Psychological Methods
2011, Vol. 16, No. 4, 373–390

1082-989X/11/$12.00 DOI: 10.1037/a0025813

Fitting Multilevel Models With Ordinal Outcomes: Performance of Alternative Specifications and Methods of Estimation

Daniel J. Bauer
University of North Carolina at Chapel Hill

Sonya K. Sterba
Vanderbilt University

Previous research has compared methods of estimation for fitting multilevel models to binary data, but there are reasons to believe that the results will not always generalize to the ordinal case. This article thus evaluates (a) whether and when fitting multilevel linear models to ordinal outcome data is justified and (b) which estimator to employ when instead fitting multilevel cumulative logit models to ordinal data, maximum likelihood (ML), or penalized quasi-likelihood (PQL). ML and PQL are compared across variations in sample size, magnitude of variance components, number of outcome categories, and distribution shape. Fitting a multilevel linear model to ordinal outcomes is shown to be inferior in virtually all circumstances. PQL performance improves markedly with the number of ordinal categories, regardless of distribution shape. In contrast to binary data, PQL often performs as well as ML when used with ordinal data. Further, the performance of PQL is typically superior to ML when the data include a small to moderate number of clusters (i.e., $\leq$ 50 clusters).

Keywords: multilevel models, random effects, ordinal, cumulative logit model, proportional odds model

Supplemental materials: http://dx.doi.org/10.1037/a0025813.supp

Psychologists, as well as researchers in allied fields of health, education, and social science, are often in the position of collecting and analyzing nested (i.e., clustered) data. Two frequently encountered types of nested data are hierarchically clustered observations, such as individuals nested within groups, and longitudinal data, or repeated measures over time. Both data structures share a common feature: dependence of observations within units (i.e., observations within clusters or repeated measures within persons). Because classical statistical models like analysis of variance and linear regression assume independence, alternative statistical models are required to analyze nested data appropriately.

In psychology, a common way to address dependence in nested data is to use a multilevel model (sometimes referred to as a unit-specific model, or conditional model). A model is specified to include cluster-level random effects to account for similarities within clusters and the observations are assumed to be independent conditional on the random effects. A random intercept captures level differences in the dependent variable across clusters (due to unobserved cluster-level covariates), whereas a random slope implies that the effect of a predictor varies over clusters (interacts with unobserved cluster-level covariates). Alternative ways to model dependence in nested data exist, including population-average (or marginal) models which are typically estimated by generalized estimating equations (GEE; Liang & Zeger, 1986). These models produce estimates of model coefficients for predictors that are averaged over clusters, while allowing residuals to correlate within clusters. Population-average models are robust to misspecification of the correlation structure of the residuals, whereas unit-specific models can be sensitive to misspecification of the random effects. However, unit-specific models are appealing to many psychologists (and others), because they allow for inference about processes that operate at the level of the group (in hierarchical data) or individual (in longitudinal data). Indeed, in a search of the PsycARTICLES database, we found that unit-specific models were used more than 15 times as often as population-average models in psychology applications published over the last 5 years.[1] To maximize relevance for psychologists, we thus focus on the unit-specific multilevel model in this article. Excellent introductions to multilevel modeling include Goldstein (2003), Hox (2010), Raudenbush and Bryk (2002), and Snijders and Bosker (1999).

Though the use of multilevel models to accommodate nesting has increased steadily in psychology over the past several decades, many psychologists appear to have restricted their attention to multilevel *linear* models. These models assume that observations within clusters are continuous and normally distributed, condi-

This article was published Online First October 31, 2011.

Daniel J. Bauer, Department of Psychology, University of North Carolina at Chapel Hill; Sonya K. Sterba, Psychology & Human Development Department, Vanderbilt University.

This research was supported by the National Institute on Drug Abuse (DA013148) and the National Science Foundation (SES-0716555). The authors would also like to thank Patrick Curran for his involvement in and support of this research.

Correspondence concerning this article should be addressed to Daniel J. Bauer, Department of Psychology, University of North Carolina, Chapel Hill, NC 27599-3270. E-mail: dbauer@email.unc.edu

[1] A full-text search of articles published in the past 5 years indicated that 211 articles included the term "multilevel model, "hierarchical linear model," "mixed model," or "random coefficient model" (all unit-specific models), whereas 14 articles included the term "generalized estimation equations" or "GEE." More general searches would be possible but this brief PsychARTICLES search gives an indication of the proportion of unit-specific to population-average applications in psychology.

tional on observed covariates. But very often psychologists measure outcomes on an ordinal scale, involving multiple discrete categories with potentially uneven spacing between categories. For instance, participants might be asked whether they "strongly disagree," "disagree," "neither disagree nor agree," "agree," or "strongly agree" with a particular statement. Although there is growing recognition that the application of linear models with such outcomes is inappropriate, it is still common to see ordinal outcomes treated as continuous so that linear models can be applied (Agresti, Booth, Hobert, & Caffo, 2000; Liu & Agresti, 2005).

Researchers may be reluctant to fit an ordinal rather than linear multilevel model for several reasons. First, researchers are generally more familiar with linear models and may be less certain how to specify and interpret the results of models for ordinal outcomes. Second, to our knowledge, no researchers have expressly examined the consequences of fitting a linear multilevel model to ordinal outcomes. Third, it may not always be apparent what estimation options exist for fitting multilevel models with ordinal outcomes, nor what the implications of choosing one option versus another might be. Indeed, there is a general lack of information on the best method of estimation for the ordinal case. Unlike the case of normal outcomes, the likelihood for ordinal outcomes involves an integral that cannot be resolved analytically, and several alternative estimation methods have been proposed to overcome this difficulty. The strengths and weaknesses of these methods under real-world data conditions are not well understood.

Our goals in writing this article were thus twofold. First, we sought to establish whether and when fitting a linear multilevel model to ordinal data may constitute an acceptable data analysis strategy. Second, we sought to evaluate the relative performance of two estimators for fitting multilevel models to discrete outcomes, namely penalized quasi-likelihood (PQL) and maximum likelihood (ML) using adaptive quadrature. These two methods of estimation were chosen for comparison because of their prevalence within applications and their availability within commonly used software (PQL is a default estimator in many software programs, such as Hierarchical Linear and Nonlinear Modeling–6 [HLM–6, Scientific Software International, or SSI, Lincolnwood, IL] and the GLIMMIX procedure in SAS, and is currently the only estimator available in SPSS; ML with adaptive quadrature is available in the GLIMMIX and NLMIXED SAS procedures as well as Mplus [Muthén & Muthén, 1998–2007], generalized linear latent and mixed models [GLLAMM; Stata Corp., College Station, TX], and SuperMix [SSI]).

We begin by presenting the two alternative model specifications, the multilevel linear model for continuous outcomes versus a multilevel model expressly formulated for ordinal outcomes. We then discuss the topic of estimation and provide a brief review of previous literature on fitting multilevel models to binary and ordinal data, focusing on gaps involving estimation in the ordinal case. Based on the literature, we develop a series of hypotheses that we test in a simulation study that compares two model specifications, linear versus ordinal, under conditions that might commonly occur in psychological research. Further, we compare the estimates of ordinal multilevel models fit via PQL versus ML with adaptive quadrature. The findings from our simulation translate directly into recommendations for current practice.

Alternative Model Specifications

Multilevel Linear Model

We first review the specification of the multilevel linear model. For exposition, let us suppose we are interested in modeling the effects of one individual-level (level-1) predictor X_{ij} and one cluster-level (level-2) predictor W_j, as well as a cross-level interaction, designated $X_{ij}W_j$, where i indexes the individual and j indexes the cluster. To account for the dependence of observations within clusters, we will include a random intercept term, designated u_{0j}, and a random slope for the effect of X_{ij}, designated u_{1j}, to allow for the possibility that the effect of this predictor varies across clusters. This model is represented as:

$$\begin{aligned}
&\text{Level 1:} \quad Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij} \\
&\text{Level 2:} \quad \beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j} \\
&\qquad\qquad \beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j} \qquad (1) \\
&\text{Combined:} \quad Y_{ij} = \gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{ij} + \gamma_{11}W_jX_{ij} + u_{0j} \\
&\qquad\qquad + u_{1j}X_{ij} + r_{ij}
\end{aligned}$$

All notation follows that of Raudenbush and Bryk (2002), with coefficients at Level 1 indicated by β, fixed effects indicated by γ, residuals at Level 1 indicated by *r,* and random effects at Level 2 indicated by *u.* Both the random effects and the residuals are assumed to be normally distributed, or

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{bmatrix} \right) \qquad (2)$$

and

$$r_{ij} \sim N(0, \sigma^2) \qquad (3)$$

An important characteristic of Equation 1 is that it is additive in the random effects and residuals. In concert with the assumptions of normality in Equations 2 and 3, this additive form implies that the conditional distribution of Y_{ij} is continuous and normal.

The use of linear models such as Equation 1 with ordinal outcomes can be questioned on several grounds (Long, 1997, p. 38-40). First, the linear model can generate impossible predicted values, below the lowest category number or above the highest category number. Second, the variability of the residuals becomes compressed as the predicted values move toward the upper or lower limits of the observed values, resulting in heteroscedasticity. Heteroscedasticity and nonnormality of the residuals cast doubt on the validity of significance tests. Third, we often view an ordinal scale as providing a coarse representation for what is really a continuous underlying variable. If we believe that this unobserved continuous variable is *linearly* related to our predictors, then our predictors will be *nonlinearly* related to the observed ordinal variable. The linear model then provides a first approximation of uncertain quality to this nonlinear function. The substitution of a linear model for one that is actually nonlinear is especially problematic for nested data when lower level predictors vary both within and between clusters (have intraclass correlations exceed-

ing zero). In this situation, estimates for random slope variances and cross-level interactions can be inflated or spurious (Bauer & Cai, 2009).

Multilevel Models for Ordinal Outcomes

In general, there are two ways to motivate models for ordinal outcomes. One motivation that is popular in psychology and the social sciences, alluded to earlier, is to conceive of the ordinal outcome as a coarsely categorized measured version of an underlying continuous latent variable. For instance, although attitudes may be measured via ordered categories ranging from "strongly disagree" to "strongly agree," we can imagine that a continuous latent variable underlies these responses. If the continuous variable had been measured directly, then the multilevel linear model in Equation 1 would be appropriate. Thus, for the continuous underlying variable, denoted Y^*_{ij}, we can stipulate the model

$$\begin{aligned}
&\text{Level 1:} \quad Y^*_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij} \\
&\text{Level 2:} \quad \beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j} \\
&\qquad\qquad \beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j} \qquad (4) \\
&\text{Combined:} \quad Y^*_{ij} = \gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{ij} + \gamma_{11}W_jX_{ij} + u_{0j} \\
&\qquad\qquad + u_{1j}X_{ij} + r_{ij}
\end{aligned}$$

To link the underlying Y^*_{ij} with the observed ordinal response Y_{ij}, we must also posit a threshold model. For Y_{ij} scored in categories $c = 1, 2, \ldots, C$, we can write the threshold model as:

$$\begin{aligned}
&Y_{ij} = 1 \text{ if } Y^*_{ij} < \nu^{(1)} \\
&Y_{ij} = 2 \text{ if } \nu^{(1)} \le Y^*_{ij} < \nu^{(2)} \\
&\qquad \vdots \\
&Y_{ij} = C \text{ if } Y^*_{ij} \ge \nu^{(C-1)} \qquad (5)
\end{aligned}$$

where $\nu^{(c)}$ is a threshold parameter and the thresholds are strictly increasing (i.e., $\nu^{(1)} < \nu^{(2)} \ldots < \nu^{(C-1)}$). In words, Equation 5 indicates that when the underlying variable Y^*_{ij} increases past a given threshold, we see a discrete jump in the observed ordinal response Y_{ij} (e.g., when Y^*_{ij} crosses the threshold $\nu^{(1)}$, Y_{ij} changes from a 1 to a 2).

Finally, to translate Equations 4 and 5 into a probability model for Y_{ij}, we must specify the distributions of the random effects and residuals. The random effects at Level 2 are conventionally assumed to be normal, just as in Equation 2. Different assumptions can be made for the Level 1 residuals. Assuming $r_{ij} \sim N(0,1)$ leads to the multilevel probit model, whereas assuming $r_{ij} \sim logistic(0,\pi^2/3)$ leads to the multilevel cumulative logit model. In both cases, the variance is fixed (at 1 for the probit specification and at $\pi^2/3$ for the logit specification) since the scale of the underlying latent variable is unobserved. Of the two specifications, we focus on the multilevel cumulative logit model because it is computationally simpler and because the estimates for the fixed effects have appealing interpretations (i.e., the exponentiated coefficients are interpretable as odds ratios).

Alternatively, the very same models can be motivated from the framework of the generalized linear model (McCullagh & Nelder, 1989), a conceptualization favored within biostatistics. Within this framework, we start by specifying the conditional distribution of our outcome. In this case, the conditional distribution of the ordinal outcome Y_{ij} is multinomial with parameters describing the probabilities of the categorical responses. By modeling these probabilities directly, we bypass the need to invoke a continuous latent variable underlying the ordinal responses.

To further explicate this approach we can define cumulative coding variables to capture the ordered-categorical nature of the observed responses. $C - 1$ coding variables are defined such that $Y^{(c)}_{ij} = 1$ if $Y_{ij} \le c$ (the cumulative coding variable for category C is omitted as it would always be scored 1). The expected value of each cumulative coding variable is then the cumulative probability that a response will be scored in category c or below, denoted as $\varphi^{(c)}_{ij} = P(Y_{ij} \le c) = P[Y^{(c)}_{ij} = 1]$.

The cumulative probabilities are predicted via the *linear predictor*, denoted η_{ij}, which is specified as a weighted linear combination of observed covariates/predictors and random effects. For our example model, the linear predictor would be specified through the equations

$$\begin{aligned}
&\text{Level 1:} \quad \eta_{ij} = \beta_{0j} + \beta_{1j}X_{ij} \\
&\text{Level 2:} \quad \beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j} \\
&\qquad\qquad \beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j} \\
&\text{Combined:} \quad \eta_{ij} = \gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{ij} + \gamma_{11}W_jX_{ij} + u_{0j} \\
&\qquad\qquad + u_{1j}X_{ij} \qquad (6)
\end{aligned}$$

where the random effects are assumed to be normally distributed as in Equation 2.

The model for the observed responses is then given as

$$Y^{(c)}_{ij} = g^{-1}[\nu^{(c)} - \eta_{ij}] + r_{ij} \qquad (7)$$

where $\nu^{(c)}$ is again a threshold parameter that allows for increasing probabilities when accumulating across categories and $g^{-1}(\cdot)$ is the *inverse link function*, a function that maps the continuous range of $[\nu^{(c)} - \eta_{ij}]$ into the bounded zero-to-one range of predicted values (model-implied cumulative probabilities) for the cumulative coding variable (Hedeker & Gibbons, 2006; Long, 1997). Any function with asymptotes of zero and one could be considered as a candidate for $g^{-1}(\cdot)$ but common choices are the cumulative density function (CDF) for the normal distribution, which produces the multilevel probit model, and the inverse logistic function,

$$g^{-1}[\nu^{(c)} - \eta_{ij}] = \varphi^{(c)}_{ij} = \frac{\exp(\nu^{(c)} - \eta_{ij})}{1 + \exp(\nu^{(c)} - \eta_{ij})} \qquad (8)$$

which produces the multilevel cumulative logit model.

Both motivations lead to equivalent models, with the selection of the link function in Equation 7 playing the same role as the choice of residual distribution in Equation 4. The two approaches thus differ only at the conceptual level. Regardless of which conception is preferred, however, a few additional features of the

model should be noted. First, the full set of thresholds and overall model intercept are not jointly identified. One can set the first threshold to zero and estimate the intercept, or set the intercept to zero and estimate all thresholds. The former choice seems to be most common, and we will use that specification in our simulations. Additionally, an assumption of the model, which can be checked empirically, is that the coefficients in the linear predictor are invariant across categories (an assumption referred to as proportional odds for the multilevel cumulative logit model). This assumption can be relaxed, for instance by specifying a partial proportional odds model. For additional details on this assumption and the partial proportional odds model, see Hedeker and Gibbons (2006).

Alternative Estimation Methods

To provide a context for comparison of estimation methods, first consider a general expression for the likelihood function for cluster *j*:

$$L_j(\boldsymbol{\theta}|\mathbf{Y}_j) = \int f(\mathbf{Y}_j|\mathbf{u}_j,\boldsymbol{\theta})h(\mathbf{u}_j|\boldsymbol{\theta})d\mathbf{u}_j \quad (9)$$

where $\boldsymbol{\theta}$ is the vector of parameters to be estimated (fixed effects and variance components), $\mathbf{Y}_j$ is the vector of all observations on Y_{ij} within cluster *j,* and $\mathbf{u}_j$ is the vector of random effects. The density function for the conditional distribution of $\mathbf{Y}_j$ is denoted $f(\cdot)$, and the density function for the random effects is denoted $h(\cdot)$, both of which implicitly depend on the parameters of the model. Integrating the likelihood over the distribution of the random effects returns the marginal likelihood for $\mathbf{Y}_j$, that is, the likelihood of $\mathbf{Y}_j$ averaging over all possible values of the random effects. This averaging is necessary because the random effects are unobserved. The overall sample likelihood is the product of the cluster-wise likelihoods, and we seek to maximize this likelihood to obtain the parameter estimates that are most consistent with our data (i.e., parameter estimates that maximize the likelihood of observing the data we in fact observed).

In the linear multilevel model, both $f(\cdot)$ and $h(\cdot)$ are assumed to be normal and in this case the integral within the likelihood resolves analytically; the marginal likelihood for $\mathbf{Y}_j$ is the multivariate normal density function (Demidenko, 2004, pp. 48–61). No such simplification arises when $f(\cdot)$ is multinomial and $h(\cdot)$ is normal, as is the case for ordinal multilevel models. Obtaining the marginal probability of $\mathbf{Y}_j$ would, in theory, require integrating over the distribution of the random effects at each iteration of the likelihood-maximization procedure, but this task is analytically intractable. One approach to circumvent this problem is to implement a quasi-likelihood estimator (linearizing the integrand at each iteration), and another is to evaluate the integral via a numerical approximation.

The idea behind quasi-likelihood estimators (PQL and marginal quasi-likelihood, or MQL) is to take the nonlinear model from Equation 7 and apply a linear approximation at each iteration. This linear model is then fit via normal-theory ML using observation weights to counteract heteroscedasticity and nonnormality of the residuals. This is an iterative process with the linear approximation improving at each step. More specifically, the linear approximation typically employed is a first-order Taylor series expansion of the nonlinear function $g^{-1}[\nu^{(c)} - \eta_{ij}]$. Algebraic manipulation of the linearized model is then used to create a "working variate" Z_{ij} which is an additive combination of the linear predictor $\nu^{(c)} - \eta_{ij}$ and a residual e_{ij} (see the Appendix in the online supplemental materials for more details). The working variate is constructed somewhat differently in MQL and PQL; it is constructed exclusively using fixed effects in the former but using both fixed effects and empirical Bayes estimates of the random effects in the latter (see Goldstein, 2003, pp. 112–114; Raudenbush & Bryk, 2002, pp. 456–459). The residual is the original Level-1 residual term scaled by a weight $e_{ij} = r_{ij}/w_{ij}$ derived from the linearization procedure to render the residual distribution approximately normal, that is, $e_{ij} \sim N(0,1/w_{ij})$. The resultant model for the "working variate," $Z_{ij} = (\nu^{(c)} - \eta_{ij}) + e_{ij}$, approximately satisfies assumptions of the multilevel linear model and can be used to construct an approximate (or quasi-) likelihood.

An alternative way to address the analytical intractability of the integral in Equation 9 is to leave the integrand intact but approximate the integral numerically. Included within this approach is ML using Gauss–Hermite quadrature, adaptive quadrature, Laplace algorithms, and simulation methods. Likewise, Bayesian estimation using Markov Chain Monte Carlo with noninformative (or diffuse) priors can be viewed as an approximation to ML that implements simulation methods to avoid integration. Here we focus specifically on ML with adaptive quadrature. With this method, the integral is approximated via a weighted sum of discrete points. The locations of these points of support (quadrature points) and their respective weights are iteratively updated (or adapted) for each cluster *j,* which has the effect of recentering and rescaling the points in a unit-specific manner (Rabe-Hesketh, Skrondal & Pickles, 2002). At each iteration, the adapted quadrature points are solved for as functions of the mean or mode and standard deviation of the posterior distribution for cluster *j*. Integral approximation improves as the number of points of support per dimension of integration increase, at the expense of computational time. Computational time also increases exponentially with the dimensions of integration, which in Equation 9 corresponds to the number of random effects. The nature of the discrete distribution employed differs across approaches (e.g. rectangular vs. trapezoidal vs. Guass–Hermite), where, for example, rectangular adaptive quadrature considers a discrete distribution of adjoining rectangles.

Prior Research

Fitting a Multilevel Linear Model by Normal-Theory ML

The practice of fitting linear models to ordinal outcomes using normal-theory methods of estimation remains common (Agresti et al., 2000; Liu & Agresti, 2005). To date, however, no research has been conducted to evaluate the performance of multilevel linear models with ordinal outcomes from which to argue against this practice. A large number of studies have, however, evaluated the use of linear regression or normal-theory structural equation modeling (SEM) with ordinal data (see Winship & Mare, 1984, and Bollen, 1989, pp. 415–448, for review). These studies are relevant here because the multilevel linear model can be considered a

generalization of linear regression and a submodel of SEM (Bauer, 2003; Curran, 2003; Mehta & Neale, 2005). Overall, this research indicates that, for ordinal outcomes, the effect estimates obtained from linear models are often attenuated, but that there are circumstances under which the bias is small enough to be tolerable. These circumstances are when there are many categories (better resembling an interval-scaled outcome) and the category distributions are not excessively nonnormal. Extrapolating from this literature, we expect multilevel linear models to perform most poorly with binary outcomes or ordinal outcomes with few categories and best with ordinal outcomes with many categories (approaching a continuum) and roughly normal distributions (Goldstein, 2003, p. 104).

Fitting a Multilevel Cumulative Logit Model by Quasi-Likelihood

Simulation research with PQL and MQL to date has focused almost exclusively on binary rather than ordinal outcomes. This research has consistently shown that PQL performs better than MQL (Breslow & Clayton, 1993; Breslow & Lin, 1995; Goldstein & Rasbash, 1996; Rodriguez & Goldman, 1995, 2001). In either case, however, the quality of the estimates depends on the adequacy of the Taylor series approximation and the extent to which the distribution of the working variate residuals is approximately normal (McCulloch, 1997). When these approximations are poor, the estimates are attenuated, particularly for variance components. In general, PQL performs best when there are many observations per cluster (Bellamy, Li, Lin, & Ryan, 2005; Ten Have & Localio, 1999; Skrondal & Rabe-Hesketh, 2004, pp. 194–197), for it is then that the provisional estimates of the random effects become most precise, yielding a better working variate. The performance of PQL deteriorates when the working variate residuals are markedly nonnormal, as is usually the case when the outcome is binary (Breslow & Clayton, 1993; Skrondal & Rabe-Hesketh, 2004, pp. 194–197). The degree of bias increases with the magnitude of the random effect variances (Breslow & Lin, 1995; McCulloch, 1997; Rodriguez & Goldman, 2001).[2]

Though it is well-known that PQL can often produce badly biased estimates when applied to binary data (Breslow & Lin, 1995; Raudenbush, Yang, & Yosef, 2000; Rodriguez & Goldman, 1995, 2001), it is presently unknown whether this bias will extend to multilevel models for ordinal outcomes. The assumption seems to be that the poor performance of PQL will indeed generalize (Agresti et al., 2000; Liu & Agresti, 2005), leading some to make blanket recommendations that quasi-likelihood estimators should not be used in practice (McCulloch, Searle, & Neuhaus, 2008, p. 198). This conclusion may, however, be premature. For instance, Saei and McGilchrist (1998) detected only slight bias for a PQL-like estimator when the outcome variable had four categories and was observed for three individuals in each of 30 clusters. Beyond the specific instance considered by Saei and McGilchrist (1998), we believe that the bias incurred from the use of PQL will diminish progressively with the number of categories of the ordinal outcome due to the increase in information with more ordered categories. To our knowledge, this hypothesis has not previously appeared in the literature on PQL, nor has the quality of PQL estimates been compared over increasing numbers of categories.

Fitting the Multilevel Cumulative Logit Model by ML With Adaptive Quadrature

ML estimation for the multilevel cumulative logit model is theoretically preferable to quasi-likelihood estimation because it produces asymptotically unbiased estimates. Moreover, a number of simulation studies have shown that ML using quadrature (or other integral approximation approaches) outperforms quasi-likelihood estimators such as PQL when used to estimate multilevel logistic models with binary outcomes (Raudenbush et al., 2000; Rodriguez & Goldman, 1995). As one would expect given its desirable asymptotic properties, ML with numerical integration performs best when there is a large number of clusters.

There are, however, still compelling reasons to compare the ML and PQL estimators for the cumulative logit model. First, although ML is an asymptotically unbiased estimator, it suffers from small sample bias (Demidenko, 2004, p. 58; Raudenbush & Bryk, 2002, p. 53). When the number of clusters is small, ML produces negatively biased variance estimates for the random effects. Additionally, this small-sample bias increases with the number of fixed effects. For ordinal outcomes, the fixed effects include $C - 1$ threshold parameters, so a higher number of categories may actually increase the bias of ML estimates. Second, Bellamy et al. (2005) showed analytically and empirically that when there is a small number of large clusters, as often occurs in group-randomized trials, the efficiency of PQL estimates can equal or exceed the efficiency of ML estimates. Third, as discussed earlier, PQL may compare more favorably to ML when the data are ordinal rather than binary, as the availability of more categories may offset PQL's particularly strong need for large clusters.

Research Hypotheses

From the literature previously reviewed, we now summarize the research hypotheses that motivated our simulation study.

Hypothesis 1: A linear modeling approach may perform adequately when the number of categories for the outcome is large (e.g., 5+) and when the distribution of category responses is roughly normal in shape, but will prove inadequate if either of these conditions is lacking.

Hypothesis 2: ML via adaptive quadrature will be unbiased and most efficient when the number of clusters is large, but these properties may not hold when there are fewer clusters. In particular, variance estimates may be negatively biased when the number of clusters is small and the number of fixed effects (including thresholds, increasing with number of categories) is large.

[2] To improve performance, Goldstein and Rabash (1996) proposed the PQL2 estimator, which uses a second-order Taylor series expansion to provide a more precise linear approximation. Rodriguez and Goldman (2001) found that PQL2 is less biased than PQL, but less efficient and somewhat less likely to converge.

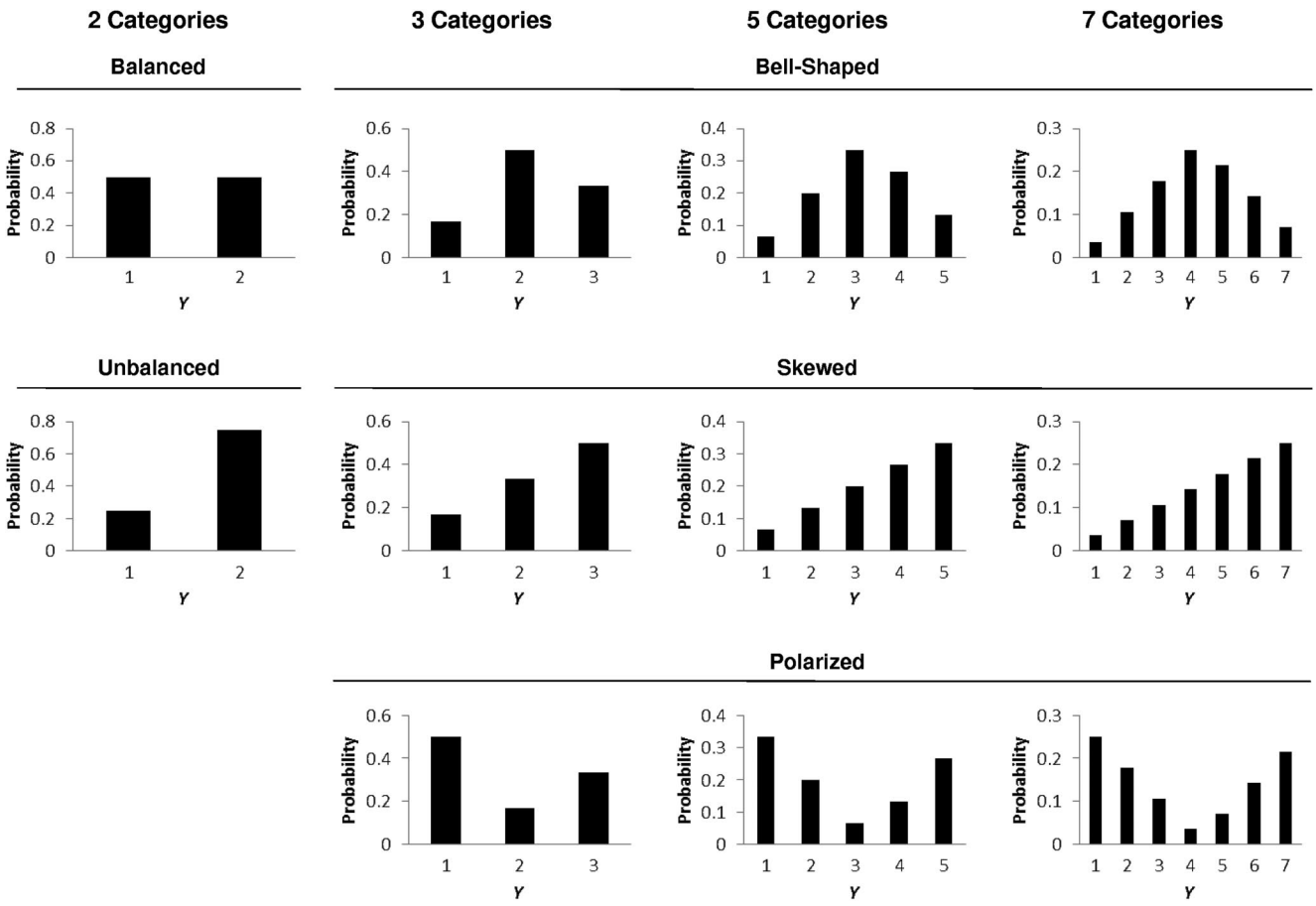

Figure 1. Marginal category distributions used in the simulation study (averaged over predictors and random effects). Within three-, five-, and seven-category outcome conditions, marginal frequencies are held constant but permuted across categories to manipulate the distribution shape (bell-shaped, skewed, or polarized) without changing sparseness. Within two-category outcome conditions, it is impossible to hold marginal frequencies constant while manipulating shape (balanced or unbalanced). Y is the outcome variable, and the numbers listed on the *x* axis indicate the category of the outcome variable.

Hypothesis 3: PQL estimates will be attenuated, especially when the variances of the random effects are large and when the cluster sizes are small. More bias will be observed for variance components than fixed effects.

Hypothesis 4: PQL will perform considerably better for ordinal outcomes as the number of categories increases. With a sufficient number of categories, PQL may have negligible bias and comparable or better efficiency than ML even when cluster sizes are small.

Of these hypotheses, no prior research has been conducted directly on Hypothesis 1, which is based on research conducted with related models (linear regression and SEM). Hypotheses 2 and 3 follow directly from research on binary outcomes. We believe Hypothesis 4 to be novel, notwithstanding the limited study of Saei and McGilchrist (1998), and it is this hypothesis that is most important to our investigation.

Simulation Study

Method

Design. To test our hypotheses, we simulated ordinal data with two, three, five, or seven categories, and we varied the number of clusters (J = 25, 50, 100, or 200), the cluster sizes (n_j = five, 10, or 20), the magnitude of the random effects, and the distribution of category responses. Our population-generating model was a multilevel cumulative logit model, with parameter values chosen to match those of Raudenbush, Yang, and Yosef (2000) and Yosef (2001), which were based on values derived by Rodriguez and Goldman (1995) from a multilevel analysis of health care in Guatemala. Whereas Rodriguez and Goldman (1995) considered three-level data with random intercepts, Raudenbush, Yang, and Yosef (2000) modified the generating model to be two levels and included a random slope for the Level-1 covariate. We in turn modified Raudenbush, Yang, and Yosef's (2000) generating model to also include a

cross-level interaction. The structure of the population-generating model was of the form specified in Equations 4 and 5 or, equivalently, Equations 6 and 7.

The fixed effects in the population model were $\gamma_{00} = 0$, $\gamma_{01} = 1$, $\gamma_{10} = 1$, $\gamma_{11} = 1$. Following Raudenbush, Yang, and Yosef (2000) and Yosef (2001), the two predictors were generated to be independent and normally distributed as $X_{ij} \sim N(.1,1)$ and $W_j \sim N(-.7,1)$.[3] In one condition, the variances of the random effects were $\tau_{00} = 1.63$, $\tau_{10} = .20$, $\tau_{11} = .25$, as in Raudenbush, Yang, and Yosef (2000). We also included smaller and larger random effects by specifying $\tau_{00} = .5$, $\tau_{10} = .03$, $\tau_{11} = .08$ and $\tau_{00} = 8.15$, $\tau_{10} = .50$, $\tau_{11} = 1.25$, respectively. According to the method described by Snijders and Bosker (1999, p. 224), these values imply residual pseudo-intraclass correlations (ICCs) of .13, .33, and .72, holding X_j at the mean. For hierarchically clustered data, an ICC of .33 is fairly large, whereas an ICC of .13 is more typical. For long-term longitudinal data (e.g., annual or biennial), an ICC of .33 might be considered moderate, whereas the larger ICC of .72 would be observed more often for closely spaced repeated measures (e.g., experience sampling data). Since typical effect sizes vary across data structures, we shall simply refer to these conditions in relative terms as *small*, *medium*, and *large*.

We varied the thresholds of the model in number and placement to determine the number of categories and shape of the category distribution for the outcome. For the binary data, thresholds were selected to yield both balanced, $P(Y = 1) = .50$, and unbalanced, $P(Y = 1) = .75$, marginal distributions. Note that for binary data, manipulating the shape of the distribution necessarily also entails manipulating category sparseness. In contrast, for ordinal data, we considered three different marginal distribution shapes—bell-shaped, skewed, and polarized—while holding sparseness constant by simply shifting which categories had high versus low probabilities. For the bell-shaped distributions, the middle categories had the highest probabilities; for the skewed distributions, the probabilities increased from low to high categories; and for the polarized distribution, the highest probabilities were placed on the end-points. The resulting distributions are shown in Figure 1.[4] As stated in Hypothesis 1, the bell-shaped distribution, approximating a normal distribution, was expected to be favorable for the linear model, although in practice skewed distributions are common in examinations of risk behaviors and polarized distributions are common with attitude data (e.g., attitudes toward abortion). The PQL and ML estimators of the multilevel cumulative logit model were not expected to be particularly sensitive to this manipulation.

SAS Version 9.1 was used for data generation, some analyses, and the compilation of results. The IML procedure was used to generate 500 sets of sample data (replications) for each of the 264 cells of the study. The linear multilevel model was fit to the data with the MIXED procedure using the normal-theory restricted maximum-likelihood (REML) estimator (maximum 500 iterations). The multilevel cumulative logit models were fit either by PQL using the GLIMMIX procedure (with residual subject-specific pseudo-likelihood, RSPL, maximum 200 iterations), or by ML with numerical integration using adaptive Gauss–Hermite quadrature with 15 quadrature points in Mplus Version 5 (with expectation-maximization algorithm, maximum 200 iterations).[5,6] The NLMIXED and GLIMMIX procedures also provide ML estimation by adaptive quadrature, but computational times were shorter with Mplus. The MIXED (REML) and GLIMMIX (PQL) procedures implement boundary constraints on variance estimates to prevent them from going below zero (no such constraint is necessary when using ML with quadrature).

Complicating comparisons of the three model-fitting approaches, results obtained from the linear and cumulative logit models are not on the same scale. To resolve this problem, we transformed linear model estimates to match the scale of the logistic model estimates. Fixed effects and standard errors were multiplied by the factor $s = \sqrt{\pi^2/3\hat{\sigma}^2}$ (where $\hat{\sigma}^2$ is the estimated Level-1 residual variance from the linear model, and $\pi^2/3$ is the variance of the logistic distribution), and variances and covariance parameter estimates were multiplied by s^2. A similar rescaling

Table 1
Top η^2_G Effect Sizes for Contrasts of Fixed-Effect Estimates Across Model Specifications and Estimators

Design factor	Fixed-effect estimates		
	$X_{ij}(\hat{\gamma}_{10})$	$W_i(\hat{\gamma}_{01})$	$X_{ij}W_i(\hat{\gamma}_{11})$
Contrast 1: Linear vs. logistic model			
Main effect	0.37	0.11	0.44
× No. of categories	0.02	0.01	0.03
× Distribution shape	0.02	<0.01	0.03
Contrast 2: PQL vs. ML logistic model			
Main effect	0.06	0.03	0.07
× Size of random effects	0.02	0.01	0.02
× No. of categories	0.01	<0.01	0.01
× Cluster size	0.01	<0.01	0.01

Note. "×" indicates an interaction of the designated between-subjects factor of the simulation design with the within-subjects contrast for method of estimation. PQL = penalized quasi-likelihood; ML = maximum likelihood.

[3] Raudenbush, Yang, and Yosef (2000) mistakenly indicated that the variances of their predictors were .07 for X_{ij} and .23 for W_j; however, Yosef (2001, p. 70) correctly indicated a variance of 1 for both. When data are generated using the lower variances of 0.07 and 0.23, both ML by adaptive quadrature and the sixth-order Laplace estimator produce estimates with *larger* root mean-square errors than PQL, opposite from the results reported in Raudenbush, Yang, and Yosef (2000). This difference is likely due to the interplay between predictor scale and effect size (i.e., a random slope variance of 0.25 for a predictor with variance 0.07 corresponds approximately to a slope variance of 3.7 for a predictor with variance 1).

[4] Information on category thresholds and the method used to determine these to produce the target marginal distributions can be obtained from the first author upon request.

[5] The Mplus implementation of adaptive quadrature iteratively updates quadrature points on the basis of mean (rather than mode) and variance of the cluster-specific posterior distribution.

[6] Several consistency checks were performed to evaluate the adequacy of the ML estimates obtained with these settings. First, nearly identical estimates were obtained using 15 versus 100 quadrature points, or using trapezoidal versus Gauss–Hermite quadrature. Second, results did not differ meaningfully between Mplus and either SAS NLMIXED or SAS GLIMMIX using adaptive quadrature (Version 9.2). Finally, the results obtained with adaptive quadrature were also consistent with those obtained via the sixth-order Laplace ML estimator in HLM–6.

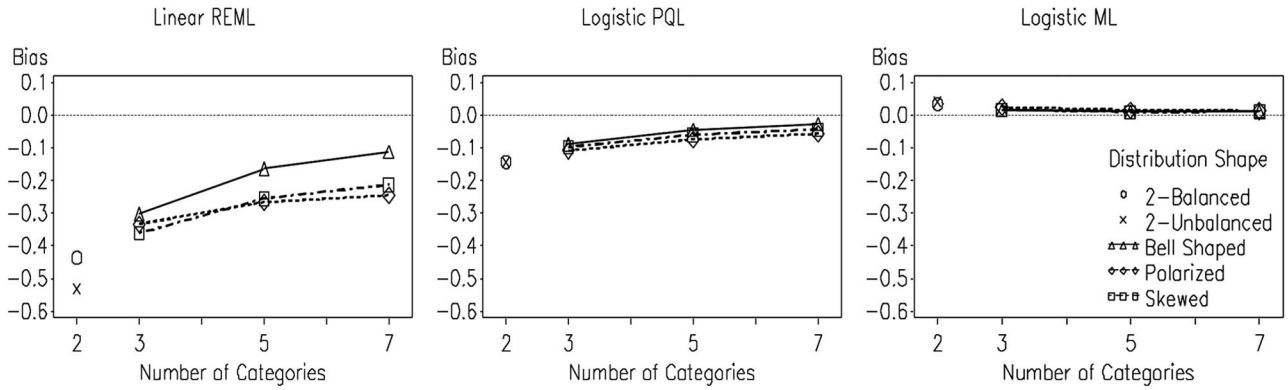

Figure 2. Average bias for the three fixed-effect estimates (excluding thresholds) across estimator, number of outcome categories, and distribution shape. The normal-theory REML (restricted maximum likelihood) estimator was used when fitting the linear multilevel model. The estimators of PQL (penalized quasi-likelihood) or ML (maximum likelihood with adaptive quadrature) were used when fitting the multilevel cumulative logit (logistic) model. Points for two-category conditions are not connected to points for the three- through seven-category conditions because their distribution shapes do not correspond. Results show that bias is large and sensitive to distribution shape when using the linear model but not when using the cumulative logit model (either estimator). Results are collapsed over the number of clusters, cluster size, and the magnitude of the random effects.

strategy has been recommended by Chinn (2000) to facilitate meta-analysis when some studies use logistic versus linear regression (see also Bauer, 2009).

Performance measures. We examined both the bias and efficiency of the estimates. Bias indicates whether a parameter tends to be over- or underestimated, and is computed as the difference between the mean of the estimates (across samples) and the true value, or

$$B = E(\hat{\theta}_r) - \theta \tag{10}$$

where θ is the parameter of interest, $\hat{\theta}_r$ is the estimate of θ for replication r, and $E(\hat{\theta}_r)$ is the mean estimate across replications. A good estimator should have bias values near zero, indicating that the sample estimates average out to equal the population value. Bias of 5%–10% is often considered tolerable (e.g., Kaplan, 1989). Likewise, to evaluate efficiency, one can examine the variance of the estimates,

$$V = E[(\hat{\theta}_r - E(\hat{\theta}_r))^2] \tag{11}$$

A good estimator will have less variance than other estimators, indicating more precision and, typically, higher power for inferential tests.

Bias and variance should be considered simultaneously when judging an estimator. For instance, an unbiased estimator with high variance is not very useful, since the estimate obtained in any single sample is likely to be quite far from the population value. Another estimator may be more biased but have low variance, so that any given estimate is usually not too far from the population value. An index that combines both bias and variance is the mean squared error (*MSE*), which is computed as the average squared difference between the estimate and the true parameter value across samples

$$MSE = E[(\hat{\theta}_r - \theta)^2]. \tag{12}$$

It can be shown that $MSE = B^2 + V$, thus *MSE* takes into account both bias and efficiency (Kendall & Stuart, 1969, Section 17.30). A low *MSE* is desirable, as it indicates that any given sample estimate is likely to be close to the population value.

Results

We first consider the estimates of the fixed effects, then the dispersion estimates for the random effects. To streamline presentation, we have provided some results in an online Appendix. In particular, bias in threshold estimates is presented in the online Appendix as thresholds are rarely of substantive interest (and are not estimated with the linear model specification). The pattern of bias in threshold estimates obtained from PQL and ML was (predictably) the mirror image of the pattern described for the other fixed effects.[7]

Fixed-effect estimates. Our first concern was with identifying factors relating to bias in the estimators. Accordingly, we fit a preliminary analysis of variance (ANOVA) model for each fixed effect, treating model-fitting approach as a within-subjects factor and all other factors as between-subjects factors, and used Helmert contrasts to (a) compare the linear model estimates to the estimates obtained from the logistic (cumulative logit) model estimates and (b) differentiate between the two logistic model estimators, PQL and ML. The

[7] Threshold bias was anticipated to be opposite in sign to the bias of other fixed effects given the sign difference of thresholds and fixed effects in the function $g^{-1}(\nu^{(c)} - \eta_{ij})$. Bias would be in the same direction had we used an alternative parameterization of the multilevel cumulative logit model that includes a unique intercept for each cumulative coding variable but no threshold parameters, e.g., $g^{-1}(-\eta_{ij}^{(c)})$ with $\eta_{ij}^{(c)} = \gamma_{00}^{(c)} + \gamma_{01}W_j + \gamma_{10}X_{ij} + \gamma_{11}W_jX_{ij} + u_{0j} + u_{1j}X_{ij}$. The intercepts obtained with this alternative parameterization and the thresholds obtained with the parameterization used in our study differ only in sign, that is, $-1(\nu^{(c)}) = \gamma_{00}^{(c)}$. Given this relationship, threshold bias results are consistent with the bias results observed for other fixed effects.

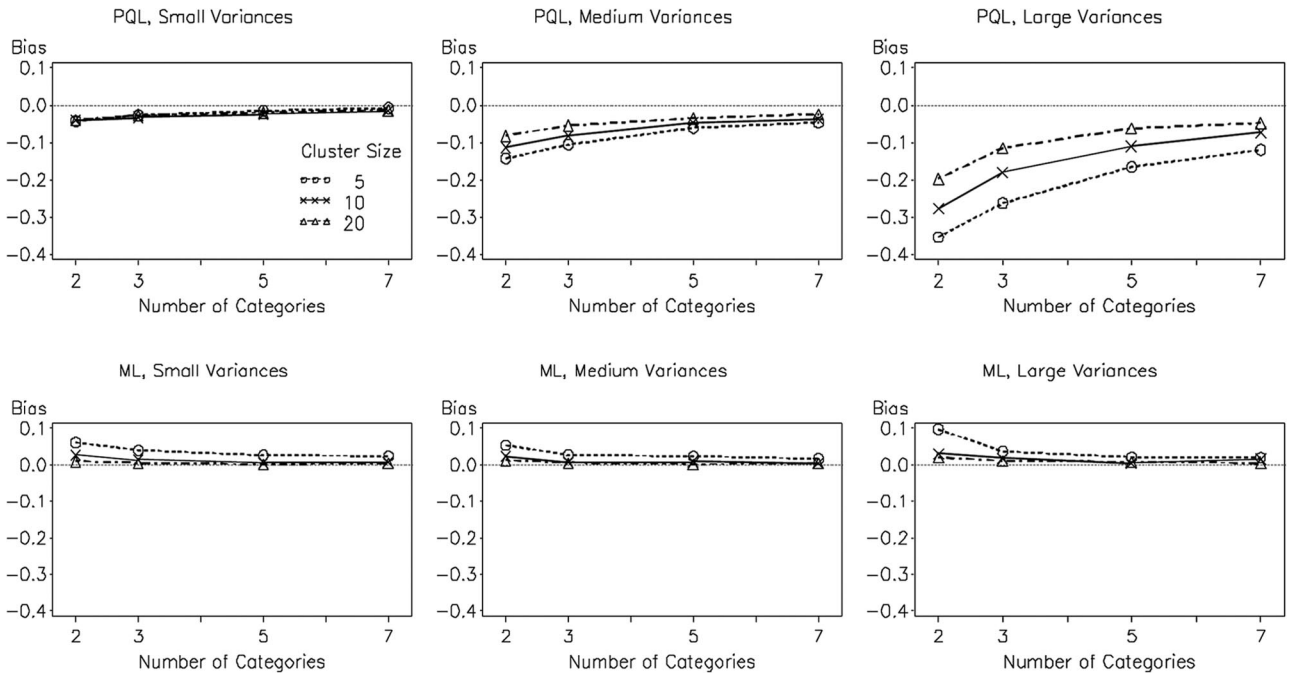

Figure 3. Average bias for the three fixed-effect estimates (excluding thresholds) across logistic estimators, number of outcome categories and cluster size. Logistic estimators were either PQL (penalized quasi-likelihood) or ML (maximum likelihood) with adaptive quadrature. Results show that PQL produces somewhat negatively biased fixed effect estimates, particularly when random effects have large variances, whereas the estimates obtained from logistic ML show small, positive bias. In both cases, bias decreases with the number of categories of the outcome. Results are collapsed over number of clusters and distribution shape and do not include linear multilevel model conditions.

three fixed-effect estimates of primary interest were the main effect of the lower level predictor X_{ij}, the main effect of the upper level predictor W_j, and the cross-level interaction $X_{ij}W_j$. Binary conditions were excluded from the ANOVAs (since the binary distribution shapes differed from the ordinal distribution shapes), but summary plots and tables nevertheless include these conditions. Given space constraints, we provide only brief summaries of the ANOVAs, focusing on the contrasts between estimators. Effect sizes were computed with the generalized eta-squared (η^2_G) statistic (Bakeman, 2005; Olejnik & Algina, 2003). η^2_G values computed for mixed designs are comparable to partial η^2_G values for fully between-subject designs. Our interpretation focuses on contrast effects with η^2_G values of .01 or higher, shown in Table 1.

The largest effect sizes were obtained for the main effect of the first Helmert contrast, comparing the estimates obtained from the linear versus cumulative logit model specifications. As hypothesized, two interaction effects involving the first contrast were identified for all three fixed effects: the number of categories and the distribution shape. Table 1 shows that effect sizes were larger for the effects of *X* and *XW* than *W*, but the pattern of differences in the estimates was similar (see online Appendix). As depicted in Figure 2, averaging over the three fixed effects, the bias of the linear REML estimator was quite severe with binary data, especially when the distribution was unbalanced. The degree of bias for this estimator diminished as the number of categories increased and was least pronounced with the bell-shaped distribution. The bias of the linear REML estimator approached tolerable levels (<10%) only with seven categories and a bell-shaped distribution. In comparison, both estimators of the multilevel cumulative logit model produced less biased estimates that demonstrated little sensitivity to the shape of the distribution.

The second Helmert contrast, comparing the PQL and ML estimates of the multilevel cumulative logit model, resulted in the second largest effect sizes. As hypothesized, the top three factors influencing differences in PQL versus ML estimates of all three fixed effects were the magnitude of the variance components, number of categories, and cluster size. Figure 3 presents the average bias of the three fixed effects as a function of these three factors (results were similar across fixed effects; see online Appendix). In general, PQL produced negatively biased estimates, whereas ML produced positively biased estimates. As expected, PQL performed particularly poorly with binary outcomes, especially when the variances of the random effects were large and the cluster sizes were small. With five to seven categories, however, PQL performed reasonably well even when the random effect variances were moderate. With very large random effects, PQL performed well only when cluster sizes were also large. In absolute terms, the bias for ML was consistently lower than PQL. Somewhat unexpectedly, ML estimates were more biased with binary outcomes than with ordinal outcomes.

To gain a fuller understanding of the differences between the PQL and ML estimators of the multilevel cumulative logit model, we plotted the *MSE*, sampling variance, and bias of the estimates in Figure 4 as a function of all design factors except distribution shape. In the figure, the overall height of each vertical line indicates the *MSE*. The *MSE* is partitioned between squared bias and sampling variance

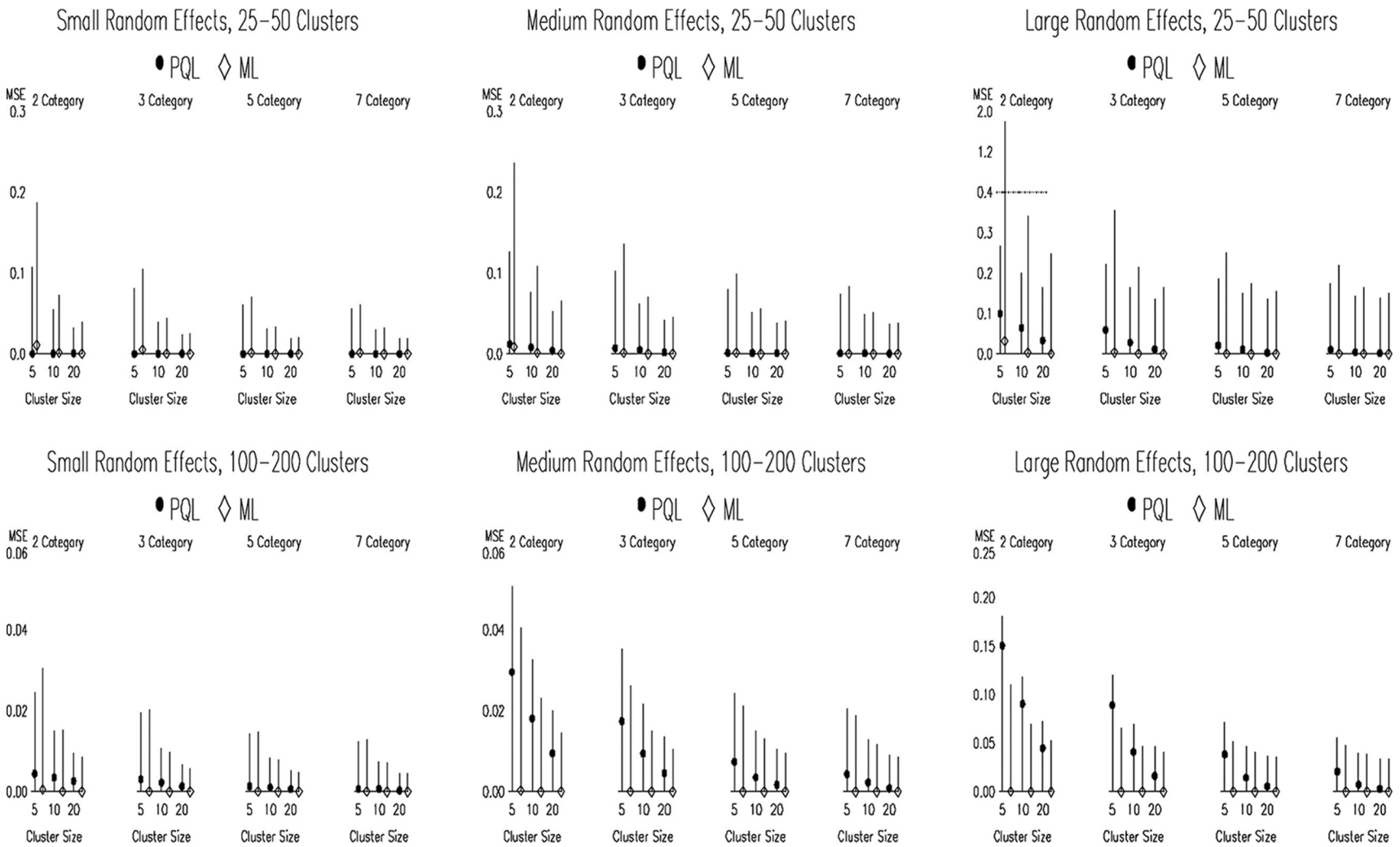

Figure 4. Mean-squared error (*MSE*) for the fixed effects (excluding thresholds) across number of outcome categories, number of clusters, and cluster size. *MSE* is indicated by the height of the vertical lines, and it is broken into components representing squared bias (portion of the line below the symbol) and sampling variance (portion of the line above the symbol). The scale differs across panels and is discontinuous in the upper right panel. *MSE* is averaged across the three fixed effects. Results are plotted for multilevel cumulative logit models; PQL denotes penalized quasi-likelihood, and ML denotes maximum likelihood with adaptive quadrature. This plot does not include linear multilevel model conditions.

by the symbol marker (dot or diamond). The distance between zero and the marker is the squared bias, whereas the distance between the marker and the top of the line is the sampling variance. Note that the scale differs between panels to account for the naturally large effect of number of clusters on the sampling variance, and the increase in sampling variance associated with larger random effects. There is also a break in the scale for the upper right panel due to exceptionally high sampling variance observed for ML with binary outcomes and few, small clusters.

Figure 4 clarifies that, in most conditions, the primary contributor to the *MSE* was the sampling variance, which tended to be lower for PQL than ML. An advantage was observed for ML only when there were many clusters and the random effects were medium or large, especially when there were also few categories and low cluster sizes. In all other conditions, PQL displayed comparable or lower *MSE*, despite generally higher bias, due to lower sampling variance. Both bias and sampling variance decreased with more categories, considerably lowering *MSE*.

Finally, we also considered the quality of inferences afforded by PQL versus ML for the fixed effects. Bias in the standard error estimates was computed for each condition as the difference between the average estimated standard error (*SE*) for an effect and that effect's empirical standard deviation across replications. Figure 5 presents the average *SE* for the fixed effects in the same format as Figure 3 (results were again similar across fixed effects; see online appendix). *SE* bias was generally minimal for both estimators except for ML with binary outcomes and small cluster size. Given the low level of *SE* bias, the quality of inferences is determined almost exclusively by point estimate bias. Indeed, confidence interval coverage rates (tabled in the online Appendix) show that ML generally maintains the nominal coverage rate, whereas PQL has lower than nominal coverage rates under conditions when PQL produces biased fixed effects.

Estimates of dispersion for the random effects. An initial examination of the variance estimates for the random effects revealed very skewed distributions, sometimes with extreme values. We thus chose to evaluate estimator performance with respect to the standard deviations of the random effects (i.e., $\sqrt{\tau_{00}}$ and $\sqrt{\tau_{11}}$), rather than their variances. Stratifying by the magnitude of the random effects, preliminary ANOVA models were fit to determine the primary sources of differences in $\sqrt{\hat{\tau}_{00}}$ and $\sqrt{\hat{\tau}_{11}}$ between the three estimators. The same two Helmert contrasts were used as described in the previous section. Effect sizes are reported in Table 2.

In all the ANOVA results for the dispersion estimates, larger random effect sizes resulted in more pronounced estimator differences and more pronounced factor effects on estimator differences. The largest effect sizes were again associated with overall differences in estimates produced by the linear model versus cumulative logit models. For $\sqrt{\hat{\tau}_{00}}$, interactions with the first contrast were detected for the number of categories of the outcome and, to a much smaller degree, cluster size. For $\sqrt{\hat{\tau}_{11}}$, no interactions with the first contrast consistently approached η^2_G values of .01.

Results for the second contrast indicated that PQL and ML estimates of dispersion also diverged with the magnitude of the random effects. The number of categories had an increasing effect on estimator differences with the magnitude of the random effects, as did cluster size. The number of clusters also had a small effect on estimator differences.

To clarify these results, Tables 3–6 display the mean and standard deviation of the dispersion estimates $\sqrt{\hat{\tau}_{00}}$ and $\sqrt{\hat{\tau}_{11}}$, respec-

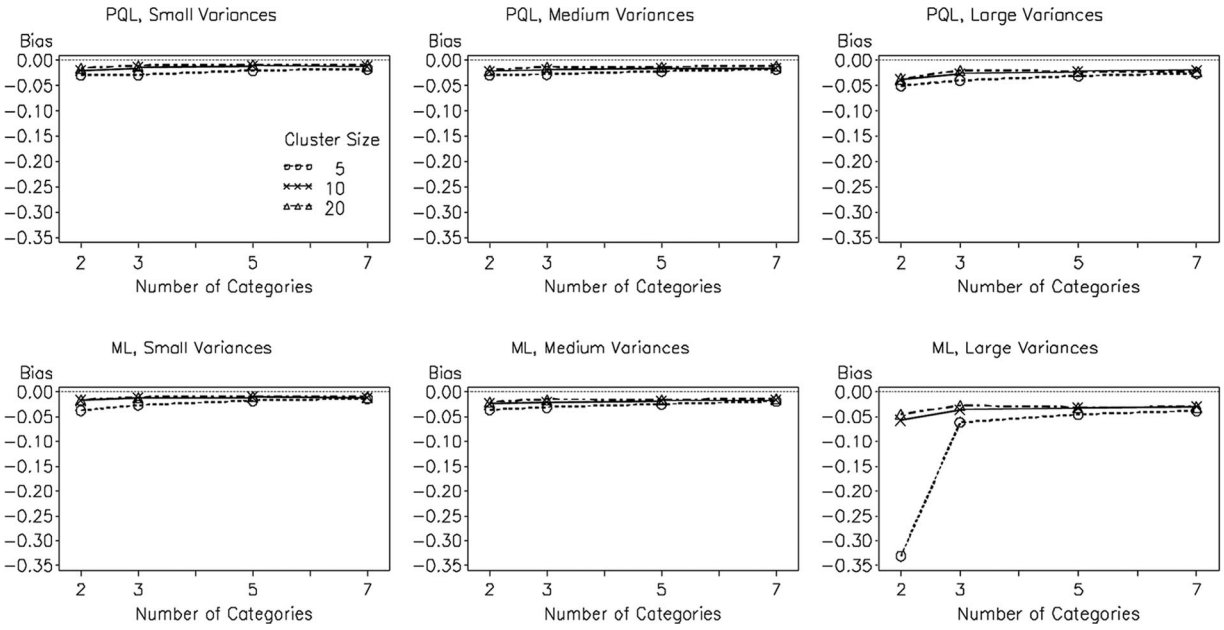

Figure 5. Average bias for the standard errors of the three fixed effect estimates (excluding thresholds) across number of outcome categories and cluster size. Results are plotted for multilevel cumulative logit models; PQL denotes penalized quasi-likelihood, and ML denotes maximum likelihood with adaptive quadrature. This plot does not include linear multilevel model conditions.

Table 2
Top η^2_G Effect Sizes for Contrasts of Random Effect Dispersion Estimates Across Model Specifications and Estimators

Design Factor	Variance components: Small $\hat{\tau}_{00}$	Small $\hat{\tau}_{11}$	Medium $\hat{\tau}_{00}$	Medium $\hat{\tau}_{11}$	Large $\hat{\tau}_{00}$	Large $\hat{\tau}_{11}$
Contrast 1: Linear vs. logistic model						
Main effect	<0.01	<0.01	0.05	0.02	0.21	0.11
× No. of categories	0.01	<0.01	0.02	<0.01	0.04	0.01
× Cluster size	<0.01	0.01	<0.01	<0.01	0.01	0.01
Contrast 2: PQL vs. ML logistic model						
Main effect	<0.01	<0.01	0.02	0.01	0.27	0.15
× No. of Categories	<0.01	<0.01	0.01	0.01	0.07	0.05
× Cluster size	<0.01	<0.01	0.01	0.01	0.05	0.05
× No. of Clusters	0.01	<0.01	0.01	<0.01	0.01	0.01

Note. "×" indicates an interaction of the designated between-subjects factor of the simulation design with the within-subjects contrast for method of estimation. PQL = penalized quasi-likelihood; ML = maximum likelihood.

tively, as a function of characteristics of the outcome variable and estimator. For both the random intercept (Table 3) and slope (Table 4), the estimates obtained from the linear model show the most bias, but they improve markedly as the number of categories increases. Like the linear model estimator, PQL performance improves markedly as the number of categories increases, whereas the estimates obtained from ML are generally less biased (but more variable) when there are fewer categories. Indeed, the ML estimates actually become negatively biased as the number of categories increases, a trend that is consistent with the known negative bias of ML dispersion estimates as a function of the number of fixed effects (with more categories requiring the addition of more thresholds).

Similarly, Tables 5 and 6 present the mean and standard deviation of the dispersion estimates $\sqrt{\hat{\tau}_{00}}$ and $\sqrt{\hat{\tau}_{11}}$, respectively, as a function of sample size. Both the linear model and PQL showed decreased levels of negative bias as the cluster sizes increased. For the linear model, the effect of cluster size was most evident with the random slope. For the random intercept, ML typically produced negatively biased dispersion estimates, attenuating as the number of clusters increased. In contrast, the bias of the PQL estimates *increased* slightly with the number of clusters. For the random slope, ML performed well when the population random effect was medium or large, but showed some positive bias when the population random effect was small, particularly at the smallest sample sizes. As anticipated, PQL was again negatively biased and generally benefited from larger cluster sizes. PQL estimates generally exhibited less sampling variability than ML estimates, with ML estimates being particularly unstable for the combination of large random effects, few clusters, and small cluster sizes.

To contextualize these differences between the PQL and ML estimators, Figures 6 and 7 present (squared) bias, variance, and *MSE* for the $\sqrt{\hat{\tau}_{00}}$ and $\sqrt{\hat{\tau}_{11}}$ estimates in the same format as Figure 4. The results generally parallel the results presented previously for the fixed effects. Although the PQL random effect dispersion estimates are more biased, their sampling variance is also often smaller. PQL thus produces lower *MSE* values than ML in many conditions. A consistent and appreciable *MSE* advantage for ML is observed only when there are many clusters (e.g., 100 or 200) and medium to large random effects. Further, this advantage diminishes as the cluster size or number of categories increases.

Discussion

Summary. An initial question we sought to address was, "When can the results of a multilevel linear model fit to an ordinal outcome be trusted?" Our results suggest the answer, "Rarely." Only when the marginal distribution of the category responses was roughly normal and the number of categories was seven did the negative bias of the linear model decrease to the acceptable level of approximately 10% for the fixed effects. The dispersion estimates of the random effects were similarly negatively biased. In almost all cells of the design, the linear model estimates were inferior to the cumulative logit model estimates (from either PQL or ML). In contrast, neither PQL nor ML estimators of the multilevel cumulative logit model demonstrated much sensitivity to the category distribution. In sum, these results argue against the practice of fitting multilevel linear models to ordinal outcomes.[8]

The second major aim of this study was to evaluate the relative performance of two estimators of the multilevel cumulative logit model, PQL versus ML with adaptive quadrature. In general, our results suggest that PQL has been somewhat unfairly maligned. While we did indeed find that PQL estimates of fixed effects, and especially dispersion parameters, were negatively biased in many conditions, PQL nevertheless often outperformed ML in terms of *MSE*. In other words, the degree of excess bias associated with using PQL was often within tolerable levels and compensated for by lower sampling variability (similar to what Bellamy et al., 2005, found for binary outcomes). As shown in other studies, PQL

[8] Indeed, the linear model estimates were generally unacceptable despite the fact that data were generated under something of a best-case scenario. Because X_{ij} was simulated with an ICC of zero, misspecification of the nonlinear relation between Y_{ij} and X_{ij} could not spuriously inflate estimates for the random slope variance or cross-level interaction (Bauer & Cai, 2009). That is, the linear model would likely have performed even more poorly had X_{ij} been simulated with an appreciable ICC.

Table 3
Mean and Standard Deviation of Random Intercept Dispersion Estimate, $\sqrt{\hat{\tau}_{11}}$, as a Function of the Number of Categories, Collapsing Over Number of Clusters, Cluster Size, and Category Distribution

	Small random-effect variance (τ_{00} = 0.50)			Medium random-effect variance (τ_{00} = 1.63)			Large random-effect variance (τ_{00} = 8.15)		
	Population $\sqrt{\tau_{00}}$ = 0.71			Population $\sqrt{\tau_{00}}$ = 1.28			Population $\sqrt{\tau_{00}}$ = 2.85		
Categories	Linear–REML	Logistic–PQL	Logistic–ML	Linear–REML	Logistic–PQL	Logistic–ML	Linear–REML	Logistic–PQL	Logistic–ML
2	0.57 (0.16)	0.63 (0.19)	0.69 (0.23)	0.94 (0.16)	1.07 (0.20)	1.26 (0.28)	1.75 (0.26)	1.95 (0.32)	2.91 (0.97)
3	0.63 (0.16)	0.66 (0.17)	0.68 (0.19)	1.07 (0.17)	1.16 (0.19)	1.25 (0.22)	2.10 (0.27)	2.27 (0.33)	2.84 (0.46)
5	0.67 (0.16)	0.69 (0.15)	0.68 (0.16)	1.17 (0.17)	1.21 (0.18)	1.25 (0.19)	2.42 (0.27)	2.52 (0.32)	2.81 (0.39)
7	0.69 (0.15)	0.69 (0.15)	0.68 (0.15)	1.21 (0.18)	1.23 (0.17)	1.24 (0.19)	2.54 (0.31)	2.62 (0.32)	2.81 (0.36)

Note. Standard deviation of estimates presented in parentheses. REML = restricted maximum likelihood; PQL = penalized quasi-likelihood; ML = maximum likelihood.

performed best when the random effects were small and the cluster sizes were large. In addition, a new finding of our study is that the performance of PQL greatly improves with the number of categories for the outcome. The ML estimator also behaved as expected. Consistent with asymptotic theory, ML was least biased and most efficient for data with 100 or 200 clusters. With 25 or 50 clusters, however, ML estimates were more variable and often had higher *MSE* than PQL estimates.

A final finding worth noting is that all of the estimators generally perform better for ordinal than binary data. Furthermore, there is a sharp reduction in *MSE* associated with increasing the number of categories available for analysis, particularly in moving from two levels to three or more. These results indicate that ordinal scales are generally preferable to binary and underscore previous pleas for researchers to abandon the practice of dichotomizing ordinal scales (Sankey and Weissfeld, 1998; Strömberg, 1996).

Limitations and directions for future research. As with all simulation studies, the conclusions we draw from our results must be limited by the range of conditions we evaluated. We discuss these limitations here as potentially fruitful directions for future research. First, we studied only one model for ordinal outcomes, the cumulative logit model. We did not evaluate model performance with alternative link functions, such as the probit. Also, as mentioned previously, the cumulative logit model imposes an assumption of invariant slopes across categories (i.e., proportional odds), which is not always tenable in practice. A generalized logit or partial proportional odds model might then be preferable. For the interested reader, Hedeker and Gibbons (2006, pp. 191–194, 202–211) have provided a useful discussion of the proportional odds assumption, how to check this assumption empirically, and models that relax this assumption.

Second, we manipulated the shape of the ordinal outcome distributions while holding category sparseness constant. Although we regard it as a strength of our design that shape and sparseness were not confounded for ordinal outcomes, these two factors are inextricably confounded for binary outcomes. Our binary outcome results should be interpreted in light of this fact. Additionally, because we did not manipulate the sparseness of the ordinal outcomes, our results do not speak to the possible effects of sparseness on model estimates.

Third, our study was limited to multilevel models with random effects. A worthy topic of future research would be a comparison of the results of models fit by PQL or ML with the results obtained using GEE. Although unit-specific and population-average model estimates differ in scale and interpretation, marginalized estimates obtained from PQL or ML are comparable to the estimates obtained from GEE (Liang & Zeger, 1986).

Fourth, there are different approaches to implementing ML with numerical integration beyond adaptive quadrature (e.g. Laplace algorithms), different versions of adaptive quadrature (e.g. quadrature points iteratively updated based on mode versus mean of posterior), and different modifications of PQL in use (e.g. PQL2; Goldstein & Rabash, 1996). The generalization of these results

Table 4
Mean and Standard Deviation of the Random Slope Dispersion Estimate $\sqrt{\hat{\tau}_{11}}$ as a Function of the Number of Categories, Collapsing Over Number of Clusters, Cluster Size, and Category Distribution

	Small random effect variance (τ_{11} = 0.08)			Medium random effect variance (τ_{11} = 0.25)			Large random effect variance (τ_{11} = 1.25)		
	Population $\sqrt{\tau_{11}}$ = 0.28)			Population $\sqrt{\tau_{11}}$ = 0.50			Population $\sqrt{\tau_{11}}$ = 1.12		
Categories	Linear–REML	Logistic–PQL	Logistic–ML	Linear–REML	Logistic–PQL	Logistic–ML	Linear–REML	Logistic–PQL	Logistic–ML
2	0.23 (0.16)	0.21 (0.20)	0.31 (0.23)	0.35 (0.17)	0.32 (0.22)	0.49 (0.27)	0.68 (0.19)	0.64 (0.30)	1.14 (0.63)
3	0.23 (0.16)	0.24 (0.19)	0.28 (0.18)	0.38 (0.17)	0.39 (0.19)	0.47 (0.20)	0.79 (0.18)	0.83 (0.25)	1.09 (0.30)
5	0.25 (0.16)	0.26 (0.18)	0.27 (0.16)	0.42 (0.17)	0.45 (0.18)	0.47 (0.17)	0.91 (0.18)	0.97 (0.20)	1.08 (0.24)
7	0.25 (0.16)	0.27 (0.18)	0.27 (0.15)	0.44 (0.17)	0.48 (0.17)	0.47 (0.17)	0.95 (0.18)	1.02 (0.19)	1.08 (0.21)

Note. Standard deviation of estimates presented in parentheses. REML = restricted maximum likelihood; PQL = penalized quasi-likelihood; ML = maximum likelihood.

Table 5
Mean and Standard Deviation of the Random Intercept Dispersion Estimate $\sqrt{\hat{\tau}_{00}}$ as a Function of the Number of Clusters and Cluster Size, Collapsing Over Number of Categories and Category Distribution

Clusters/cluster size	Small random effect variance (τ_{00} = 0.50)			Medium random effect variance (τ_{00} = 1.63)			Large random effect variance (τ_{00} = 8.15)		
	Population $\sqrt{\tau_{00}}$ = 0.71			Population $\sqrt{\tau_{00}}$ = 1.28			Population $\sqrt{\tau_{00}}$ = 2.85		
	Linear–REML	Logistic–PQL	Logistic–ML	Linear–REML	Logistic–PQL	Logistic–ML	Linear–REML	Logistic–PQL	Logistic–ML
25 clusters									
5 members	0.63 (0.33)	0.68 (0.33)	0.66 (0.35)	1.11 (0.34)	1.17 (0.34)	1.22 (0.42)	2.28 (0.57)	2.30 (0.54)	2.92 (1.41)
10 members	0.64 (0.21)	0.68 (0.22)	0.65 (0.23)	1.11 (0.25)	1.20 (0.26)	1.21 (0.28)	2.26 (0.50)	2.45 (0.49)	2.80 (0.62)
20 members	0.64 (0.15)	0.68 (0.16)	0.66 (0.16)	1.11 (0.22)	1.22 (0.22)	1.21 (0.23)	2.24 (0.47)	2.56 (0.45)	2.76 (0.53)
50 clusters									
5 members	0.64 (0.23)	0.65 (0.23)	0.67 (0.25)	1.12 (0.24)	1.14 (0.24)	1.26 (0.28)	2.26 (0.44)	2.21 (0.42)	2.87 (0.57)
10 members	0.65 (0.14)	0.67 (0.15)	0.68 (0.16)	1.11 (0.19)	1.18 (0.18)	1.25 (0.20)	2.45 (0.40)	2.40 (0.38)	2.83 (0.42)
20 members	0.65 (0.11)	0.68 (0.11)	0.68 (0.11)	1.11 (0.17)	1.21 (0.15)	1.25 (0.16)	2.23 (0.39)	2.53 (0.34)	2.81 (0.36)
100 clusters									
5 members	0.65 (0.15)	0.65 (0.16)	0.69 (0.17)	1.11 (0.18)	1.12 (0.17)	1.26 (0.19)	2.24 (0.38)	2.16 (0.37)	2.85 (0.37)
10 members	0.65 (0.10)	0.67 (0.10)	0.69 (0.11)	1.11 (0.15)	1.17 (0.13)	1.26 (0.14)	2.23 (0.36)	2.37 (0.33)	2.83 (0.29)
20 members	0.65 (0.08)	0.68 (0.08)	0.70 (0.08)	1.11 (0.14)	1.21 (0.11)	1.26 (0.12)	2.23 (0.34)	2.52 (0.28)	2.83 (0.25)
200 clusters									
5 members	0.65 (0.11)	0.65 (0.11)	0.70 (0.12)	1.11 (0.14)	1.12 (0.13)	1.27 (0.13)	2.23 (0.34)	2.14 (0.34)	2.84 (0.26)
10 members	0.65 (0.08)	0.67 (0.07)	0.70 (0.08)	1.11 (0.13)	1.17 (0.10)	1.27 (0.10)	2.24 (0.33)	2.35 (0.30)	2.85 (0.20)
20 members	0.65 (0.07)	0.68 (0.05)	0.70 (0.05)	1.11 (0.12)	1.21 (0.08)	1.27 (0.08)	2.23 (0.32)	2.51 (0.25)	2.84 (0.18)

Note. Standard deviation of estimates presented in parentheses. REML = restricted maximum likelihood; PQL = penalized quasi-likelihood; ML = maximum likelihood.

across these other estimation algorithms cannot be fully guaranteed.

Recommendations. Notwithstanding the limitations noted, we believe that our results can be used to better inform the analysis of ordinal outcomes in nested data. As noted, our results clearly indicate that use of a linear model with ordinal outcomes should be avoided. With our selection of the multilevel cumulative logit model as more appropriate for ordinal outcomes, the central ques-

Table 6
Mean and Standard Deviation of the Random Slope Dispersion Estimate $\sqrt{\hat{\tau}}_{11}$ as a Function of the Number of Clusters and Cluster Size, Collapsing Over Number of Categories and Category Distribution

Clusters/cluster size	Small random effect variance (τ_{11} = 0.08)			Medium random effect variance (τ_{11}0.25)			Large random effect variance (τ_{11} = 1.25)		
	Population $\sqrt{\tau_{11}}$ = 0.28			Population $\sqrt{\tau_{11}}$ = 0.50			Population $\sqrt{\tau_{11}}$ = 1.12		
	Linear–REML	Logistic–PQL	Logistic–ML	Linear–REML	Logistic–PQL	Logistic–ML	Linear–REML	Logistic–PQL	Logistic–ML
25 clusters									
5 members	0.26 (0.28)	0.34 (0.33)	0.37 (0.31)	0.37 (0.32)	0.44 (0.35)	0.51 (0.37)	0.82 (0.39)	0.80 (0.42)	1.12 (0.94)
10 members	0.24 (0.20)	0.26 (0.23)	0.28 (0.20)	0.39 (0.22)	0.42 (0.25)	0.44 (0.24)	0.84 (0.24)	0.92 (0.30)	1.05 (0.36)
20 members	0.25 (0.15)	0.25 (0.17)	0.24 (0.14)	0.42 (0.14)	0.45 (0.18)	0.45 (0.18)	0.85 (0.19)	1.00 (0.23)	1.06 (0.26)
50 clusters									
5 members	0.22 (0.22)	0.26 (0.25)	0.32 (0.23)	0.36 (0.25)	0.38 (0.27)	0.48 (0.28)	0.83 (0.28)	0.74 (0.34)	1.09 (0.41)
10 members	0.24 (0.16)	0.24 (0.18)	0.27 (0.16)	0.40 (0.16)	0.41 (0.19)	0.45 (0.19)	0.86 (0.18)	0.91 (0.22)	1.10 (0.24)
20 members	0.26 (0.11)	0.24 (0.13)	0.25 (0.12)	0.43 (0.10)	0.45 (0.12)	0.47 (0.13)	0.86 (0.15)	1.00 (0.16)	1.09 (0.17)
100 clusters									
5 members	0.21 (0.18)	0.23 (0.20)	0.29 (0.19)	0.36 (0.20)	0.35 (0.22)	0.47 (0.23)	0.83 (0.21)	0.73 (0.29)	1.10 (0.28)
10 members	0.24 (0.13)	0.23 (0.15)	0.26 (0.14)	0.42 (0.11)	0.42 (0.14)	0.48 (0.14)	0.85 (0.15)	0.90 (0.18)	1.10 (0.17)
20 members	0.27 (0.08)	0.25 (0.10)	0.26 (0.09)	0.43 (0.07)	0.46 (0.08)	0.49 (0.09)	0.86 (0.13)	1.00 (0.13)	1.11 (0.12)
200 clusters									
5 members	0.21 (0.15)	0.21 (0.17)	0.28 (0.16)	0.37 (0.15)	0.35 (0.19)	0.47 (0.17)	0.84 (0.15)	0.73 (0.26)	1.11 (0.19)
10 members	0.25 (0.10)	0.23 (0.12)	0.26 (0.11)	0.42 (0.08)	0.42 (0.11)	0.49 (0.10)	0.86 (0.13)	0.90 (0.15)	1.11 (0.12)
20 members	0.28 (0.06)	0.26 (0.07)	0.27 (0.07)	0.44 (0.06)	0.46 (0.06)	0.49 (0.06)	0.86 (0.12)	0.99 (0.11)	1.11 (0.09)

Note. Standard deviation of estimates presented in parentheses. REML = restricted maximum likelihood; PQL = penalized quasi-likelihood; ML = maximum likelihood.

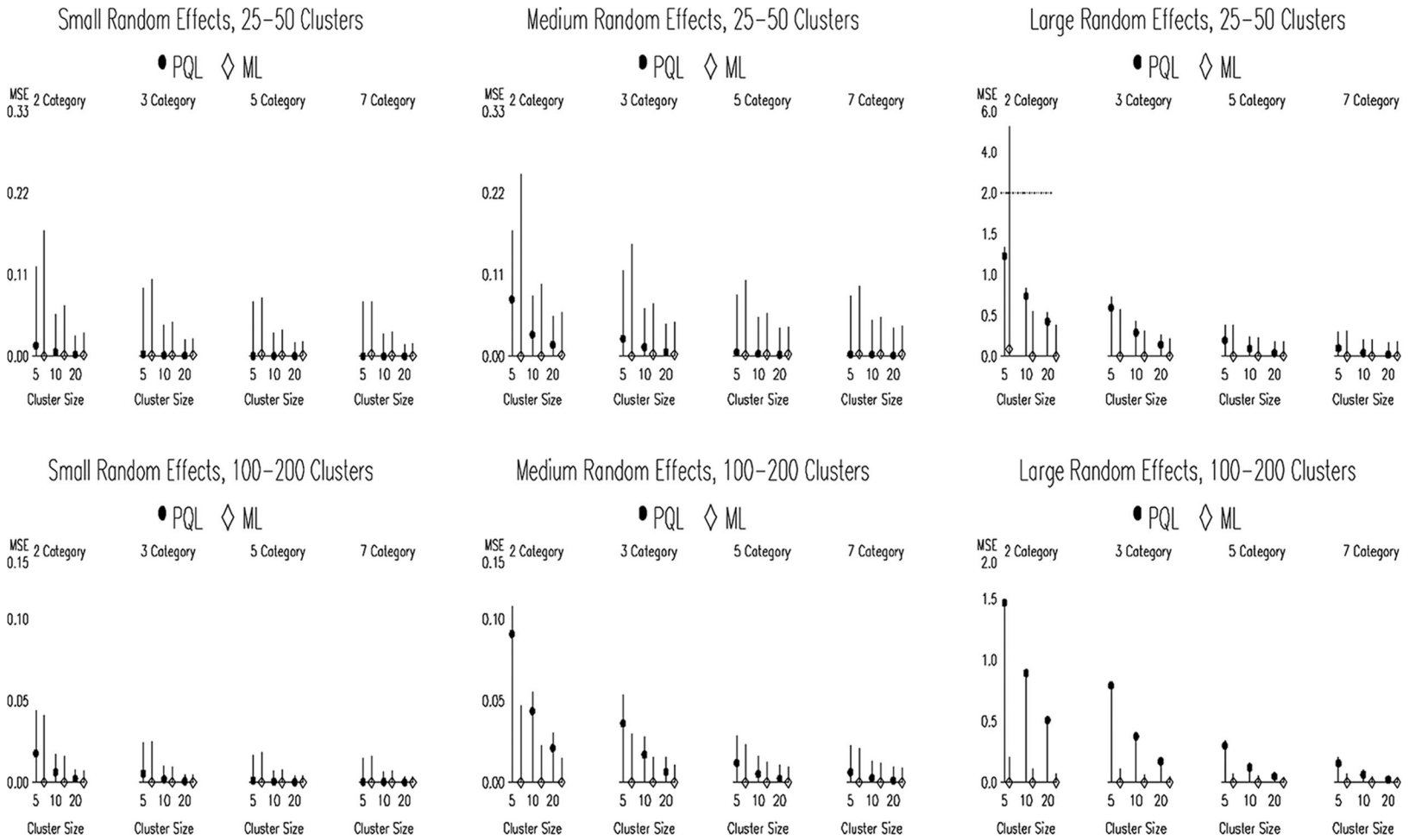

Figure 6. Mean-squared error for the standard deviation of the random intercept, across number of outcome categories, number of clusters, and cluster size. The scale differs across panels and is discontinuous in the upper right panel. See Figure 4 caption for definition of quantities in this plot. PQL denotes penalized quasi-likelihood, and ML denotes maximum likelihood with adaptive quadrature.

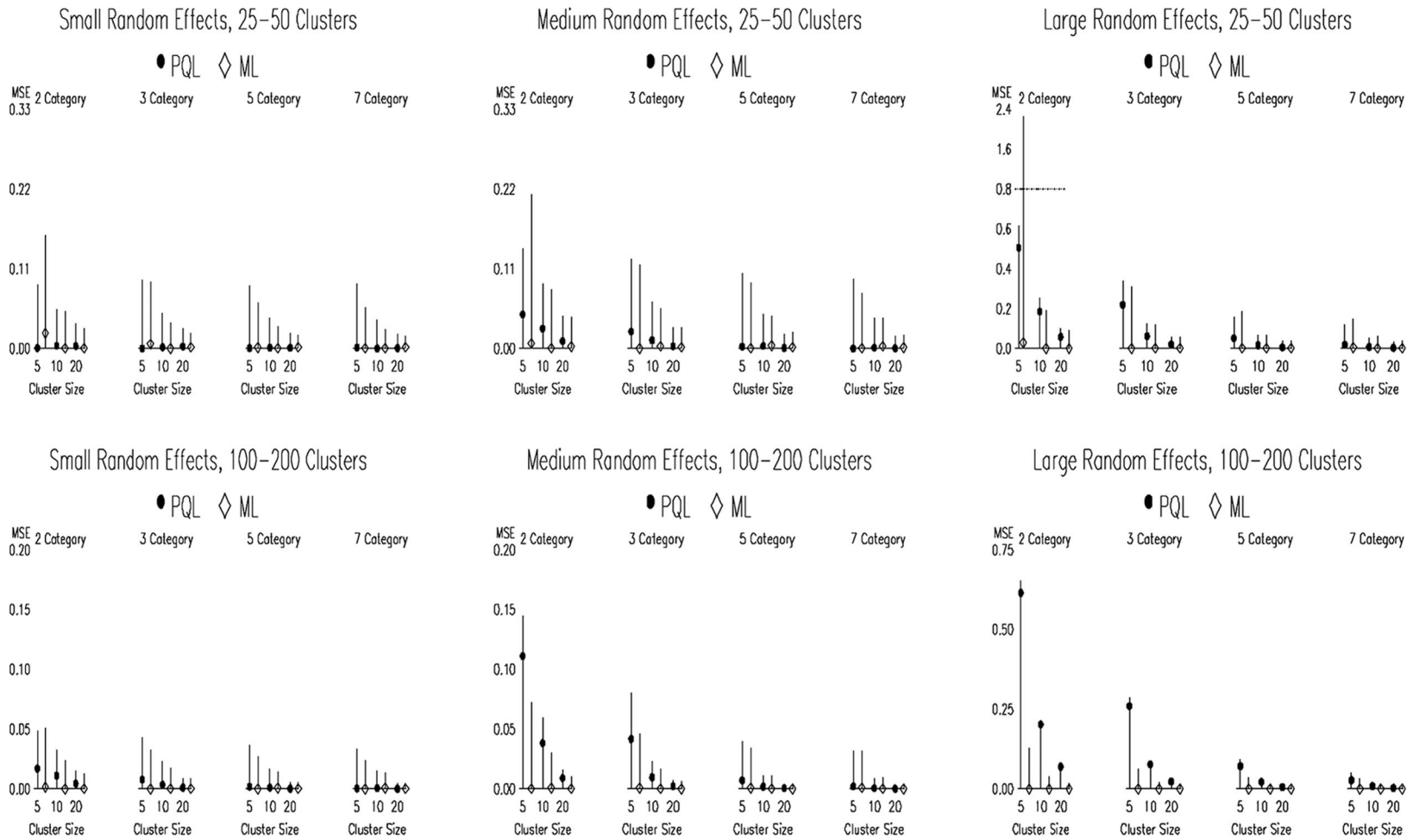

Figure 7. Mean-squared error for the standard deviation of the random slope, across number of outcome categories, number of clusters, and cluster size. The scale differs across panels and is discontinuous in the upper right panel. See Figure 4 caption for definition of quantities in this plot. PQL denotes penalized quasi-likelihood, and ML denotes maximum likelihood with adaptive quadrature.

Table 7
Working Recommendations for Estimator Choice When Fitting Multilevel Cumulative Logit Models in Practice

Variable	Small variance components (e.g. ICC = .13)		Large variance components (e.g. ICC = .71)	
	DV has many categories (5+)	DV has few categories (2)	DV has many categories (5+)	DV has few categories (2)
Many clusters (e.g. 100–200)				
Large clusters (e.g., 20)	PQL, ML–AQ	PQL, ML–AQ	PQL, ML–AQ	ML–AQ
Small clusters (e.g., 5)	PQL, ML–AQ	PQL, ML–AQ	ML–AQ	ML–AQ
Few clusters (e.g. 25–50)				
Large clusters (e.g., 20)	PQL, ML–AQ	PQL, ML–AQ	PQL, ML–AQ	PQL, ML–AQ
Small clusters (e.g., 5)	PQL, ML–AQ	PQL	PQL, ML–AQ	

Note. ICC = intraclass correlation; DV = dependent variable; PQL = penalized quasi-likelihood; ML–AQ = maximum likelihood with adaptive quadrature. Situations under which ML with adaptive quadrature or PQL perform similarly are denoted "PQL, ML–AQ." The selection of an estimator in these conditions should depend on the investigator's focus (fixed effects vs. variance components) or other factors (such as computational speed or the desire to perform nested model comparisons). Situations under which PQL performs consistently better are denoted "PQL" and situations under which ML–AQ performs consistently better are denoted "ML–AQ." Even in these conditions, however, the magnitude of performance differences is not always large. Consult Figures 2–7 and Tables 3–6 for more detailed information on estimator differences.

tion then is which estimator is to be preferred, PQL or ML with adaptive quadrature?

The answer to this question depends not only on the bias and sampling variability of the estimates, but also on other factors. For instance, one issue that must be considered when choosing between PQL and ML is whether one wishes to evaluate the relative fit of competing models. Because PQL uses a quasi-likelihood, rather than a true likelihood, it does not produce a deviance statistic that can be used for model selection (e.g., by likelihood ratio test or penalized information criteria). This is a significant limitation of PQL that is not shared by ML. If comparison of competing models is a key goal of the analysis, then ML may be preferred to PQL on these grounds alone. Another factor that might influence estimator selection is computational efficiency. PQL is much faster, particularly when the number of random effects (dimensions of integration) is large. Finally, a third factor related to estimator selection is model complexity. Some models may only be feasible with one estimator or the other. For instance, PQL readily allows the incorporation of serial correlation structures for the Level-1 residuals.

Beyond these factors, our simulation results suggest that the preferred choice between PQL and ML depends on the characteristics of the data. If data are obtained on 100 or more clusters, cluster sizes are small, dispersion across clusters is anticipated to be moderate to large, and the outcome variable has only two or three categories, then ML is the best choice. Under virtually all other conditions, however, PQL is a viable, often superior alternative. In particular, if data are available on 50 clusters or fewer, PQL will generally have lower *MSE*—even with just two- or three-category outcomes. The bias of the PQL estimates is also tolerable when either cluster sizes are large or outcomes have five or more categories.

Table 7 translates our results for PQL and ML into a table of working recommendations for fitting multilevel cumulative logit models (primarily based on *MSE* but also considering bias). These are gross recommendations, and we encourage researchers to consider the more detailed results of our simulation before making a final selection. Situations under which ML with adaptive quadrature (AQ) or PQL perform similarly (and thus either could be chosen) are denoted with the table entry "PQL, ML–AQ." Situations under which PQL is preferable are denoted "PQL" and situations where ML–AQ is clearly preferable are denoted "ML–AQ." Note that the cell of Table 7 corresponding to few clusters, small cluster size, binary outcomes, and large random effects is empty because the performance of both estimators was unacceptable (PQL showed excessive bias, whereas ML showed excessive sampling variability). For this situation, researchers will need to look outside the two estimators studied here (e.g., Markov chain Monte Carlo might perform better through the implementation of mildly informative priors that prevent estimates from becoming excessively large).

To see how Table 7 might be used in practice, we will consider two common situations. First, many samples of hierarchical data consist of a relatively small number of groups but a fairly large number of individuals in each group. For instance, a study might sample 30 students from each of 30 schools. In this instance, the variance components are likely to be on the smaller side, and PQL can be expected to perform as well or better than ML regardless of the number of categories of the outcome. Second, many experience sampling studies include a modest number of participants, say 25–50, but many repeated measures per person. Experience suggests that variance components are often sizeable in such studies. If our outcome is binary, we might choose ML due to the higher bias of PQL (despite similar *MSE*). Alternatively, if our outcome is a five-level ordinal variable then PQL becomes a more attractive option: the bias of PQL will then be within tolerable levels, and PQL will have lower *MSE* than ML. One additional factor that might tip the balance in favor of PQL is that PQL easily incorporates serial correlation structures for the residuals at Level 1, and serial correlation is often present with experience sampling data.

In conclusion, although further research on the estimation of multilevel models with ordinal data is warranted, it is our hope that the results of the present study can help analysts to make better-informed choices when fitting multilevel models to ordinal outcomes.

References

Agresti, A., Booth, J. G., Hobert, J. P., & Caffo, B. (2000). Random-effects modeling of categorical response data. *Sociological Methodology, 30,* 27–80. doi:10.1111/0081-1750.t01–1-00075

Bakeman, R. (2005). Recommended effect size statistics for repeated

measures designs. *Behavior Research Methods, 37,* 379–384. doi: 10.3758/BF03192707

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics, 28,* 135–167. doi:10.3102/10769986028002135

Bauer, D. J. (2009). A note on comparing the estimates of models for cluster-correlated or longitudinal data with binary or ordinal outcomes. *Psychometrika, 74,* 97–105. doi:10.1007/s11336-008-9080-1

Bauer, D. J., & Cai, L. (2009). Consequences of unmodeled nonlinear effects in multilevel models. *Journal of Educational and Behavioral Statistics, 34,* 97–114. doi:10.3102/1076998607310504

Bellamy, S. L., Li, Y., Lin, X., & Ryan, L. M. (2005). Quantifying PQL bias in estimating cluster-level covariate effects in generalized linear mixed models for group-randomized trials. *Statistica Sinica, 15,* 1015–1032.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York, NY: Wiley.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88,* 9–25. doi:10.2307/2290687

Breslow, N. E., & Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika, 82,* 81–91. doi:10.1093/biomet/82.1.81

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine, 19,* 3127–3131. doi:10.1002/1097-0258(20001130)19:22<3127::AID-SIM784>3.0.CO;2-M

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research, 38,* 529–569. doi:10.1207/s15327906mbr3804_5

Demidenko, E. (2004). *Mixed models: Theory and applications*. Hoboken, NJ: Wiley. doi:10.1002/0471728438

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.; Kendall's Library of Statistics No. 3). London, England: Arnold.

Goldstein, H., & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A: Statistics in Society, 159,* 505–513. doi:10.2307/2983328

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis.* New York, NY: Wiley.

Hox, J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

Kaplan, D. (1989). A study of the sampling variability and *z* values of parameter estimates from misspecified structural equation models. *Multivariate Behavioral Research, 24,* 41–57. doi:10.1207/s15327906mbr2401_3

Kendall, M. G., & Stuart, A. (1969). *The advanced theory of statistics.* London, England: Griffin.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73,* 13–22. doi:10.1093/biomet/73.1.13

Liu, I., & Agresti, A. (2005). The analysis of ordered categorical data: An overview and a survey of recent developments. *Test, 14,* 1–73. doi: 10.1007/BF02595397

Long, J. S. (1997). *Regression models for categorical and limited dependent variables.* Thousand Oaks, CA: Sage.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models.* New York, NY: Chapman & Hall.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association, 92,* 162–170. doi:10.2307/2291460

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, linear, and mixed models.* New York, NY: Wiley.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods, 10,* 259–284. doi:10.1037/1082-989X.10.3.259

Muthén, L. K., & Muthén, B. O. (1998–2007). *Mplus statistical analysis with latent variables: User's guide*. Los Angeles, CA: Authors.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8,* 434–447. doi:10.1037/1082-989X.8.4.434

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2002). Reliable estimation of generalized linear models using adaptive quadrature. *The Stata Journal, 2,* 1–21.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Raudenbush, S. W., Yang, M.-L., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics, 9,* 141–157. doi:10.2307/1390617

Rodriguez, G., & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A: Statistics in Society, 158,* 73–89. doi:10.2307/2983404

Rodriguez, G., & Goldman, N. (2001). Improved estimation procedures for multilevel models with binary response: A case study. *Journal of the Royal Statistical Society, Series A: Statistics in Society, 164,* 339–355. doi:10.1111/1467-985X.00206

Saei, A., & McGilchrist, C. A. (1998). Longitudinal threshold models with random components. *The Statistician, 47,* 365–375. doi:10.1111/1467-9884.00137

Sankey, S. S., & Weissfeld, L. A. (1998). A study of the effect of dichotomizing ordinal data upon modeling. *Communications in Statistics—Simulation and Computation, 27,* 871–887. doi:10.1080/03610919808813515

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models.* New York, NY: Chapman & Hall/CRC. doi:10.1201/9780203489437

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London, England: Sage.

Strömberg, U. (1996). Collapsing ordered outcome categories: A note of concern. *American Journal of Epidemiology, 144,* 421–424.

Ten Have, T. R., & Localio, A. R. (1999). Empirical Bayes estimation of random effects parameters in mixed effects logistic regression models. *Biometrics, 55,* 1022–1029. doi:10.1111/j.0006-341X.1999.01022.x

Winship, C., & Mare, R. D. (1984). Regression models with ordinal variables. *American Sociological Review, 49,* 512–525. doi:10.2307/2095465

Yosef, M. (2001). *A comparison of alternative approximations to maximum likelihood estimation for hierarchical generalized linear models: The logistic-normal model case* (Unpublished doctoral dissertation). Michigan State University, East Lansing,

Received September 1, 2009
Revision received April 26, 2011
Accepted July 21, 2011 ■