2

RANDOM-EFFECTS MODELING OF CATEGORICAL RESPONSE DATA

*Alan Agresti**
*James G. Booth**
*James P. Hobert**
*Brian Caffo**

In many applications observations have some type of clustering, with observations within clusters tending to be correlated. A common instance of this occurs when each subject in the sample undergoes repeated measurement, in which case a cluster consists of the set of observations for the subject. One approach to modeling clustered data introduces cluster-level random effects into the model. The use of random effects in linear models for normal responses is well established. By contrast, random effects have only recently seen much use in models for categorical data. This chapter surveys a variety of potential social science applications of random effects modeling of categorical data. Applications discussed include repeated measurement for binary or ordinal responses, shrinkage to improve multiparameter estimation of a set of proportions or rates, multivariate latent variable modeling, hierarchically structured modeling, and cluster sampling. The models discussed belong to the class of generalized linear mixed models (GLMMs), an extension of ordinary linear models that permits nonnormal response variables and both fixed and random effects in the predictor term. The models are GLMMs for either binomial or Pois-

This work was partially supported by a grant from the National Science Foundation. The authors appreciate comments from Brent Coull, Russ Wolfinger, and two referees. They also thank Jonathan Hartzel for advice on computing and the use of his program for the nonparametric random-effects approach.

*University of Florida

son response variables, although we also present extensions to multicategory (nominal or ordinal) responses. We also summarize some of the technical issues of model-fitting that complicate the fitting of GLMMs even with existing software.

1. INTRODUCTION

Response variables in social science studies, particularly those dealing with opinions and attitudes, are often measured on categorical scales. For many years, to model such data it was common to use ordinary least-squares methods, either for the original scale or some transformation for which the variance tends to be more stable. Recently it has become more common to use models designed specifically for categorical variables, such as logistic regression for binomial responses, generalized logit models for multinomial responses, and log-linear models for Poisson responses.

In many applications, however, the dependence structure is more complex than the independent observations assumed by ordinary models for categorical or continuous variables. In particular, observations often exhibit a clustering, with observations within clusters being correlated. A common instance of this occurs with repeated measurement on each subject in the sample, in which case a cluster consists of the set of observations for a given subject and those observations are typically positively correlated.

For continuous variables, the multivariate normal distribution provides considerable flexibility for describing dependencies. For categorical variables, there is no natural analog of the multivariate normal distribution, which makes the specification of models somewhat awkward. One solution to this introduces cluster-level terms into the model. These terms are unobserved and, varying randomly among a sample of clusters, are called *random effects*. For instance, with linear models for repeated measurement, it is often effective to add a random effect u_i to the predictor of the response for cluster i. If u_i is positive, the observations within that cluster have a larger mean than otherwise, whereas if u_i is negative, the observations within that cluster have a smaller mean than otherwise. Considered over clusters, this induces a positive within-cluster association.

1.1. *Generalized Linear Mixed Models*

Parameters in ordinary linear models are said to be *fixed effects*. They apply to *all* the levels of interest (e.g., gender, race, political party affili-

ation), whereas random effects apply to a *sample* of all the possible clusters. The use of random effects in linear models for normal responses is well established (e.g., see Searle, Casella, and McCulloch 1992). By contrast, only recently have random effects seen much use in models for categorical responses. We survey here a variety of potential applications of random-effects modeling for social science research.

The class of *generalized linear models* (GLMs) extends ordinary regression models in two ways: (1) it allows for nonnormal responses, and (2) it allows modeling a function of the mean rather than the mean itself. This extension is important for categorical data. For such data, one assumes a binomial or Poisson distribution for the response rather than normal. Also, one usually models the logit of a probability or the logarithm of an expected count instead of the probability or expected count itself (e.g., so that predictions are necessarily on the proper scale, and so that an additive effects model is more likely to fit well). The *generalized linear mixed model*, which we denote by GLMM, is a further extension that permits both fixed and random effects in the predictor rather than only fixed effects. The models discussed here are GLMMs for either binomial (or multinomial) or Poisson response variables.

1.2. *Applications of GLMMs*

Early applications of GLMMs occurred in the psychometrics literature, in the context of *item-response models*, generalizing the *Rasch model* (Rasch 1961). For a set of subjects and test items, the Rasch model states that the probability π_{ij} that subject i makes the correct response on question j satisfies

$$\text{logit}(\pi_{ij}) = u_i + \beta_j. \tag{1}$$

In estimating $\{\beta_j\}$, Rasch promoted the fixed-effects approach of *conditional maximum likelihood*. This method eliminates $\{u_i\}$ from the analysis by conditioning on their sufficient statistics, yielding a likelihood function that depends only on $\{\beta_j\}$. Later authors used a random-effects approach with this model and the corresponding probit model by treating $\{u_i\}$ as having a normal distribution (e.g., see Bock and Aitkin 1981; Stiratelli, Laird, and Ware 1984).

As mentioned above, random effects can represent clustering in a sample. Section 3 shows several examples of this type. Often the clusters

result from repeated measurement, representing a set of observations on the same subject at different times or on different components of a response variable. In Section 3.2, for instance, subjects indicate whether or not they support legalized abortion in each of three situations; in Section 3.7, subjects indicate whether government spending should increase, stay the same, or decrease, on items related to environment, health, law enforcement, and education. In other cases the clusters may result from clustering in a multistage sample, as shown in the Section 3.8 example regarding a survey about household satisfaction. Random-effects modeling is also useful when parameters such as proportions or rates for a large number of geographical areas may share some common features. In estimating the parameters with a random-effects model, one effect is shrinkage of separate sample values toward a common value, which can result in dramatically improved estimators in an overall sense. Section 3.1 illustrates this with the use of survey data to estimate simultaneously the proportion favoring a presidential candidate in each of the 50 states. We also show applications such as hierarchically structured modeling (Section 3.4), and handling effects (such as clinics, schools, hospitals) for which the data may have only a sample of the possible levels (Section 3.3). We also present extensions to multicategory (nominal or ordinal) responses.

Random effects are sometimes regarded as unobserved responses on variables. In model (1), for instance, u_i might represent the unknown "ability" of subject i in answering the test items. More generally, models that have unobserved variables of a variety of types are, in essence, random-effects models. For instance, random-effects terms have been used in models to represent omitted explanatory variables and random measurement error in the explanatory variables (e.g., Follmann and Lambert 1989). Related to this, random effects also provide a mechanism for explaining *overdispersion* (e.g., see Breslow and Clayton 1993)—the presence of greater variability in the data than the sampling model predicts. In fact, another strand of literature on random effects for count data developed as a way of handling overdispersion. This literature discussed alternative mixture models such as the *beta-binomial* as well as related *quasi-likelihood* methods that allowed greater variance than the standard sampling models but without assuming a particular parametric distribution (e.g., Williams 1982).

In recent years GLMMs have become increasingly popular in applications in fields such as education (e.g., Goldstein 1991, 1995; Bryk and Raudenbush 1992) and medicine (e.g., Daniels and Gatsonis 1999). They have also been receiving increased attention in social science re-

search. Published work using GLMMs includes the following: Nee (1996) used GLMMs to analyze data from a multistage, multilevel nationwide social survey of households in rural China; Sampson, Raudenbush, and Earls (1997) employed these models to construct and evaluate measures of neighborhood social organization in a study of the relationship between social cohesion among neighbors and crime; Langford (1998) used a GLMM to model an individual's "willingness to pay" to prevent saline flooding in the East Anglian region of England as a function of the cost; and Murphy and Wang (1998) used a GLMM in a discrete-time hazards context to handle cluster effects of children sampled having the same mother. Other references that used a GLMM include Raudenbush, Rowan, and Kang (1991); Langford (1994); McArdle and Hamagami (1994); Congdon (1996); Murray, Moskowitz, and Dent (1996); Saunderson and Langford (1996); Jones, Gould, and Watt (1998); Hedeker, Gibbons, and Flay (1994); Gibbons and Hedeker (1994); Gibbons, Hedeker, Charle, and Frisch (1994); Enberg, Gottschalk, and Wolf (1990); Daniels and Gastonis (1997); Akin, Guilkey, and Sickles (1979); Henretta, Hill, Li, Soldo, and Wolf (1997); Montgomery, Richards, and Braun (1986); Tsutakawa (1988); Wong and Mason (1985); Albert (1992); and Anderson and Aitkin (1985). We hope that the examples shown here will stimulate further uses of random-effects modeling in the social sciences.

1.3. *Scope of This Article*

Section 2 describes the general form of GLMMs for categorical response variables, with focus on binomial or Poisson responses. Section 3 is the heart of the paper, showing a variety of applications of random-effects modeling. Maximum likelihood (ML) is used for all of the model fitting in our examples. The basic ideas underlying ML fitting and inference are given in Section 4 along with a description of some alternative (approximate) fitting methods.

In some applications the random effects part of the model is a mechanism for representing how correlation occurs between observations within a cluster, yet the main interest is in estimating fixed effects, in which case the parameters pertaining to the random effects are *nuisance parameters*. Often though, those parameters are themselves of interest, for instance to characterize the degree of heterogeneity of a population. More generally one may be interested in combinations of fixed and random effects, in order to predict responses, as illustrated in Section 3.1. Some details re-

garding the prediction of random effects are given in Section 4.5. The GLMM discussed in this paper leads to conditional (i.e., *subject-specific*) interpretations of the regression parameters. Section 5 presents a discussion of a related class of models for the estimation of marginal (i.e., *population-averaged*) effects.

Software for GLMMs is still limited. The results given here were obtained using PROC NLMIXED in SAS (available beginning in version 7), which uses numerical integration for approximating the likelihood function. Section 6 discusses available software and cautions one should follow in using it. Finally, our conclusions are stated in Section 7.

2. RANDOM-EFFECTS MODELS FOR CATEGORICAL DATA

We first introduce some general notation that applies to the examples presented here. Let y_{ij} denote the jth response in cluster i, $i = 1, \ldots, I$, $j = 1, \ldots, n_i$. Let $\mathbf{x}_{ij}$ denote a column vector of values of a set of explanatory variables for the jth response in cluster i, which serve as coefficients of fixed effects in the model. Let $\mathbf{z}_{ij}$ denote a corresponding vector of coefficients of random effects. Note that these sets of coefficients need not be identical for all observations in a cluster. Let $\mathbf{u}_i$ denote the vector of random effect values for cluster i. Conditional on $\mathbf{u}_i$, a GLMM resembles an ordinary GLM. Let $\mu_{ij} = E(y_{ij}|\mathbf{u}_i)$ denote the mean of the conditional distribution of y_{ij} given $\mathbf{u}_i$. Denote the variance of the conditional distribution by $\mathrm{Var}(y_{ij}|\mathbf{u}_i) = \phi_{ij} v(\mu_{ij})$, where typically $\phi_{ij} = \phi/w_{ij}$ with the w_{ij}'s being known "weights" and ϕ being an unknown dispersion parameter, and the function v is called the *variance function*.

The linear predictor for a GLMM has the form

$$g(\mu_{ij}) = \mathbf{x}_{ij}^t \boldsymbol{\beta} + \mathbf{z}_{ij}^t \mathbf{u}_i, \tag{2}$$

where $g(\cdot)$ is a link function (such as the logit for binary data and the log for count data), a t superscript denotes the transpose of a vector, and $\mathbf{u}_i$ is usually assumed to have a normal distribution. In (2), the random effect enters the model on the same scale as the predictor terms. This is convenient but also natural for the many applications in which the random effect partly represents unmodeled heterogeneity caused by not including certain important explanatory variables.

2.1. *A GLMM for Two Dependent Binomial Samples*

We illustrate this general expression for a GLMM using perhaps the simplest example of a random-effects model for categorical response data: binary matched pairs, yielding two dependent binomial samples. For cluster i, let (y_{i1}, y_{i2}) denote the two responses in the matched pair. For instance, y_{i1} might denote a binary response measured at a particular time, and y_{i2} a response on the same outcome scale at a later time, such as when a sample of subjects is asked at two dates about whether the president is doing a good job. Suppose that subject i at time j has $y_{ij} = 1$ (a "success") or 0 (a "failure"), $j = 1, 2$. Then μ_{ij} is the probability of success. For binary data, the link function g is most often the logit transform.

Table 1, from the General Social Survey of 1994, is an example of matched-pairs data. Subjects were asked "Do you think a person has the right to end his or her own life if this person has an incurable disease?" and "When a person has a disease that cannot be cured, do you think doctors should be allowed to end the patient's life by some painless means if the patient and his family request it?" The table, which refers to these variables as "suicide" and "let patient die," reports the numbers of "yes" and "no" responses for each question. The two responses for each subject form a matched pair. For these matched pairs, consider the model

$$\text{logit}(\mu_{i1}) = \alpha + u_i, \qquad \text{logit}(\mu_{i2}) = \alpha + \beta + u_i,$$

where u_i is the value of a random effect for subject i, with $E(u_i) = 0$. This model allows heterogeneity among the probabilities for each question but assumes that the logit shifts uniformly for each subject by β for the two

TABLE 1
Opinions About Suicide and Letting an Incurable Patient Die

	Let Patient Die		
Suicide	Yes	No	Total
Yes	1097	90	1187
No	203	435	638
Total	1300	525	1825

Source: General Social Survey (1994).

questions. Thus, for each subject the odds of a "yes" response on letting the patient die equal $\exp(\beta)$ times the odds of a "yes" response on suicide. This is the special case of model (2) in which $\boldsymbol{\beta}^t = (\alpha, \beta)$, $\mathbf{x}_{i1}^t = (1,0)$ and $\mathbf{x}_{i2}^t = (1,1)$ for all i and $z_{ij} = 1$ for all i and j.

The usual random-effects model assumes that $\{u_i\}$ are independent from a $N(0, \sigma^2)$ distribution, with σ unknown, and that conditionally on the $\{u_i\}$ the $\{y_{ij}\}$ are independent. Unconditionally, variability among $\{u_i\}$ reflects subject heterogeneity, whereby different subjects have different probabilities of making a particular response. This variability induces a positive association between the binomial responses that form the margins of Table 1, manifested by a nonnegative log-odds ratio for the true cell probabilities underlying Table 1. This special case of a GLMM in which a cluster's random effect affects only the intercept of the model is called a *random intercept model*.

This model with logit link provides one of the rare instances in which a closed-form ML estimate is available for the effect β. When the sample log-odds ratio in tables that have the structure of Table 1 is nonnegative, it follows from Neuhaus, Kalbfleisch, and Hauck (1994) that $\hat{\beta}$ equals the log ratio of counts falling off the main diagonal; here, log(203/90) = .813. For each subject, the estimated odds of a "yes" response on letting the patient die are exp(.813) = 2.26 times the estimated odds of a "yes" response on suicide.

With this analysis for the logit model, the estimate of the treatment effect and subsequent inference does not depend on the other two cell counts. Matched-pairs data usually display a positive association, with the majority of the observations falling in these two cells. In Table 1, for instance, 1532 of the 1825 observations make no contribution to the analysis. The model provides justification for McNemar's test of equality of matched proportions, which uses only the same counts as does $\hat{\beta}$ (e.g., see Agresti and Finlay 1997:231). McNemar's test is based on a normal approximation for the probability that a binomial random variable with 90 + 203 = 293 trials and parameter .5 takes a value of at most 90 or at least 203.

2.2. *Cluster-Specific Effects for Random-Effects Models*

Note that the effect β in the model of Section 2.1 refers to the *conditional* log-odds ratio, given the random effect. Thus it has a *subject-specific* interpretation, being the change in the logit for a given subject. This is not the same as the marginal (so-called *population-averaged*) effect, because

of the nonlinearity of the logit link. For instance, in the matched-pairs model,

$$E(y_{ij}) = E[E(y_{ij}|u_i)] = E\left(\frac{\exp[\mathbf{x}_{ij}^t\boldsymbol{\beta} + u_i]}{1 + \exp[\mathbf{x}_{ij}^t\boldsymbol{\beta} + u_i]}\right),$$

and this expectation does not have the form $\exp(\mathbf{x}_{ij}^t\boldsymbol{\beta})/[1 + \exp(\mathbf{x}_{ij}^t\boldsymbol{\beta})]$ except when u_i has a degenerate distribution with a variance of 0. The estimate of β from the marginal (unconditional) model is typically smaller in absolute value than the estimate from the conditional model. The discrepancy between the two increases as σ, and hence the correlation between observations within a cluster, increases. For Table 1, the estimated marginal log-odds ratio uses the sample marginal distributions, equaling $\log[(1300/525)/(1187/638)] = .286$, compared to the conditional estimate of .813. Neuhaus, Kalbfleisch, and Hauck (1991) and Zeger, Liang, and Albert (1988) provided approximate relationships between the two types of estimate.

Similar remarks apply to any GLMM of form (2). For the probit link with binary data, however, the conditional probit model with normal random effects does imply a marginal model of probit form. In the case of a univariate random intercept, the conditional model has effect equal to the marginal effect multiplied by $[1 + \sigma^2]^{1/2}$ (Searle et al. 1992:377). For count data, such as those that occur in estimating rates or expected frequencies in a contingency table, the link function g is usually the log transform. In that case, the conditional loglinear model with normal random effect also implies a marginal model of log-linear form. The marginal model has the same effect but has intercept equal to the conditional one multiplied by $\exp(\sigma^2/2)$.

2.3. *Varieties of Ways of Handling Random Effects*

In any GLMM, a possibly controversial aspect is the assumption of a particular parametric distribution for the random effect. By far the most common choice is normality. For logit random-intercept models, evidence indicates that other choices usually provide similar results for the regression effects, with skewed actual distributions for the random effects possibly having some effect on the intercept estimate (Neuhaus, Hauck, and Kalbfleisch 1992). When the random effect relates more directly to the characteristic estimated, the choice of distribution can be more crucial

(Heckman and Singer 1984). An important advantage of the normal family is its extension to models with several correlated random effects, in which case the multivariate normal family is both flexible and simple.

The research literature now has a considerable variety of ways of handling clustering of various sorts. At first, *conjugate* mixture models received most of the attention. These are models that assume a particular parametric distribution but with the parameter itself coming from a distribution such that the marginal distribution has closed form. For binary data, one assumes that given the parameter, the response has a binomial distribution, and the parameters follow a beta distribution. This leads to the *beta binomial* model (Crowder 1978). For count data, one assumes that given the parameter the responses are Poisson, and the parameters follow a gamma distribution. This leads to the *negative binomial model* (Lawless 1987; McCullagh and Nelder 1989). A disadvantage of the conjugate approach is the lack of generality and flexibility, requiring a different mixture distribution for each type of problem. In addition, the extra variability does not enter on the same scale as the ordinary predictors. Lee and Nelder (1996) generalized these models to hierarchical models of GLMM form but in which the random effect need not be normal.

Finally, there is also some work on a nonparametric random effects approach, in which instead of choosing a parametric family one uses a discrete mixture determined empirically (Heckman and Singer 1984; Aitkin 1996). We discuss this further in Section 3.5.

3. EXAMPLES OF RANDOM-EFFECTS MODELING

This section, the heart of the paper, shows examples of random-effects modeling for a variety of types of examples.

3.1. *Example 1: Shrinkage of Proportions*

This example involves estimating a large number of proportions with only small to moderate sample sizes for each proportion. A common application is *small-area estimation*, in which relatively few observations occur in each of many geographical areas. For each area, for instance, one might want to estimate the unemployment rate or the proportion of families not having health insurance coverage or the proportion of children living in single-parent families. Random effects models can serve as a mechanism for improving on the sample proportions in estimating the true area-

specific proportions. In assuming that those true proportions vary according to some distribution, one can use information from all the areas to estimate the proportion in any given area (Ghosh and Rao 1994).

Let π_i denote the true proportion in area i, $i = 1, \ldots, I$. Let $\{Y_i\}$ denote independent binomial variates with sample size indices $\{n_i\}$ and parameters $\{\pi_i\}$; that is, $Y_i = \sum_{j=1}^{n_i} y_{ij}$, where $\{y_{ij}, j = 1, \ldots, n_i\}$ are independent with $P(y_{ij} = 1) = \pi_i$ and $P(y_{ij} = 0) = 1 - \pi_i$. The sample proportions $\{p_i = Y_i/n_i\}$ are the ML estimates of $\{\pi_i\}$ for the fixed-effects model

$$\text{logit}(\pi_i) = \alpha + \beta_i, \quad i = 1, \ldots, I,$$

where identifiability requires a constraint such as $\sum \beta_i = 0$ or $\beta_I = 0$. This model is saturated, having I nonredundant parameters for the I binomial observations. For small samples, the sample estimates often display much more variability than the true values, and when $\{\beta_i\}$ are similar it can be helpful to shrink the sample proportions toward the overall mean. One can accomplish this using the random effects model

$$\text{logit}(\pi_i) = \alpha + u_i, \tag{3}$$

where $\{u_i\}$ are assumed to be independent from a $N(0, \sigma^2)$ distribution. After estimating α and σ, one estimates $\text{logit}(\pi_i)$ using $[\hat{\alpha} + \hat{u}_i]$, where $\hat{u}_i$ is the predicted random effect based on the observed data. Section 4.5 outlines the standard method for computing $\hat{u}_i$. This is an example of an *empirical Bayes analysis*, which assumes a prior distribution for unknown parameters and uses the data to estimate parameters of that distribution. This paradigm has been in use for some time—for instance, using normal approximations for distributions of sample proportions (Efron and Morris 1975).

To motivate why the random-effects approach can be highly beneficial, we first use an artificial example. For I coins, let π_i denote the probability of a "head" in a single flip of coin i, and suppose $\{Y_i\}$ are counts of heads based on $\{n_i = 10\}$ flips of each coin. Probably $\{\pi_i\}$ are all close to .50 and the sample data would yield estimates for model (3) close to $\hat{\alpha} = 0$ and $\hat{\sigma} = 0$ (especially if I is large). In fact, if $\hat{\sigma} = 0$ the model fit simplifies to that for the simpler model $\text{logit}(\pi_i) = \alpha$, and $\{\hat{\pi}_i = (\sum_h Y_h)/(\sum_h n_h)\}$, the overall sample proportion of heads. This estimate, or estimates that are very close to it (which occur when $\hat{\sigma} > 0$

but is small) tend to be much better than the separate sample proportions $\{p_i = Y_i/n_i\}$. Generally, "borrowing from the whole" provides the advantage of smoothing the sample estimates and effectively basing the resulting estimates on much larger sample sizes than using the data for the separate samples on their own. For instance, in the extreme case that $\sigma = 0$ and $\hat{\sigma} = 0$, the random effects estimate of the common probability of a head uses I times as many observations as the separate sample proportions. Generally, the random-effects estimates provide shrinkage of the separate estimates toward the overall sample proportion. The amount of shrinkage decreases as $\hat{\sigma}$ or n_i grow.

For an illustration based on a more realistic problem, we simulated a sample to mimic a poll taken before the 1996 U.S. presidential election. For a sample of size n_i in state i ($i = 1, \ldots, 51$ with $i = 51$ being the District of Columbia), we generated Y_i as a binomial variate with π_i equal to the actual proportion of votes in state i for Bill Clinton in the 1996 election, conditional on voting for Clinton or Dole. We set n_i proportional to the population size in that state, subject to $\sum n_i = 2000$. The $\{n_i\}$ ranged from 4 to 240. Table 2 shows $\{n_i\}$, $\{\pi_i\}$ and $\{p_i = Y_i/n_i\}$.

The ML fit of the random-effects model (3) (using PROC NLMIXED in SAS) provides $\hat{\alpha} = .164$ and $\hat{\sigma} = .29$. For that model, the predicted random-effects values (also estimated using NLMIXED) yield the corresponding proportion estimates $\{\hat{\pi}_i^{(1)} = \exp(\hat{\alpha} + \hat{u}_i)/[1 + \exp(\hat{\alpha} + \hat{u}_i)]\}$ (also shown in Table 2). Since the sample sizes are mostly small and since $\hat{\sigma}$ is relatively small, the amount of shrinkage for these estimates is considerable. They vary between only .468 (Texas) and .696 (New York), whereas the sample proportions vary between .111 (for Idaho) and 1.0 (DC). Estimates based on fewer observations, such as DC, tend to receive greater shrinkage. Although the random-effects estimates are relatively homogeneous, they do tend to be closer than the sample proportions to the true values. For instance, $\sum|p_i - \pi_i|/51 = .079$ and $\sum|\hat{\pi}_i^{(1)} - \pi_i|/51 = .053$.

To check whether these results are typical, we simulated 10,000 studies with these sample sizes and probabilities. Overall, the mean distance from true probabilities was .060 for the random-effects estimates and .091 for the sample proportions. In 99.6 percent of the studies, the mean distance was smaller for the random-effects estimates.

Although the random-effects estimates tend to be closer than the sample proportions to the true proportions, the amount of shrinkage can be excessive, given other information we know about presidential elections.

TABLE 2
Estimates of Proportion of Vote for Clinton, Conditional on Voting for Clinton or Dole in 1996 U.S. Presidential Election

State	n_i	π_i	p_i	$\hat{\pi}_i^{(1)}$	$\hat{\pi}_i^{(2)}$	State	n_i	π_i	p_i	$\hat{\pi}_i^{(1)}$	$\hat{\pi}_i^{(2)}$
AK	5	0.394	0.200	0.508	0.438	MT	7	0.483	0.429	0.526	0.528
AL	32	0.463	0.500	0.524	0.484	NC	55	0.475	0.455	0.494	0.492
AR	19	0.594	0.526	0.537	0.604	ND	5	0.461	0.600	0.546	0.444
AZ	34	0.512	0.618	0.573	0.531	NE	13	0.395	0.462	0.524	0.408
CA	240	0.572	0.538	0.538	0.557	NH	9	0.567	0.556	0.543	0.527
CO	29	0.492	0.586	0.558	0.553	NJ	60	0.600	0.667	0.611	0.579
CT	25	0.604	0.720	0.602	0.588	NM	13	0.540	0.462	0.524	0.556
DC	4	0.903	1.000	0.576	0.909	NV	12	0.506	0.500	0.533	0.530
DE	5	0.586	0.400	0.527	0.561	NY	137	0.660	0.752	0.696	0.686
FL	108	0.532	0.602	0.583	0.553	OH	84	0.536	0.488	0.507	0.510
GA	56	0.494	0.554	0.548	0.531	OK	23	0.456	0.478	0.520	0.463
HI	9	0.643	0.556	0.543	0.580	OR	24	0.547	0.625	0.569	0.589
IA	22	0.557	0.500	0.528	0.544	PA	90	0.552	0.567	0.558	0.569
ID	9	0.391	0.111	0.472	0.395	RI	7	0.689	0.571	0.545	0.629
IL	89	0.596	0.539	0.540	0.574	SC	28	0.469	0.571	0.552	0.491
IN	44	0.468	0.432	0.488	0.464	SD	6	0.479	0.667	0.555	0.502
KS	19	0.400	0.316	0.477	0.455	TN	40	0.513	0.500	0.522	0.531
KY	29	0.506	0.448	0.506	0.516	TX	144	0.473	0.444	0.468	0.465
LA	33	0.566	0.667	0.592	0.571	UT	15	0.380	0.333	0.490	0.372
MA	46	0.686	0.739	0.637	0.665	VA	51	0.489	0.412	0.473	0.465
MD	38	0.586	0.474	0.511	0.566	VT	4	0.633	0.500	0.538	0.615
ME	9	0.627	0.778	0.578	0.591	WA	42	0.572	0.619	0.578	0.599
MI	73	0.573	0.589	0.570	0.573	WI	39	0.559	0.487	0.517	0.529
MN	35	0.594	0.571	0.554	0.588	WV	14	0.584	0.571	0.548	0.591
MO	41	0.535	0.561	0.550	0.575	WY	4	0.426	0.250	0.518	0.470
MS	21	0.472	0.333	0.477	0.445						

Note: π_i = true, p_i = sample, $\hat{\pi}_i^{(1)}$ = random effects, $\hat{\pi}_i^{(2)}$ = random effects with shrinkage toward 1992 election.

For instance, in 16.1 percent of the simulations $\hat{\sigma}$ was so small that these estimates predicted a Clinton victory in every state. Rather than assuming a common mean for the random effects, one might instead use supplementary information that should improve the predictions. For instance, let q_i denote the true proportion of votes for Clinton in state i in the 1992 election, conditional on voting for Clinton or Bush. This is known information for polls taken in 1996, and one could fit the model

$$\text{logit}(\pi_i) = \text{logit}(q_i) + \alpha + u_i,$$

where $\{q_i\}$ are known and $\{u_i\}$ are independent from a $N(0, \sigma^2)$ distribution. Known terms in the linear predictor, such as $\text{logit}(q_i)$, are referred to as *offsets*. Rearranging the previous equation we obtain

$$\log \frac{\pi_i/(1-\pi_i)}{q_i/(1-q_i)} = \alpha + u_i. \tag{4}$$

Thus, $\alpha + u_i$ represents the log-odds ratio for the ith state of voting for Clinton versus Dole in 1996 relative to voting for Clinton versus Bush in 1992.

Table 2 also shows the resulting estimates $\{\hat{\pi}_i^{(2)}\}$ of $\{\pi_i\}$. Here, $\hat{\sigma} = .19$ and the estimates shrink considerably toward the prior values from 1992, with a slight upward adjustment since $\hat{\alpha} = .205$. For model (4), when $\hat{\sigma} = 0$, $\hat{\pi}_i^{(2)} = q_i \exp(\hat{\alpha})/[1 - q_i + q_i \exp(\hat{\alpha})]$, and when also $\hat{\alpha} = 0$, $\hat{\pi}_i^{(2)} = q_i$. Otherwise when $\hat{\sigma} = 0$, compared to the previous election results, the estimates shift up or down on the logit scale depending on how the overall Democratic vote compares in the current poll to the previous election (i.e., depending on $\hat{\alpha}$).

With model (4), the random effects estimates vary between .372 (for Utah) and .909 (for DC), whereas the true values vary between .380 (for Utah) and .903 (for DC). Now, these random-effect estimates tend to be much closer than the sample proportions to the true values, with, $\sum|p_i - \pi_i|/51 = .079$ and $\sum|\hat{\pi}_i^{(2)} - \pi_i|/51 = .024$. Over 10,000 simulations, the mean distance values were .091 and .027, and the mean distance was smaller for the random-effects estimates in 100 percent of the cases. In 27.5 percent of these cases $\hat{\sigma} = 0$, but in none of them did the random-effects estimates predict a Clinton victory in each state. Figure 1 displays the values of $(\pi_i, \hat{\pi}_i, q_i, \hat{\pi}_i^{(2)})$ for the data in Table 2, with the states ordered by their values of $\{\pi_i\}$.

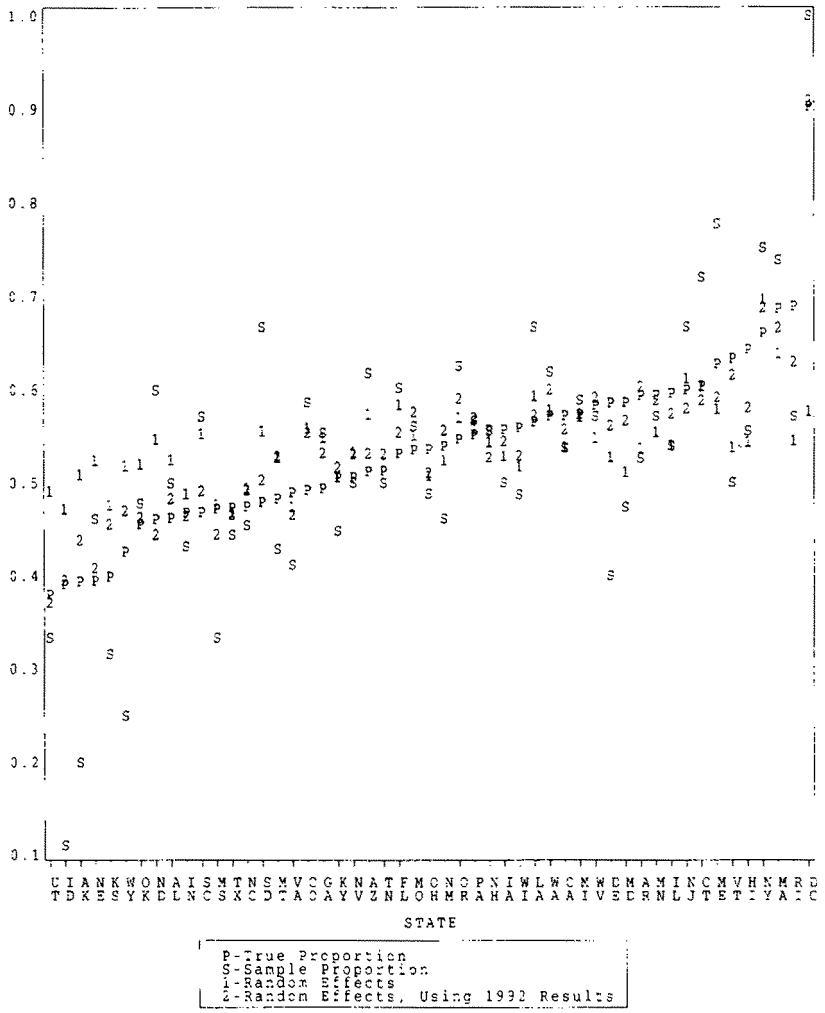

FIGURE 1. Estimates of statewide proportions planning to vote for Clinton in 1996 presidential election (conditional on voting for Clinton or Dole), based on random sample of 2000 voters.

3.2. *Example 2: Modeling Repeated Binary Measurement*

Section 2.1 introduced a random-effects model for repeated categorical measurement with the simple case of binary matched pairs. The model then has the form

$$\text{logit}[P(y_{i1} = 1)] = \alpha + u_i, \qquad \text{logit}[P(y_{i2} = 1)] = \alpha + \beta + u_i,$$

where $\{u_i\}$ are an independent sample from a $N(0, \sigma^2)$ distribution. Let n_{ab} denote the number of observations for which $(y_{i1}, y_{i2}) = (a, b)$, $a = 0,1$, $b = 0,1$. The counts can be summarized in a table such as Table 1. Let $Y_1 = \sum y_{i1}$ and $Y_2 = \sum y_{i2}$. Unconditionally, Y_1 is a binomial random variable with parameter $E\{\exp(\alpha + u)/[1 + \exp(\alpha + u)]\}$, and Y_2 is a binomial random variable with parameter $E\{\exp(\alpha + \beta + u)/[1 + \exp(\alpha + \beta + u)]\}$, where the expectation is taken with respect to u, a $N(0, \sigma^2)$ random variable.

This model implies a nonnegative correlation between the binomial variates Y_1 and Y_2, with greater association resulting from greater heterogeneity (i.e., larger σ). Under this model, Y_1 and Y_2 are independent only if $\sigma = 0$. When the sample data are consistent with this model, in the sense that $\log(n_{00}n_{11}/n_{10}n_{01}) \geq 0$, then the perhaps surprising result (Neuhaus et al. 1994) occurs that $\hat{\beta} = \log(n_{01}/n_{10})$. This is also the estimate with the conditional ML approach, eliminating $\{u_i\}$ by conditioning on their sufficient statistics.

This random-intercept model extends to more than two repeated measurements, with covariates that themselves may or may not vary. We illustrate using Table 3, which contains data from the 1994 General Social Survey. The subjects indicate whether they support legalizing abortion in three situations: (1) if the family has a very low income and cannot afford any more children, (2) when the woman is not married and does not want to marry the father, and (3) if the woman wants to terminate the pregnancy

TABLE 3
Cross Classification of Support for Legalizing Abortion in Three Cases, by Gender

	Sequence of Responses on the Three Items*							
Gender	(1,1,1)†	(1,1,2)	(2,1,1)	(2,1,2)	(1,2,1)	(1,2,2)	(2,2,1)	(2,2,2)
Male	342	26	6	21	11	32	19	356
Female	440	25	14	18	14	47	22	457

Source: Data from *General Social Survey* (1994).

*Respondents were asked about their support for legalizing abortion (1) if the family has a very low income and cannot afford any more children, (2) if the woman is not married and does not want to marry the father, and (3) if the woman wants to terminate the pregnancy for any reason.

†Response 1 is "yes" and 2 is "no."

for any reason. Table 3 also classifies the subjects by gender. Let y_{ij} denote the response for subject i on item j, with a value of 1 representing support. We consider the model

$$\text{logit}[P(y_{ij} = 1)] = \alpha + \beta_j + \gamma x + u_i, \tag{5}$$

where $x = 1$ for females and 0 for males, $\{u_i\}$ are independent from a $N(0, \sigma^2)$ distribution, and $\{\beta_j\}$ satisfy a constraint such as $\beta_3 = 0$. (Equivalently, one could delete α from the model and then remove the constraint on $\{\beta_j\}$ or allow a nonzero mean for $\{u_i\}$.)

For the ML fit of this model, contrasts of $\{\hat{\beta}_j\}$ provide evidence of greater support for legalized abortion when the family has a low income and cannot afford any more children than in the other two instances ($\hat{\beta}_1 - \hat{\beta}_3 = .83$, $se = .16$, $\hat{\beta}_1 - \hat{\beta}_2 = .54$, $se = .16$) and slight evidence of greater support when the woman is not married and does not want to marry the man than when the woman wants the abortion for any reason ($\hat{\beta}_2 - \hat{\beta}_3 = .29$, $se = .16$). Also, $\hat{\gamma} = .01$ ($se = .49$). The estimates have log-odds ratio interpretations. For each item, for instance, the estimated odds of females supporting legalized abortion equal $\exp(.01) = 1.01$ times the estimated odds for males. For these data, subjects are highly heterogeneous ($\hat{\sigma} = 8.8$, with $se = .54$), resulting in strong associations among the items as reflected by 1595 of the 1850 observations falling in the four cells where subjects made the same response on all three items. We also considered the interaction model having different $\{\beta_j\}$ for men and women, but it did not provide an improved fit (likelihood-ratio statistic = 1.0 with $df = 2$ for testing that the extra parameters equal 0). Essentially, there is no difference between males and females in this study.

3.3. *Example 3: Summarizing Results from Several 2-by-2 Tables*

Many applications refer to comparing two groups on a binary response when data are stratified according to levels of a third variable. The data then take the form of several 2-by-2 contingency tables. The strata are sometimes themselves a sample—for example, schools or medical centers; or they may be levels of a control variable, such as age or severity of the condition being treated, or combinations of levels of several control variables; or, they may be different studies of the same sort evaluated in a meta-analysis. The main concerns for data of this sort relate to investigat-

ing the "average" level of association and the degree of variability about that average (i.e., the treatment-by-center interaction); for instance, see DerSimonian and Laird (1986).

When the strata are sampled, a random-effects approach is natural. One then has a structure for extending inferences to the population of strata sampled. Moreover, the random-effects model can provide a simple summary such as an estimated mean and standard deviation of log-odds ratios for the population of centers. It can also provide predicted odds ratios for separate strata that have the benefits of shrinkage, especially when sample sizes in some of the strata are small. Even when the strata are not a sample, the model can be beneficial for these two purposes.

We illustrate using Table 4, which shows the results of admissions decisions for applicants to graduate school in departments of the College of Liberal Arts and Sciences at the University of Florida during the 1997–1998 academic year. Stratifying by the department to which the student applied, the table compares males and females on whether admitted. Overall 983 men applied with 35.9 percent accepted, and 1093 females applied with 34.4 percent accepted.

For a subject of gender j ($j = 1$, males, $j = 2$, females) applying to department i, let π_{ij} denote the probability of being admitted. One possible model is the logit-normal model

$$\text{logit}(\pi_{i1}) = \alpha + \beta/2 + u_i, \qquad \text{logit}(\pi_{i2}) = \alpha - \beta/2 + u_i, \tag{6}$$

assuming that $\{u_i\}$ are independent from a $N(0, \sigma^2)$ distribution. This model assumes that the log-odds ratio β between gender and being admitted is constant over departments. A logit-normal model permitting interaction is

$$\text{logit}(\pi_{i1}) = \alpha + b_i/2 + u_i, \qquad \text{logit}(\pi_{i2}) = \alpha - b_i/2 + u_i, \tag{7}$$

where $\{u_i\}$ are independent from a $N(0, \sigma_u^2)$ distribution, $\{b_i\}$ are independent from a $N(\beta, \sigma_b^2)$ distribution, β is the expected value of study-specific log-odds ratios and σ_b describes the variability in those log-odds ratios. The model parameters are then $(\alpha, \beta, \sigma_u, \sigma_b)$.

For these data, a standard fixed-effects analysis for the lack of interaction model has ML estimate of the gender effect $\hat{\beta} = -.173$ ($se = .112$). The corresponding model (6) in which departments are a random effect also has $\hat{\beta} = -.163$ ($se = .111$). The model (7) permitting interaction but assuming that $\{u_i\}$ are independent of $\{b_i\}$ has $\hat{\beta} = -.18$ ($se = .13$), with $\hat{\sigma}_b = .20$ ($se = .35$). Similar answers result from allowing

TABLE 4
1997–1998 Admissions Decisions to Graduate School at the University of Florida

	Males		Females				Males		Females		
Department	Yes*	No	Yes	No	OR**	Department	Yes*	No	Yes	No	OR**
Anthropology	21	41	32	81	1.30	Linguistics	7	8	21	10	.42
Astronomy	3	8	6	0	0	Mathematics	31	37	25	18	0.60
Chemistry	34	110	12	43	1.11	Philosophy	9	6	3	0	0
Classics	4	0	3	1	∞	Physics	25	53	10	11	.52
Communicative Processes	5	10	52	149	1.43	Political Science	39	49	25	34	1.08
Computer Science	6	12	8	7	.44	Psychology	4	41	2	123	6.00
English	30	112	35	100	.77	Religion	0	2	3	3	0
Geography	11	11	9	1	.11	Romance Languages	6	3	29	13	.90
Geology	15	6	6	3	1.25	Sociology	7	17	16	33	.85
Germanic/Slavic	4	1	17	0	0	Statistics	36	14	23	9	1.01
History	21	19	9	9	1.11	Zoology	10	54	4	62	2.87
Latin American Studies	25	16	26	7	0.42						

*Yes = accept, no = reject; **OR = sample odds ratios.

$\{u_i\}$ and $\{b_i\}$ to be correlated. For all these models the estimated gender effect is not large. For instance, for the interaction model the estimated mean log-odds ratio of $-.176$ corresponds to an odds ratio of .84. The gender effect is not significant. For instance, the likelihood-ratio statistic for testing $H_0: \beta = 0$ is 2.4 ($df = 1$) for the fixed-effects model and 2.0 for the random-effects model allowing interaction. Note that because of the extra source of variability in the interaction models, the standard error of $\hat{\beta}$ is slightly larger than with the other analyses.

As in Example 1, one can obtain smoothed estimates, this time of the department-specific odds ratios. Table 4 shows that the sample odds ratios varied between 0 (for Astronomy, Germanic and Slavic Languages, Philosophy, Religion) and ∞ (Classics); by contrast, the estimates of $\exp(b_i)$ for model (7) vary between .75 (for Astronomy) and .96 (for Zoology). There is not much variability in these predictions, since the estimated variance component for the interaction is so small. Interestingly, Simpson's paradox occurs here, as the marginal sample odds ratio relating gender to whether admitted equals 1.07, whereas the predicted odds ratio for every department is less than 1.0.

Incidentally, in passing we mention that one needs to be careful about implications of the model formula expression and the random-effects structure. For instance, model (7) with uncorrelated $\{b_i\}$ and $\{u_i\}$ is different from the model

$$\text{logit}(\pi_{i1}) = \alpha + b_i + u_i, \qquad \text{logit}(\pi_{i2}) = \alpha + u_i,$$

with uncorrelated $\{b_i\}$ and $\{u_i\}$. Generally, if the interaction model has form $\text{logit}(\pi_{ij}) = \alpha + b_i x_j + u_i$ where x_j is a dummy variable (with, e.g., $x_1 - x_2 = 1$), and if we let $z_j = x_j + c$ for some constant c, then the model is also $\text{logit}(\pi_{ij}) = \alpha + b_i(z_j - c) + u_i = \alpha + b_i z_j + v_i$, where $v_i = u_i - cb_i$. Thus (b_i, v_i) are correlated even if (b_i, u_i) are not.

3.4. *Example 4: Hierarchical Modeling*

Hierarchical data, in which units are grouped at different levels, is common in the social sciences. Models for data in which hierarchical grouping or clustering occurs are often referred to as *multilevel models*. These models fall within the GLMM framework that is the focus of this paper. For example, in a study of factors that affect school performance, the level 1 units might be students, the level 2 units schools, and the level 3 units the

county, region, or school district. Clearly, socioeconomic factors that affect school choice will often cause students within the same schools to have correlated responses. Similarly, variation between locations may induce additional correlation at a regional level. Correlations caused by additional sources of variability are accounted for in a multilevel model through the inclusion of random effects at each stage of the hierarchy. The variances of these effects are estimated as part of the model-fitting process, and they measure the amount of variability not explained by fixed effects at each level. For early descriptions of the use of random-effects modeling with binary responses in the context of educational assessment, see Aitkin, Anderson, and Hinde (1981) and Aitkin and Longford (1986). For a recent application of multilevel modeling to social observation of neighborhoods, see Raudenbush and Sampson (1999). Other general references include Bryk and Raudenbush (1992), Goldstein (1995), Plewis (1997), and Muthen (1997).

Goldstein (1995, sec. 7.3) provided an example from a survey of voting behavior in the United Kingdom with a similar multilevel structure. The data in this case were obtained from a series of surveys carried out in Britain following elections held in 1983, 1987, and 1992. Respondents were grouped according to year and by the parliamentary constituency in which they lived at the time. A binary response variable of interest is whether or not an individual voted for the Conservative party as opposed to the Liberal or Labour parties. Some constituencies were sampled in all three years and others in only one or two years, resulting in a multilevel structure with respondent at level 1 and year by constituency combinations at level 2. Let π_{ijk} denote the probability that respondent k in year j from constituency i says that he or she voted for the Conservatives. Then a potential GLMM for this data is

$$\text{logit}(\pi_{ijk}) = \mathbf{x}_{ijk}^t \boldsymbol{\beta} + u_{ij}, \tag{8}$$

where $\mathbf{x}_{ijk}^t \boldsymbol{\beta}$ models fixed effects such as socioeconomic status and (u_{i1}, u_{i2}, u_{i3}) is a trivariate normal random variable representing the effects of constituency i in the three election years. If we assume that between-constituency variation is the same in each year and that the correlations between pairs of years are all equal, then (8) is equivalent to the model:

$$\eta_{ijk} \equiv \text{logit}(\pi_{ijk}) = \mathbf{x}_{ijk}^t \boldsymbol{\beta} + u_{ij} + v_i, \tag{9}$$

where now u_{i1}, u_{i2}, u_{i3} and v_i are independent normal variables, with $u_{ij} \sim N(0, \sigma_u^2)$ and $v_i \sim N(0, \sigma_v^2)$. This model formally has a three-level struc-

ture. Level 1 variation between respondents in the same year and constituency (and the same fixed factors) is Bernoulli. This is combined with a normal random effect (u_{ij}) which accounts for year-to-year (level 2) variation in log-odds ratios for respondents in the same constituency. Finally, level 3 variation between respondents in different constituencies is a combination of Bernoulli and two independent normal random effects (u_{ij} and v_i). Contrasting models (8) and (9) illustrates an important point about multilevel modeling. Incorporating an additional level of hierarchy in (9) led to a more parsimonious model requiring only two parameters (σ_u and σ_v) to describe the random-effects distribution compared with six in the two-level model (8).

In the survey of voting behavior data, Goldstein obtained the estimates $\sigma_u^2 = 0.05$ and $\sigma_v^2 = 0.38$. This implies a correlation between the log-odds of a respondent saying they voted Conservative in two different election years of

$$\text{cor}(\eta_{ijk}, \eta_{ij'k}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2} = \frac{0.05}{0.05 + 0.38} = 0.12.$$

An interesting feature of the model is that this correlation is the same for two different subjects—i.e., $\text{cor}(\eta_{ijk}, \eta_{ij'k}) = \text{cor}(\eta_{ijk}, \eta_{ij'k'})$. In fact an unattractive property of the model is that equality also holds for the correlation between the corresponding pairs of binary responses:

$$\text{cor}(Y_{ijk}, Y_{ij'k}) = \text{cor}(Y_{ijk}, Y_{ij'k'}).$$

Intuitively, one would expect a higher correlation between responses on the same subject. In principle this "deficiency" could be overcome by adding an additional random effect for each respondent. However, the amount of information for estimating the additional variance component is limited in this example, since there are a maximum of three repeated measurements on each respondent. In addition, Goldstein found little evidence of lack of fit for the simpler model.

3.5. *Example 5: Nonparametric Random-Effects Approach*

The examples discussed so far have assumed a normal distribution for random-effects distributions. An alternative, nonparametric, random-effects approach uses instead a distribution on a finite set of mass points having location that is estimated empirically (Heckman and Singer 1984; Lind-

say, Clogg, and Grego 1991; Aitkin 1996, 1999). With latent-class models, one specifies the number of mass points for this mixture distribution. The general approach does not specify the number of mass points, but one treats it as fixed and sequentially increases the sample value until the likelihood is maximized. In fact, maximizing the likelihood usually requires relatively few mass points. Aitkin (1996, 1999) presented examples of the general nonparametric approach. Heckman and Singer (1984) noted that this approach is primarily useful when the mixing distribution is a nuisance parameter rather than of direct interest, since the nonparametric estimate of that distribution may be poor even for large samples.

Follmann and Lambert (1989) provided an interesting example in which the number of mass points was prespecified. They analyzed data on the effect of the dosage of a poison on the death rate of a protozoan of a particular genus, assuming two varieties of that genus. For their model, the probability of death at a particular dosage level x equals $\rho\pi_1(x) + (1-\rho)\pi_2(x)$, where $\text{logit}[\pi_i(x)] = \alpha_i + \beta x$ and the mixing proportion ρ is unknown. The fit of this model was much better than that of a single logistic regression model.

In a similar spirit, Lindsay et al. (1991) studied models of the Rasch form (1) but in which the subject term can assume only an unknown finite number of values. They showed that with k items, the likelihood is maximized when the subject parameter takes at most $(k+1)/2$ values. In related work, Tjur (1982) showed that with a distribution-free approach for the subject term, the fit of the Rasch model satisfies the quasi-symmetry log-linear model; see Conaway (1989), Darroch (1981), Hatzinger (1989), Kelderman (1984), and Agresti (1993, 1995, 1997) for related work and details.

We illustrate the nonparametric approach by fitting model (5) to the attitudes about abortion data in Table 3, now using a finite mixture distribution rather than a normal distribution for the random effect u_i. The likelihood is then maximized with a two-point mixture. The results are very similar to those obtained with the normal mixture model. For instance, with the nonparametric approach $\hat{\beta}_1 - \hat{\beta}_3 = .833$ ($se = .16$) and $\hat{\beta}_2 - \hat{\beta}_3 = .304$ ($se = .16$), compared to $\hat{\beta}_1 - \hat{\beta}_3 = .834$ ($se = .16$) and $\hat{\beta}_2 - \hat{\beta}_3 = .292$ ($se = .16$) with the normal approach.

It follows from the papers cited above that one can also estimate the within-subject comparisons of items $\beta_h - \beta_j$ in model (5) by fitting a quasi-symmetric log-linear model. Let $\mu_g(h_1, h_2, h_3)$ denote the expected frequency for gender g making response h_j to item j, $j = 1,2,3$, where $g = 1$ for

female and 0 otherwise and where $h_j = 1$ for approval of legalized abortion for item j and 0 otherwise. This model is

$$\log \mu_g(h_1, h_2, h_3) = \beta_1 I(h_1 = 1) + \beta_2 I(h_2 = 1) + \beta_3 I(h_3 = 1) + \tau I(g = 1) + \sum_{k=0}^{3} \lambda_k I(h_1 + h_2 + h_3 = k), \quad (10)$$

where $I(\cdot)$ is the indicator function. Here, λ_k is a parameter referring to all cells in which subjects voiced approval in k of the three items, $k = 0, 1, 2, 3$. Fitting this model with ordinary software for GLMs (such as PROC GENMOD in SAS), we obtain $\hat{\beta}_1 - \hat{\beta}_2 = .521$ ($se = .154$), $\hat{\beta}_1 - \hat{\beta}_3 = .828$ ($se = .160$), $\hat{\beta}_2 - \hat{\beta}_3 = .307$ ($se = .161$), very similar to the normal random effects and nonparametric random-effects estimates. In fact, it follows from Tjur (1982) that these estimates also equal those obtained using conditional ML with model (5), treating the subject terms as fixed. With this approach (and with conditional ML), however, one cannot estimate between-groups effects, such as the gender effect in model (5).

3.6. *Example 6: Matched Pairs with a Bivariate Binary Response*

Examples discussed so far have had univariate random effects or independent random effects. We next show an example in which a multivariate, correlated random-effects structure is natural. We use Table 5, taken from Coleman (1964) and analyzed in several papers by Leo Goodman—e.g., Goodman (1974). A sample of schoolboys were interviewed twice, several months apart, and asked about their self-perceived membership in the "leading crowd" and about whether one must sometimes go against their principles to be part of that leading crowd. Thus there are two variables, which we refer to as membership (M) and attitude (A), measured at each of two interview times for each subject. Table 5 labels the categories for attitude as positive and negative, where positive refers to disagreeing with the statement that students must go against their principles.

For subject i, let π_{ijv} be the probability of response in category 1 for variable v at interview time j. We consider the multivariate logit model

$$\text{logit}(\pi_{ijv}) = \beta_{jv} + u_{iv}, \quad (11)$$

in which (u_{i1}, u_{i2}) describes subject heterogeneity for membership and attitude and $\beta_{2v} - \beta_{1v}$ describes the change in the response distribution for

TABLE 5
Cross Classification Illustrating Matched Pairs with a Bivariate Binary Response

(M, A)* for first interview		(M, A) for second interview				
		(Yes, Positive)	(Yes, Negative)	(No, Positive)	(No, Negative)	Total
Yes	Positive	458	140	110	49	757
		(451.3)	(145.5)	(121.9)	(48.8)	
Yes	Negative	171	182	56	87	496
		(173.5)	(180.5)	(58.2)	(73.4)	
No	Positive	184	75	531	281	1071
		(178.0)	(71.3)	(531.9)	(279.4)	
No	Negative	85	97	338	554	1074
		(85.0)	(107.2)	(333.2)	(558.8)	
Total		898	494	1035	971	3398

Source: From Coleman (1964).
*Membership (M) and attitude (A) toward the "leading crowd"; fitted values for model (11) are in parentheses.

variable v between interview times 2 and 1; that is, there are two nondegenerate random effects, one for attitude and one for membership. We fitted this model assuming that $\{(u_{i1}, u_{i2})\}$ are a random sample from a bivariate normal distribution.

Let M denote the membership variable and A the attitude variable. The ML fit of the bivariate random-effects model yields $\hat{\beta}_{2M} - \hat{\beta}_{1M} = .366$ (std. error = .073) and $\hat{\beta}_{2A} - \hat{\beta}_{1A} = .176$ (std. error = .058). For both variables, the probability of the first response is higher at the second interview; for instance, for each subject the odds of self-perceived membership in the leading crowd at time 2 are estimated to be $\exp(.366) = 1.44$ times the odds at time 1. The values in the estimated covariance matrix (not reported here) suggest that there is more heterogeneity with respect to membership than with respect to attitude, and the estimated correlation between the random effects is 0.33. The model fits well, with likelihood-ratio statistic (deviance) comparing the observed counts to the fitted values equal to $G^2 = 5.5$, based on $df = 8$. The likelihood-ratio test comparing this model to the one that constrains $\beta_{2M} - \beta_{1M} = 0$ and $\beta_{2A} - \beta_{1A} = 0$ has test statistic 35.2 based on $df = 2$. The model constraining the random effects to be uncorrelated also fits poorly ($G^2 = 97.5$, $df = 9$), as does the model with perfectly correlated random effects ($G^2 = 655.5$, $df = 10$).

For this model, Agresti (1997) used a nonparametric approach, whereby a lack of assumption about the distribution of (u_{i1}, u_{i2}) motivates a quasi-symmetric log-linear model. This yields the same estimates as obtained with a conditional ML (CML) approach that eliminates (u_{i1}, u_{i2}) by conditioning on their sufficient statistics. The results reported here are nearly identical to those obtained using that approach (see Agresti 1997, but the attitude labels and the sign of the estimates are incorrectly stated there). There are a few basic differences, however. For instance, the nonparametric/CML approach necessarily provides estimates of $\beta_{2M} - \beta_{1M}$ and $\beta_{2A} - \beta_{1A}$ that are identical to the CML estimates from two separate univariate analyses. In essence, because that approach makes no assumption about the joint distribution of the random effects, the multivariate form of the data does not affect the analysis. This does not happen for the bivariate normal random effects analysis, although the estimates are very close to those obtained with univariate analyses.

The approach described above with correlated normal random effects is a continuous analog to discrete latent-class models proposed by Goodman (1974), based on two associated binary latent variables. Results

are similar for the two approaches, although advantages of the random-effects model are that it is more parsimonious and it directly provides estimates $\beta_{2M} - \beta_{1M}$ and $\beta_{2A} - \beta_{1A}$ that compare the margins of the observed classifications. For additional examples of multivariate random effects analyses, see Coull and Agresti (2000).

3.7. *Example 7: Extensions to Ordinal/Nominal Response Data*

Random-effects models for binary data extend to handle multinomial responses, whether measured on ordinal or nominal scales. For instance, let y_{ij} denote the jth response in cluster i, where the possible values for y_{ij} are the response category outcomes $1,2,\ldots,K$. For ordinal responses, GLMMs have been formulated for the $K-1$ cumulative logits,

$$\text{logit}[P(y_{ij} \le k)] = \alpha_k + \mathbf{x}_{ij}^t \boldsymbol{\beta} + \mathbf{z}_{ij}^t \mathbf{u}_i, \quad k = 1,\ldots,K-1, \tag{12}$$

(e.g., see Ezzet and Whitehead 1991; Hedeker and Gibbons 1994; Tutz and Hennevogl 1996). This model has the simple *proportional odds* form whereby fixed and random effects are the same for all cumulative probabilities, that is for all ways of collapsing the K categories to a binary response.

For nominal response variables, cumulative probabilities are not meaningful. One can then formulate an ordinary binary model by pairing each category with a baseline (e.g., category K) and fit these $K-1$ models simultaneously while allowing separate effects for each. This necessitates using a vector of random effects, one for each logit. This case has received little attention in the literature.

To illustrate a model of form (12), we analyze data from the 1994 General Social Survey on subjects' opinions on four items (the environment, health, law enforcement, education) relating to whether they believe that government spending on each item should increase, stay the same, or decrease. Subjects are also classified by their gender and their race. (The contingency table has 486 cells and is not shown here.) For subject i, let $G_i = 1$ for females and 0 for males, let $R_{1i} = 1$ for whites and 0 otherwise, $R_{2i} = 1$ for blacks and 0 otherwise. Let y_{ij} denote the response for subject i on spending item j, where outcomes (1, 2, 3) represent (increase, stay the same, decrease). Consider first the random-intercept model

$$\text{logit}[P(y_{ij} \le k)] = \alpha_k + \beta_j + \beta_g G_i + \beta_{r1} R_{i1} + \beta_{r2} R_{2i} + u_i, \quad k = 1,\ldots,K-1. \tag{13}$$

Using NLMIXED in SAS with constraint $\beta_4 = 0$, we obtained ML estimates (−.551, −.603, −.486, 0) of the item parameters $\{\beta_j\}$. The first three estimates have absolute values greater than five standard errors, providing strong evidence of greater support for increased government spending on education than on the other items.

However, substantial evidence of interaction exists. For instance, the deviance drops by 33.4 with the addition of a race-by-item interaction term. For that model, Table 6 shows the ML estimates and standard errors. Each race shows relatively more support for spending on education than the other items, with blacks also giving relatively high support for spending on health. To help show how to interpret these estimates, Table 7 shows the linear predictor estimates for males for the logit of the probability of supporting increased spending (category 1 contrasted with the other two). For instance, for white subjects with the environment item, the estimated linear predictor equals $1.065 - .055 - .357 - .170 = .483$, so for a white male at the mean of the random-effects distribution, the estimated probability of supporting increased spending is $e^{.483}/[1 + e^{.483}] = .62$. The

TABLE 6
Parameter Estimates and Standard Errors for Cumulative Logit Model on Government Spending, with Random Subject Intercept, Permitting Item-by-Race Interaction

Variable	Estimate	Standard Error
Intercept-1	1.065	.391
Intercept-2	1.919	.051
Gender	.409	.088
Race1-w	−.055	.397
Race2-b	.434	.452
Item1-envir	−.357	.539
Item2-health	−.319	.493
Item3-crime	−.585	.480
Race1∗Item1	−.170	.549
Race1∗Item2	−.387	.503
Race1∗Item3	.197	.491
Race2∗Item1	−.452	.606
Race2∗Item2	.454	.598
Race2∗Item3	−.518	.560

Source: General Social Survey (1994).
Note: Coding 0 for Item 4 (educ.) and race 3 (other).

TABLE 7
Linear Predictor Estimates for Logit Probability of Males Preferring Increased Spending*

	Race		
Item	White	Black	Other
Environment	.48	.69	.71
Health	.30	1.64	.74
Crime	.62	.40	.48
Education	1.01	1.50	1.06

*For model with item-by-race interaction with government spending data (Table 6); this increases by .41 for females and by 1.92 for logit probability of increased or the same spending.

linear predictor values increase by 1.919 for the cumulative probability for the second category—that is, the probability of response in categories "increasing" or "staying the same." For this model, $\hat{\beta}_g = .409$; for females, the (subject-specific) odds of supporting increased spending instead of the same or lower spending, and the odds of supporting increased or the same spending instead of lower spending, are estimated to be $\exp(.409) = 1.51$ times the corresponding odds for males.

Some evidence exists of additional interactions, but the race-by-item interaction provides the strongest departure from the main effects model. For this model, the estimated standard deviation of the random intercept equals 1.0, representing a considerable positive association among repeated responses by each subject.

3.8. *Example 8: Cluster Sampling*

The use of cluster sampling methods has traditionally presented a stumbling block for categorical data methodology. Although numerous methods have been proposed, few are reported in the social science literature or have been adopted by leading software packages. Standard errors based on simple random sampling are too small, and the usual chi-squared test statistics have weighted sums of chi-squared, not chi-squared null distributions. For instance, see Rao and Thomas (1988) for a survey of ways of adjusting standard inferences to take into account complex sampling methods in the analysis and modeling of categorical data.

When the sampling scheme uses a random sample of clusters, with independent observations within each cluster, one can account for the clustering by using random effects for the clusters. To illustrate, we analyze data from Brier (1980), who reported 96 observations taken from 20 neighborhoods (the clusters) on $Y =$ *satisfaction with home* and $X =$ *satisfaction with neighborhood as a whole*. Each variable was measured with the ordinal scale (*unsatisfied, satisfied, very satisfied*). Brier's (1980) analysis adjusted for the clustering by reducing the usual Pearson statistic for testing independence in the 3×3 contingency table relating X and Y from 17.9 to 15.7 ($df = 4$).

Again, let y_{ij} denote the jth response in cluster i, and consider the model

$$\text{logit}[P(y_{ij} \leq k)] = \alpha_k + x_{ij}\beta + u_i, \tag{14}$$

where we use scores (1, 2, 3) for the satisfaction levels of x_{ij}. Assuming a $N(0, \sigma^2)$ distribution for u_i and using NLMIXED in SAS, we obtained $\hat{\beta} = -1.201$, with standard error of .407, and $\hat{\sigma} = .92$ ($se = .37$). By contrast, the analysis treating the 96 observations as a random sample corresponds to this model forcing $\sigma = 0$; it has $\hat{\beta} = -1.226$, with standard error of .370. As in the Brier (1980) analysis, there is a slight reduction in significance from taking the clustering into account. The ratio of $\text{Var}(\hat{\beta})$ in the clustered to unclustered analysis is 1.14, as is the ratio of Brier's Pearson statistics in the unclustered to clustered analyses. It is a bit surprising that the cluster-specific $\hat{\beta}$ estimate is not larger (in absolute value) than the unclustered one. A referee has indicated that this may reflect the fact that asymptotics may not apply well with a relatively small number of clusters (20 in this case) or that the cluster factor is confounded with the satisfaction with neighborhood covariate (Neuhaus and Kalbfleisch 1998; Berlin et al. 1999).

3.9. *Example 9: Capture-Recapture Data*

This section has presented a variety of data sets and applications to illustrate the potential use of random-effects modeling with categorical response data in the social sciences. Some alternative forms of such models that have been used in other scientific disciplines also have the potential for social science applications.

An example is random-effects models as employed in capture-recapture problems. These methods have repeated measurement over time, with scale (sampled, not sampled) at each time. Observations are completely missing for the cell corresponding to those subjects not sampled for every list. Such methods have traditionally been used to estimate animal abundance in some habitat. However, they have increasingly been applied to estimate population size in census and public health settings. For instance, Davies, Cormack, and Richardson (1999) estimated population prevalence of injecting drug use and HIV infection in Glasgow, and Darroch et al. (1993) used a three-sample multiple-recapture approach in census population estimation. Another possible application is to estimate the number of files on the World Wide Web relating to some subject by taking samples using several search engines (Fienberg, Johnson, and Junker 1999).

For capture-recapture modeling, Coull and Agresti (1999) recently used a logit model with a random-effects term to represent heterogeneity among subjects in their probability of capture at any given time. This allowance for heterogeneity results in wider prediction intervals for the population size than ordinary methods provide, indicating that intervals based on a possibly unrealistic assumption of homogeneity among subjects may be overly optimistic.

3.10. *Extensions to Discrete Data*

The focus of this paper is on random-effects models for categorical response data. More generally, GLMMs are useful for other types of discrete data as well. For instance, consider Poisson regression modeling of count data. A severe limitation of the Poisson model is that the variance must be identical to the mean; hence, at a fixed mean there is not the potential for the variance to decrease as predictors are added to the model. In particular, count data often show overdispersion, with the variance exceeding the mean.

A flexible way to account for overdispersion with count data is with a mixture model. Traditionally this is done by assuming that, given the mean, the distribution is Poisson, but the mean itself varies according to a gamma distribution. The mixture distribution is then the negative binomial. There are two versions of the negative binomial model, depending on how the gamma is parameterized; one version has variance that is a constant multiple of the mean, and the other has variance that is a quadratic function of the mean (McCullagh and Nelder 1989). ML estimation for the

latter case is available with PROC GENMOD in SAS (starting with Version 7).

Alternatively, one can use the GLMM structure (2), typically with the log-link function and a normal random effect. For the log link with random intercept, for instance, the model for the mean μ_{ij} for the jth response in cluster i is

$$\log(\mu_{ij}) = \mathbf{x}_{ij}^t \boldsymbol{\beta} + u_i,$$

where u_i has a $N(0, \sigma^2)$ distribution. This model is an appealing way to account for overdispersion due to important unobserved explanatory variables. The implication about the marginal distribution (averaging out the random effect) is that

$$E(y_{ij}) = E[E(y_{ij}|u_i)] = E[e^{\mathbf{x}_{ij}^t \boldsymbol{\beta} + u_i}] = e^{\mathbf{x}_{ij}^t \boldsymbol{\beta} + \sigma^2/2}$$

since (by its moment-generating function) a $N(0, \sigma^2)$ variate u_i has $E(e^{tu_i}) = e^{t^2\sigma^2/2}$. That is, if this model holds, then the log of the mean unconditionally equals $\mathbf{x}_{ij}^t \boldsymbol{\beta} + \sigma^2/2$, so the cluster-specific effects of the explanatory variables are the same as the marginal effects but the intercept is offset (Zeger et al. 1988). Similarly, the marginal distribution has

$$\begin{aligned}
\mathrm{Var}(y_{ij}) &= E[\mathrm{Var}(y_{ij}|u_i)] + \mathrm{Var}[E(y_{ij}|u_i)] \\
&= E[e^{\mathbf{x}_{ij}^t \boldsymbol{\beta} + u_i}] + e^{2\mathbf{x}_{ij}^t \boldsymbol{\beta}} \mathrm{Var}(e^{u_i}) \\
&= e^{\mathbf{x}_{ij}^t \boldsymbol{\beta} + \sigma^2/2} + e^{2\mathbf{x}_{ij}^t \boldsymbol{\beta}} (e^{2\sigma^2} - e^{\sigma^2}) \\
&= E(y_{ij}) + [E(y_{ij})]^2 (e^{\sigma^2} - 1).
\end{aligned}$$

That is, the unconditional variance is a quadratic function of the mean. The ordinary Poisson model results from $\sigma^2 = 0$, and the extent to which the variance exceeds the mean increases as σ^2 increases.

Note that one can obtain the negative binomial model with log link from a GLMM construction by letting $\exp(u_i)$ have a gamma distribution with a mean of 1. The GLMM with normal random effect has the advantage, relative to the negative binomial model, of providing a way of permitting multiple random effects and multilevel models. Land, McCall, and Nagin (1996) discussed a semiparametric version of the GLMM that treats the random effect in a nonparametric manner. This is in the same spirit as the work of Aitkin (1996, 1999) for nonparametric fitting of GLMMs mentioned in Section 3.6.

We illustrate a situation in which it is important to allow for a random effect with count data using a simple data set from the 1990 General Social Survey. We look at one question asked subjects: "Within the past 12 months, how many people have you known personally that were victims of homicide?" We consider this response here for the white and black categories of race. The mean for the 159 blacks who responded was .522 with a variance of 1.150; the mean for the 1149 whites who responded was .092 with a variance of .155. The ratio of the variance to the mean for each race provides evidence of overdispersion for a Poisson model. It is plausible that, for each race, the expected value of the response would vary according to various unmeasured factors such as demographic variables and the location of one's residence.

For the ordinary Poisson model with log link, the estimated difference of 1.733 between the log mean for blacks and the log mean for whites has an estimated standard error of .147. However, it is much more natural to use a model permitting subject heterogeneity. Adding a parameter by using the negative binomial approach with quadratic variance function (using ML fitting in SAS with PROC GENMOD), the log-likelihood increases by 61.1 (deviance decreases by 122.2). The estimated difference is still 1.733 between the log means, since for this case both models provide fitted means equal to the observed ones (and $\log(.522/.092) = 1.733$), but now the estimated standard error increases to .238. The Wald 95 percent confidence interval for the ratio of means for blacks and whites goes from $\exp[1.733 \pm 1.96(.147)] = (4.2, 7.5)$ for the ordinary Poisson model to $\exp[1.733 \pm 1.96(.238)] = (3.5, 9.0)$ for the negative binomial model.

Other examples of applications of models that add random effects to Poisson regression include the analysis of cancer maps in epidemiology (Breslow and Clayton 1993) and modeling variability in bacteria counts (Aitchison and Ho 1989).

4. MODEL FITTING AND PREDICTION

Specification of a parametric GLMM is done in two stages. First, conditional on the random effects $\mathbf{u}$, the data $\mathbf{y}$ are assumed to follow a probability distribution in the exponential family. This is a broad family of probability distributions that includes the normal, binomial, and Poisson. Let $f(\mathbf{y}|\mathbf{u};\boldsymbol{\beta})$ represent the conditional density (or mass) function of $\mathbf{y}$ given $\mathbf{u}$, where $\boldsymbol{\beta}$ is as in (2). For example, consider the model of Section 2.1. In this case, $\mathbf{y} = (y_{11}, y_{12}, \ldots, y_{I1}, y_{I2})^t$, $\mathbf{u} = (u_1, \ldots, u_I)^t$, $\boldsymbol{\beta} = (\alpha, \beta)^t$,

and we have

$$f(\mathbf{y}|\mathbf{u};\alpha,\beta) = \prod_{i=1}^{I} \frac{\exp\{y_{i1}(\alpha+u_i)\}}{(1+\exp\{\alpha+u_i\})} \frac{\exp\{y_{i2}(\alpha+\beta+u_i)\}}{(1+\exp\{\alpha+\beta+u_i\})}.$$

The second part of the specification involves making an assumption about the distribution of the random effects, $\mathbf{u}$. Typically, $\mathbf{u}$ is assumed to be multivariate normal with mean zero and covariance matrix $\mathbf{V}$. Often $\mathbf{V}$ is known up to a vector of *variance components*, $\boldsymbol{\sigma}^2$. Let $f(\mathbf{u};\boldsymbol{\sigma})$ denote the probability density function of $\mathbf{u}$. In the model of Section 2.1, the components of $\mathbf{u}$ are assumed to be independent $N(0,\sigma^2)$, which means that $\boldsymbol{\sigma}$ has only one component and

$$f(\mathbf{u};\sigma) = \prod_{i=1}^{I} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\, u_i^2\right\}.$$

In this section we discuss estimation of $\boldsymbol{\psi} = (\boldsymbol{\beta}^t, \boldsymbol{\sigma}^t)^t$, the vector of unknown parameters in our model, using exact ML estimation as well as two approximate ML techniques: one based on an analytical approximation of the likelihood integrand and the other on Bayes methods with diffuse priors.

4.1. *Exact Maximum Likelihood*

As Searle et al. (1992:232) point out, maximum likelihood is a "well-established and well-respected method of estimation that has a variety of optimality properties." As such, ML estimation is usually the default technique for estimating parameters. In general, the GLMM likelihood function is the marginal mass function of the *observed* data, $\mathbf{y}$, viewed as a function of the parameters; that is,

$$L(\boldsymbol{\beta},\boldsymbol{\sigma}|\mathbf{y}) = \int f(\mathbf{y}|\mathbf{u};\boldsymbol{\beta}) f(\mathbf{u};\boldsymbol{\sigma})\, d\mathbf{u}. \tag{15}$$

This expression nearly always involves intractable integrals whose dimension depends on the structure of the random effects. For example, the likelihood function for the model of Section 2.1 is given by

$$L(\alpha,\beta,\sigma|\mathbf{y}) = \prod_{i=1}^{I} \int_{-\infty}^{\infty} \frac{\exp\{y_{i1}(\alpha+u_i)\}}{(1+\exp\{\alpha+u_i\})} \frac{\exp\{y_{i2}(\alpha+\beta+u_i)\}}{(1+\exp\{\alpha+\beta+u_i\})} \times \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{u_i^2}{2\sigma^2}\right\} du_i$$

which has no closed-form solution. When the dimension of the intractable integrals is small, numerical integration can be used to closely approximate the likelihood (Crouch and Spiegelman 1990), as is done in SAS's NLMIXED. However, the error induced by replacing the intractable integral with a finite sum (as is done in Gauss-Hermite quadrature methods) becomes more and more difficult to control as the dimension of the integral increases.

Recently developed Monte Carlo methods for finding the exact maximum-likelihood estimate provide an alternative to numerical integration. These iterative methods can handle high-dimensional integrals better than numerical integration. Unfortunately, they require fairly sophisticated computer programs, and, as of now, there is no general software available. The Monte Carlo-based method that has received the most attention is the Monte Carlo EM (MCEM) algorithm, which is now described.

The EM algorithm (Dempster, Laird, and Rubin 1977) is a popular method of finding ML estimates in normal theory mixed models (Searle et al. 1992, ch. 8). Consider application of the EM algorithm in the GLMM setting with $\mathbf{u}$ assuming the role of *missing data*. The E-step of the EM algorithm requires calculation of

$$E\{\log f(\mathbf{y},\mathbf{u};\boldsymbol{\psi})|\mathbf{y};\boldsymbol{\psi}^{(r)}\}, \tag{16}$$

where $f(\mathbf{y},\mathbf{u};\boldsymbol{\psi}) = f(\mathbf{y}|\mathbf{u};\boldsymbol{\beta})f(\mathbf{u};\boldsymbol{\sigma})$ is the density of the *complete data* and $\boldsymbol{\psi}^{(r)}$ denotes the value of $\boldsymbol{\psi}$ from the rth iteration of EM. As the notation suggests, the expectation in (16) is with respect to the conditional distribution of $\mathbf{u}$ given $\mathbf{y}$ with parameter value set equal to $\boldsymbol{\psi}^{(r)}$, whose density we write as $f(\mathbf{u}|\mathbf{y};\boldsymbol{\psi}^{(r)})$. Unfortunately, analytical evaluation of (16) is also impossible, because (15) cannot be written in closed form.

The MCEM algorithm, introduced by Wei and Tanner (1990), circumvents this difficulty by replacing the intractable expectation with a Monte Carlo approximation. There are (at least) three different methods of constructing a Monte Carlo estimate of (16) in the GLMM context. The most obvious method is to use independent simulations from $f(\mathbf{u}|\mathbf{y};\boldsymbol{\psi}^{(r)})$.

Booth and Hobert (1999) explained how to obtain such a sample through rejection sampling and also how to form a different estimate using importance sampling. The third method is Markov chain Monte Carlo (MCMC). McCulloch (1994) and Chan and Kuk (1997) showed how to use the Gibbs sampler for some specific binary data models, while McCulloch (1997) gave a general Hastings-Metropolis algorithm that will, in theory, work for any GLMM; see also Liao (1999).

Of course, there is no free lunch. While the use of MCEM circumvents a complicated expectation at each E-step, it requires a method for choosing the Monte Carlo sample size at each MCE-step. Booth and Hobert (1999) and Levine and Casella (1998) discussed methods for choosing an appropriate Monte Carlo sample size at each iteration.

In stating that the methods of this subsection provide "exact" ML estimates, we mean that the approximations converge to the ML estimates as they are applied more finely—for instance, as the number of quadrature points increases in an appropriate manner for numerical integration and as the Monte Carlo sample size increases in the MCEM method. This is in contrast to the approximate methods of the next subsection, which may potentially yield values far from the ML estimates no matter how applied.

4.2. *Penalized Quasi Likelihood*

If one is willing to sacrifice exactness for ease of implementation, there are approximate ML methods that maximize an analytical approximation of the likelihood function instead of the likelihood function itself. The main approaches involve integrating a first-order Taylor series expansion of the likelihood integrand around the approximate posterior modes of the random effects (Goldstein 1991; Schall 1991; Breslow and Clayton 1993; Wolfinger and O'Connell 1993; Longford 1993; Longford 1994; McGilchrist 1994). In particular, Breslow and Clayton's (1993) algorithm, which is motivated using a penalized quasi-likelihood (PQL) argument, is essentially the same as the algorithm proposed by Wolfinger and O'Connell (1993), based on the idea of pseudo-likelihood. This algorithm involves iterative fitting of normal theory linear mixed models and can be implemented using the %GLIMMIX macro in SAS (see Littell et al. 1996). Even though this method is iterative, it involves no numerical integration or Monte Carlo approximation and so is much simpler to program than the exact ML methods. Here is a brief description:

Each iteration contains two steps; the first updates $\boldsymbol{\beta}$ and the second updates $\boldsymbol{\sigma}$. Suppose that $(\boldsymbol{\beta}^{(r)}, \boldsymbol{\sigma}^{(r)})$ are the values after r iterations.

The $\boldsymbol{\beta}$ update can be motivated using Henderson's mixed-model equations (Henderson et al. 1959). To this end, consider the normal theory mixed model—that is, the model in which both $f(\mathbf{y}|\mathbf{u};\boldsymbol{\beta})$ and $f(\mathbf{u};\boldsymbol{\sigma})$ are normal densities. Let $\hat{\boldsymbol{\sigma}}$ be the ML or restricted ML (REML) estimate of $\boldsymbol{\sigma}$. It is well known (Searle et al. 1992, sec. 7.6) that the values of $\boldsymbol{\beta}$ and $\mathbf{u}$ that jointly maximize the function $f(\mathbf{y}|\mathbf{u};\boldsymbol{\beta})f(\mathbf{u};\hat{\boldsymbol{\sigma}})$ are the ML estimate of $\boldsymbol{\beta}$ and the estimated best linear unbiased predictor (EBLUP) of $\mathbf{u}$. This motivates the following update for $\boldsymbol{\beta}$ in the GLMM context. Given $\boldsymbol{\sigma}^{(r)}$, the function $f(\mathbf{y}|\mathbf{u};\boldsymbol{\beta})f(\mathbf{u};\boldsymbol{\sigma}^{(r)})$ is maximized with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$ and $\boldsymbol{\beta}^{(r+1)}$ is assigned the maximizing value of $\boldsymbol{\beta}$. This maximization is not trivial and will almost always require an iterative technique such as Newton-Raphson.

The $\boldsymbol{\sigma}$ update is based on another normal approximation. Given $\boldsymbol{\beta}^{(r)}$ and $\mathbf{u}^{(r)}$, a *working dependent variable*, $\mathbf{z}$, is constructed just as it is in the usual iteratively reweighted least-squares algorithm for fitting generalized linear models (McCullagh and Nelder 1989:40). This working dependent variable is then assumed to follow a normal linear mixed model and ML or REML is used to estimate the variance components (e.g., see Searle et al. 1992, ch. 6). The components of $\boldsymbol{\sigma}^{(r)}$ are then assigned the values of the corresponding ML or REML estimates.

The main advantage of PQL is its relative simplicity, avoiding numerical integration and being computationally feasible for very large data sets and complex multilevel models that may not be feasible with the methods of Section 4.1. However, this iterative scheme does *not* yield the ML estimate of $\boldsymbol{\psi} = (\boldsymbol{\beta}^t, \boldsymbol{\sigma}^t)^t$. Indeed, McCulloch (1997) uses some analytical arguments in conjunction with simulation to show that this method can perform quite poorly relative to ML; see also Booth and Hobert (1999, sec. 7.3) for an analysis of a data set that illustrates the potential for large differences between ML and PQL, and Breslow and Lin (1995) and Lin and Breslow (1996) for "bias-corrected" versions of PQL. Generally, the PQL approach deteriorates as the data depart from normal (e.g., binary) and as the variance components increase. To illustrate, consider the opinion about abortion data of Section 3.2. For the parameterization setting $\beta_3 = 0$, the ML estimates for the random-effects model are $\hat{\beta}_1 = .83$ $(se = .16)$, $\hat{\beta}_2 = .29$ $(se = .16)$ with $\hat{\sigma} = 8.8$ reflecting a very strong within-subject dependence. By contrast, the PQL estimates (obtained using the %GLIMMIX macro in SAS) are $\hat{\beta}_1 = .87$ $(se = .07)$, $\hat{\beta}_2 = .31$ $(se = .07)$

with $\hat{\sigma} = 4.3$. The PQL approximations to the ML estimates are decent for $\{\beta_j\}$, but the standard errors and the estimate of σ are only about half of what they should be. In fact, when true variance components are large, PQL ordinarily tends to produce variance component estimates that have substantial negative bias (Breslow and Lin 1995). The PQL approach provided a good approximation for the ML estimates for the other binary data examples presented here.

Generally speaking, PQL is a good approximation to ML, provided the random-effects variances are relatively small (that is, when the fixed effects dominate the model) or the response is approximately normal. Improvements of such approximations have been proposed for cases in which they may behave poorly (e.g., Breslow and Lin 1995; Lin and Breslow 1996; Goldstein and Rasbash 1996). However, we recommend that analysts attempt to use exact ML, rather than possibly poor approximations such as PQL. We have briefly described the approximate methods above because most current software for GLMMs uses them rather than ML and because of the scope of their computational feasibility. Over time, however, as computational methods continue to be refined, we believe that ML fitting of GLMMs will become more commonplace and the approximate methods will lose their current appeal.

4.3. *A Bayesian Model with a Diffuse Prior*

In a Bayesian version of our model $\boldsymbol{\psi}$ is treated as a random variable with prior density $\pi(\boldsymbol{\psi})$. The posterior density is given by

$$\pi(\mathbf{u}, \boldsymbol{\psi} | \mathbf{y}) = \frac{f(\mathbf{y}, \mathbf{u}; \boldsymbol{\psi}) \pi(\boldsymbol{\psi})}{c(\mathbf{y})}$$

where

$$c(\mathbf{y}) = \iint f(\mathbf{y}, \mathbf{u}; \boldsymbol{\psi}) \pi(\boldsymbol{\psi}) \, d\mathbf{u} \, d\boldsymbol{\psi},$$

which is typically not available in closed form because of the same intractable integrals that cause trouble in the likelihood function. A flat prior, $\pi(\boldsymbol{\psi}) = 1$, results in a posterior (for $\boldsymbol{\psi}$) that is simply a constant multiple of the likelihood function (15). Therefore, if the resulting posterior is *proper*, MCMC methods that can also be used to study intractable posterior distributions (e.g., the Gibbs sampler) can be used to study the likelihood

function. While flat priors typically result in proper posteriors in normal theory mixed models (Hobert and Casella 1996), they lead to improper posteriors for many of the models considered in this paper (Natarajan and McCulloch 1995).

One way to ensure a proper posterior is to use a proper prior. If a "diffuse" but proper prior is used in place of a flat prior, we might hope that the resulting posterior is close to the likelihood function. Furthermore, we can legitimately use MCMC methods to find the posterior mode, which we might hope is a reasonable approximation to the ML estimate. However, the use of a proper diffuse prior need not result in a posterior mode that is close to the ML estimate, especially when the data contain little information about the variance components (Kass and Wasserman 1996). Moreover, the simulation results of Natarajan and McCulloch (1998) showed that using a diffuse prior can lead to Markov chains that converge very slowly. Thus, even if the likelihood and posterior are similar, MCMC techniques may be of no practical use because of slow mixing. In our opinion, this approximate Bayes method is the least attractive of all the approximate methods.

4.4. *Inference for Model Parameters*

After fitting the model, the next step is usually inference about the components of $\boldsymbol{\psi}$. We first consider inference about $\boldsymbol{\beta}$. The asymptotic normality of the ML estimate of $\boldsymbol{\beta}$ can be used to form approximate confidence sets in the usual way (McCullagh and Nelder 1989, app. A). Furthermore, hypotheses involving $\boldsymbol{\beta}$ can be tested using asymptotic likelihood-ratio tests (LRTs); that is, using the fact that minus twice the log of the likelihood-ratio statistic $(-2 \log \lambda)$ has an asymptotic χ^2 distribution under the null (McCullagh and Nelder 1989, app. A). If an MCEM program for fitting the model is available, the necessary evaluations of the likelihood and derivatives of the likelihood at the ML estimate can be performed via Monte Carlo with little additional programming; for example, see Booth and Hobert (1999, sec. 6).

Regarding the testing of variance components, it is unfortunate that $-2 \log \lambda$ does not necessarily have an asymptotic χ^2 distribution under the null when the hypothesis involves parameters on the boundary of the parameter space—e.g., when testing that a variance component is equal to zero (Self and Liang 1987). (This difficulty has nothing to do with the categorical nature of the data; indeed, the same problem arises in normal

linear mixed models for tests about the variance components (Miller 1977).) While calculation of the true asymptotic distribution can in general be quite difficult, there are several important special cases for which it is known. In particular, suppose that the model contains a single variance component, σ^2, and that we wish to test $H_0 : \sigma^2 = 0$ versus $H_1 : \sigma^2 > 0$. Self and Liang (1987, case 5) show that, in this situation, the asymptotic distribution of $-2 \log \lambda$ under the null is a 50:50 mixture of χ^2_0 and χ^2_1 random variables. (A χ^2_0 is a point mass at 0 and corresponds to $\hat{\sigma} = 0$, for which the maximized likelihoods are identical under H_0 and H_1, and hence their ratio $\lambda = 1$ and $-2 \log \lambda = 0$.) Thus, when $\hat{\sigma} > 0$ and $t = -2 \log \lambda > 0$, the P-value for this large-sample test is $(1/2)P(\chi^2_1 > t)$, half the P-value that applies for χ^2_1 asymptotic tests (such as tests about components of $\boldsymbol{\beta}$).

Recently, Lin (1997) has shown that the score test (McCullagh and Nelder 1989, app. A) is a flexible alternative to the LRT for testing that one or all of the variance components in the model are equal to zero. Again, in the MCEM context, it is straightforward to form Monte Carlo approximations of the likelihood derivatives that comprise the score statistic.

4.5. *Prediction of Random Effects*

In Section 4.1 we discussed a method for forecasting election results by predicting random effects associated with 50 states and the District of Columbia. In that setting the procedure involved estimating/predicting a sum of the form $\alpha + u_i$, where α represented a fixed but unknown nationwide propensity to vote for Clinton and u_i was a random effect associated with the ith state. More generally, when random effects are used to measure variation among a relatively large number of small areas or domains, it is often of interest to estimate/predict mixed linear combinations of fixed and random effects of the form $\eta = \mathbf{x}^t\boldsymbol{\beta} + \mathbf{z}^t\mathbf{u}$ specific to the domains of interest. For example, in educational surveys, the domains might be schools for which a rating is desired or even individual children whose ability is being predicted.

Except for the presence of the unknown fixed effects parameter $\boldsymbol{\beta}$ and the variance components parameter $\boldsymbol{\sigma}$, the GLMM provides a complete description of the joint distribution of the observable data $\mathbf{y}$ and the unobservable random effect $\mathbf{u}$. After the data have been collected (i.e., observed), all the information about the random effects is contained in the

conditional distribution of $\mathbf{u}$ given $\mathbf{y}$. This distribution is implicitly defined by the assumed GLMM via the relationship $f(\mathbf{u}|\mathbf{y}) \propto f(\mathbf{y},\mathbf{u})$. Thus, for example, a point prediction for $\mathbf{u}$ is given by the mean of this conditional distribution, $E(\mathbf{u}|\mathbf{y})$. This predictor is "best" in the sense that its mean squared error is less than that of any other predictor (Searle et al. 1992, sec. 7.2).

Two practical issues that arise with the use of $E(\mathbf{u}|\mathbf{y})$ as a predictor for $\mathbf{u}$ are that the conditional expectation (1) depends on the unknown parameters $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$ and (2) is usually not available in closed form. The first of these difficulties is overcome in practice by simply plugging in estimates in place of the unknown parameters. In the case of the normal theory linear mixed model, substitution of the ML estimate for $\boldsymbol{\beta}$ results in the best linear unbiased predictor or BLUP while further substitution for $\boldsymbol{\sigma}$, if necessary, results in the so-called "empirical" BLUP (Searle et al. 1992, sec. 9.3). The second complication can be dealt with by numerically calculating the desired expectation, either exactly (using numerical integration or Monte Carlo methods) or approximately. In particular, a minor side benefit of the PQL algorithm used in the SAS macro GLIMMIX is that it automatically produces approximations for the predicted random effects. The predictor $\hat{\mathbf{u}}$ obtained by plugging parameter estimates into the conditional expectation $E(\mathbf{u}|\mathbf{y})$ is often referred to as the *empirical Bayes predictor*; for a detailed discussion of this approach, see Carlin and Louis (1996).

The more data that are available from a particular domain, the more accurately random effects associated with that domain can be predicted. Suppose that $\mathbf{y}_i$ denotes the data collected from the ith domain with associated random effect u_i. The amount of uncertainty about u_i is measured by the conditional variance, $\mathrm{Var}(u_i|\mathbf{y}_i)$, or by the corresponding standard deviation. These *standard errors of prediction* can also be computed or approximated using the conditional distribution implied by the assumed GLMM and by substituting estimates of unknown parameters where necessary. A common criticism of this method is that no adjustment is made for the sampling variability in the parameter estimates. The parameters are effectively treated as though they are known, and hence the amount of uncertainty about the random effects tends to be underestimated. Booth and Hobert (1998) discussed this issue and proposed a method for correcting the "naive" standard errors. However, their adjustments are complicated and difficult to compute. Moreover, unless the total amount of data is very limited (leading to very unreliable parameter estimates), corrections

to the naive standard errors are often relatively small and of little practical significance.

5. MARGINALLY SPECIFIED MODELS

Section 2.2 highlighted the conditional (i.e., *subject-specific*) interpretation of the regression parameters in a GLMM. In some instances, however, the marginal (i.e., *population averaged*) effects are of primary interest. For example, consider a survey of opinions of different ethnic groups in which responses on a variety of social issues are measured on a binary scale (yes/no, favor/oppose, etc.). Quantities of primary interest in such a setting would typically include between-group odds ratios among marginal probabilities for the different ethnic groups. In such cases it is more convenient to parameterize the model in such a way that the regression parameters have a direct marginal interpretation.

One popular approach to modeling marginal effects is to use generalized estimating equations (GEE). A brief description of this approach follows. As before let y_{ij} denote the jth response in domain or cluster i and let $\mathbf{x}_{ij}$ denote a vector of associated explanatory variable values, $i = 1, \ldots, I$ and $j = 1, \ldots, n_i$. Now suppose that the marginal mean of the (i, j)th response, $m_{ij} = E(y_{ij})$, is described by the linear model

$$g(m_{ij}) = \mathbf{x}_{ij}^t \boldsymbol{\beta}, \tag{17}$$

where g is a link function. In addition, suppose that the variance of y_{ij} is given by $\mathrm{Var}(y_{ij}) = \phi_{ij} v(m_{ij})$, where $\phi_{ij} = \phi / w_{ij}$. If the assumed variance function corresponds to an exponential family model, this assumed structure for the mean and variance is exactly that of a GLM. However, unlike in a GLM, the GEE method does not require the distribution of the responses to be fully specified. Furthermore, GEE allows for dependence between responses from the same cluster via a *working correlation matrix*. The terminology "working" is used because an adjustment to the standard errors of the regression parameters is usually made using a *sandwich variance* formula to account for misspecification of this part of the correlation structure (Liang and Zeger 1986). For example, for the two-level sampling design considered in this section and in much of the paper, it is often reasonable to assume that responses within the same cluster are equicorrelated. On the other hand, if the responses consist of repeated measurements taken at different times within each cluster then a working correlation

matrix incorporating an autocorrelation structure might be more reasonable. For a more detailed description of the GEE approach, see Liang and Zeger (1986), Diggle et al. (1994, ch. 8), and Liang et al. (1992).

A drawback of the GEE approach is that it does not explicitly model random effects and therefore does not allow these effects to be estimated. In addition, likelihood-based inferences are not possible because the joint distribution of the responses is not fully specified. A promising recent proposal by Heagerty (1999) attempts to overcome these deficiencies by defining a marginally specified GLMM. Heagerty notes that the traditional conditionally specified GLMM implicitly determines the relationship between the covariates and the marginal mean through the relation $m_{ij} = E(\mu_{ij})$. For example, we pointed out in Section 2.2 that with binary responses the assumption of a linear relationship between the covariates and the conditional logits implied a nonlinear relationship for the marginal logits. Conversely, if a marginal model for the mean of the form (17) is assumed, then this implicitly determines the form of the fixed portion of η_{ij} in the conditional model. That is, the linear predictor, $\eta_{ij} = \mathbf{x}_{ij}^t \boldsymbol{\beta} + \mathbf{z}_{ij}^t \mathbf{u}_i$, in the conditional GLMM is replaced by $\eta_{ij} = \Delta_{ij} + \mathbf{z}_{ij}^t \mathbf{u}_i$, where Δ_{ij} is a function of $(\boldsymbol{\beta}, \boldsymbol{\sigma})$ implicitly defined by the relation between the marginal and conditional means. Heagerty's idea is to specify the model for the conditional mean or the marginal mean depending upon whether a subject-specific or population-averaged interpretation is more relevant.

6. SOFTWARE

Applications of generalized linear mixed models have undoubtedly been hindered by the lack of adequate software. In recent years perhaps the most popular software has been the GLIMMIX macro provided by SAS. This macro provides estimates based on approximating the likelihood using methods of Breslow and Clayton (1993) and Wolfinger and O'Connell (1993). Having the GLM framework, it can fit models for a variety of response distributions and link functions, assuming normal random effects. Version 7 of SAS has introduced PROC NLMIXED, which can also use an adaptive version of Gauss-Hermite quadrature to approximate the likelihood. This is substantially better when variance components are large or data are far from normal. The availability of this procedure (or its successors) will likely make fitting of models with random effects much more common in the future. However, it has limitations. For instance, quadrature methods are computationally feasible only for integrals of small di-

mensions, and NLMIXED currently cannot accommodate nested random effects.

Table 8 shows the use of NLMIXED for the shrinkage analyses of Table 2 described in Section 3.1. Although it is easiest to use NLMIXED with standard univariate response distributions such as Poisson and binomial, it is also possible to use it with multinomial models. Table 9 shows its use for the ordinal cumulative logit analyses of the government spending data described in Section 3.7. First, one must define the two linear predictors (one for each cumulative probability) and the relationship between each multinomial probability and the linear predictors. For these data, the response is recoded as a vector, (y_1, y_2, y_3) taking three possible values, (0,0,1), (0,1,0) or (1,0,0), corresponding to the three possible responses, 1, 2, or 3. These response vectors are multinomials

TABLE 8
SAS (PROC NLMIXED) Code for Analyses of Table 2

```
data new;
input x n offset;
sub = _n_;
datalines;
1   5   −0.40898
16  32  −0.29363
10  19  0.26574
21  34  −0.19275
129 240 0.20312
....
8   14  0.17246
1   4   −0.29066
;

proc nlmixed;
 parms alpha = .2 sigma = .04;
 eta = alpha + u + offset;
 p = exp(eta) / (1 + exp(eta));
 model x ~ binomial(n,p);
 random u ~ normal(0,sigma*sigma) subject = sub;
 predict p out = new2;
run;

proc print data=new2;
run;
```

TABLE 9
SAS (PROC NLMIXED) Code for Analyses of Government Spending Data

```
data new;
input subject gender race1 race2 item1 item2 item3 y1 y2 y3 count;
datalines;
/*Subject                  Gender     Race1     Race2     Item1     Item2     Item3     Y1     Y2     Y3     Count*/
                    1         0         1         0         1         0         0      1      0      0       107
                    1         0         1         0         0         1         0      1      0      0       107
                    1         0         1         0         0         0         1      1      0      0       107
                    1         0         1         0         0         0         0      1      0      0       107
                    2         1         0         1         1         0         0      1      0      0        20
                    2         1         0         1         0         1         0      0      0      1        20
                    2         1         0         1         0         0         1      0      1      0        20
                    2         1         0         1         0         0         0      0      0      1        20
                    .
                    .
                    .
;
proc nlmixed data=new;
   bounds i2 > 0;
   eta1 = i1 + gender * beta1 + race1 * beta2 + race2 * beta3
          + item1 * beta4 + item2 * beta5 + item3 * beta6 + u;
   eta2 = i1 + i2 + gender * beta1 + race1 * beta2 + race2 * beta3
          + item1 * beta4 + item2 * beta5 + item3 * beta6 +  u;
   p1 = 1/(1 + exp(-eta1));
   p2 = 1/(1 + exp(-eta2)) - 1/(1 + exp(-eta1));
   p3 = 1 - 1/(1 + exp(-eta2));
   z = (p1**y1)*(p2**y2)*(p3**y3);
   if (z > 1e-8) then 11 = log(z);
      else ll=-1e100;
  model y1 ~ general(ll);
  estimate 'thresh2' i1+i2;
  random u ~ normal(0,su * su) subject = subject;
  replicate count;
run;
```

with sample sizes of 1. For outcome probabilities, p_1, p_2, and p_3, the contribution to the multinomial log likelihood is $p_1^{y_1}p_2^{y_2}p_3^{y_3}$. NLMIXED allows the user to code general likelihoods, as we defined it with the statement $z = (p1{**}y1){*}(p2{**}y2){*}(p3{**}y3)$. This likelihood is checked to see if it is numerically too close to zero, then converted to the log likelihood (the statement $ll = log(z)$). The statement $y_1 \sim general(ll)$ tells SAS that *ll* gives the value of the log likelihood. (Since the likelihood is a function of the parameters, it does not matter if y_1, y_2, or y_3 is used for that statement). Finally, an estimate statement is used to obtain an estimate of the second threshold.

A variety of other programs are currently in general circulation. For instance, EGRET (now distributed by Cytel Software, in Cambridge, Massachusetts) can fit certain mixed logit models, approximating the likelihood with Gauss-Hermite quadrature or replacing the normal random-effects distribution by a binomial distribution. Hedeker and Gibbons (1994) supplied a FORTRAN program MIXOR for ML fitting of proportional odds models with random effects. Harvey Goldstein and colleagues at the Institute of Education in London provide a general-purpose program for multilevel modeling called MLn (www.ioe.ac.uk/multilevel/), that can fit the model using an improved version of PQL. One can use a fully Bayesian approach using MCMC with BUGS, available from the MRC Biostatistics Unit at Cambridge (www.mrc-bsu.cam.ac.uk/bugs). Other programs include HLM (Scientific Software International, Chicago), written by A. Bryk, S. Raudenbush, and R. Congdon and which also uses an improved version of PQL, LogXact (from Cytel Software) for the conditional ML approach to eliminating cluster terms, and a GLIM macro for parametric and nonparametric fitting of GLMMs (Aitkin and Francis 1995); see Zhou, Perkins, and Hui (1999) for a description of some software for multilevel models.

The numerical approximations necessary to fit GLMMs require careful use even with software such as NLMIXED in SAS. With quadrature-based software, one should use a sufficient number of quadrature points to obtain simultaneously close approximations to the maximized log-likelihood and to ML estimates of the fixed effects, the standard errors of the fixed effects, the variance components, and the standard errors of the variance components. NLMIXED determines the number of quadrature points adaptively, and the default number selected is often quite low (at least, this is the case for versions 7 and 8 of SAS). Usually this is sufficient for the fixed effects but not the random effects part of the model; obtaining

sufficiently precise approximations for the standard errors and for the variance components usually requires considerably more points. In NLMIXED, we recommend checking these estimates while increasing the number of quadrature points (using the **qpoints**= option), to be confident that results have stabilized. We also recommend specifying the variance component in terms of the standard deviation in the model code. This helps in estimating variance components very close to 0, and also the standard deviation is usually preferred over the variance for interpretation.

Using large numbers of quadrature points may require long computing times. This is also true of data with a large number of clusters or several fixed and random effects within a cluster. Accurate starting values help speed convergence. Starting values can be obtained using the faster nonadaptive quadrature—for instance, by specifying *noad* and *noadscale* in the NLMIXED options or using the GLIMMIX macro. Since starting values need not be too accurate, milder convergence bounds can be used to obtain them.

7. CONCLUDING REMARKS

This article has shown a variety of social-science-related applications of generalized linear models for categorical data that contain random effects. Although introductory in nature, the models discussed have had relatively simple random-effects structure. However, there are many situations, especially in multilevel or multivariate settings (Catalano and Ryan 1992; Gueorguieva and Agresti 2000), where more complex models are appropriate. There is also continuing methodological research on random-effects models, such as developing ways of efficiently obtaining ML estimates and ways of checking goodness-of-fit of models.

Although random effects provide a natural way of handling many social science applications, as with any advanced statistical method there is the potential for misuse or inadequate use. For instance, with added complexity of models it can be more difficult to obtain ML estimates, and some algorithms may provide poor approximations for them. There is still much work to be done on the development of model-fitting methodology, as numerical integration is generally infeasible for complex models in which obtaining the likelihood involves high-dimensional integrals. The Bayesian paradigm (e.g., using the software BUGS) is becoming increasingly popular, but again with complex models there is the greater danger of

inappropriate choices of priors (e.g., improper priors leading to improper posteriors that are not detected by MCMC methods).

In addition, little work has been done on model checking (e.g., goodness-of-fit tests) and model diagnostics for GLMMs, even in the normal theory case. Also, model comparison of GLMMs can be difficult. As we have seen, in some cases standard methods of comparing likelihoods fail because, under the null hypothesis, certain parameters (e.g., variance components) fall on the boundary of the parameter space, thus violating standard assumptions required to generate the usual asymptotic distributions.

Finally, choice of form of an appropriate model is still an issue. There is a controversy among some statisticians about whether the effects generated in marginal models are more or less relevant than the conditional effects resulting from random-effects models (e.g., Lindsey 1999). Most of the discussion of this has been with relation to biomedical and epidemiological issues, and it is time to consider the practical implications of these matters for social science applications. In particular, it is a challenge for methodologists even to explain to practitioners why marginal and conditional effects differ when one uses a nonlinear link function.

Even with these cautions in mind, we think that the random-effects approach provides a potentially very useful extension of standard generalized linear models for social science applications. We hope that this article contributes toward helping methodologists understand their use.

REFERENCES

Agresti, Alan. 1993. "Distribution-free Fitting of Logit Models with Random Effects of Repeated Categorical Responses." *Statistics in Medicine* 12:1969–87.

———. 1995. "Logit Models and Related Quasi-symmetric Loglinear Models for Comparing Responses to Similar Items in a Survey." *Sociological Methods and Research* 24:68–95.

———. 1997. "A Model for Repeated Measurements of a Multivariate Binary Response." *Journal of the American Statistical Association* 22:315–21.

Agresti, Alan, and Barbara Finlay. 1997. *Statistical Methods for the Social Sciences*, 3rd ed. Upper Saddle River, NJ: Prentice Hall.

Aitchison, John, and C. H. Ho. 1989. "The Multivariate Poisson-log Normal Distribution." *Biometrika* 76:643–53.

Aitkin, Murray. 1996. "A General Maximum Likelihood Analysis of Overdispersion in Generalized Linear Models." *Statistics and Computing* 6:251–62.

———. 1999. "A General Maximum Likelihood Analysis of Variance Components in Generalized Linear Models." *Biometrics* 55:117–28.

Aitkin, Murray, and Brian J. Francis. 1995. "Fitting Overdispersed Generalized Linear Models by Non-parametric Maximum Likelihood." *The GLIM Newsletter* 25:37–45.

Aitkin, Murray, and Nicholas Longford. 1986. "Statistical Modelling in School Effectiveness Studies" (with discussion). *Journal of the Royal Statistical Society*, ser. A, 149:1–43.

Aitkin, Murray, Dorothy Anderson, and John Hinde. 1981. "Statistical Modelling of Data on Teaching Styles" (with discussion). *Journal of the Royal Statistical Society*, ser. A, General 144:419–61.

Akin, John S., David K. Guilkey, and Robin Sickles. 1979. "A Random Coefficient Probit Model with an Application to a Study of Migration." *Journal of Econometrics* 11:233–46.

Albert, James. 1992. "A Bayesian Analysis of a Poisson Random Effects Model for Home Run Hitters." *The American Statistician* 46:246–53.

Anderson, Dorothy A., and Murray Aitkin. 1985. "Variance Component Models with Binary Response: Interviewer Variability." *Journal of the Royal Statistical Society*, ser. B, 47:203–10.

Berlin, Jesse A., Stephen E. Kimmel, Thomas R. Ten Have, and Mary D. Sammel. 1999. "An Empirical Comparison of Several Clustered Data Approaches Under Confounding Due to Cluster Effects in the Analysis of Complications of Coronary Angioplasty." *Biometrics* 55:470–76.

Bock, Darrell R., and Murray Aitkin. 1981. "Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm." *Psychometrika* 46:443–59.

Booth, James G., and James P. Hobert. 1998. "Standard Errors of Prediction in Generalized Linear Mixed Models." *Journal of the American Statistical Association* 93:262–72.

———. 1999. "Maximizing Generalized Linear Mixed Model Likelihoods with an Automated Monte Carlo EM Algorithm." *Journal of the Royal Statistical Society*, ser. B, 61:265–85.

Breslow, Norman E., and David G. Clayton. 1993. "Approximate Inference in Generalized Linear Mixed Models." *Journal of the American Statistical Association* 88:9–25.

Breslow, Norman E., and Xihong Lin. 1995. "Bias Correction in Generalized Linear Mixed Models with a Single Component of Dispersion." *Biometrika* 82:81–91.

Brier, Stephen S. 1980. "Analysis of Contingency Tables Under Cluster Sampling." *Biometrika* 67:591–96.

Bryk, Anthony S., and Stephen W. Raudenbush. 1992. *Hierarchical Linear Models*. Thousand Oaks, CA: Sage.

Carlin, Bradley P., and Thomas A. Louis. 1996. *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman and Hall.

Catalano, Paul J., and Louise M. Ryan. 1992. "Bivariate Latent Variable Models for Clustered Discrete and Continuous Outcomes." *Journal of the American Statistical Association* 87:651–58.

Chan, Jennifer S. K., and Anthony Y. C. Kuk. 1997. "Maximum Likelihood Estimation for Probit-linear Mixed Models with Correlated Random Effects." *Biometrics* 53:86–97.

Coleman, James S. 1964. *Introduction to Mathematical Sociology*. London: Free Press of Glencoe.

Conaway, Mark R. 1989. "Analysis of Repeated Categorical Measurements with Conditional Likelihood Methods." *Journal of the American Statistical Association* 84:53–62.

Congdon, Peter. 1996. "General Linear Gravity Models for the Impact of Casualty Unit Closures." *Urban Studies* 33:1707–28.

Coull, Brent A., and Alan Agresti. 1999. "The Use of Mixed Logit Models to Reflect Heterogeneity in Capture-recapture Studies." *Biometrics* 55:294–301.

———. 2000. "Random Effects Modeling of Multiple Binomial Responses Using the Multivariate Binomial Logit-normal Distribution." *Biometrics* 56:73–80.

Crouch, Edmund A. C., and Donna Spiegelman. 1990. "The Evaluation of Integrals of the Form $\int_{-\infty}^{+\infty} f(t)\exp(-t^2)dt$: Application to Logistic-normal Models." *Journal of the American Statistical Association* 85:464–69.

Crowder, Martin J. 1978. "Beta-binomial ANOVA for Proportions." *Applied Statistics* 27:34–37.

Daniels, Michael J., and Constantine Gatsonis. 1997. "Hierarchical Polytomous Regression Models with Applications to Health Services Research." *Statistics in Medicine* 16:2311–26.

———. 1999. "Hierarchical Generalized Linear Models in the Analysis of Variations in Health Care Utilization." *Journal of the American Statistical Association* 94: 29–42.

Darroch, John N. 1981. "The Mantel-Haenszel Test and Tests of Marginal Symmetry; Fixed-effects and Mixed Models for a Categorical Response." *International Statistical Review* 49:285–307.

Darroch, John N., Stephen E. Fienberg, Gary F. V. Glonek, and Brian W. Junker. 1993. "A Three-sample Multiple-recapture Approach to Census Population Estimation with Heterogeneous Catchability." *Journal of the American Statistical Association* 88:1137–48.

Davies, A. G., Richard M. Cormack, and A. M. Richardson. 1999. "Estimation of Injecting Drug Users in the City of Edinburgh, Scotland, and Number Infected with Human Immunodeficiency Virus." *International Journal of Epidemiology* 28:117–21.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. "Maximum Likelihood from Incomplete Data Via the EM Algorithm" (with discussion). *Journal of the Royal Statistical Society*, ser. B, 39:1–38.

DerSimonian, Rebecca, and Nan Laird. 1986. "Meta-analysis in Clinical Trials." *Controlled Clinical Trials* 7:177–88.

Diggle, Peter J., Kung-Yee Liang, and Scott L. Zeger. 1994. *Analysis of Longitudinal Data*. Oxford, England: Clarendon Press.

Efron, Bradley, and Carl N. Morris. 1975. "Data Analysis Using Stein's Estimator and its Generalizations." *Journal of the American Statistical Association* 70:311–19.

Enberg, John, Peter Gottschalk, and Douglas Wolf. 1990. "A Random-effects Logit Model of Work-welfare Transitions." *Journal of Econometrics* 43:63–75.

Ezzet, Farkad, and John Whitehead. 1991. "A Random Effects Model for Ordinal Responses from a Crossover Trial" (with discussion). *Statistics in Medicine* 10:901–906.

Fienberg, Stephen E., Matthew S. Johnson, and Brian W. Junker. 1999. "Classical Multi-level and Bayesian Approaches to Population Size Estimation Using Multiple Lists." *Journal of the Royal Statistical Society*, ser. A, 162:383–406.

Follmann, Dean A., and Diane Lambert. 1989. "Generalizing Logistic Regression by Nonparametric Mixing." *Journal of the American Statistical Association* 84:295–300.

Ghosh, Malay, and J. N. K. Rao. 1994. "Small Area Estimation: An Appraisal." *Statistical Science* 9:55–76.

Gibbons, Robert D., and Donald Hedeker. 1994. "Application of Random-effects Probit Regression Models." *Journal of Consulting and Clinical Psychology* 62:285–96.

Gibbons, Robert D., Donald Hedeker, Sara C. Charles, and Paul Frisch. 1994. "A Random-effects Probit Model for Predicting Medical Malpractice Claims." *Journal of the American Statistical Association* 89:760–67.

Goldstein, Harvey. 1991. "Nonlinear Multilevel Models, with an Application to Discrete Response Data." *Biometrika* 78:45–51.

———. 1995. *Multilevel Statistical Models*, 2nd ed. London: Arnold.

Goldstein, Harvey, and Jon Rasbash. 1996. "Improved Approximations for Multilevel Models with Binary Responses." *Journal of the Royal Statistical Society*, ser. A, General 159:505–13.

Goodman, Leo A. 1974. "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models." *Biometrika* 61:215–31.

Gueorguieva, Ralitza V., and Alan Agresti. 2000. "A Correlated Probit Model for Multivariate Repeated Measures of Mixtures of Binary and Continuous Responses." Technical report, University of Florida.

Hatzinger, Reinhold. 1989. "The Rasch Model, Some Extensions and Their Relation to the Class of Generalized Linear Models." Pp. 172–79 in *Statistical Modelling*, edited by A. Decarli, B. J. Francis, R. Gilchrist, and G. V. H. Seeber. New York: Springer-Verlag.

Heagerty, Patrick. 1999. "Marginally Specified Logistic-normal Models for Longitudinal Binary Data." *Biometrics* 55:688–98.

Heckman, James, and Burton Singer. 1984. "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data." *Econometrica* 52:271–320.

Hedeker, Donald, and Robert D. Gibbons. 1994. "A Random-effects Ordinal Regression Model for Multilevel Analysis." *Biometrics* 50:933–44.

Hedeker, Donald, Robert D. Gibbons, and B. R. Flay. 1994. "Random-effects Regression Models for Clustered Data with an Example from Smoking Prevention Research." *Journal of Consulting and Clinical Psychology* 62:757–65.

Henderson, Charles R., Oscar Kempthorne, Shayle R. Searle, and C. N. VonKrosig. 1959. "Estimation of Environmental and Genetic Trends from Records Subject to Culling." *Biometrics* 15:192–218.

Henretta, John, Martha S. Hill, Wei Li, Beth J. Soldo, and Douglas A. Wolf. 1997. "Selection of Children to Provide Care: The Effect of Earlier Parental Transfers." *Journals of Gerontology*, ser. B, 52:110–19.

Hobert, James P., and George Casella. 1996. "The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models." *Journal of the American Statistical Association* 91:1461–73.

Jones, K., M. I. Gould, and R. Watt. 1998. "Multiple Contexts as Cross-classified Models: The Labor Vote in the British General Election of 1992." *Geographical Analysis* 30:65–93.

Kass, Robert E., and Larry Wasserman. 1996. "The Selection of Prior Distributions by Formal Rules." *Journal of the American Statistical Association* 91: 1343–70.

Kelderman, Henk. 1984. "Loglinear Rasch Model Tests." *Psychometrika* 49:223–45.

Land, Kenneth C., Patricia L. McCall, and Daniel S. Nagin. 1996. "A Comparison of Poisson, Negative Binomial, and Semiparametric Mixed Poisson Regression Models—with Empirical Applications to Criminal Careers Data." *Sociological Methods and Research* 24:387–442.

Langford, Ian H. 1994. "Using a Generalized Linear Mixed Model to Analyze Dichotomous Choice Contingent Valuation Data." *Land Economics* 70:507–14.

———. 1998. "Improved Estimation of Willingness to Pay in Dichotomous Choice Contingent Valuation Studies." *Land Economics* 74:65–75.

Lawless, Jerald F. 1987. "Negative Binomial and Mixed Poisson Regression." *The Canadian Journal of Statistics* 15:209–25.

Lee, Youngjo, and John A. Nelder. 1996. "Hierarchical Generalized Linear Models" (with discussion). *Journal of the Royal Statistical Society*, ser. B, 58:619–78.

Levine, Richard A., and George Casella. 1998. "Implementations of the Monte Carlo EM Algorithm," Technical report, University of California, Davis.

Liang, Kung-Yee, and Scott L. Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73:13–22.

Liang, Kung-Yee, Scott L. Zeger, and Bahjat Qaqish. 1992. "Multivariate Regression Analysis for Categorical Data" (with discussion). *Journal of the Royal Statistical Society*, ser. B, 54:3–40.

Liao, Jiangang G. 1999. "Maximum Likelihood Estimation in Generalized Linear Mixed Models," Technical report, University of South Florida.

Lin, Xihong. 1997. "Variance Component Testing in Generalised Linear Models with Random Effects." *Biometrika* 84:309–25.

Lin, Xihong and Norman E. Breslow. 1996. "Bias Correction in Generalised Linear Mixed Models with Multiple Components of Dispersion." *Journal of the American Statistical Association* 91:1007–16.

Lindsay, Bruce, Clifford Clogg, and John Grego. 1991. "Semiparametric Estimation in the Rasch Model and Related Exponential Response Models, Including a Simple Latent Class Model for Item Analysis." *Journal of the American Statistical Association* 86:96–107.

Lindsey, James K. 1999. *Models for Repeated Measurements*, 2nd ed. Oxford, England: Oxford University Press.

Littell, Ramon C., George A. Milliken, Walter W. Stroup, and Russell D. Wolfinger. 1996. *SAS System for Mixed Models*, SAS Institute Inc., Cary, NC.

Longford, Nicholas T. 1993. *Random Coefficient Models*. Oxford, England: Oxford University Press.

———. 1994. "Logistic Regression with Random Coefficients." *Computational Statistics and Data Analysis* 17:1–15.

McArdle, John J., and Fumiaki Hamagami. 1994. "Logit and Multilevel Logit Model-

ing of College Graduation for 1984–1985 Freshman Student-athletes." *Journal of the American Statistical Association* 89:1107–23.

McCullagh, Peter, and John A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.

McCulloch, Charles E. 1994. "Maximum Likelihood Variance Components Estimation for Binary Data." *Journal of the American Statistical Association* 89:330–35.

———. 1997. "Maximum Likelihood Algorithms for Generalized Linear Mixed Models." *Journal of the American Statistical Association* 92:162–70.

McGilchrist, C. A. 1994. "Estimation in Generalized Mixed Models." *Journal of the Royal Statistical Society*, ser. B, 56:61–69.

Miller, John J. 1977. "Asymptotic Properties of Maximum Likelihood Estimates in the Mixed Model of the Analysis of Variance." *The Annals of Statistics* 5:746–62.

Montgomery, Mark R., Toni Richards, and Henry I. Braun. 1986. "Child Health, Breastfeeding and Survival in Malaysia: A Random-effects Logit Approach." *Journal of the American Statistical Association* 81:297–309.

Murphy, Mike, and Duolao Wang. 1998. "Family and Sociodemographic Influences on Patterns of Leaving Home in Postwar Britain." *Demography* 35:293–305.

Murray, David M., Joel M. Moskowitz, and Clyde W. Dent. 1996. "Design and Analysis Issues in Community-based Drug Abuse Prevention." *American Behavioral Scientist* 39:853–67.

Muthen, Bengt. 1997. "Longitudinal and Multilevel Modeling: Latent Variable Modeling of Longitudinal and Multilevel Data." *Sociological Methodology* 27:453–80.

Natarajan, Ranjini, and Charles E. McCulloch. 1995. "A Note on the Existence of the Posterior Distribution for a Class of Mixed Models for Binomial Responses." *Biometrika* 82:638–43.

———. 1998. "Gibbs Sampling with Diffuse Priors: A Valid Approach to Data-driven Inference?" *Journal of Computational and Graphical Statistics* 7:267–77.

Nee, V. 1996. "The Emergence of a Market Society: Changing Mechanisms of Stratification in China." *American Journal of Sociology* 101:908–49.

Neuhaus, John M., and John D. Kalbfleisch. 1998. "Between- and Within-cluster Covariate Effects in the Analysis of Clustered Data." *Biometrics* 54:638–45.

Neuhaus, John M., Walter W. Hauck, and John D. Kalbfleisch. 1992. "The Effects of Mixture Distribution Misspecification when Fitting Mixed-effects Logistic Models." *Biometrika* 79:755–62.

Neuhaus, John M., John D. Kalbfleisch, and Walter W. Hauck. 1991. "A Comparison of Cluster-specific and Population-averaged Approaches for Analyzing Correlated Binary Data." *International Statistical Review* 59:25–35.

———. 1994. "Conditions for Consistent Estimation in Mixed-effects Models for Binary Matched-pairs Data." *Canadian Journal of Statistics* 22:139–48.

Plewis, Ian. 1997. *Statistics in Education*. London: Arnold.

Rao, J. N. K., and D. Roland Thomas. 1988. "The Analysis of Cross-classified Categorical Data from Complex Sample Surveys." *Sociological Methodology* 18: 213–69.

Rasch, Georg. 1961. "On General Laws and the Meaning of Measurement in Psychology." *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability* 4:321–33.

Raudenbush, Stephen W., Brian Rowan, and Sang Jin Kang. 1991. "A Multilevel, Multivariate Model for Studying School Climate with Estimation via the EM Algorithm and Application to U.S. High-School Data." *Journal of Educational Statistics* 16:295–330.

Raudenbush, Stephen W., and Robert J. Sampson. 1999. "Econometrics: Toward a Science of Assessing Ecological Settings, with Application to the Systematic Social Observation of Neighborhoods." Pp. 1–41 in *Sociological Methodology 1999*, edited by Michael E. Sobel. Cambridge, MA: Blackwell Publishers.

Sampson, Robert J., Stephen W. Raudenbush, and Felton Earls. 1997. "Neighborhoods and Violent Crime: A Multilevel Study of Collective Efficacy." *Science* 277:918–24.

Saunderson, T. R., and Ian H. Langford. 1996. "A Study of the Geographical Distribution of Suicide Rates in England and Wales 1989–92 Using Empirical Bayes Estimates." *Social Science and Medicine* 43:489–502.

Schall, Robert. 1991. "Estimation in Generalized Linear Models with Random Effects." *Biometrika* 78:719–27.

Searle, Shayle R., George Casella, and Charles E. McCulloch. 1992. *Variance Components*. New York: Wiley.

Self, Steven G., and Kung-Yee Liang. 1987. "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions." *Journal of the American Statistical Association* 82:605–10.

Stiratelli, Robert, Nan Laird, and James H. Ware. 1984. "Random-effects Models for Serial Observations with Binary Response." *Biometrics* 40:961–71.

Tjur, Tue. 1982. "A Connection Between Rasch's Item Analysis Model and a Multiplicative Poisson Model." *Scandinavian Journal of Statistics* 9:23–30.

Tsutakawa, Robert K. 1988. "Mixed Model for Analyzing Geographic Variability in Mortality Rates." *Journal of the American Statistical Association* 83:37–42.

Tutz, Gerhard, and Wolfgang Hennevogl. 1996. "Random Effects in Ordinal Regression Models." *Computational Statistics and Data Analysis* 22:537–57.

Wei, Greg C. G., and Martin A. Tanner. 1990. "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms." *Journal of the American Statistical Association* 85:699–704.

Williams, D. A. 1982. "Extra-binomial Variation in Logistic Linear Models." *Applied Statistics* 31:144–48.

Wolfinger, Russell, and Michael O'Connell. 1993. "Generalized Linear Mixed Models: A Pseudo-likelihood Approach." *Journal of Statistical Computation and Simulation* 48:233–43.

Wong, George Y., and William M. Mason. 1985. "The Hierarchical Logistic Regression Model for Multilevel Analysis." *Journal of the American Statistical Association* 80:513–24.

Zeger, Scott L., Kung-Yee Liang, and Paul S. Albert. 1988. "Models for Longitudinal Data: A Generalized Estimating Equation Approach." *Biometrics* 44:1049–60.

Zhou, Xiao-Hua, Anthony J. Perkins, and Siu L. Hui. 1999. "Comparisons of Software Packages for Generalized Linear Multilevel Models." *American Statistician* 53: 282–90.