal models are simpler than in the binomial-normal case because the expressions for the derivatives $m_{ij}^{(k)}$ are simpler (see (4.2) and following paragraph).

5. ILLUSTRATIVE EXAMPLE

How close are the estimates produced by sixth order Laplace approximation ("Laplace6") to the desired ML estimates? One way to answer this question is to compare results from Laplace6 with those based on Gauss–Hermite quadrature (Hedeker and Gibbons 1994) using a large number of quadrature points. As the number of quadrature points increases, the Gauss–Hermite approximation to the integral becomes increasingly

Table 1. Descriptive Statistics of Thailand Data

Variable name	*N*	*mean*	*sd*	*min*	*max*
Repetition	7877	.14	.35	.00 (= no)	1.00 (= yes)
SES	7877	.00	.69	−1.76	3.48
Sex	7877	.51	.50	.00 (= female)	1.00 (= male)
Dialect	7877	.47	.50	.00 (= no)	1.00 (= yes)
Breakfast	7877	.84	.37	.00 (= no)	1.00 (= yes)
Pre-primary	7877	.50	.50	.00 (= no)	1.00 (= yes)
MeanSES	376	.00	.45	−.93	2.01
Enrollment	376	.01	.86	−1.77	2.60
Meantxt	376	.00	1.84	−5.91	2.63

accurate. We therefore compared Laplace6 to the Gauss–Hermite approach using from 10 to 40 quadrature points ("Gauss-10,"..., "Gauss-40") using software developed by Hedeker and Gibbons (1994) that maximizes the resulting likelihood using a Fisher scoring algorithm. We chose the binomial-normal case for the comparison because this is the case for which simpler approximations have proven most problematic (Rodriguez and Goldman 1995; Breslow and Lin 1995). Both Laplace6 and the Gauss approaches are compared to penalized quasi-likelihood ("PQL") as described by Breslow and Clayton (1993) and implemented in the software developed by Bryk, Raudenbush, and Congdon (1996).

For the purpose of this comparison we selected a large dataset characterized by a large between-cluster variance, a binomial denominator of 1.0, and somewhat asymmetric probabilities (i.e., probabilities of success concentrated away from .50). This is the case for which PQL can be expected to perform rather poorly and which poses the greatest challenge for the other methods.

The data are from a national survey of primary education in Thailand in 1988 (see Raudenbush and Bhumirat 1992 for details), and include 7,877 sixth grade students clustered within a nationally representative sample of 376 primary schools. Our interest focuses on grade repetition, which occurs when a student makes unsatisfactory academic progress and is required to repeat a given grade. Specifically, y_{ij}, the repetition outcome of student i in school j, takes on a value of unity if a student repeats a grade during the primary years and zero if not. Of special interest are whether the instructional resources available to a school, as indicated by the availability of textbooks ("textbooks"), are associated with a reduced risk of repetition in that school, and whether a child's preprimary school experience ("pre-primary") reduces the risk of repetition. However, we also wish to control for important covariates. Two of these vary at the school level: school size ("enrollment"), and the socioeconomic level of students attending the school ("meanSES"). Child-level covariates include sex ("sex"), socioeconomic status ("SES"), child dialect ("dialect"), and an indicator of nutritional status ("breakfast"). The scale of these covariates and descriptive statistics are provided in Table 1. We found evidence of significant variation across clusters in the intercept with no evidence of variation across clusters in regression coefficients. We therefore estimated the model

$$\begin{aligned}\eta_{ij} = \beta_{00} + \beta_{01}(\text{textbooks})_j + \beta_{02}(\text{enrollment})_j + \beta_{03}(\text{meanSES})_j + \beta_{10}(\text{pre-primary})_{ij} \\ + \beta_{20}(\text{SES})_{ij} + \beta_{30}(\text{sex})_{ij} + \beta_{40}(\text{dialect})_{ij} + \beta_{50}(\text{breakfast})_{ij} + b_j,\end{aligned}$$

Table 2. Estimates Based on Thailand Data[a]

	PQL	*Gauss-10*	*Gauss-20*	*Gauss-30*	*Gauss-40*	*Laplace6*
β_{00}	−1.9757	−2.1661	−2.1632	−2.1608	−2.1614	−2.1554
	(.1356)	(.1339)	(.1338)	(.1340)	(.1340)	(.1340)
β_{01}	.0059	.0338	.0036	.0065	.0064	.0063
	(.0411)	(.0433)	(.0436)	(.0439)	(.0438)	(.0438)
β_{02}	−.1999	−.2030	−.2007	−.1996	−.2002	−.1992
	(.1173)	(.1410)	(.1410)	(.1420)	(.1419)	(.1419)
β_{03}	−.6717	−.8709	−.7649	−.7759	−.7744	−.7692
	(.2638)	(.3246)	(.3249)	(.3256)	(.3255)	(.3251)
β_{10}	−.4132	−.4587	−.4450	−.4474	−.4471	−.4466
	(.0943)	(.0966)	(.0983)	(.0980)	(.0981)	(.0980)
β_{20}	−.3716	−.3874	−.3862	−.3860	−.3861	−.3866
	(.0917)	(.0939)	(.0947)	(.0946)	(.0946)	(.0946)
β_{30}	.5486	.5763	.5760	.5761	.5761	.5763
	(.0725)	(.0704)	(.0699)	(.0698)	(.0698)	(.0698)
β_{40}	.2444	.2679	.2784	.2771	.2773	.2759
	(.1280)	(.1227)	(.1258)	(.1261)	(.1261)	(.1260)
β_{50}	−.3663	−.3776	−.3818	−.3809	−.3812	−.3811
	(.1019)	(.1021)	(.1021)	(.1023)	(.1022)	(.1022)
D_{00}	1.0757	1.454	1.389	1.394	1.394	1.383

[a]: standard error estimates are in parentheses.

with $b_j \sim N(0, D_{00})$.

The results across methods of estimation are similar in broad outline: Accessibility of textbooks is not significantly related to repetition, but preprimary experience is associated with a lower risk of repetition, given the covariates. Covariates meanSES, SES, sex, dialect, and breakfast are associated with repetition in the expected directions.

The comparison of the Laplace6 results with those of Gauss–Hermite method leads us to believe Laplace6 is a very accurate approximation. As shown in Table 2, most differences between Gauss-10 and Gauss-20 are in the second or third decimal place. Gauss-20 and Gauss-30 typically differ in the third decimal place. Gauss-30 and Gauss-40 tend to differ at the fourth decimal place. Most Laplace6 estimates differ from Gauss-40 in the third place or in the fourth place. Laplace6 results are generally closer to Gauss-30 and Gauss-40 than to Gauss-10 and Gauss-20. The standard errors of the estimates produced by Laplace6 are nearly identical to those of Gauss-30 or Gauss-40. Note that PQL, as expected, consistently gives smaller estimates (in absolute values) for all parameters than do the better approximations. Of course, these results are based on only one dataset and include only one random effect per cluster.

Table 3. Averages of the Estimates for Simulated Data

	PQL	*Gauss-10*	*Gauss-20*	*Laplace6*
$D_{00} = 1.625$	1.2752	1.6532	1.6546	1.6352
$D_{01} = .100$	.0538	.1003	.0995	.0960
$D_{11} = .250$	.1614	.2575	.2562	.2667
$\beta_{00} = -1.200$	−1.0904	−1.1977	−1.2045	−1.2007
$\beta_{01} = 1.000$	.9004	1.0081	1.0148	1.0029
$\beta_{10} = 1.000$	.9114	.9971	.9976	.9975

Table 4. Mean Squared Errors for Simulated Data

	PQL	*Gauss-10*	*Gauss-20*	*Laplace6*
$D_{00} = 1.625$	.1522	.0737	.0633	.0563
$D_{01} = .100$	.0080	.0115	.0120	.0108
$D_{11} = .250$	.0113	.0073	.0072	.0075
$\beta_{00} = -1.200$	.0271	.0231	.0196	.0190
$\beta_{01} = 1.000$	.0236	.0193	.0175	.0164
$\beta_{10} = 1.000$	.0116	.0051	.0053	.0051

6. SIMULATION STUDY

6.1 Comparison to PQL and Gauss-Hermite Quadrature

To evaluate the performance of the algorithm proposed in Section 4 in the case of dependent random effects, we first simulated 100 datasets and compared its results (Laplace6) with those from PQL, Gauss-10 and Gauss-20. The model produced data with asymmetric probabilities having an average conditional expectation equal to .14; that is, $\mu_{ij}^{(0)} = \Pr(y_{ij} = 1 | b_j = 0)$ is, on average, .14. The structure of the datasets follows Rodriguez and Goldman (1995). These datasets involve 20 hypothetical children nested within each of 200 hypothetical communities, with 4,000 children overall. The model reflects the belief that the outcome y_{ij} depends on a child-level covariate ("childcov") and a community-level covariate ("commucov"), and that intercepts and slopes varying across "communities." [Note: The level-1 covariate, childcov, was sampled from a normal distribution with mean .0955621, and variance .0676, while the level-2 predictor, commucov, was sampled from a normal distribution with mean -.6857591 and variance .2304.] Thus, we have a model for person i in group j

$$\eta_{ij} = \beta_{00} + \beta_{01}(\text{commu cov})_j + \beta_{10}(\text{child cov})_{ij} + b_{0j} + b_{1j}(\text{child cov})_{ij}$$

with b_{0j} and b_{1j} together forming a bivariate normal distribution with means 0, variances D_{00}, D_{11}, respectively, and covariance D_{01}, where $D_{00} = 1.625$, $D_{11} = .25$, and $D_{01} = .1$. The large components of dispersion tend to render simple approximations inaccurate.

Tables 3 and 4 present averages and mean squared errors of the estimates across 100 replications. As expected, PQL estimates are biased toward zero for all parameters. The underestimation of the variance components range from 22% to 46%, while that of the β's are around 9%. The advantage of Gauss-20 over Gauss-10 is not very clear from the tables, since the averages from the two are very similar, though mean squared errors tend to be a bit smaller for Gauss-20 than for Gauss-10. In general, the Gauss-10, Gauss-20, and Laplace6 are similar, though the Laplace6 results are generally closer to

Table 5. Averages of the Estimates for Simulated Data[a]

	Laplace6 (n = 100)	*Laplace6 (n = 97)*	*AGQ (n = 97)*
$D_{00} = 1.625$	1.5915	1.5826	1.6009
$D_{01} = .100$	.0916	.0903	.0922
$D_{11} = .250$	.2460	.2468	.2374
$\beta_{00} = -1.200$	−1.1948	−1.1955	−1.1978
$\beta_{01} = 1.000$	.9866	.9844	.9949
$\beta_{10} = 1.000$	.9929	.9918	.9920

[a] AGQ did not produce estimates for three of the 100 datasets. Therefore, the comparison between Laplace6 and AGQ was based on the 97 datasets for which both algorithms obtain convergence.

the Gauss-20 than the Gauss-10 results. Mean squared errors tend to be a bit smaller under Laplace6 than under Gauss-10 or Gauss-20 in these results. In particular, when Gauss-20 appears nontrivially better than Gauss-10, Laplace replicates the advantage or possibly improves upon it.

In summary, for these simulated data with bivariate random effects, Laplace6 seems to do as well as or better than Gauss using 10 or 20 quadrature points. To assess computational efficiency, we timed analyses on six randomly selected data sets using a Pentium 233mHz, and found the analysis took on average 35 seconds per dataset using Laplace6. These analyses were found to take, on average, 180 seconds for Gauss-10 and 720 seconds for Gauss-20.

6.2 COMPARISON TO ADAPTIVE QUADRATURE

Adaptive Gauss–Hermite quadrature (AGQ) (Pinheiro and Bates 1995) might well provide a better comparison to our Laplace approach than does Gauss–Hermite quadrature (GQ). Our Laplace approach expands the variable of integration around its approximate posterior mode, an approach that should perform better than an expansion around the marginal mean of zero, especially when the random effects dispersion is large. Similarly, AGQ centers the variable of integration around the approximate posterior mode, and is therefore similar in construction to our Laplace method, while nonadaptive GQ centers the variable of integration around the marginal mean of zero.

We therefore replicated the simulation study above using an experimental version of SAS PROC NLMIXED that computes AGQ estimates for nonlinear mixed regression models (Wolfinger 1999). The SAS algorithm selects the number of quadrature points

Table 6. Mean Squared Errors for Simulated Data

	Laplace6 (n = 100)	*Laplace6 (n = 97)*	*AGQ (n = 97)*
$D_{00} = 1.625$	.0847	.0848	.0925
$D_{01} = .100$	.0094	.0094	.0103
$D_{11} = .250$	.0083	.0084	.0086
$\beta_{00} = -1.200$	.0122	.0125	.0126
$\beta_{01} = 1.000$	.0108	.0110	.0112
$\beta_{10} = 1.000$	.0047	.0047	.0048

empirically. For our data, seven quadrature points were used per simulation. The results (Tables 5 and 6) showed that AGQ and Laplace6 produced similar results. However, mean squared errors of Laplace6 never exceeded those from AGQ. The average run time for the AGQ analyses was 725 seconds, as compared to 35 seconds for Laplace6.

7. CONCLUSION

Highly accurate approximation of difficult integrals via high-order, multivariate Laplace approximation appears to be a promising strategy for evaluating likelihoods in generalized linear models with nested random effects. The approach can be extended to the degree of accuracy required, and can be written for multivariate random effects of arbitrary dimension per cluster. We have developed computational formulas for two-level generalized linear models with canonical link and multivariate normal random effects. When applied to the logistic regression model with random coefficients, computations for sixth-order Laplace approximation were much faster than those required for Gauss–Hermite quadrature, both adaptive and nonadaptive, and results were as good or better than those produced by the Gauss–Hermite method with 20 quadrature points or the adaptive method with seven quadrature points.

Given the current algorithms, the quadrature approach is superior in that one may easily choose the number of quadrature points to obtain needed accuracy in approximation to ML estimates. In contrast, more work will be required before it becomes easy to specify the degree of polynomial in the Taylor series that governs the accuracy of the Laplace method. Yet the Laplace method appears to have significant potential as a highly accurate and fast approximation to ML for hierarchical models. Much more research is needed: (1) on the behavior of this approach in binomial-normal models; (2) in other generalized linear models with random effects; (3) in a broader class of hierarchical models; and (4) in increasing the degree of accuracy by increasing the degree of the polynomial in the Taylor series. We shall also be interested in applications in other difficult integration problems, for example, those that confront Bayesian inference.

APPENDIX

In this appendix we sketch a proof for Theorem 2. Since T_k is a scalar

$$\begin{aligned} E(T_k) &= E\left[(1/k!)\left[\overset{k-1}{\otimes}(b-\tilde{b})^T\right] h^{(k)}(\tilde{b})(b-\tilde{b})\right] \\ &= (1/k!)\mathrm{vec}^T\left(E\left\{\left[\overset{k-1}{\otimes}(b-\tilde{b})\right](b-\tilde{b})^T\right\}\right)\mathrm{vec}\left\{h^{(k)}(\tilde{b})\right\} \end{aligned}$$

(see Magnus and Neudecker 1988, p. 30, eq. (3)).

Now, $E\{[\overset{k-1}{\otimes}(b-\tilde{b})](b-\tilde{b})^T\} = \mu_{(k)}$ is the kth centered moment of the q-variate normal distribution with covariance matrix V. For k odd, $\mu_{(k)} = 0$. Using the moment-generating function for a multivariate normal distribution, Yang (1998) has shown that for k even, $\mu_{(k)}$ is composed of commutations of Kronecker products of $(k-1)(k-3)\ldots 3$ matrices having the form V or vec$V$. However, we are not currently interested in $\mu_{(k)}$

itself but in the scalar $\text{vec}^T(\mu_{(k)})\text{vec}(S)$, where S is any conformable matrix of constants. Yang (1998) showed that we can ignore the commutations since what we are interested in is the trace, a scalar. For example,

$$E(T_4) = \frac{1}{4!}\text{vec}^T(\mu_{(4)})\text{vec}(h^{(4)}(\tilde{b})) = \frac{3}{4!}\text{vec}^T\left\{V \otimes V\right\}\text{vec}\left[h^{(4)}(\tilde{b})\right],$$

despite the fact that

$$\mu_{(4)} = (V \otimes \text{vec}V) + (\text{vec}V \otimes V) + \left(K_{qq} \otimes I_q\right)(V \otimes \text{vec}V),$$

where K_{qq} is the $q^2 \times q^2$ commutation matrix (Magnus and Neudecker 1988) and I_q is the $q \times q$ identity matrix. In general, then, $\text{vec}^T(\mu_{(k)})\text{vec}(S) = (k-1)(k-3)\ldots 3\ \text{vec}^T\left(\overset{k/2}{\otimes} V\right)\text{vec}(S)$. Substituting $S = \text{vec}(h^{(k)}(\tilde{b}))$ produces Equation (3.3). The proof for $E(T_kT_l)$ is similar.

ACKNOWLEDGMENTS

The research reported here was supported by the Project on Human Development in Chicago Neighborhoods (PHDCN) with funding from the MacArthur Foundation, the National Institute of Justice, and the National Institute of Mental Health. The authors thank Richard Congdon for the programming of all computations.

[Received June 1998. Revised June 1999.]

REFERENCES

Aitken, M., and Longford, N. (1986), "Statistical Modeling Issues in School Effectiveness Studies," *Journal of Royal Statistical Society*, Ser. A, 149, 1–43.

Anderson, D. A., and Aitkin, M. (1985), "Variance Component Models with Binary Response: Interviewer Variability," *Journal of the Royal Statistical Society*, Ser. B, 47, 204–210.

Belin, T., Diffendal, J., Mack, S., Rubin, D., Schafer, J., and Zazlavsky, A. (1993), "Hierarchical Logistic Regression Models for Imputation of Unresolved Enumeration Status in Undercount Estimation," *Journal of the American Statistical Association*, 88, 1149–1159.

Berkey, D. C., Hoaglin, F., Mosteller, F., and Colditz, G. A. (1995), "A Random Effects Regression Model for Meta-Analysis," *Statistics in Medicine*, 14, 395–411.

Bock, R. D. (1989), *Multilevel Analysis of Educational Data*, New York: Academic Press.

Breslow, N. E., and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9–25.

Breslow, N. E., and Lin, X. (1995), "Bias Correction in Generalised Linear Mixed Models With a Single Component of Dispersion," *Biometrika*, 82, 81–91.

Bryk, A., and Raudenbush, S. (1992), *Hierarchical Linear Models for Social and Behavioral Research: Applications and Data Analysis Methods*, Newbury Park, CA: Sage.

Bryk, A. S., Raudenbush, S. W., and Congdon, R. (1996), *HLM: Hierarchical Linear and Nonlinear Modeling with the HLM/2L and HLM/3L Programs*, Chicago: Scientific Software International, Inc.

Chan, W. C. (1994), "Toward a Multilevel Model: The Case for Poisson Distributed Data," unpublished Ph.D. dissertation, Michigan State University, Department of Counseling, Educational Psychology, and Special Education.

Cox, D., and Hinkley, D. (1974), *Theoretical Statistics*, Boca Raton, FL: CRC Press.

DeLeeuw, J., and Kreft, I. (1986), "Random Coefficient Models for Multilevel Analysis," *Journal of Educational Statistics*, 11, 57–85.

Dempster, A., Laird, N., and Rubin, D. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1–8.

Dempster, A. P., Rubin, D. B., and Tsutakawa, R. K. (1981), "Estimation in Covariance Components Models," *Journal of American Statistical Association*, 76, 341–353.

DerSimonian, R., and Laird, N. (1986), "Meta-Analysis in Clinical Trials," *Controlled Clinical Trials*, 7, 177–188.

Diggle, P. J., Liang, K. Y., and Zeger, S. L. (1994), *Analysis of Longitudinal Data*, Oxford: Clarendon Press.

Gilks, W. R. (1987), "Some Applications of Hierarchical Models in Kidney Transplantation," *The Statistician*, 36, 127–136.

Goldstein, H. (1986), "Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares," *Biometrika*, 73, 43–56.

——— (1991), "Nonlinear Multilevel Models, With an Application to Discrete Response Data," *Biometrika*, 78, 45–51.

——— (1995), *Multilevel Statistical Models* (2nd ed.), New York: Wiley.

Goldstein, H., and Rasbash, J. (1996), "Improved Approximations for Multilevel Models with Binary Responses," *Journal of the Royal Statistical Society*, Ser. B, 159, 505–513.

Green, P. J. (1984), "Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and Some Robust and Resistant Alternatives," *Journal of the Royal Society*, Ser. B, 46, 149–192.

Hedeker, D., and Gibbons, R. D. (1994), "A Random Effects Ordinal Regression Model for Multilevel Analysis," *Biometrics*, 50, 933–944.

——— (1996), "MIXOR: A Computer Program for Mixed-Effects Ordinal Probit and Logistic Regression Analysis," *Computer Methods and Programs in Biomedicine*, 49, 157–176.

Karim, M. R. (1991), "Generalized Linear Models with Random Effects," unpublished Ph.D. dissertation, Department of Biostatistics, Johns Hopkins University.

Kass, R. E., Tierney, L., and Kadane, J. B. (1990), "The Validity of Posterior Expansions Based on Laplace's Method," in *Bayesian and Likelihood Methods in Statistics and Econometrics*, eds. S. Geisser, J. S. Hodges, S. J. Press, and A. Zellner, North-Holland: Elsevier.

Laird, N., and Ware, J. (1982), "Random Effects Models for Longitudinal Data," *Biometrika*, 65, 581–590.

Lee, Y., and Nelder, J. A. (1996), "Hierarchical Generalized Linear Models," *Journal of the Royal Statistical Society*, Series B, 58, 619–678.

Lin, X., and Breslow, N. E. (1996), "Bias Correction in Generalized Linear Mixed Models with Multiple Components of Dispersion," *Journal of the American Statistical Association*, 91, 1007–1016.

Lindley, D. (1980), "Approximate Bayesian Methods," in *Bayesian Statistics*, eds. J. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Valencia: Valencia University Press, 223–245.

Lindley, D., and Smith, A. (1972), "Bayes Estimates for the Linear Model," *Journal of the Royal Statistical Society*, Ser. B, 34, 1–41.

Lindstrom, M. J., and Bates, D. M. (1988), "Newton–Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data," *Journal of the American Statistical Association*, 83, 1014–1022.

Liu, Q., and Pierce, D. A. (1993), "Heterogeneity in Mantel–Haenszel-type Models," *Biometrika*, 80, 543–556.

Longford, N. (1987), "A Fast Scoring Algorithm for Maximum Likelihood Estimation in Unbalanced Mixed Models With Nested Random Effects," *Biometrika*, 74, 817–827.

——— (1993), *Random Coefficient Models*, Oxford: Clarendon Press.

Magnus, J. R., and Neudecker, H. (1988), *Matrix Differential Calculus with Applications in Statistics and Econometrics*, New York: Wiley.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

McGilchrist, C. E. (1994), "Estimation in Generalized Mixed Models," *Journal of the Royal Statistical Society*, Ser. B, 56, 61–69.

Pinheiro, J. C., and Bates, D. M. (1995), "Approximations to the Log-Likelihood Function in the Non-Linear

Mixed-Effects Model," *Journal of Computational and Graphical Statistics*, 4, 12–35.

Raudenbush, S. W., and Bryk, A .S. (1985), "Empirical Bayes Meta-Analysis," *Journal of Educational Statistics*, 10, 75–98.

——— (1986), "A Hierarchical Model for Studying School Effects," *Sociology of Education*, 59, 1–17.

Raudenbush, S. W., and Bhumirat, C. (1992), "The Distribution of Resources for Primary Education and its Consequences for Educational Achievement in Thailand," *International Journal of Educational Research*, 143–164.

Rodriguez, G., and Goldman, N. (1995), "An Assessment of Estimation Procedures for Multilevel Models with Binary Responses," *Journal of the Royal Statistical Society*, Ser. A, 158, 73–89.

Rubin, D. (1981), "Estimation in Parallel Randomized Experiments," *Journal of Educational Statistics*, 6, 337–401.

Schall, R. (1991), "Estimation in Generalized Linear Models With Random Effects," *Biometrika*, 40, 719–727.

Seltzer, M. H. (1993), "Sensitivity Analysis for Fixed Effects in Hierarchical Model: A Gibbs Sampling Approach," *Journal of Educational Statistics*, 18, 207–235.

Shun, Z. (1997), "Another Look at the Salamander Mating Data: A Modified Laplace Approximation Approach," *Journal of American Statistical Association*, 92, 341–349.

Shun, Z., and McCullagh, P. (1995), "Laplace Approximation of High Dimensional Integrals," *Journal of the Royal Statistical Society*, Ser. B, 57, 749–760.

Solomon, P. J., and Cox, D. R. (1992), "Nonlinear Components of Variance Models," *Biometrika*, 79, 1–11.

Strenio, J., Weisberg, H., and Bryk, A. (1983), "Empirical Bayes Estimation of Individual Growth Curve Parameters and Their Relationship to Covariates," *Biometrics*, 39, 71–86.

Stiratelli, R., Laird, N., and Ware, J. (1984), "Random Effects Models for Serial Observations With Binary Responses," *Biometrics*, 40, 961–971.

Thum, Y. (1997), "Hierarchical Linear Models for Multivariate Outcomes," *Journal of Educational and Behavioral Statistics*, 22, 77–108.

Tsutakawa, R. K. (1985), "Estimation of Cancer Mortality Rates: A Bayesian Analysis of Small Frequencies," *Biometrics*, 41, 69–79.

Wei, G. C. G., and Tanner, M. A. (1990), "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Augmentation Algorithms," *Journal of the American Statistical Association*, 86, 669–704.

Wolfinger, R. (1993), "Laplace's Approximation for Nonlinear Mixed Models," *Biometrika*, 80, 791–795.

——— (1999), "Nonlinear Mixed Models: A Future Direction," presented at the Annual Interface Conference, Shaumberg, IL, June 10, 1999.

Wong, G. Y., and Mason, W. M. (1985), "The Hierarchical Logistic Regression Model for Multilevel Analysis," *Journal of the American Statistical Association*, 80, 513–524.

Yang, M. (1998), "Increasing the Efficiency in Estimating Multilevel Bernoulli Models," unpublished Ph.D. dissertation, Michigan State University, Department of Counseling, Educational Psychology, and Special Education.