A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data

Maren K. OLSEN and Joseph L. SCHAFER

A semicontinuous variable has a portion of responses equal to a single value (typically 0) and a continuous, often skewed, distribution among the remaining values. In cross-sectional analyses, variables of this type may be described by a pair of regression models; for example, a logistic model for the probability of nonzero response and a conditional linear model for the mean response given that it is nonzero. We extend this two-part regression approach to longitudinal settings by introducing random coefficients into both the logistic and the linear stages. Fitting a two-part random-effects model poses computational challenges similar to those found with generalized linear mixed models. We obtain maximum likelihood estimates for the fixed coefficients and variance components by an approximate Fisher scoring procedure based on high-order Laplace approximations. To illustrate, we apply the technique to data from the Adolescent Alcohol Prevention Trial, examining reported recent alcohol use for students in grades 7–11 and its relationships to parental monitoring and rebelliousness.

KEY WORDS: Generalized linear mixed model; Integral approximation; Laplace method; Variance components.

1. INTRODUCTION

1.1 Semicontinuous Variables

A semicontinuous random variable combines a continuous distribution with point masses at one or more locations. In this article we discuss a particular type of semicontinuous variable frequently encountered in practice, a mixture of 0's and continuously distributed positive values. These methods were motivated partly by studies of adolescent substance use in which researchers seek to model self-reported usage for alcohol, tobacco, or marijuana. Semicontinuous variables are also common in economic surveys. For example, in the Consumer Expenditure Survey conducted by the U.S. Bureau of Labor Statistics, subjects report dollar amounts spent on goods and services in such finely detailed categories as automobile repair and maintenance, footwear, and major household appliances. A semicontinuous variable is quite different from one that has been left-censored or truncated, because the 0's are valid self-representing data values, not proxies for negative or missing responses. It is natural to view a semicontinuous response as the result of two processes, one determining whether the response is 0 and the other determining the actual level if it is non-0. The two processes are qualitatively distinct and may be influenced by covariates in different ways. In substance use research, for example, investigators may wish to understand what characteristics distinguish the adolescents who used alcohol in the past 30 days from those who did not, and also what characteristics influence the amount consumed by users. An added complication is that these two processes may be related, particularly if the semicontinuous response is observed at multiple time points; a high level of use on one occasion may affect the probability of use on another occasion.

1.2 Motivating Example

The data in Figure 1 were drawn from the Adolescent Alcohol Prevention Trial (AAPT), a school-based longitudinal study of adolescents in the Los Angeles area (Hansen and Graham 1991). The four graphs show levels of reported recent alcohol use in grades 7–11 for four typical subjects. Individual (a) reports a moderate level of use initially, a high level of use by grade 9, then lower use by grade 11. Individual (c) reports no use at any occasion. Individuals (b) and (d) report a mixture of no use and use with missing values at some time points. Approximately 49% of the AAPT subjects reported no use for any time point at which they were measured. Conventional linear growth models, which assume normally distributed random coefficients and normal residuals (e.g., Laird and Ware 1982), would fit these data poorly because of the preponderance of 0's. Semiparametric approaches based on generalized estimating equations (Zeger, Liang, and Albert 1988) would also fail to recognize the qualititative distinctions between 0 and non-0 responses. Characterizing adolescent substance use as dual processes, one binary and one continuous, is theoretically appealing and provides a richer description than a model with a single mean function.

1.3 Related Work

Methods for estimating the moments of positive random variables with discrete probability mass at the origin were investigated by Aitchison (1955). Two-part regression models for variables of this type have appeared in econometric analyses for nearly two decades. Duan, Manning, Morris, and Newhouse (1983) and Manning et al. (1981) described expenditures for medical care by a pair of regression equations, one for the logit probability of expenditure and one for the conditional mean of log expenditure (given that one exists). In cross-sectional applications, the two models may be fit separately with standard logit and linear regression software.

Sample selection models, including the Tobit model (Amemiya 1984; Tobin 1958) and Heckman's selection model

Maren K. Olsen is Biostatistician, Division of Health Services Research and Development, Durham Veterans Affairs Medical Center, and Assistant Research Professor, Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC. Joseph L. Schafer is Associate Professor, Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA. The authors thank John Graham for providing data from the Adolescent Alcohol Prevention Trial and advice on their analysis. This research was supported by the National Institute on Drug Abuse grant 1-P50-DA10075 and by the American Statistical Association/National Science Foundation/Bureau of Labor Statistics Research Fellow program. The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs.

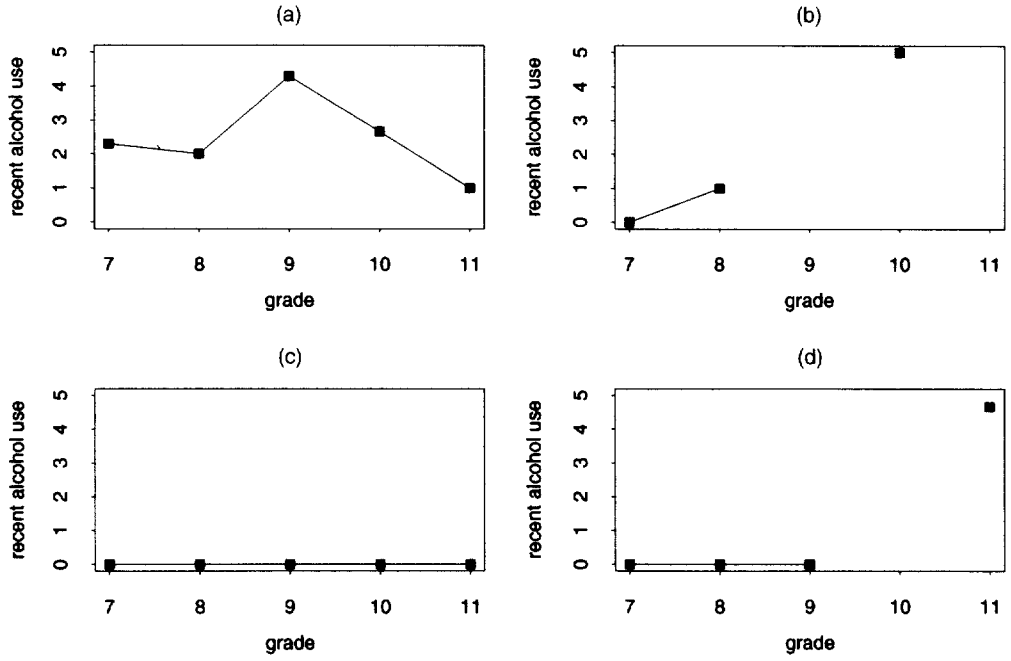

Figure 1. Reported Level of Recent Alcohol Use in Grades 7–11 for Four Subjects [(a)–(d)] in the AAPT.

(Heckman 1974, 1976), are often applied to limited or censored dependent variables. These posit an underlying normal random variable that is censored by a random mechanism; the mean of the underlying variable and the probability of censoring are jointly modeled as functions of covariates. As noted by Duan et al. (1983), two-part models are easier to interpret than selection models when 0's represent actual data, because the meaning of the underlying normal variable becomes dubious when 0 is a valid response rather than a proxy for a negative or missing value. In contrast, the second part of a two-part regression describes the *conditional* mean of the response given that it is non-0, a quantity that is highly meaningful. The Tobit and Heckman selection models have been extended to panel data (Cowles, Carlin, and Connett 1996; Hajivassiliou 1994; Kyriazidou 1997); our work, on the other hand, is a natural extension of the two-part regression of Duan et al. (1983) and Manning et al. (1981).

A related body of work pertains to excess 0's in count data that may arise from a combination of overdispersion and true 0 inflation. Two-part regressions for counts include 0-inflated Poisson (Lambert 1992) and 0-altered Poisson (Heilbron 1989) models. Greene (1994) defined a sample selection model for count data analogous to Heckman's model in which the underlying precensored response is described by a Poisson or negative binomial regression.

1.4 Scope of This Article

In the remainder of this article, we develop a new two-part random-effects model for semicontinuous longitudinal data. In Section 2 we introduce the model and discuss strategies for parameter estimation. We derive a closed-form approximation to the log-likelihood based on high-order multivariate Laplace expansion recently developed by Raudenbush, Yang, and Yosef (2000) and maximize the function by an approximate Fisher scoring procedure. In addition, we present a method for calculating empirical Bayes estimates of the random coefficients by importance sampling. In Section 3 we illustrate the use of these procedures on data from the AAPT, investigating relationships among reported recent alcohol use, parental monitoring, and adolescent rebelliousness. We give simulation results demonstrating the good performance of the estimation method in Section 4.

2. A MODEL FOR SEMICONTINUOUS LONGITUDINAL DATA

2.1 Model and Notation

Let Y_{ij} denote a semicontinuous response for subject $i = 1, \ldots, m$ at occasion $j = 1, \ldots, n_i$. This response can be recoded as two variables,

$$U_{ij} = \begin{cases} 1 \text{ if } Y_{ij} \neq 0 \\ 0 \text{ if } Y_{ij} = 0 \end{cases}$$

and

$$V_{ij} = \begin{cases} g(Y_{ij}) & \text{if } Y_{ij} \neq 0 \\ \text{irrelevant if } Y_{ij} = 0, \end{cases}$$

where g is a monotone increasing function (e.g., log) that will make V_{ij} approximately Gaussian. We model these responses by a pair of correlated random-effects models, one for the logit probability of $U_{ij} = 1$ and one for the mean conditional response $E(V_{ij} \mid U_{ij} = 1)$. The logit model is

$$\eta_i = X_i\beta + Z_i c_i, \tag{1}$$

where $\pi_{ij} = P(U_{ij} = 1)$, η_i is the vector with elements $\eta_{ij} = \log \pi_{ij}/(1 - \pi_{ij})$, $j = 1, \ldots, n_i$, and X_i $(n_i \times q_c)$ and Z_i $(n_i \times p_c)$ are matrices of covariates pertaining to the fixed and random effects. Times of measurement may be included in X_i

and possibly Z_i, allowing intercepts, slopes, and so on to vary by subject. The model for the continuous response is

$$V_i = X_i^* \gamma + Z_i^* d_i + \epsilon_i, \tag{2}$$

where V_i is the vector of length n_i^* containing all relevant values of V_{ij} for subject i, the values corresponding to $U_{ij} = 1$. The residuals ϵ_i are assumed to be distributed as $N(0, \sigma^2 I)$, and X_i^* $(n_i^* \times q_d)$ and Z_i^* $(n_i^* \times p_d)$ are matrices of covariates. The random coefficients from the two parts are assumed to be jointly normal and possibly correlated,

$$b_i = \begin{pmatrix} c_i \\ d_i \end{pmatrix} \sim N\left(0, \psi = \begin{pmatrix} \psi_{cc} & \psi_{cd} \\ \psi_{dc} & \psi_{dd} \end{pmatrix}\right). \tag{3}$$

It is possible for an individual to report no use at any time point, resulting in $n_i^* = 0$; such individuals contribute little to the estimation of γ, σ^2, ψ_{dd}, and ψ_{cd}. If $\psi_{cd} = 0$, then the two parts of the model separate, making U_{ij} independent of $V_{ij'}$ for all $j \neq j'$. In the context of substance use, separability would imply that the presence or absence of use at one occasion has no influence on the amount of use, if any, at other occasions. In our analyses we have found that this condition typically does not hold; random effects from two parts are usually correlated, often strongly.

In our model the same set of covariates may appear in the logit and linear parts, but this is not required. Even if the same covariates are used in both parts, it will not generally be true that $X_i^* = X_i$ and $Z_i^* = Z_i$, because part (2) applies only to those occasions where $Y_{ij} \neq 0$. Intercepts and slopes for either curve may be fixed or random, and additional static or time-varying covariates may be included in either one. Responses need not be recorded at the same set of time points for all individuals; the data may be unbalanced by design or have ignorably missing values. In many situations, the hypotheses of primary interest will focus on β and γ. It also may be useful to examine the covariances ψ among the subject-specific features and test hypotheses about them—for example, the hypothesis that the models are separable ($\psi_{cd} = 0$).

2.2 Computational Strategies

Estimation of β, γ, σ^2, and ψ under our model shares many of the computational challenges associated with generalized linear mixed models (GLMMs) (e.g., Stiratelli, Laird, and Ware 1984). One method for fitting GLMMs is Bayesian simulation using Markov chain Monte Carlo (MCMC) (Clayton 1996; Zeger and Karim 1991). A Gibbs sampler (Gelfand and Smith 1990) for our two-part model may be constructed by partitioning the joint distribution of $(\beta, \gamma, \sigma^2, b_1, \ldots, b_m, \psi)$ into conditional distributions. With convenient priors, the conditional distributions for (σ^2, γ) and ψ are simple to draw from, but the distributions for β and b_i are nonstandard and require additional steps. In preliminary work, we implemented a Gibbs sampler in Fortran embedded with steps of a Metropolis–Hastings algorithm (Hastings 1970; Tierney 1994) to sample from the conditional distributions of β and b_i. Despite attempts to optimize our program, when we applied it to data from the AAPT study we found that iterations were time-consuming and that the algorithm converged so slowly as to make the method rather impractical; runs of 24 hours or more were required for fitting even the simplest models. Slow convergence of Gibbs samplers for random-effects models has been noted by others (e.g., Carlin 1996); the worst performance occurs when m is large and the individual random effects are not well estimated. With smaller datasets, Bayesian procedures for this model could be implemented using the general-purpose MCMC programs WinBUGS (Spiegelhalter, Thomas, and Best 1999) or Bassist (Toivonen, Mannila, Seppanen, and Vasko 1999).

In earlier work, we also experimented with EM algorithms. EM is stable and straightforward to apply in normal linear random-effects models (Laird and Ware 1982), but with nonnormal responses both the E and the M steps become difficult. The E step requires the expectation of nonlinear quantities with respect to a nonstandard distribution, and the M step cannot be expressed in closed form. We implemented an EM procedure that approximates the E step by importance sampling (see Olsen and Schafer 1998 for details), a method similar to the Monte Carlo EM for GLMMs described by McCulloch (1997). Although the method produced accurate estimates, we again found that it was far too slow to be practicable for modeling data from the AAPT study.

GLMMs have also been fit by penalized quasi-likelihood (PQL), an extension of quasi-likelihood that incorporates random coefficients (Breslow and Clayton, 1993). When applied to binomial responses, PQL may severely underestimate both the fixed effects and the variance component parameters when the binomial denominator is small (McCulloch 1997, Rodriguez and Goldman 1995). Even recent corrections (Breslow and Lin 1995; Goldstein and Rasbash 1996; Lin and Breslow 1996) are unsatisfactory when the variance components are large; that is, when the variance among random intercepts is .5 or larger. Because of the large intersubject variation in adolescent substance use patterns, we were hesitant to apply PQL to the AAPT data, and instead focused on methods for approximating and maximizing the full likelihood function, which is sometimes called the "marginal likelihood" because the unobserved random effects have been integrated out.

Except for special cases in which the link function is the identity, the marginal likelihood for a GLMM involves an intractable integral that cannot be evaluated directly. Various strategies for approximating this integral have been proposed, including Gauss–Hermite quadrature (Anderson and Aitkin 1985), adaptive Gaussian quadrature (Pinheiro and Bates 1995), and Laplace approximations (Tierney and Kadane 1986). Hedeker and Gibbons (1994) used Gauss–Hermite quadrature and an approximate Fisher scoring algorithm to obtaining maximum likelihood (ML) estimates for an ordinal random-effects regression model. Raudenbush et al. (2000) approximated the log-likelihood for GLMMs by a sixth-order Laplace expansion. The Laplace method appears to have some important advantages; through simulation, Raudenbush et al. found it to be as accurate as quadrature but considerably faster, reducing computational time by about 95%. Moreover, the form of the programming required (e.g., the dimensionality of arrays) does not need to change as the number of random effects increases. The required derivatives simplify and are not especially cumbersome to compute.

We have extended the method of Raudenbush et al. to our two-part model, implementing an approximate scoring procedure to calculate ML estimates and standard errors. On convergence, we calculate empirical Bayes estimates of the random coefficients by an importance sampling method analogous to the E step of EM. Whereas MCMC and EM algorithms took more than a day to produce accurate estimates for the AAPT data, our Laplace procedure required than less than 1 minute on a 400-MhZ Pentium II computer.

2.3 The Likelihood

The marginal likelihood for the model defined by (1)–(3) can be expressed as

$$L \propto \prod_{i=1}^{m} \int \exp\{l_{U_i}\} \exp\{l_{V_i}\} p(b_i) db_i, \tag{4}$$

where $l_{U_i} = \sum_{j=1}^{n_i} (U_{ij}\eta_{ij} + \log(1-\pi_{ij}))$ comes from the logistic regression,

$$l_{V_i} = -\frac{n_i^*}{2}(\log\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n_i^*} \left(V_{ij} - (X_{ij}^*\gamma + Z_{ij}^* d_i)\right)^T \times \left(V_{ij} - (X_{ij}^*\gamma + Z_{ij}^* d_i)\right)$$

comes from the linear regression, and $p(b_i) = |\psi|^{-1/2} \exp\{(-1/2) b_i^T \psi^{-1} b_i\}$ arises from the joint normal distribution applied to the random effects. This third part captures possible correlations between the model's logit and linear parts.

For both conceptual and computational reasons, it is convenient to factor the joint distribution of random effects $p(b_i) = p(c_i, d_i)$ into the marginal distribution of c_i and the conditional distribution of d_i given c_i, yielding

$$L \propto \prod_{i=1}^{m} \int \exp\{l_{U_i}\} p(c_i) \left(\int \exp\{l_{V_i}\} p(d_i|c_i) dd_i \right) dc_i. \tag{5}$$

Because $V_i - X_i^*\gamma | Z_i^* d_i, \sigma^2 \sim N(Z_i^* d_i, \sigma^2 I)$ and $Z_i^* d_i | c_i, \psi \sim N(Z_i^* \psi_{dc} \psi_{cc}^{-1} c_i, Z_i^* H Z_i^{*T})$ for $n_i^* > 0$, it follows that $d_i | V_i, c_i, \psi, \gamma, \sigma^2 \sim N(\hat{d}_i, \sigma^2 B_i^{-1})$, where

$$\hat{d}_i = \psi_{dc}\psi_{cc}^{-1} c_i + B_i^{-1}\left(Z_i^{*T} A_i - Z_i^{*T} Z_i^* \psi_{dc}\psi_{cc}^{-1} c_i\right), \tag{6}$$

$H = \psi_{dd} - \psi_{dc}\psi_{cc}^{-1}\psi_{cd}$, $B_i = Z_i^{*T} Z_i^* + \sigma^2 H^{-1}$, and $A_i = V_i - X_i^*\gamma$. The inner integral in (5) can be evaluated directly by adding and subtracting $Z_i^* \hat{d}_i$ to l_{V_i} and $\hat{d}_i$ to $p(d_i|c_i)$. After algebraic simplification, the likelihood becomes

$$\begin{aligned} L \propto \prod_{i=1}^{m} & |\psi_{cc}|^{-\frac{1}{2}} |H|^{-\frac{1}{2}} \left(\sigma^2\right)^{-\frac{n_i^*}{2}} |\sigma^2 B_i^{-1}|^{\frac{1}{2}} \\ & \times \exp\left\{ -\frac{1}{2\sigma^2}(A_i^T A_i - A_i^T Z_i^* B_i^{-1} Z_i^{*T} A_i) \right\} \\ & \times \exp\left\{ \frac{1}{2}(E_i^{*T}(\psi_{cc}^{-1} + D_i^*)^{-1} E_i^*) \right\} \\ & \times \int \exp\left\{ l_{U_i} - \frac{1}{2}\left(c_i - (\psi_{cc}^{-1} + D_i^*)^{-1} E_i^*\right)^T \right. \\ & \left. \times \left(\psi_{cc}^{-1} + D_i^*\right)\left(c_i - (\psi_{cc}^{-1} + D_i^*)^{-1} E_i^*\right) \right\} dc_i, \end{aligned} \tag{7}$$

where $D_i^* = (\psi_{dc}\psi_{cc}^{-1})^T H^{-1} B_i^{-1} Z_i^{*T} Z_i^* (\psi_{dc}\psi_{cc}^{-1})$ and $E_i^* = (\psi_{dc}\psi_{cc}^{-1})^T H^{-1} B_i^{-1} Z_i^{*T} A_i$.

The remaining integral in (7) is intractable and has the same form as the likelihood for a logit model with normal random effects. The only difference is that the distribution of random effects in the integrand is shifted and scaled from what is usually expected. For those subjects with $n_i^* = 0$, the normal distribution of the random effects reduces to the standard of mean 0 and covariance matrix ψ_{cc}, and the terms relating to the model's linear part cancel out.

2.4 Laplace Approximation

As discussed in Section 2.2, various strategies for evaluating the integral in (7) are available. We apply a high-order multivariate Laplace approximation described by Raudenbush et al. (2000). For GLMMs, they found that a sixth-order approximation compared favorably to Gauss–Hermite quadrature, PQL, and adaptive Gaussian quadrature. We calculated values for l_i in (7) using both 20-point Gauss–Hermite quadrature and sixth-order Laplace and found the two comparable. The Laplace method seems easier to implement in general-purpose software, because unlike in quadrature, the nature of the calculations does not change dramatically with the changing dimension of c_i.

The Laplace method approximates an integral of the form $\int \exp\{f(c)\} dc$ by expanding the exponent $f(c)$ in a Taylor series about a value $\tilde{c}$ such that $f'(\tilde{c}) = 0$. The exponent in the integrand of (7),

$$f(c_i) = l_{U_i} - \frac{1}{2}\left(c_i - (\psi_{cc}^{-1} + D_i^*)^{-1} E_i^*\right)^T \times \left(\psi_{cc}^{-1} + D_i^*\right)\left(c_i - (\psi_{cc}^{-1} + D_i^*)^{-1} E_i^*\right),$$

has the form of a log-posterior density function for the coefficients of a logistic regression model with a normal prior. The mode $\tilde{c}_i$ of this log density can be found by iteratively solving the equation

$$\tilde{c}_i = (\psi_{cc}^{-1} + D_i^* + Z_i^T \tilde{W}_i Z_i)^{-1} (Z_i^T \tilde{W}_i (U_i^* - X_i\beta) + E_i^*), \tag{8}$$

where $U_i^* = \tilde{W}_i^{-1}(U_i - \tilde{\pi}_i) + \tilde{\eta}_i$, W_i is a diagonal matrix with elements $\pi_{ij}(1-\pi_{ij})$, and π_i is a vector with elements π_{ij}, $j = 1, \ldots, n_i$. In these expressions, any symbol with a tilde (e.g., $\tilde{W}_i$) is evaluated at $c_i = \tilde{c}_i$.

Using a notation developed by Raudenbush et al. (2000), the multivariate Taylor expansion of $f(c_i)$ about $\tilde{c}_i$ is

$$\begin{aligned} & \int \exp\{f(c_i)\} dc_i \\ & = \int \exp\left\{ f(\tilde{c}_i) + f^{(1)}(\tilde{c}_i)(c_i - \tilde{c}_i) \right. \\ & \quad + \frac{1}{2!}(c_i - \tilde{c}_i)^T f^{(2)}(\tilde{c}_i)(c_i - \tilde{c}_i) \\ & \quad + \cdots + \frac{1}{k!}\left[\otimes^{(k-1)}(c_i - \tilde{c}_i)^T\right] \\ & \quad \left. \times f^{(k)}(\tilde{c}_i)(c_i - \tilde{c}_i) + \ldots \right\} dc_i, \end{aligned} \tag{9}$$

where the derivatives are defined recursively as $f^{(k)}(c_i) = \partial \text{vec} f^{(k-1)}/\partial c_i^T$ and $\otimes^k c_i = c_i \otimes c_i \otimes \ldots \otimes c_i$ is an iterated Kronecker product with k terms. Because $\tilde{c}_i$ is the mode, $f^{(1)}(\tilde{c}_i) = 0$, and (9) reduces to

$$\int \exp\{f(c_i)\}dc_i = \exp(f(\tilde{c}_i))\left|-f^{(2)}(\tilde{c}_i)\right|^{-1/2} E(\exp(R_i)),$$

where the expectation is taken with respect to the normal distribution $c_i - \tilde{c}_i \sim N(0, -[f^{(2)}(\tilde{c}_i)]^{-1})$ and the remainder is $R_i = \sum_{k=3}^{\infty} T_{ik}$ with

$$T_{ik} = \frac{1}{k!}[\otimes^{(k-1)}(c_i - \tilde{c}_i)^T] f^{(k)}(\tilde{c}_i)(c_i - \tilde{c}_i).$$

If R_i is close to 0, then $\exp(R_i)$ is well approximated by the initial terms of its Maclaurin expansion $1 + R_i + R_i^2/2 + \cdots$. Substituting $R_i = \sum_{k=3}^{\infty} T_{ik}$ into this series gives

$$\exp(R_i) \approx 1 + T_{i3} + T_{i4} + T_{i5} + T_{i6} + \frac{1}{2}T_{i3}^2, \tag{10}$$

and taking expectations yields

$$\int \exp\{f(c_i)\} \approx \exp(f(\tilde{c}_i))\left|-[f^{(2)}(\tilde{c}_i)]^{-1}\right|^{\frac{1}{2}} \times \left(1 + E(T_{i4}) + E(T_{i6}) + \frac{1}{2}E(T_{i3}^2)\right), \tag{11}$$

because $E(T_{ik}) = 0$ for odd values of k. Simple expressions for the even moments can be derived as follows. The fourth-order term is

$$E[T_{i4}] = \frac{(4-3)(4-1)}{4!} \times \text{vec}^T\left(G_i^{-1} \otimes \{\text{vec } G_i^{-1}\}\right)\text{vec}\left[f^{(4)}(\tilde{c}_i)\right], \tag{12}$$

where $G_i = -f^{(2)}(\tilde{c}_i) = Z_i^T \tilde{W}_i Z_i + \psi_{cc}^{-1} + D_i^*$. But Raudenbush et al. showed that for $k \geq 3$,

$$f^{(k)}(\tilde{c}_i) = -\sum_{j=1}^{n_i} (\otimes^{k-1} Z_{ij}^T)\tilde{m}_{ij}^{(k)} Z_{ij},$$

where $m_{ij}^{(k)}$ is the $(k-1)$th derivative of π_{ij} with respect to η_{ij} evaluated at $\tilde{c}_i$. For example, if $\tilde{w}_{ij} = \tilde{\pi}_{ij}(1 - \tilde{\pi}_{ij})$, then $\tilde{m}_{ij}^{(4)} = \tilde{w}_{ij}(1 - 6\tilde{w}_{ij})$. Applying this result, (12) simplifies to

$$E[T_{i4}] = -\frac{1}{8}\sum_{i=1}^{n_i} \tilde{m}_{ij}^{(4)} \left(Z_{ij}^T G_i^{-1} Z_{ij}\right)^2.$$

The remaining moments, which we state without proof, are

$$E(T_{i6}) = -\frac{1}{48}\sum_{i=1}^{n_i} \tilde{m}_{ij}^{(6)} \left(Z_{ij} G_i^{-1} Z_{ij}^T\right)^3$$

and

$$E(T_{i3}^2) = \frac{15}{72}\left(\sum_{j=1}^{n_i} Z_{ij}^T \tilde{m}_{ij}^{(3)} Z_{ij} G_i^{-1} Z_{ij}^T\right)^T [-2pt] \times G_i^{-1}\left(\sum_{j=1}^{n_i} Z_{ij}^T \tilde{m}_{ij}^{(3)} Z_{ij} G_i^{-1} Z_{ij}^T\right).$$

Substituting these expressions into (11) yields our final expression for the approximate log-likelihood,

$$\begin{aligned} l \approx & -\frac{m}{2}\log|H| - \frac{m}{2}\log|\psi_{cc}| - \frac{N^*}{2}\log\sigma^2 \\ & + \frac{1}{2}\sum_{i=1}^{m}\log\left|-f^{(2)}(\tilde{c}_i)\right| + \frac{1}{2}\sum_{i=1}^{m}\log\left|\sigma^2 B_i^{-1}\right| \\ & - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(A_i^T A_i - A_i^T Z_i^* B_i^{-1} Z_i^{*T} A_i) \\ & + \frac{1}{2}\sum_{i=1}^{m}(E_i^{*T}(\psi_{cc}^{-1} + D_i^*)^{-1} E_i^*) \\ & + \sum_{i=1}^{m} f(\tilde{c}_i) + \sum_{i=1}^{m}\log A_i^*, \end{aligned} \tag{13}$$

where $N^* = \sum_{i=1}^{m} n_i^*$ and

$$\begin{aligned} A_i^* = & 1 - \frac{1}{8}\sum_{j=1}^{n_i}\tilde{m}_{ij}^{(4)}\left(Z_{ij}G_i^{-1}Z_{ij}^T\right)^2 - \frac{1}{48}\sum_{j=1}^{n_i}\tilde{m}_{ij}^{(6)}\left(Z_{ij}G_i^{-1}Z_{ij}^T\right)^3 \\ & + \frac{15}{72}\left(\sum_{j=1}^{n_i} Z_{ij}^T \tilde{m}_{ij}^{(3)} Z_{ij} G_i^{-1} Z_{ij}^T\right)^T \\ & \times G_i^{-1}\left(\sum_{j=1}^{n_i} Z_{ij}^T \tilde{m}_{ij}^{(3)} Z_{ij} G_i^{-1} Z_{ij}^T\right). \end{aligned}$$

The accuracy of this approximation hinges on (10). In an asymptotic sequence where n_i increases, $c_i - \tilde{c}_i = O_p(n_i^{-1/2})$ and the terms omitted from (10) are $O_p(n_i^{-5/2})$ or smaller. In panel studies with attrition and nonresponse, the value of n_i may be as small as 1 for some individuals, and these asymptotic arguments have little value. But even for small n_i, the approximation (10) would still be reasonable if $f(c_i)$ resembled a low-order polynomial. The presence of a normal log-density kernel in $f(c_i)$ does seem to help in this regard, contributing a quadratic component whose influence is greatest when n_i is small. Additional terms beyond the sixth order could be included in (10) for greater precision, but empirical evidence suggests that this is not necessary. Through simulation, Raudenbush et al. concluded that sixth-order terms substantially improve the accuracy but eighth-order terms do not. In our data examples, we have found that the sixth-order terms contribute between .5% and 5% of the approximate log-likelihood l_i for any subject.

2.5 Approximate Fisher Scoring Algorithm

In the Newton–Raphson procedure, a function $l(\theta)$ is maximized by repeated application of $\theta^{(t+1)} = \theta^{(t)} + C^{-1}S_\theta$, where $C = -\partial^2 l/\partial\theta\partial\theta^T$, $S_\theta = \partial l/\partial\theta$, and the derivatives are evaluated at $\theta = \theta^{(t)}$. In well-behaved statistical applications where l is a log-likelihood and θ represents unknown parameters, Newton–Raphson converges to an ML estimate $\hat{\theta}$, and first-order asymptotic theory allows the final value of C^{-1} to be used as an estimate of $V(\hat{\theta} - \theta)$. In practice, it is not necessary to use the exact second derivatives of l; the asymptotic properties still hold if $C = -\partial^2 l/\partial\theta\partial\theta^T$ is replaced by $C = -\partial^2 l/\partial\theta\partial\theta^T + R$, provided that $R = o_p(n)$, where n is proportional to the sample size (e.g., Cox and Hinkley 1974,

Chap. 9). When $C = -E(\partial^2 l/\partial\theta\partial\theta^T)$, the technique is called Fisher scoring. In situations where the second derivatives of the log-likelihood are troublesome to calculate, the well-known identity $E(\partial^2 l/\partial\theta\partial\theta^T) = -E[(\partial l/\partial\theta)(\partial l/\partial\theta)^T]$ suggests an approximate scoring procedure with $C = \sum_i S_{\theta i} S_{\theta i}^T$, where $S_{\theta i} = \partial l_i/\partial\theta$ is the ith subject's contribution to the score vector (Bock and Lieberman 1970; Hedeker and Gibbons 1994; Raudenbush et al. 2000).

In Section 2.3 the covariance matrix ψ was reexpressed as ψ_{cc}, $\psi_{dc}\psi_{cc}^{-1}$, and $H = \psi_{dd} - \psi_{dc}\psi_{cc}^{-1}\psi_{cd}$. We have found it convenient to apply the scoring method to $\psi_{dc}\psi_{cc}^{-1}$ and the free elements (upper triangles) of ψ_{cc}^{-1} and H. Consequently, we have $\theta = (\beta^T, \gamma^T, \sigma^2, \text{vec}^T(\psi_{dc}\psi_{cc}^{-1}), \phi_{cc}^T, \phi_H^T)^T$, where ϕ_{cc} and ϕ_H represent the vectorized upper triangles of ψ_{cc}^{-1} and H^{-1}. Differentiating the approximate log-likelihood is somewhat tedious, because we must take into account the dependence of $f(\tilde{c}_i)$ on $\tilde{c}_i$ as shown in (8). Expressions for the components of the score vector for the ith subject, $S_{\theta i} = (S_{\beta i}^T, S_{\gamma i}^T, S_{\sigma^2 i}^T, S_{\psi_{dc}\psi_{cc}^{-1} i}^T, S_{\phi_{cc} i}^T S_{\phi_H i}^T)^T$, are given in the Appendix. It is possible for scoring-updated estimates to stray outside the parameter space. At each iteration, we check the new values of σ^2, ψ_{cc}^{-1}, and H to see whether they are positive; if any are not, then a step-halving procedure is applied to bring them back to allowable values.

This approximate scoring procedure has been implemented in a Fortran-90 program that is available at http://methcenter.psu.edu. In this program, a default starting value for β is generated by logistic regression of U_i on X_i, and starting values for γ, σ^2, and ψ_{dd} are estimated using a quick procedure for linear mixed models (Schafer 1998), which would give exact ML estimates if $\hat{\psi}_{cd}$ were 0. For simplicity, the starting values for ψ_{cc} and ψ_{cd} are taken to be I and 0. This program provides an option for fitting a restricted model with $\psi_{cd} = 0$; under this condition, the logit and linear parts separate.

2.6 Empirical Bayes Estimates

Viewing (3) as a prior distribution for b_i, approximate inferences for b_i can be obtained by calculating a posterior mean $E(b_i \mid \beta, \gamma, \sigma^2, \psi, Y_i)$ and covariance matrix $V(b_i \mid \beta, \gamma, \sigma^2, \psi, Y_i)$ with the unknown β, γ, σ^2, and ψ replaced by ML estimates. These empirical Bayes (EB) estimates are not an automatic byproduct of the scoring procedure. The integrals required for posterior moments could be approximated by Laplace's method, but the calculations would be somewhat different from those in Section 3.3. The integrand for a posterior moment contains additional terms involving c_i, and the form is no longer that of a posterior density function for the coefficients of a logistic model with a normal prior. In lieu of another set of Laplace-type expansions, we approximate EB estimates by importance sampling.

Consider first the posterior expectations of c_i and $c_i c_i^T$, which are of the form

$$\begin{aligned} &E(h(c_i)|\beta, \gamma, \sigma^2, \psi, Y_i) \\ &\quad = \int h(c_i) p(b_i|\beta, \gamma, \sigma^2, \psi, Y_i) dc_i \\ &\quad = \int \Big(h(c_i) p(c_i|\beta, \psi, Y_i) \\ &\qquad \times \int p(d_i|c_i, \beta, \gamma, \sigma^2, \psi, Y_i) dd_i \Big) dc_i \\ &\quad \propto \int h(c_i) q(c_i|\beta, \gamma, \sigma^2, \psi, Y_i) dc_i. \end{aligned}$$

The marginal density $q(c_i|\beta, \gamma, \sigma^2, \psi, Y_i)$ is unnormalized and nonstandard, so we apply importance sampling for unnormalized densities (e.g., Gelman, Carlin, Stern, and Rubin 1995). If $c_i^{(1)}, \ldots, c_i^{(K)}$ is a random sample from a density $g(c_i)$, then

$$E(h(c_i)) \approx \frac{\frac{1}{K}\sum_{k=1}^K h(c_i^{(k)}) w(c_i^{(k)})}{\frac{1}{K}\sum_{k=1}^K w(c_i^{(k)})},$$

where $w(c_i^{(k)}) = q(c_i^{(k)}|\beta, \gamma, \sigma^2, \psi, Y_i)/g(c_i^{(k)}|\beta, \gamma, \sigma^2, \psi, Y_i)$ are the importance ratios. The density g should be chosen to approximate the shape of q so that the importance ratios will be fairly constant. Following the advice of Gelman et al., we apply a multivariate t distribution with 4 degrees of freedom, centered at the mode $\tilde{c}_i$ with covariance matrix proportional to G_i^{-1}. Once the sample of c_i values is drawn, we obtain estimated means not only for c_i and $c_i c_i^T$, but also for d_i, $c_i d_i^T$, and $d_i d_i^T$. The latter follow by the results of Section 3.2 pertaining to the conditional normal posterior distribution of d_i given c_i. For example, $E(d_i) = E(E(d_i|c_i))$, so an approximate posterior mean for d_i is

$$\frac{\frac{1}{K}\sum_{k=1}^K \hat{d}_i^{(k)} w(c_i^{(k)})}{\frac{1}{K}\sum_{k=1}^K w(c_i^{(k)})}, \tag{14}$$

where $\hat{d}_i$ is defined in (6) as a function of c_i. For those subjects with $n_i^* = 0$, $E(d_i)$ reduces to $\psi_{dc}\psi_{cc}^{-1}E(c_i)$.

3. EXAMPLES

3.1 Description of Data

The data for these examples come from one panel of the Adolescent Alcohol Prevention Trial (AAPT), a longitudinal study of middle school and high school students in Los Angeles and Orange Counties in California (Hansen and Graham 1991). In this panel students responded to questionnaires in grade 7 and yearly thereafter until grade 11. These questionnaires contained items pertaining to recent and lifetime use of alcohol and other controlled substances, attitudes toward school, personality traits, and family characteristics. Using our two-part random-effects model, we explore the relationships between reported recent alcohol use over grades 7–11 and two variables measured in grade 7: degree of parental monitoring and rebelliousness. Low monitoring and high rebelliousness are commonly cited as potential risk factors for illicit substance use. It may be, however, that a risk factor operates quite differently on the probability of alcohol use and the amount of alcohol consumption when it occurs. Because adolescent boys and girls are known to differ appreciably in their patterns of substance use, gender plays an important role in each model.

To reduce burden on respondents, the AAPT used a multiform design in which groups of items were randomly omitted from some of the questionnaires. Items pertaining to monitoring and rebelliousness were placed in the same group and

were missing together by design for one-third of the sample. Removing subjects with missing values for these left us with $m = 1,961$ individuals with complete grade 7 data for the response variable and the potential risk factors.

Our response is a composite measure formed by taking the mean of three items: (1) "How many alcoholic drinks have you had in the past month (30 days)?"; (2) "How many alcoholic drinks have you had in the past week (7 days)?"; (3) and "How many days in the past month (30 days) have you had alcohol to drink? (Do not count religious service.)". The first item has nine response categories ranging from "none" and "only sips for religious service" to "more than 100." The second has eight categories ranging from "none" and "only sips for religious service" to "11 or more." The third has six categories ranging from "none" to "15–30." In our composite measure, $Y_{ij} = 0$ represents no use or sips for religious purposes only. The distribution of the response at each grade for $m = 1,961$ subjects is shown in Figure 2; the missing values are denoted by NA. To help reduce skewness, we applied a log transformation, taking $V_{ij} = \log Y_{ij}$ if $Y_{ij} > 0$.

The missing values in the response arise from a combination of absenteeism, attrition, failure to complete the questionnaire on time, and other factors. Analyses based on the likelihood function (4) are appropriate if these missing values are ignorably missing or missing at random (MAR) in the sense defined by Rubin (1976), which means that probabilities of missingness are unrelated to missing values. In longitudinal studies with substantial rates of attrition, the reasonableness of MAR should be carefully investigated. Most of the attrition in the AAPT came from students moving to another school district or to another, non-AAPT school within the district. Followup evaluations described by Graham, Hofer, and Piccinin (1994) suggest that in the vast majority of cases, data were missing for reasons that could be at most only weakly correlated with substance use, lending support to the plausibility of MAR. The missing-data patterns seen in the sample and the percentage of subjects in each pattern are displayed in Table 1. In this table "1" denotes an observed response and "0" denotes a missing response. Approximately one-third of the subjects responded at all five time points, one-third dropped out of the study prior to the 11th grade, and the remaining one-third were missing at one or more grades but returned in subsequent years.

3.2 Parental Monitoring

Our measure of parental monitoring is a standardized composite of three items recorded in grade 7: (1) "When you go out with your friends, how often do your parents tell you what time to be home?"; (2) "How often do your parents refuse to let you go places and do things with other people your age?"; (3) "How often do your parents ask where you are going when you leave the house?" In this model the columns of X_i and X_i^* are identical and represent an intercept; the main effects of sex, time, and parental monitoring; all possible two-way interactions among them; and the three-way interaction. Time is coded from 0=grade 7 to 4=grade 11, and gender is coded as 0=female and 1=male. The resulting intercept is a mean response for girls at grade 7, and the coefficient for time is an average yearly increase for girls. The mean response for boys at grade 7 is obtained by adding the intercept to the main effect for gender, and the average yearly increase for boys is the sum of the main effect for time and the time-by-gender interaction. Higher values for the monitoring variable denote greater levels of supervision. Including two- and three-way interactions

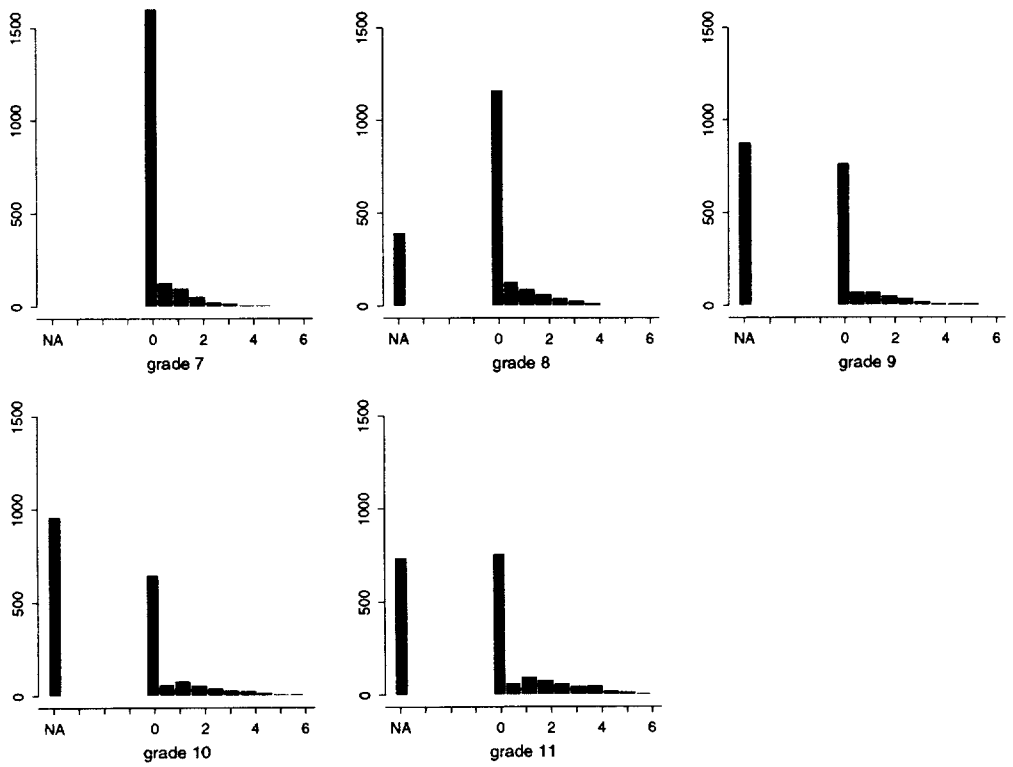

Figure 2. Histograms of Reported Recent Alcohol Use.

Table 1. Missing Data Patterns for Recent Alcohol Use in Grades 7–11

Grade 7	Grade 8	Grade 9	Grade 10	Grade 11	%
1	1	1	1	1	34.8
1	0	1	1	1	3.5
1	1	0	1	1	5.7
1	0	0	1	1	1.1
1	1	1	0	1	6.1
1	0	1	0	1	1.5
1	1	0	0	1	7.5
1	0	0	0	1	2.0
1	1	1	1	0	3.0
1	0	1	1	0	.5
1	1	0	1	0	1.9
1	0	0	1	0	.3
1	1	1	0	0	4.5
1	0	1	0	0	1.0
1	1	0	0	0	16.2
1	0	0	0	0	10.3

allows us to see whether the effect of early parental monitoring on adolescent alcohol use evolves differently over time for boys and girls.

With a maximum of only $n_i = 5$ time points per subject, the data contain little information about the time slopes for individuals. We tried models that allowed the slopes to vary by subject, both in the logit and linear parts, but in each case the estimated variances of the slopes approached 0. Our final model allows only intercepts to vary by subject, so that Z_i and Z_i^* are simply vectors of 1's. The scoring procedure converged in 38 iterations to a maximum relative parameter change of .001. Execution on a Pentium II 400-Mhz workstation, including calculation of starting values, took 52 seconds. Estimates and standard errors for β and γ are shown in Table 2, and estimates for the unique elements of ψ are $\hat{\psi}_{cc} = 3.492$, $\hat{\psi}_{cd} = .648$, and $\hat{\psi}_{dd} = .214$.

To better understand the role of parental monitoring, we calculated the estimated effect of a one–standard unit decrease in monitoring at each grade and gender. For the logit model, we expressed the effect as an odds ratio; for the linear model, the effect is a mean difference. These two sets of effects are plotted in Figure 3. Consider first the plot of the odds ratio in Figure 3(a). For girls, low monitoring in grade 7 appears to substantially increase the odds of alcohol use in grade 7 (odds ratio $\approx$ 1.5), but the effect diminishes rapidly over time;

Table 2. Parameter Estimates and Standard Errors of the Fixed Effects for Model With Grade 7 Parental Monitoring as the Predictor of Interest

	$\hat{\beta}$	$SE(\hat{\beta})$	$\hat{\gamma}$	$SE(\hat{\gamma})$
Intercept	-1.941	.101	-.282	.050
Sex	-.353	.145	.0417	.061
Parental Monitoring	-.415	.136	-.057	.051
Time	.351	.029	.171	.014
Parental Monitoring × Sex	.461	.206	-.045	.081
Time × Sex	.063	.044	.014	.022
Time × Parental Monitoring	.161	.042	-.010	.020
Time × Parental Monitoring × Sex	-.220	.065	.037	.030

by grade 11, the girls with less monitoring actually had *lower* odds of use than their highly supervised counterparts (odds ratio $\approx$ .8). For boys, however, the effects are nearly opposite. Low monitoring for boys in grade 7 appears to have no effect on the rate of alcohol use in grade 7, but by grade 11 the unsupervised boys are using at substantially higher rates than the highly supervised boys (odds ratio $\approx$ 1.2). Because a nonlinear link function has been applied, the curves shown represent subject-specific rather than population-averaged odds of use (e.g. Diggle, Liang, and Zeger 1994).

Figure 3(b), which shows the effect of low monitoring on reported amounts of consumption, tells quite a different story. In grade 7, reduced monitoring appears to increase the amount of alcohol consumption when it occurs for both boys and girls. But for girls this effect increases over time, whereas for boys it vanishes by grade 11. This example clearly demonstrates the utility of two-part modeling; the discrepancies between Figures 3(a) and 3(b) show substantial differences in the effect of monitoring on probability of alcohol use and the amount of use over time.

3.3 Rebelliousness

In this second example, our predictor of interest is a standardized composite of three items from grade 7: (1) "How much of what you learn in school is a waste of time?"; (2) "Is it worth getting into trouble if you have fun?"; (3) "How often do you do things you've been told not to do?" Higher values of this predictor denote increasing degrees of rebelliousness. The matrices X_i and X_i^* are coded as in the previous example, except that rebelliousness has now replaced monitoring. With this model, we were able to estimate random slopes for the linear part but not for the logit part. Consequently, each Z_i was

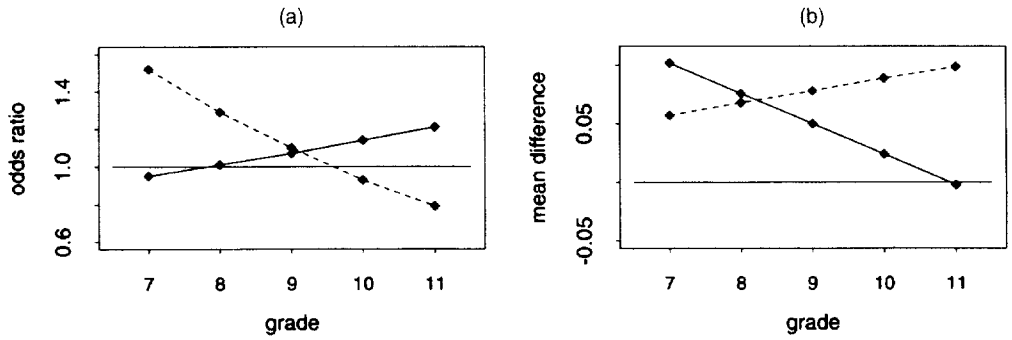

Figure 3. The Effects of Low Grade 7 Parental Monitoring on the Odds of Use (a) and the Mean of Log Amount of Use (b) for Girls (- - -) and Boys (—).

a vector of 1's, and each Z_i^* contained a column of 1's and a column for time. Intra-individual variation is described by three random effects (two intercepts and one slope), and ψ is 3×3. Estimation took 46 iterations and 70 seconds on a Pentium II 400-Mhz workstation. Estimates and standard errors for β and γ are shown in Table 3.

The estimate for ψ in this model is

$$\hat{\psi} = \begin{pmatrix} 2.61 & .328 & .004 \\ .328 & .233 & -.0032 \\ .004 & -.0032 & .0015 \end{pmatrix}.$$

The variance for the random slope appears small, suggesting that a model with only random intercepts might fit nearly as well. We can test the null hypothesis that only random intercepts are needed by fitting the reduced model and examining the decrease in log-likelihood. The log-likelihood values for the larger and smaller models are $-1,966.4045$ and $-1,971.4345$, so the likelihood ratio (LR) test statistic is 10.06. The models differ by three parameters, because under the null hypothesis $\psi_{13} = \psi_{23} = \psi_{33} = 0$. It is important to note, however, that null hypothesis lies on the boundary of the parameter space defined by the alternative, so the standard chi-squared approximation for the LR test does not apply. Under the null hypothesis, the limiting distribution of this LR test statistic is actually a 50 : 50 mixture of χ_3^2 and χ_2^2 (Self and Liang 1987; Stram and Lee 1994). In practice, data analysts rarely make decisions about random effects solely on the basis of a test statistic; they also consider whether the estimate under the larger model lies on the boundary. If it does, then they typically accept the null hypothesis; if it does not, then they may examine the log-likelihood difference. Under this modified testing procedure, the appropriate reference distribution becomes a 50 : 50 mixture of χ_3^2 and $\chi_0^2 = 0$. Adopting this rule, our p value becomes $P(\chi_3^2 \geq 10.06)/2 \approx .01$, so it appears that random slopes are supported.

It is also useful to test whether the logit and linear parts are separable ($\psi_{cd} = 0$) to see whether a joint fitting procedure is necessary. This null hypothesis does not lie on the boundary, so the test is standard. Our estimate $\hat{\psi}_{cd} = (.328, .004)$ is highly significant and shows that a subject's probability of use at one occasion is positively correlated with his or her level of use at other occasions. To illustrate, we calculated EB estimates for individuals' log odds of use at grade 7 and expected amount of use, if any, at grade 8. A scatterplot of these two quantities is shown in Figure 4. In this plot the points that cluster along lines represent EB estimates for boys and girls having no observed use at any occasion ($n_i^* = 0$). Because no data are available to estimate $E(V_{ij} \mid U_{ij} = 1)$ for these subjects specifically, the EB estimates for d_i revert to population-level regression predictions based on their estimates for c_i, and the standard errors for these estimates of d_i are quite large. Based on this plot, it appears that individuals with low propensities to use alcohol also tend to consume less when they do use it. Ignoring this relationship by fitting the logit and linear parts separately could introduce substantial bias into the estimated coefficients, particularly for γ.

Table 3. Parameter Estimates and Standard Errors of the Fixed Effects for the Model With Grade 7 Rebelliousness as the Predictor of Interest

	$\hat{\beta}$	$SE(\hat{\beta})$	$\hat{\gamma}$	$SE(\hat{\gamma})$
Intercept	-1.937	.098	-.266	.069
Sex	-.462	.142	.035	.071
Rebel	1.381	.131	.431	.062
Time	.366	.029	.169	.024
Rebelliousness × Sex	-.133	.182	-.087	.089
Time × Sex	.072	.045	.012	.026
Time × Rebelliousness	-.216	.041	-.065	.022
Time × Rebelliousness × Sex	.142	.061	.033	.032

3.4 Assessing Model Adequacy

Our models make a number of assumptions that could be investigated: normality of b_i and the elements of ϵ_i, linear relationships between covariates and the logit-probability of use, and linear relationships between covariates and the mean log amount of use. For normal mixed-effects models, only a few formal diagnostics have been developed (Lange and Ryan 1989), and practitioners often rely on informal techniques such as normal quantile plots of the estimated random effects (e.g. Pinheiro and Bates 2000). Diagnostics for GLMMs are even more scarce.

To detect any large discrepancies in fit, one anonymous reviewer suggested comparing the observed values for $U_{i\cdot} = \sum_{j=1}^{n_i} U_{ij}$ and $V_{i\cdot} = \sum_{j=1}^{n_i^*} V_{ij}$ with their predicted values $\hat{U}_{i\cdot}$ and $\hat{V}_{i\cdot}$ obtained by substituting ML estimates for β and γ and EB estimates for c_i and d_i. Figure 5 plots $V_{i\cdot}$ versus $\hat{V}_{i\cdot}$ from the parental monitoring model. We see that the $\hat{V}_{i\cdot}$ tend to lie above the observed values at the low end of the scale and below the observed values at the high end; this is to be expected, because EB procedures tend to smooth estimates toward central values. The scaled residuals $(V_{i\cdot} - \hat{V}_{i\cdot})/\sqrt{n_i^* \hat{\sigma}^2}$ all lie between -2.5 and 2.5, suggesting no extreme outliers. But the curved pattern in Figure 5 may indicate that some transformation other than the log might be more suitable for $V_{ij} = g(Y_{ij})$, or that the increases in the mean of V_{ij} over time might be nonlinear. A plot created from simulated data (not shown), following the model's distributional assumptions and using parameter values similiar to those estimated here, had an appearance similar to that of Figure 5 but with slightly less curvature. The plot of $U_{i\cdot}$ versus $\hat{U}_{i\cdot}$, displayed in Figure 6, also shows a pattern consistent with EB smoothing, overestimating the probability of use for individuals reporting no use at any occasion and underestimating for those reporting use at every occasion. The scaled residuals $(U_{i\cdot} - \hat{U}_{i\cdot})/\sqrt{\hat{U}_{i\cdot}(n_i - \hat{U}_{i\cdot})/n_i}$ all have magnitude below 2.5. In principle, one could check the adequacy of the logit link function using a goodness-of-link test for binary data as described by Collett (1991), but a significant result would not necessarily provide sensible suggestions for alternative links. Formal and informal procedures for diagnosing lack of fit in this model are an important topic for future research.

4. SIMULATIONS

4.1 Performance of Estimated Fixed Effects

Because we are maximizing an approximate log-likelihood, it is not clear how close our estimates lie to those that would

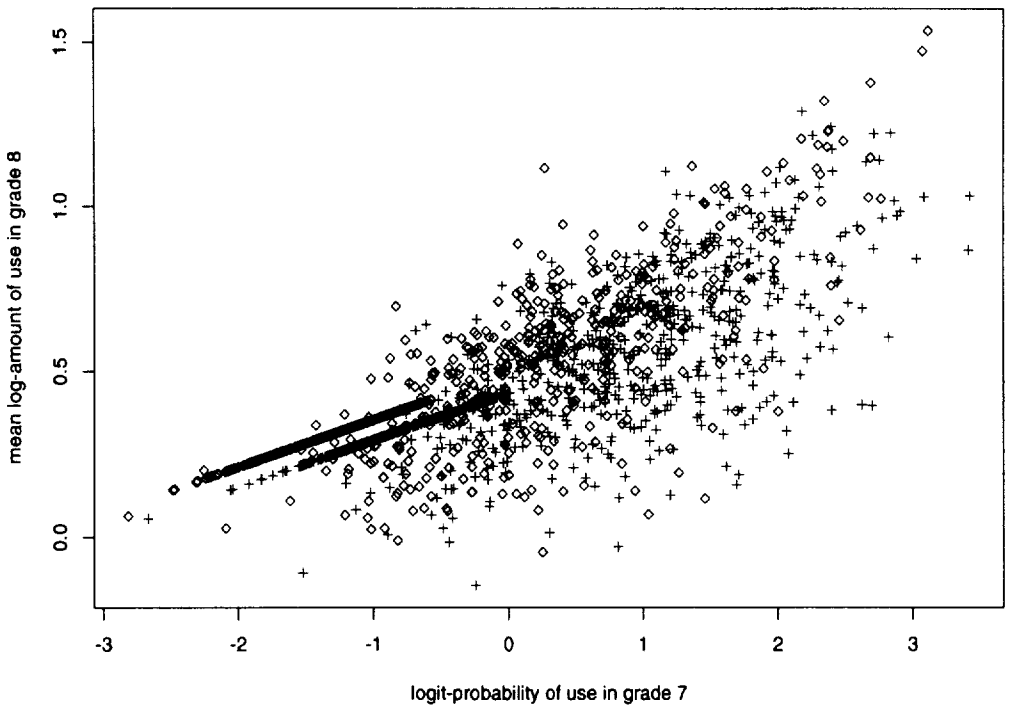

Figure 4. Individual Levels of Mean Log Amount of Use Versus Logit Probability of Use for Boys (◇) and Girls (+) at a High Level of Grade 7 Rebelliousness.

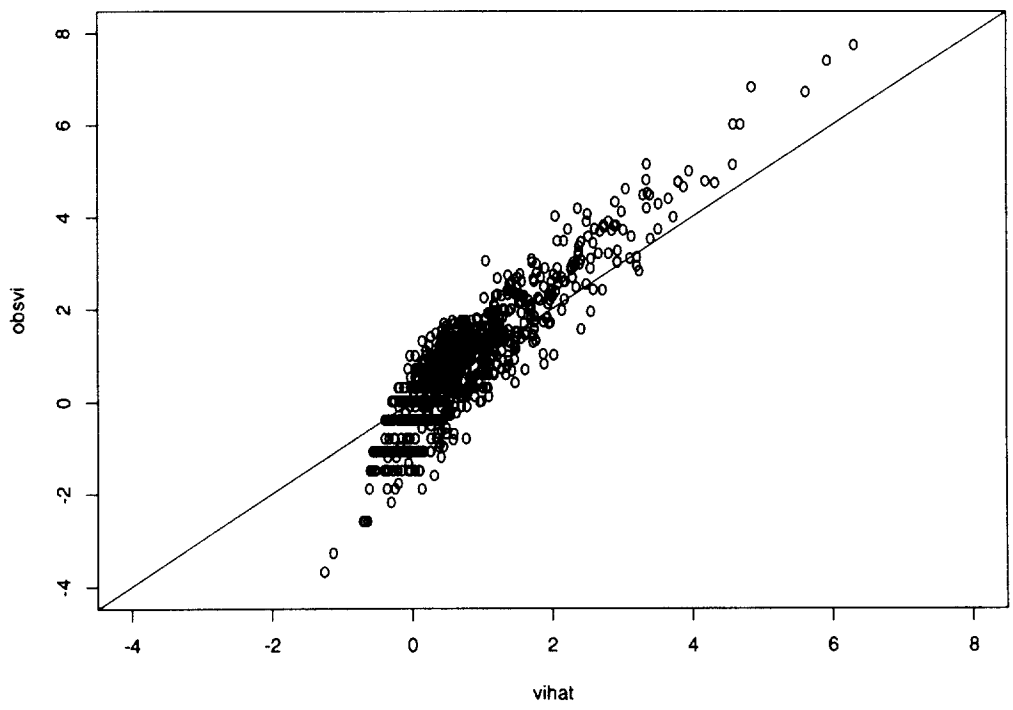

Figure 5. Observed Log Amount of Use Versus Expected Log Amount of Use.

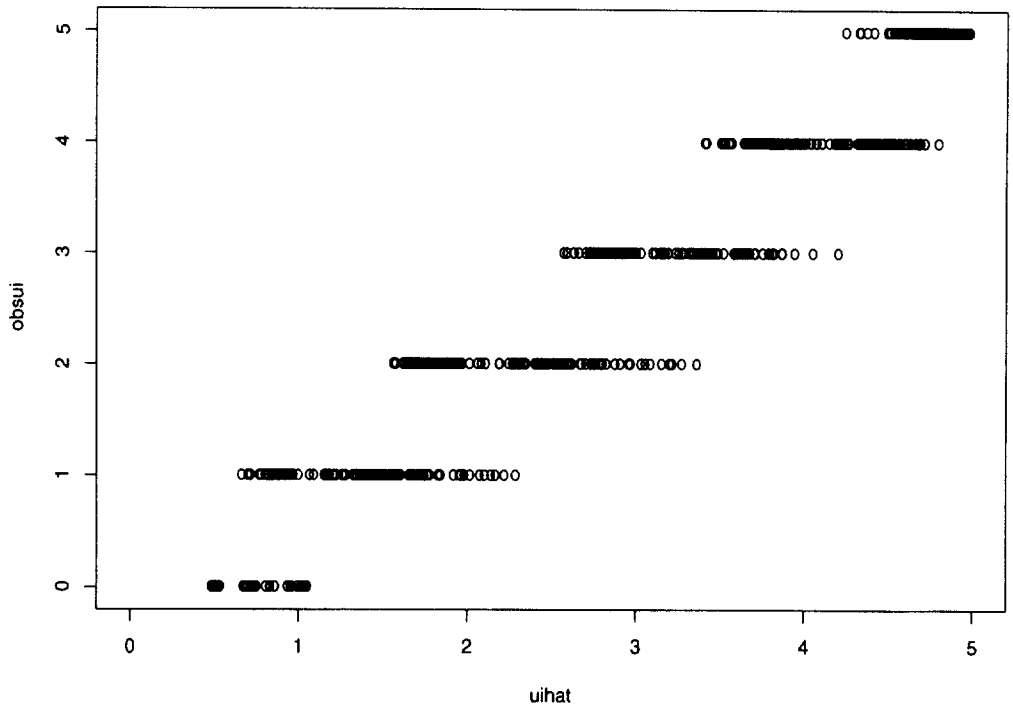

Figure 6. Observed Number of Times Reporting Use Versus Expected Probability of Use.

be obtained from the true log-likelihood. Investigating how errors in Laplace approximation propagate to the estimates would be theoretically interesting, but these errors would impact the quality of inferences only if they approached the same order of magnitude as the sampling errors of the ML estimates themselves. From a user's standpoint, a more salient issue is how the mode of the *approximate* log-likelihood behaves as an estimate of the population parameters. Other important questions focus on the quality of the standard errors obtained from the approximate scoring procedure and on the behavior of LR test statistics calculated from the approximate log-likelihood.

To investigate these issues, we simulated data from a population that could be regarded as typical of potential growth modeling applications. For each subject, X_i and X_i^* were constructed with three columns: a constant equal to 1, a dummy indicator for a non-time-varying covariate (e.g., sex) drawn from Bernoulli (.5), and a time-varying covariate (e.g., time) taking values $0, 1, \ldots, n-1$, where n is the number of occasions in the hypothetical study. The matrices Z_i and Z_i^* were each set to be columns of 1's, so that intercepts would randomly vary by subject in both parts of the model. The population fixed effects were set to $\beta = (-1, -.5, .4)^T$ and $\gamma = (-.3, .1, .4)^T$, and the variance parameters were $\psi_{cc} = 1$, $\psi_{cd} = .2$, $\psi_{dd} = .5$, and $\sigma^2 = .5$. Note that ψ_{cc} is smaller than the estimates from the AAPT examples but large enough so that PQL methods would perform poorly.

In our simulation the number of subjects and the number of occasions were varied in a 2×2 design with $m = 200$ or 1,000 and $n = 5$ or 10. To simulate a process of nonresponse and attrition, we removed measurements within subjects completely at random, with probabilities increasing linearly from 0% at wave 1 to 50% at wave n, so that $1 \leq n_i \leq n$ and $0 \leq n_i^* \leq n$ for each subject. Within each of the four sample-size conditions, we generated 1,000 datasets and applied our scoring procedure to each one. Under $m = 1,000$, the method converged to an interior point for every sample in an average of five or six steps. Under $m = 200$, the procedure failed to converge by 50 steps for 29 of the samples with $n = 10$ and for 287 of the samples with $n = 5$; our summaries for those conditions are based on the remaining samples. Failure to converge with smaller samples, a problem common to Newton-type procedures for random-effects models, suggests that the log-likelihood is oddly shaped and that ML estimates are nonidentified or near a boundary; analysts encountering this behavior will typically abandon the model and try a simpler one.

The behavior of the estimates for β and γ is summarized in Table 4. For each condition, this table lists the average and standard deviation (SD) of the point estimates, the average size of the standard error (SE), the standardized bias (as a percent of the SD), and the coverage rate of the nominal 95% confidence intervals (estimate ± 2 SE). With 1,000 replications, any standardized bias larger than 6.3% is statistically significant at the .05 level, but a useful rule of thumb is that biases do not have a substantial negative impact on inferences (e.g., by impairing the coverage of confidence intervals) unless they exceed 40%; by this rule, all of the bias estimates in the table are inconsequential. In most cases the average SE is close to the SD of the estimates, indicating that the standard errors from the approximate scoring algorithm are effective measures of variation. For the least favorable condition ($m = 200$, $n = 5$), the SD of the estimates is somewhat smaller than the average SE, showing that in the 711 well-behaved samples the estimates tend to be closer to the population values than the usual theory of ML would suggest; this phenomenon leads to slightly higher-than-nominal coverage rates.

Table 4. Simulation Results for Estimated Fixed Effects

	$-\beta_1$	$-\beta_2$	β_3	$-\gamma_1$	γ_2	γ_3
Population	1.00	.500	.400	.300	.100	.400
$m = 1,000$, $n = 10$						
Average estimate	1.00	.503	.400	.298	.096	.400
SD estimate	.070	.083	.012	.043	.051	.004
Average SE	.072	.086	.012	.043	.052	.004
Standard bias (%)	−2.0	4.0	−0.9	−4.3	−6.8	−1.8
Coverage (%)	95.9	96.1	96.1	95.1	95.6	96.3
$m = 1,000$, $n = 5$						
Average estimate	1.00	.501	.401	.300	.102	.401
SD estimate	.089	.105	.030	.059	.069	.016
Average SE	.088	.103	.030	.060	.067	.016
Standard bias (%)	3.6	.5	1.7	−.7	2.5	3.8
Coverage (%)	95.1	94.7	94.9	94.9	95.4	95.3
$m = 200$, $n = 10$						
Average estimate	.995	.512	.402	.297	.097	.400
SD estimate	.165	.188	.028	.097	.117	.010
Average SE	.164	.195	.028	.098	.119	.010
Standard bias (%)	−3.2	6.5	5.4	−3.1	−2.1	1.0
Coverage (%)	95.4	95.4	95.7	95.9	95.4	96.1
$m = 200$, $n = 5$						
Average estimate	.993	.497	.399	.303	.100	.400
SD estimate	.199	.219	.067	.133	.145	.034
Average SE	.200	.234	.068	.139	.154	.037
Standard bias (%)	−3.7	−1.5	−1.4	2.0	−.2	−.6
Coverage (%)	96.3	97.2	94.5	96.1	96.2	96.9

Under this nonresponse mechanism for $n = 5$, the average number of observed responses per subject is $n_i = 3.6$, and 1.2% of subjects have data for only a single time point ($n_i = 1$). This is somewhat more optimistic than the AAPT scenario, where more than 10% of the subjects had $n_i = 1$. To mimic the situation in the AAPT, we generated artificial data under the model of Section 3.2, setting the parameters equal to the ML estimates from the AAPT. We imposed missing values at random according to the patterns and rates shown in Table 1. We drew 1,000 samples of $m = 1,000$ subjects each and ran the estimation procedure on each sample. The algorithm failed to converge for 609 of the 1,000 samples; among the remaining 391, the estimation procedure showed excellent performance similar to the results in Table 4. This leads us to believe that when the algorithm does converge, the estimated coefficients and standard errors are quite reliable.

4.2 Performance of Estimated Variance Components

We now turn our attention to simulation results for σ^2 and ψ. The average values of the point estimates are shown in Table 5. No substantial biases are found under any of the sample-size conditions. In many analyses these tend to be regarded as nuisance parameters, and their estimates are not of primary interest. In practice, however, analysts will often test hypotheses regarding these parameters to help choose an appropriate model. For example, a user may want to know whether a model with random intercepts only provides nearly the same fit as a model with random intercepts and slopes. These tests are usually carried out by examining changes in the achieved log-likelihood.

To examine LR testing procedures for ψ, we carried out two additional sets of simulations. The first set focuses on the behavior of a test for which the null hypothesis lies on the boundary. We drew samples from a random-intercepts population like the one used in Section 4.1 with $\beta = (-1.5, 0, .5)^T$, $\gamma = (-1, .5, -.3)$, $\psi_{cc} = .5$, $\psi_{cd} = .4$, $\psi_{dd} = .1$, and $\sigma^2 = .2$. As before, we used a 2×2 design with $m = 1,000$ or 200 and $n = 10$ or 5. Under each sample-size condition, we drew 1,000 datasets and imposed missing values randomly with probability increasing to 50% by the final wave. We then attempted to fit two models to each dataset: a null model with random intercepts for both the logit and linear parts, and an expanded model that includes a random slope in the linear part.

Asymptotic theory suggests that the ML estimate from the expanded model will lie on the boundary about half the time, and when it does not, the LR statistic will be approximately χ^2_3 (Stram and Lee 1994). We found that scoring failed to converge to an interior point more often than not: 561 times for $m = 1,000$, $n = 10$; 665 times for $m = 200$, $n = 10$; 895 times for $m = 1,000$, $n = 5$; and every time for $m = 200$, $n = 5$.

Table 5. Average Point Estimates for Variance Component Parameters

	ψ_{cc}	ψ_{cd}	ψ_{dd}	σ^2
Population	1.00	.200	.500	.500
$m = 1,000$, $n = 10$	1.00	.199	.498	.500
$m = 1,000$, $n = 5$	1.01	.201	.500	.499
$m = 200$, $n = 10$	.99	.196	.495	.501
$m = 200$, $n = 5$	1.00	.211	.494	.499

Table 6. Percentage of Likelihood Ratio Statistics Exceeding the $100(1-\alpha)$ Percentile of χ^2_3

α	.10	.05	.01
$m = 1,000,\ n = 10$	9.0	4.3	.6
$m = 1,000,\ n = 5$	3.8	1.9	1.0
$m = 200,\ n = 10$	6.2	1.3	0
$m = 200,\ n = 5$			

In some of these failed runs, the estimates did appear to be approaching a boundary, but in other cases the reason for failure was unclear. Note that this simulation places a heavy burden on the algorithm; to obtain a test statistic, the procedure must converge not only under the true null model, but also under an expanded model with unnecessary random slopes. For the successful runs, we calculated the LR statistics to the $100(1-\alpha)$ percentiles of χ^2_3 for $\alpha = .10$, .05, and .01. The percentage of observed test statistics exceeding these cutoff values are displayed in Table 6. For the $m = 1,000$, $n = 10$ condition the percentages bear some similarity to α, but for the smaller samples they tend to be lower. This suggests that comparing the LR statistic to a 50 : 50 mixture of χ^2_3 and $\chi^2_0 = 0$ is a conservative procedure, because the algorithm fails to produce estimates under the larger model much more than half the time, and when it does, the test statistics tend to be smaller than χ^2_3. In smaller samples, simply obtaining any positive estimate for the variance of a random effect may be fairly strong evidence against the null.

Our final set of simulations focuses on the test of separability ($\psi_{cd} = 0$), which does not involve a boundary. We drew 1,000 samples from a population with $\beta = (1, .5, -1)^T$, $\gamma = (.4, -.7, -.4)$, $\psi_{cc} = 2$, $\psi_{cd} = 0$, $\psi_{dd} = .8$, and $\sigma^2 = .9$ under each of the sample-size conditions and the same nonresponse mechanism as before. We fit the null model with $\psi_{cd} = 0$ and the alternative with ψ_{cd} unspecified to each dataset. The algorithm failed to converge for one or both models 0 times for $m = 1,000$, $n = 10$; 6 times for $m = 1,000$, $n = 5$; 293 times for $m = 200$, $n = 10$; and 223 times for $m = 200$, $n = 5$. The percentage of observed test statistics exceeding the $100(1-\alpha)$ percentiles of χ^2_1 are shown in Table 7. For this nonboundary test, the chi-squared approximation appears to work well.

5. DISCUSSION

We have chosen a fully parametric, subject-specific approach to define our model (1)–(3) and estimate parameters. Consequently, our logistic coefficients β are not equivalent to population-averaged log-odds ratios unless little individual heterogeneity is present (e.g. Diggle et al. 1994). Differences between these two types of estimates have been explored by Goldberg, Hedeker, Flay, and Pentz (1998), Neuhaus, Kalbfleisch, and Hauck (1991), and Zeger et al. (1988).

Table 7. Percentage of Likelihood Ratio Statistics Exceeding the $100(1-\alpha)$ Percentile of χ^2_1

α	.10	.05	.01
$m = 1,000,\ n = 10$	10.3	4.7	1.2
$m = 1,000,\ n = 5$	9.7	5.0	1.1
$m = 200,\ n = 10$	11.0	5.2	1.6
$m = 200,\ n = 5$	10.3	4.9	.5

An interesting and somewhat paradoxical feature of our model is that we are able to estimate the mean level of reported use, if any, $E(V_{ij} \mid U_{ij})$, for all subjects and occasions, including those for which no reported use occurred ($U_{ij} = 0$). The information for estimating $E(V_{ij} \mid U_{ij} = 1)$ for an occasion with $U_{ij} = 0$ comes from two sources: observed values of V_{ij} from the same subject at occasions where use occurred and observed values of V_{ij} from other subjects with similar covariates who did experience use. If no use occurred ($U_{ij} = 0$) at an occasion for which the estimated probability of use, $P(U_{ij} = 1)$, is moderately large, then imagining what the level of use might have been if some use had occurred is sensible. But if no use occurred at an occasion for which $P(U_{ij} = 1)$ is very small, then the meaning of $E(V_{ij} \mid U_{ij} = 1)$ becomes tenuous, because the counterfactual event $U_{ij} = 1$ is difficult to imagine. Moreover, as $P(U_{ij} = 1)$ becomes small, an estimate of $E(V_{ij} \mid U_{ij} = 1)$ becomes less trustworthy, because it involves extrapolation beyond the outer regions of observed data. The resulting predictions would have large standard errors and may be sensitive to departures from the assumed model.

In longitudinal studies of substance use, trajectories like the one shown in Figure 1(c) suggest that some individuals may be "teetotalers" with essentially no probability of use at any time point. Teetotalers may violate the assumption (3) of normally distributed random effects by having actual logit probabilities of use of $-\infty$. One could allow for teetotalers by specifying a two-part mixture for the distribution of c_i. The status of any individual as a teetotaler or a potential user would be unobserved, producing a latent classification similar to the mover–stayer models used by social scientists (Blumen, Kogan, and McCarthy 1955). Methods for incorporating latent classes into random-effects models have been considered by Muthén and Shedden (1999). For the AAPT examples of Section 3, some exploratory work suggests that the presence of teetotalers would have little impact on our substantive conclusions. We removed individuals with no reported use from the analysis and found that the estimated effects for parental monitoring and rebelliousness were essentially the same; the only real change was a rise in the mean intercept in the logit part of the model, indicating higher overall probabilities of use in the sample.

Although our model may be quite useful for describing recent alcohol use, it would not be appropriate for a nondecreasing quantity such as lifetime alcohol use, which represents a total response accumulated over time. Once the first use has occurred, transition back to a state of never having used is not possible. Transitions to first use are more appropriately described by methods of survival analysis.

In the future, we plan to introduce additional levels of random effects into our model to describe clustering of sampled individuals (e.g., students nested within schools). Clustered longitudinal data are quite common in prevention trials, and treatments are often applied at the school or cluster level. In the AAPT study, for example, each participating school was randomly assigned to one of four substance use prevention programs, and primary research questions focused on the differences in substance use patterns among the four treatment

groups. When assessing the statistical significance of treatment group differences or other cluster-level effects, failing to account for intracluster heterogeneity may lead one to substantially overstate the actual precision of the estimates.

APPENDIX: THE SCORE VECTORS

Here we give the equations for the components of $S_{\theta i}$, derived using the standard techniques of matrix differention and matrix algebra presented by Schott (1997):

$$\begin{aligned}S_{\beta i} &= X_i^T W_i (U_i^* - X_i\beta - Z_i\tilde{c}_i)\\ &\quad + \sum_{j=1}^{n_i}\left(-\frac{1}{2}\tilde{m}_{ij}^{3}(Z_{ij}G_i^{-1}Z_{ij}^T)\right)(X_{ij}^T - X_i^T W_i Z_i G_i^{-1} Z_{ij}^T)\\ &\quad + \frac{1}{A_i^*}\sum_{j=1}^{n_i} R_{ij}^*(X_{ij}^T - X_i^T W_i Z_i G_i^{-1} Z_{ij}^T);\end{aligned}$$

$$\begin{aligned}S_{\gamma i} &= \frac{1}{\sigma^2}\left[X_i^{*T}V_i - X_i^{*T}Z_i^* B_i^{-1} Z_i^{*T} V_i\right.\\ &\qquad \left. - X_i^{*T}X_i^*\gamma + X_i^{*T}Z_i^* B_i^{-1} Z_i^{*T} X_i^*\gamma\right]\\ &\quad - X_i^{*T} Z_i^* B_i^{-1} H^{-1}\psi_{dc}\psi_{cc}^{-1}\tilde{c}_i + t_i^T E_{ij};\end{aligned}$$

$$\begin{aligned}S_{\sigma^2 i} &= -\frac{n_i^*}{2\sigma^2} + \frac{1}{2\sigma^2} tr\left(B_i^{-1}Z_i^{*T}Z_i^*\right)\\ &\quad + \frac{1}{2(\sigma^2)^2}\left(A_i^T A_i - A_i^T Z_i^*(B_i^{-1}\right.\\ &\qquad \left. + \sigma^2 B_i^{-1}H^{-1}B_i^{-1})Z_i^{*T}A_i\right)\\ &\quad + \frac{1}{2}\tilde{c}_i^T(\psi_{dc}\psi_{cc}^{-1})^T(H^{-1}B_i^{-1})^2 Z_i^{*T}Z_i^*\psi_{dc}\psi_{cc}^{-1}\tilde{c}_i\\ &\quad - \tilde{c}_i^T(\psi_{dc}\psi_{cc}^{-1})^T(H^{-1}B_i^{-1})^2 Z_i^{*T}A_i + P_i^T E_{ij}\\ &\quad - tr\left[((\psi_{dc}\psi_{cc}^{-1})^T(H^{-1}B_i^{-1})^2 Z_i^{*T}Z_i^*\psi_{dc}\psi_{cc}^{-1})\right.\\ &\qquad \left.\times\left(-\frac{1}{2}G_i^{-1} + \frac{1}{A_i^*}\left(\frac{1}{4}F_i + \frac{1}{16}f_i - \frac{15}{36}h_i\right.\right.\right.\\ &\qquad \left.\left.\left. - \frac{15}{72}G_i^{-1}K_iK_i^TG_i^{-1}\right)\right)\right];\end{aligned}$$

$$\begin{aligned}S_{\phi_{cc}} &= \left(\frac{\partial \mathrm{vec}(\psi_{cc}^{-1})}{\partial\phi_{cc}^T}\right)^T\\ &\quad \left(\frac{1}{2}\mathrm{vec}(\psi_{cc} - G_i^{-1} - \tilde{c}_i\tilde{c}_i^T) + D_i^T E_{ij}\right.\\ &\qquad + \frac{1}{A_i^*}\left\{\mathrm{vec}\left(G_i^{-1}Z_{ij}^T Z_{ij}G_i^{-1}\right)F_{ij}^*\right.\\ &\qquad \left.\left. - \frac{15}{72}\mathrm{vec}(G_i^{-1}K_iK_i^TG_i^{-1})\right\}\right);\end{aligned}$$

$$\begin{aligned}S_{\psi_{dc}\psi_{cc}^{-1}i} &= -\left(I_{pc}\otimes H_i^*\right)\left(\mathrm{vec}(G_i^{-1})\right.\\ &\quad \left. + \mathrm{vec}(\tilde{c}_i\tilde{c}_i^T)\right) + \mathrm{vec}(H^{-1}B_i^{-1}Z_i^{*T}A_i\tilde{c}_i^T)\\ &\quad + S_i^{*T}E_{ij} + \frac{2}{A_i^*}\left\{\sum_{j=1}^{n_i}\mathrm{vec}\left(H_i^*G_i^{-1}Z_{ij}^TZ_{ij}G_i^{-1}\right)F_{ij}^*\right.\\ &\qquad \left. - \frac{15}{36}\mathrm{vec}\left(H_i^*G_i^{-1}K_iK_i^TG_i^{-1}\right)\right\};\end{aligned}$$

and

$$\begin{aligned}S_{\phi_H} &= \left(\frac{\partial \mathrm{vec}(H^{-1})}{\partial\phi_H^T}\right)^T\left(-\frac{1}{2}((B_i^{-1}Z_i^{*T}Z_i^*\psi_{dc}\psi_{cc}^{-1}\otimes\psi_{dc}\psi_{cc}^{-1})\right.\\ &\quad -\sigma^2(B_i^{-1}H^{-1}\psi_{dc}\psi_{cc}^{-1}\otimes B_i^{-1}Z_i^{*T}Z_i^*\psi_{dc}\psi_{cc}^{-1}))\mathrm{vec}(G_i^{-1})\\ &\quad +\frac{1}{2}\mathrm{vec}[H - \sigma^2 B_i^{-1} - B_i^{-1}Z_i^{*T}A_iA_i^TZ_i^*B_i^{-1}\\ &\quad -\psi_{dc}\psi_{cc}^{-1}\tilde{c}_i\tilde{c}_i^T(\psi_{dc}\psi_{cc}^{-1})^TZ_i^{*T}Z_i^*B_i^{-1}\\ &\quad +2B_i^{-1}Z_i^{*T}A_i\tilde{c}_i^T(\psi_{dc}\psi_{cc}^{-1})^T\\ &\quad +\sigma^2(B_i^{-1}H^{-1}\psi_{dc}\psi_{cc}^{-1}\tilde{c}_i\tilde{c}_i^T(\psi_{dc}\psi_{cc}^{-1})^TZ_i^{*T}Z_i^*)\\ &\quad -2\sigma^2B_i^{-1}Z_i^{*T}A_i\tilde{c}_i^T(\psi_{dc}\psi_{cc}^{-1})^TH^{-1}B_i^{-1}]\\ &\quad +M_i^{*T}E_{ij} + \frac{1}{A_i^*}\left\{\sum_{j=1}^{n_i}\mathrm{vec}\left((I_{pd} - \sigma^2B_i^{-1}H^{-1})\psi_{dc}\right.\right.\\ &\qquad \times\psi_{cc}^{-1}G_i^{-1}Z_{ij}^TZ_{ij}G_i^{-1}(\psi_{dc}\psi_{cc}^{-1})^TZ_i^{*T}Z_i^*B_i^{-1})F_{ij}^*\\ &\quad -\frac{15}{72}\mathrm{vec}\left((I_{pd} - \sigma^2B_i^{-1}H^{-1})\psi_{dc}\right.\\ &\qquad \left.\left.\left.\times\psi_{cc}^{-1}G_i^{-1}K_iK_i^TG_i^{-1}(\psi_{dc}\psi_{cc}^{-1})^TZ_i^{*T}Z_i^*B_i^{-1})\right\}\right),\end{aligned}$$

where

$$\begin{aligned}P_i &= \frac{\partial\tilde{c}_i}{\partial\sigma^2} = G_i^{-1}(\psi_{dc}\psi_{cc}^{-1})^T(H^{-1}B_i^{-1})^2\\ &\qquad \times\left[Z_i^{*T}Z_i^*\psi_{dc}\psi_{cc}^{-1}G_i^{-1}Y_i^* - Z_i^{*T}A_i\right],\end{aligned}$$

$$\begin{aligned}M_i^* &= (\partial\tilde{c}_i)/(\partial\mathrm{vec}^T(H^{-1}))\\ &= -\left(Y_i^{*T}G_i^{-1}(\psi_{dc}\psi_{cc}^{-1})^TZ_i^{*T}Z_i^*B_i^{-1}\otimes G_i^{-1}(\psi_{dc}\psi_{cc}^{-1})^T\right)\\ &\quad +(A_i^TZ_i^*B_i^{-1}\otimes G_i^{-1}(\psi_{dc}\psi_{cc}^{-1})^T)\\ &\quad -\sigma^2(A_i^TZ_i^*B_i^{-1}\otimes G_i^{-1}(\psi_{dc}\psi_{cc}^{-1})^TH^{-1}B_i^{-1})\\ &\quad +\sigma^2(Y_i^{*T}G_i^{-1}(\psi_{dc}\psi_{cc}^{-1})^TZ_i^{*T}Z_i^*B_i^{-1}\\ &\qquad \otimes G_i^{-1}(\psi_{dc}\psi_{cc}^{-1})^TH^{-1}B_i^{-1}),\end{aligned}$$

$$\begin{aligned}D_i &= (\partial\tilde{c}_i)/(\partial\mathrm{vec}^T(\psi_{cc}^{-1})) = -\left(Y_i^{*T}G_i^{-1}\otimes G_i^{-1}\right),\\ t_i &= (\partial\tilde{c}_i)/(\partial\gamma^T) = -G_i^{-1}(\psi_{dc}\psi_{cc}^{-1})^TH^{-1}B_i^{-1}Z_i^{*T}X_i^*,\end{aligned}$$

$$\begin{aligned}S_i^* &= (\partial\tilde{c}_i)/(\partial\mathrm{vec}^T(\psi_{dc}\psi_{cc}^{-1})) = \left((A_i^TZ_i^*B_i^{-1}H^{-1})\otimes G_i^{-1}\right)\\ &\quad -\left(G_i^{-1}\otimes Y_i^{*T}G_i^{-1}H_i^{*T}\right) - \left(Y_i^{*T}G_i^{-1}\otimes G_i^{-1}H_i^{*T}\right),\end{aligned}$$

$$\begin{aligned}E_{ij} &= E_i^* - (\psi_{cc}^{-1} + D_i^*)\tilde{c}_i\\ &\quad +\sum_{j=1}^{n_i}Z_{ij}^T\left\{-\frac{1}{2}\tilde{m}_{ij}^{(3)}(Z_{ij}G_i^{-1}Z_{ij}^T) + (U_{ij} - \tilde{\pi}_{ij}) + \frac{1}{A_i^*}R_{ij}^*\right\},\end{aligned}$$

$$\begin{aligned}H_i^* &= Z_i^{*T}Z_i^*B_i^{-1}H^{-1}\psi_{dc}\psi_{cc}^{-1},\\ Y_i^* &= \left(Z_i^T\tilde{W}_i(U_i^* - X_i\beta) + E_i^*\right),\end{aligned}$$

$$\begin{aligned}F_{ij}^* &= \frac{1}{4}\tilde{m}_{ij}^{(4)}(Z_{ij}G_i^{-1}Z_{ij}^T) + \frac{1}{16}\tilde{m}_{ij}^{(6)}(Z_{ij}G_i^{-1}Z_{ij}^T)^2\\ &\qquad - \frac{15}{36}\tilde{m}_{ij}^{(3)}(Z_{ij}G_i^{-1}K_i),\end{aligned}$$

$$R_{ij}^{*} = -\frac{1}{8}\tilde{m}_{ij}^{(5)}\left(Z_{ij}G_i^{-1}Z_{ij}^T\right)^2 + \frac{1}{4}\tilde{m}_{ij}^{(3)}Z_{ij}F_iZ_{ij}^T - \frac{1}{48}\tilde{m}_{ij}^{(7)}\left(Z_{ij}G_i^{-1}Z_{ij}^T\right)^3 + \frac{1}{16}\tilde{m}_{ij}^{(3)}Z_{ij}f_iZ_{ij}^T - \frac{15}{72}\tilde{m}_{ij}^{(3)}\left(Z_{ij}G_i^{-1}K_i\right)^2 + \frac{15}{36}\tilde{m}_{ij}^{(4)}(K_i^TG_i^{-1}Z_{ij}^T)(Z_{ij}G_i^{-1}Z_{ij}^T) - \frac{15}{36}\tilde{m}_{ij}^{(3)}Z_{ij}h_iZ_{ij}^T,$$

$$F_i = \sum_{j=1}^{n_i}\tilde{m}_{ij}^{(4)}G_i^{-1}Z_{ij}^T(Z_{ij}G_i^{-1}Z_{ij}^T)Z_{ij}G_i^{-1},$$

$$f_i = \sum_{j=1}^{n_i}\tilde{m}_{ij}^{(6)}G_i^{-1}Z_{ij}^T(Z_{ij}G_i^{-1}Z_{ij}^T)^2Z_{ij}G_i^{-1},$$

$$h_i = \sum_{j=1}^{n_i}\tilde{m}_{ij}^{(3)}G_i^{-1}Z_{ij}^TZ_{ij}G_i^{-1}(Z_{ij}G_i^{-1}K_i),$$

and

$$K_i = \sum_{j=1}^{n_i}Z_{ij}^T\tilde{m}_{ij}^{(3)}Z_{ij}G_i^{-1}Z_{ij}^T.$$

[Received September 1999. Revised November 2000.]

REFERENCES

Aitchison, J. (1955), "On the Distribution of a Positive Random Variable Having a Discrete Probability Mass at the Origin," *Journal of the American Statistical Association*, 50, 901–908.

Amemiya, T. (1984), "Tobit Models: A Survey," *Journal of Econometrics*, 24, 3–61.

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.

Blumen, J., Kogan, M., and McCarthy, P. J. (1955), *The Industrial Mobility of Labor as a Probability Process*, Ithaca, NY: Cornell University Press.

Bock, R. D., and Lieberman, M. (1970), "Fitting a Response Model for n Dichotomously Scored Items," *Psychometrika*, 35, 179–197.

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," *Biometrics*, 82, 81–91.

Carlin, B. P. (1996), "Hierarchical Longitudinal Modelling," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. London: Chapman & Hall. pp. 303–319..

Clayton, D. G. (1996), "Generalized Linear Mixed Models," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. New York: Chapman & Hall.

Collett, D. (1991), *Modelling Binary Data*. London: Chapman & Hall.

Cowles, M. K., Carlin, B. P., and Connett, J. E. (1996), "Bayesian Tobit Modeling of Longitudinal Ordinal Clinical Trial Compliance Data With Nonignorable Missingness," *Journal of the American Statistical Association*, 91, 86–98.

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, New York: Chapman & Hall.

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

Duan, N., Manning, W. G., Morris, C. N., and Newhouse, J. P. (1983), "A Comparison of Alternative Models for the Demand for Medical Care," *Journal of Business and Economic Statistics*, 1, 115–126.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995), *Bayesian Data Analysis*. New York: Chapman and Hall.

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

Graham, J. W., Hofer, S. M., and Piccinin, A. M. (1994), "Analysis With Missing Data in Drug Prevention Research," in *National Institute on Drug Abuse Research Monograph Series*, Vol. 142, eds. L. Collins and L. Seitz, Washington, DC: National Institute on Drug Abuse.

Greene, W. H. (1994), "Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models," Working Paper #94-10, New York University, Stern School of Business.

Hajivassiliou, V. A. (1994), "A Simulation Estimation Analysis of the External Debt Crises of Developing Countries," *Journal of Applied Econometrics*, 9, 109–131.

Hansen, W. B., and Graham, J. W. (1991), "Preventing Alcohol, Marijuana, and Cigarette Use Among Adolescents: Peer Pressure Resistance Training Versus Establishing Conservative Norms," *Preventive Medicine*, 20, 414–430.

Hastings, W. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109.

Heckman, J. (1974), "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42, 679–674.

———, (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables, and a Sample Estimator for Such Models," *The Annals of Economic and Social Measurement*, 5, 475–592.

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

Heilbron, D. C. (1989), "Generalized Linear Models for Altered Zero Probabilities and Overdispersion in Count Data," technical report, University of California, San Francisco, Dept. Epidemiology and Biostatistics.

Hu, F. B., Goldberg, J., Hedeker, D., Flay, B. R., and Pentz, M. (1998), "Comparison of Population-Averaged and Subject-Specific Approaches for Analyzing Repeated Binary Outcomes," *American Journal of Epidemiology*, 147, 694–703.

Kyriazidou, E. (1997), "Estimation of Panel Data Sample Selection Model," *Econometrics*, 65, 1335–1364.

Laird, N. M., and Ware, J. H. (1982), "Random-Effects Models for Longitudinal Data," *Biometrics*, 38, 963–974.

Lambert, D. (1992), "Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing," *Technometrics*, 34, 1–14.

Lange, N., and Ryan, L. (1989), "Assessing Normality in Random Effects Models," *The Annals of Statistics*, 17, 624–642.

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

Manning, W., Morris, C. N., Newhouse, J. P., Orr, L. L., Duan, N., Keeler, E. B., Leibowitz, A., Marquis, K. H., Marquis, M. S., and Phelps, C. E. (1981), "A Two-Part Model of the Demand for Medical Care: Preliminary Results From the Health Insurance Experiment," in *Health, Economics, and Health Economics*, eds. J. van der Gaag and M. Perlman, Amsterdam: North-Holland, pp. 103–104.

McCulloch, C. E. (1997), "Maximum Likelihood Algorithms for Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 92, 162–170.

Muthén, B., and Shedden, K. (1999), "Finite Mixture Modeling With Mixture Outcomes Using the EM Algorithm," *Biometrics*, 55, 463–469.

Neuhaus, J. M., Kalbfleisch, J. D., and Hauck, W. W. (1991), "A Comparison of Cluster-Specific and Population-Averaged Approaches for Analyzing Correlated Binary Data," *International Statistical Review*, 59, 25–35.

Olsen, M. K., and Schafer, J. L. (1998), "Parameter Estimates for Semicontinuous Longitudinal Data Using an EM Algorithm," Technical Report #98-31, Pennsylvania State University, The Methodology Center.

Pinheiro, J. C., and Bates, D. M. (1995), "Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model," *Journal of Computational and Graphical Statistics*, 4, 12–35.

———, (2000), *Mixed-Effects Models in S and S-PLUS*, New York: Springer-Verlag.

Raudenbush, S. W., Yang, M., and Yosef, M. (2000), "Maximum Likelihood for Generalized Linear Models With Nested Random Effects via High-Order, Multivariate Laplace Approximation," *Journal of Computational and Graphical Statistics*, 9, 141–157.

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. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592.

Schafer, J. L. (1998), "Some Improved Procedures for Linear Mixed Models," Technical Report #98-27, Pennsylvania State University, The Methodology Center.

Schott, J. R. (1997), *Matrix Analysis for Statistics*. New York: Wiley.

Self, S. G., and Liang, K. (1987), "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions," *Journal of the American Statistical Association*, 82, 605–610.

Spiegelhalter, D. J., Thomas, A., and Best, N. G. (1999), *WinBUGS Version 1.2 User Manual*. MRC Biostatistics Unit, Cambridge, UK.

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

Stram, D. O., and Lee, J. W. (1994), "Variance Components Testing in the Longitudinal Mixed Effects Model," Biometrics, 50, 1171–1177.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions" (with discussion), *The Annals of Statistics*, 22, 1701–1762.

Tierney, L., and Kadane, J. B. (1986), "Accurate Approximations for Posterior Moments and Densities," *Journal of the American Statistical Association*, 81, 82–86.

Tobin, J. (1958), "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, 26, 24–36.

Toivonen, H., Mannila, H., Seppanen, J., and Vasko, K. (1999), "Bassist User's Guide," Technical Report C-1999-36, University of Helsinki, Dept. of Computer Science.

Zeger, S. L., and Karim, M. R. (1991), "Generalized Linear Models With Random Effects: A Gibbs Sampling Approach," *Journal of the American Statistical Association*, 86, 79–86.

Zeger, S. L., Liang, K. Y., and Albert, P. S. (1988), "Models for Longitudinal Data: A Generalized Estimating Equation Approach," *Biometrics*, 44, 1049–1060.