Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

British Journal of Mathematical and Statistical Psychology (2006), 59, 225–255
© 2006 The British Psychological Society



The British Psychological Society

www.bpsjournals.co.uk

Statistical inference in generalized linear mixed models: A review

Francis Tuerlinckx[1]*, Frank Rijmen[1], Geert Verbeke[2] and Paul De Boeck[1]

[1]Department of Psychology, University of Leuven, Belgium
[2]Biostatistical Centre, University of Leuven, Belgium

We present a review of statistical inference in generalized linear mixed models (GLMMs). GLMMs are an extension of generalized linear models and are suitable for the analysis of non-normal data with a clustered structure. A GLMM contains parameters common to all clusters (fixed regression effects and variance components) and cluster-specific parameters. The latter parameters are assumed to be randomly drawn from a population distribution. The parameters of this population distribution (the variance components) have to be estimated together with the fixed effects. We focus on the case in which the cluster-specific parameters are normally distributed. The cluster-specific effects are integrated out of the likelihood so that the fixed effects and variance components can be estimated. Unfortunately, the integral over the cluster-specific effects is intractable for most GLMMs with a normal mixing distribution. Within a classical statistical framework, we distinguish between two broad classes of methods to handle this intractable integral: methods that rely on a numerical approximation to the integral and methods that use an analytical approximation to the integrand. Finally, we present an overview of available methods for testing hypotheses about the parameters of GLMMs.

1. Introduction

In many studies in the social sciences, researchers are faced with data having a clustered structure, thereby violating the usual independence assumption underlying familiar off-the-shelf statistical methods such as multiple linear regression and simple analysis-of-variance models. Clustered data arise, for example, in educational measurement applications when several test items are administered to pupils; in longitudinal studies of personality where the occurrence of an emotion is repeatedly assessed over time for a sample of persons; in survey research where the political preference of household members is questioned; or in experimental psychology where the experimental design contains within-subject variables. In these and other examples, we can distinguish

* *Correspondence should be addressed to Francis Tuerlinckx, Department of Psychology, University of Leuven, Tiensestraat 102, B-3000 Leuven, Belgium (e-mail: francis.tuerlinckx@psy.kuleuven.be).*

DOI:10.1348/000711005X79857

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

at least two levels in the data: *measurements* or level 1 units (e.g. items, time points, household members) nested within *clusters* or level 2 units (e.g. persons, households). Henceforth, we will use the terms 'measurement' and 'cluster' to refer to level 1 and level 2 units, respectively. This framework is readily extended to data structures with more than two hierarchical levels of nesting (see Goldstein, 2003; Snijders & Bosker, 1999), but for simplicity we will consider in this paper only two-level data.

The measurements within a cluster are likely to be dependent because they tend to be more homogeneous than measurements from different clusters, giving rise to within-cluster dependence. By the same token, there is between-cluster heterogeneity because clusters tend to differ systematically from each other. Thus, both within-cluster dependence and between-cluster heterogeneity emanate from the clustered data structure (Collett, 1991).

How should one analyse such clustered data? On the one hand, one could perform an analysis that ignores the between-cluster heterogeneity or within-cluster dependence and treats all measurements as independent. As a consequence, the model would contain only parameters common to all clusters from which general effects of predictors can be derived. A major shortcoming of this approach is that parameter estimates will probably be biased and that the uncertainty measures of the common parameter estimates will presumably be incorrect because the observations belonging to the same cluster do not contribute independent bits of information, contrary to the model assumption. Moreover, the model does not provide information about the extent of between-cluster heterogeneity.

On the other side of the spectrum lies the possibility of performing a separate analysis for each cluster. By definition, all parameters are now cluster-specific and none of the parameters are common to the clusters. Such an analysis takes into account the between-cluster heterogeneity but scientific questions of interest concerning common effects of variables cannot be addressed. Moreover, for clusters with only a few measurements, some parameter estimates may have large standard errors or it may even be impossible to estimate these parameters.

As a compromise between these two extreme positions, the data analyst could build a statistical model for the clustered data containing both ordinary regression parameters common to all clusters and cluster-specific parameters. Our main focus is on so-called *mixed models*, which assume that the cluster-specific effects are a random sample from a population distribution. Mixed models cope in a natural way with between-cluster heterogeneity. They provide common parameter estimates with adequate levels of uncertainty and different cluster sizes are no longer a problem. Other names for mixed models are *multi-level models* (Goldstein, 2003), *hierarchical models* (Raudenbush & Bryk, 2002) and *random coefficient models* (Longford, 1993). The cluster-specific parameters in mixed models are also called *random effects* or, as in the literature on item response theory (IRT), *latent traits* or *latent variables*. For a thorough introduction to mixed models, we refer the reader to Davidian and Giltinan (1995), Diggle, Heagerty, Liang, and Zeger (2002), Goldstein (2003), Fahrmeir and Tutz (2001), Longford (1993), McCulloch and Searle (2001) and Verbeke and Molenberghs (2000).

The aim of this paper is to give non-specialists an overview of issues of statistical inference for a special but important class of mixed models for non-normal data: *generalized linear mixed models* (GLMMs). Our motivation for this is twofold. First, many measurements in the social sciences are categorical (binary, ordered, discrete choice data), and so unsuitable for linear modelling. For non-normal data without clustering, generalized linear models (GLMs: McCullagh & Nelder, 1989) are an

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

appropriate alternative to linear models. For clustered data, the GLM can be extended easily to mixed models, leading to a GLMM. Recently, GLMMs have become important tools for analysing clustered data. Many models from IRT (Legler & Ryan, 1997; Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003) and multi-level models for non-normal data (Snijders & Bosker, 1999) are special cases of GLMMs.

The second motivation is that the research on statistical inference for these models is well under way and has spawned a large number of methods and procedures, all coming with specific advantages and disadvantages. But much of the literature on these methods is scattered among specialized biometrical and statistical journals and often not easily accessible to social scientists. Moreover, software packages have implemented different methods or slightly different variants of similar methods, and they explain only very briefly the characteristics of the algorithms. Therefore, there is a need for an integrated review of available statistical inference methods in GLMMs for social scientists.[1]

Linear mixed models are a special case of GLMMs, but our attention will not be directed towards these models. Statistical inference for linear mixed models is well developed and the references quoted above are for the most part devoted to the linear case (see also Raudenbush & Bryk, 2002; Kreft & De Leeuw, 1998; Searle, Casella, & McCulloch, 1992).

The remainder of the paper is organized as follows. Section 2 introduces GLMs with cluster-specific parameters and the GLMM. Most of the rest of the paper is then devoted to the estimation of the parameters in a GLMM. A separate section deals with hypothesis and goodness-of-fit testing. Subsequently, we also discuss briefly alternative approaches to handling clustered data.

2. Generalized linear models with cluster-specific effects

In this section, we start by introducing the GLM and then move on to the variant with cluster-specific effects. With cluster-specific effects, several perspectives are possible, and this has consequences on the inferential methods. Three different approaches will be discussed, one of which leads to the GLMM.

2.1. The model

Let y_{ni} denote the ith measurement on cluster n, where $i = 1, \ldots, I_n$ and $n = 1, \ldots, N$. For example, y_{ni} can stand for the binary scored response of person n on item i of a cognitive test: $y_{ni} = 1$ if the response is correct and $y_{ni} = 0$ otherwise. The number of measurements can differ over clusters, but for simplicity we will deal only with the case where $I_n = I$, for all n. However, all results presented are easily extended to the more general case.

Suppose first that there is only one person ($N = 1$) in the sample, so that we can drop the index n. In a GLM, given the predictors, the I measurements are independent realizations from an exponential family distribution. A density $f(y_i)$ belongs to the exponential family if it can be expressed as:

$$f(y_i) = f(y_i|\omega_i) = \exp\{y_i\omega_i - b(\omega_i) + e(y_i)\}, \quad (1)$$

[1] *In 2004, after the submission of this paper, Skrondal and Rabe-Hesketh (2004) also provided an extensive overview of inferential methods in GLMMs.*

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

where ω_i is called the *natural parameter*, $b(\omega_i)$ is a function whose form depends on the specific distribution and $e(y_i)$ is a function of the data that does not depend on the natural parameter of the model. Typical examples of distributions of the exponential family are the normal, Bernoulli, binomial, Poisson and gamma distributions.[2]

The Bernoulli distribution is a member of the exponential family:

$$\pi_i^{y_i}(1-\pi_i)^{1-y_i} = \exp\left\{y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i)\right\}$$
$$= \exp\{y_i\omega_i - \log(1+e^{\omega_i})\}, \quad (2)$$

so that $\omega_i = \log(\frac{\pi_i}{1-\pi_i})$, $b(\omega_i) = \log(1+e^{\omega_i})$, and $e(y_i) = 0$.

It can be shown (McCullagh & Nelder, 1989) that the mean of an exponential family distribution equals $E(y_i) = \mu_i = \frac{db(\omega_i)}{d\omega_i} = b'(\omega_i)$. The variance of an observation y_i is denoted as $v(\mu_i)$ to acknowledge that it may depend on the mean. It is derived as follows: $\frac{d^2b(\omega_i)}{d\omega_i^2} = v(\mu_i)$. It is straightforward to check that these properties hold for the Bernoulli model in our example.

In the most common GLMs (and the only ones we consider here), the natural parameter ω_i is a linear function of predictors and parameters (called the *linear predictor* η_i):

$$\omega_i = b'^{-1}(\mu_i) = g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is a P-dimensional vector of regression coefficients and $\mathbf{x}_i$ is the P-dimensional design vector for the ith observation. The first component of $\mathbf{x}_i$, x_{i1}, is usually equal to 1, representing the intercept. The function $g(\cdot)$ is called the (canonical or natural) *link function* because it relates the mean of the distribution through a transformation (the particular transformation being dependent on the distribution) to the predictors. The inverse of $g(\cdot)$, equal to $h(\cdot) = b'(\cdot)$, is called the *response function* and maps the linear predictor $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$ onto the mean of the observations. In our Bernoulli example, the mean $\mu_i = \pi_i$ is transformed into the natural parameter ω_i, which is then regressed linearly on the predictors. Thus, the logit transform is the link function. Then it follows that the probability of observing 1 equals:

$$\Pr(y_i = 1|\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i^T\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^T\boldsymbol{\beta})}, \quad (3)$$

and this is the equation for the logistic regression model. The normal linear model is also a GLM in which the data have a normal distribution for the observations and the link function is the identity function so that the mean is directly regressed on the predictors.

If there is more than one cluster in the sample ($N > 1$), we need to take the within-cluster dependence or between-cluster differences into account, an aspect that is lacking in the standard GLM. Therefore, we will introduce a Q-dimensional vector of cluster-specific parameters, $\boldsymbol{\theta}_n = (\theta_{n1}, \ldots, \theta_{nQ})$. This vector is indexed by n, indicating that it is associated with cluster n. A set of predictors corresponding to these effects is denoted by $\mathbf{z}_{ni}$. In a GLM with cluster-specific variables, the

[2] *Often there is also a dispersion parameter φ present in the exponential model formulation (Fahrmeir & Tutz, 2001; McCullagh & Nelder, 1989). The dispersion parameter is crucial in normal models where it is the variance. However, because normal models are not the focus of this paper, we will not include the dispersion parameter in our equations.*

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

transformed mean of a single observation, $\mu_{ni} = E(y_{ni})$, is regressed on the predictors as follows:

$$g(\mu_{ni}) = \mathbf{x}_{ni}^T\boldsymbol{\beta} + \mathbf{z}_{ni}^T\boldsymbol{\theta}_n. \tag{4}$$

The simplest extension of the logistic regression model to a model with cluster-specific parameters includes a separate intercept θ_n for each cluster. In that case, $\mathbf{z}_{ni} = 1$ for all n and i and the probability of observing 1 is equal to:

$$\Pr(y_{ni} = 1|\boldsymbol{\beta}, \theta_n) = \frac{\exp(\theta_n + \mathbf{x}_{ni}^T\boldsymbol{\beta})}{1 + \exp(\theta_n + \mathbf{x}_{ni}^T\boldsymbol{\beta})}.$$

In the remainder of the paper, we will denote the probability or density of a single observation, conditional upon time parameters in the model, by $p(y_{ni}|\boldsymbol{\beta}, \boldsymbol{\theta}_n)$. The joint probability or density of the response pattern for cluster n is expressed as $f(\mathbf{y}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n) = \Pi_{i=1}^I p(y_{ni}|\boldsymbol{\beta}, \boldsymbol{\theta}_n)$. The latter equality follows from the conditional or local stochastic independence assumption, which means that, given the cluster-specific effects, the measurements are independent.

2.2. *Status of the cluster-specific effects*

Cluster-specific effects can be regarded in three ways. First, they can be considered as fixed effects parameters. In such a model, the parameter vector with cluster-specific effects, $\boldsymbol{\theta}_n$, has the same status as the fixed regression coefficients $\boldsymbol{\beta}$. Parameter estimates are found by maximizing the joint likelihood:

$$L_{\text{JML}} = \prod_{n=1}^{N} f(\mathbf{y}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n), \tag{5}$$

hence the term *joint maximum likelihood* (JML) estimation. However, these maximum likelihood estimators obtained from JML for the fixed effects parameters $\boldsymbol{\beta}$ are not consistent (Neyman & Scott, 1948) because the number of parameters grows at the same rate as the sample size N (each new observation brings along a new set of parameters). Inconsistency of the estimators of $\boldsymbol{\beta}$ is unwanted, certainly if one is specifically interested in the common parameters and not in the cluster-specific parameters that are seen as nuisance or incidental parameters. Moreover, as shown by Neyman and Scott, the fixed effects estimators in a JML framework are not guaranteed to be asymptotically the most efficient ones. Two alternatives are possible to get rid of the nuisance parameters: conditioning or marginalizing.

In the conditional inference approach, the cluster-specific parameters disappear from the likelihood that is optimized by conditioning on their sufficient statistics. It can be shown that in a GLM with the natural link function, sufficient statistics exist for all parameters (Andersen, 1980; McCullagh & Nelder, 1989). The vector of sufficient statistics for $\boldsymbol{\theta}_n$ is defined as $\mathbf{s}_n = \sum_{i=1}^I \mathbf{z}_n y_{ni}$ and let $r(\mathbf{s}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n)$ be the density of the sufficient statistics $\mathbf{s}_n$. Then the likelihood L_{JML} from equation (5) can be factorized as follows:

$$L_{\text{JML}} = \prod_{n=1}^{N} f(\mathbf{y}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n) = \prod_{n=1}^{N} f(\mathbf{y}_n|\boldsymbol{\beta}, \mathbf{s}_n) r(\mathbf{s}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n) = L_{\text{CML}} \prod_{n=1}^{N} r(\mathbf{s}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n).$$

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

When doing conditional inference, the conditional likelihood L_{CML} is maximized (hence *conditional maximum likelihood* or CML). Through conditioning on the sufficient statistics, L_{CML} is free of the cluster-specific parameters so that the CML estimators are consistent and asymptotically normally distributed (Andersen, 1970). On the other hand, the distribution of the sufficient statistics may contain information about the fixed effects parameters, which is discarded when maximizing L_{CML}. Except for some cases (see Andersen, 1970), the CML estimators are not efficient.

Besides this, CML has two other disadvantages. First, one may condition out some of the fixed effects when conditioning on sufficient statistics for the cluster-specific effects (Diggle *et al.*, 2002; Verbeke, Spiessens, & Lesaffre, 2001). Suppose a cluster-specific intercept is conditioned out of model. Then all fixed effects parameters of predictors that are constant within clusters (background variables applying to all measurements of the cluster, such as gender or educational level) are also conditioned out and cannot be estimated. The reason is that such predictors provide information about between-cluster comparisons but, by the conditioning operation, we restrict ourselves to within-cluster comparisons only. Second, some clusters do not contribute to the conditional likelihood because they result in values for the sufficient statistics that are uniquely related to their pattern of measurements. Thus, conditioning on the sufficient statistics gives a probability of 1 of observing the pattern of measurements and this has no influence on the likelihood L_{CML}. For example, persons who fail or who solve all items in a cognitive test do not contribute to the conditional likelihood for an IRT model.

In the third approach, the nuisance parameters are considered as a random sample from a population. The GLM with cluster-specific effects is thus supplemented with an assumption about their population distribution, leading to a mixed model. The cluster-specific effects are called *random effects*. The likelihood for this model is called the *marginal likelihood* and the estimation procedure is *marginal maximum likelihood*.

Three types of mixed models can be distinguished, depending on the assumptions made about the unobserved mixing distribution. The most general approach is to leave the mixing distribution entirely unspecified. In such a case, the estimated distribution function will be a step function with a finite number of steps (Laird, 1978). This method is called *non-parametric maximum likelihood estimation* (NPMLE) or fully semi-parametric estimation (Heinen, 1996). In the context of IRT models, De Leeuw and Verhelst (1986) and Lindsay, Clogg, and Grego (1991) have studied when the CML estimator and the NPMLE are equivalent. A more restricted model assumes that the locations of the steps (mass points) of the non-parametric distribution are known, and only its masses have to be estimated. This method is called *semi-parametric estimation* by Heinen. Finally, it can be assumed that the cluster-specific effects come from a (continuous) parametric distribution, with one or a few free parameters that have to be estimated.

In the remainder of the paper, we focus on the GLMM with a (multivariate) normal mixing distribution for the random effects. This is the model that is most often applied in practice. In this model, the marginal probability of a response pattern $\mathbf{y}_n$ is obtained as:

$$p(\mathbf{y}_n|\boldsymbol{\beta}) = \int_{-\infty}^{+\infty} \prod_{i=1}^{I} p(y_{ni}|\boldsymbol{\beta}, \boldsymbol{\theta}_n)\phi(\boldsymbol{\theta}_n|\mathbf{0}, \boldsymbol{\Sigma})\mathrm{d}\boldsymbol{\theta}_n = \int_{-\infty}^{+\infty} f(\mathbf{y}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n)\phi(\boldsymbol{\theta}_n|\mathbf{0}, \boldsymbol{\Sigma})\mathrm{d}\boldsymbol{\theta}_n \quad (6)$$

where $\phi(\boldsymbol{\theta}_n|\mathbf{0}, \boldsymbol{\Sigma})$ is the multivariate normal distribution of dimension Q with mean vector $\mathbf{0}$ (of length Q) and covariance matrix $\boldsymbol{\Sigma}$ (dimension $Q \times Q$).

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

The normal distribution has mean vector **0** because the means can always be considered as fixed effects and can be made part of the fixed effects part without changing the model. The limits of integration will henceforth be dropped.

Assuming a normal mixing distribution for the random effects extends the model with the assumption that the cluster-specific effects are a random sample from a normal distribution. It has been shown that the maximum likelihood estimators are relatively robust against the misspecification of the population distribution (Neuhaus, Hauck, & Kalbfleisch, 1992). Other model misspecifications may have larger consequences. For instance, Heagerty and Zeger (2000) show that the assumption of a constant variance–covariance matrix $\boldsymbol{\Sigma}$ for all level 2 units of the population can lead to serious biases in the regression coefficients for predictors that vary only between clusters (see also Daniels & Zhao, 2003). Extensions to accommodate model misspecifications of this type are not considered in this paper.

For some GLMMs the integral in equation (6) has a closed-form solution due to the particular combination of a probability distribution and a distribution for the random effects (i.e. the random effects distribution is conjugate to the probability distribution of the data). Examples are the obvious linear mixed model (Verbeke & Molenberghs, 2000), the beta-binomial model (Crowder, 1978; Kleinman, 1973; Skellam, 1948; Williams, 1975, 1982), the model proposed by Conaway (1990) and the gamma-Poisson model (e.g. Jansen, 1994). However, for the most common model, the logistic-normal model, there is no analytical solution.

3. Fitting a GLMM with a normal mixing distribution

The estimation of the parameters in a GLMM with a normal mixing distribution is not a trivial task and several methods have been proposed in the literature. In this section, we review the most common methods, discuss their advantages and disadvantages and briefly mention the software packages in which they are implemented.

The parameters in $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ in the GLMM are estimated by maximizing the following likelihood:

$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma} | \mathbf{y}_1, \ldots, \mathbf{y}_N) = \prod_{n=1}^{N} L_n(\boldsymbol{\beta}, \boldsymbol{\Sigma} | \mathbf{y}_n) = \prod_{n=1}^{N} p(\mathbf{y}_n | \boldsymbol{\beta}). \tag{7}$$

The likelihood in this equation is the marginal likelihood and the contribution of cluster n to this marginal likelihood is denoted as $L_n(\boldsymbol{\beta}, \boldsymbol{\Sigma} | \mathbf{y}_n)$.

The intractable integral in (7), appearing in most GLMMs with normally distributed random effects, is the main impediment to applying GLMMS. Therefore, methods for approximating the integral will be the main theme of this section. There are two general types of solution. The first is to approximate the integral numerically, so that the marginal likelihood can be computed and optimized. The second is to approximate the integrand, so that the integral of the approximation has closed form. Either approximation leads to a standard optimization problem. For these, introductory texts are widely available (Bunday, 1984; Everitt, 1987; Gill, Murray, & Wright, 1981; Lange, 2004), and so we do not discuss them here.

3.1. Approximation to the integral

In a full likelihood analysis, an approximation to the marginal likelihood in equation (7) is maximized. There are four methods to tackle the problem, and they can be classified

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

in a 2×2 table. One factor distinguishes between a direct and indirect maximization of the marginal likelihood, and the other between a deterministic (i.e. non-stochastic) and a Monte Carlo based (i.e. stochastic) numerical approximation to the intractable integral. In this subsection, we will first discuss direct maximization of the marginal likelihood and then indirect maximization; the issue of deterministic versus Monte Carlo approximations is contained within each of these.

3.1.1. Direct maximization

In both deterministic and stochastic versions of the direct maximization method, the integral is replaced by a finite sum and then maximized. We begin by discussing Gaussian quadratures (i.e. non-stochastic) and then consider Monte Carlo integration and some general optimization methods.

Gaussian quadratures. When the random effects are assumed to be normally distributed, the non-stochastic numerical approximation is usually done via Gauss–Hermite (GH) quadrature (e.g. Naylor & Smith, 1982). With a GH quadrature, unidimensional integrals of the form $\int u(t)e^{-t^2}\mathrm{d}t$ are approximated as:

$$\int u(t)e^{-t^2}\mathrm{d}t \approx \sum_{b=1}^{m} u(t_b)v_b, \tag{8}$$

where t_b and v_b are the GH quadrature nodes and weights that can be found (up to $m = 20$) in Abramowitz and Stegun (1974). (For $m > 20$, the standard nodes and weights can be computed with an algorithm described by Golub and Welsch, 1969.) The theory on GH quadratures states that the nodes are optimally spaced and weighted so that, with m nodes, the quadrature will be exact if $u(t)$ is a polynomial of degree $2m - 1$.

If the weight function e^{-t^2} in equation (8) is replaced by a normal density with mean μ and standard deviation σ, then the weights of the regular GH quadrature have to be linearly transformed (Naylor & Smith, 1982). This can be derived easily as follows:

$$\begin{aligned}\int u(t)\phi(t|\mu,\sigma^2)\mathrm{d}t &= \frac{1}{\sqrt{2\pi}\sigma}\int u(t)\exp\left[-\tfrac{1}{2}\tfrac{(t-\mu)^2}{\sigma^2}\right]\mathrm{d}t \\ &= \frac{\sqrt{2}\sigma}{\sqrt{2\pi}\sigma}\int u(\mu+\sigma\sqrt{2}r)e^{-r^2}\mathrm{d}r \\ &\approx \sum_{b=1}^{m} u(\mu+\sigma\sqrt{2}t_b)\frac{v_b}{\sqrt{\pi}},\end{aligned} \tag{9}$$

where from the second to the third expression, a change-of-variable operation is carried out with $r = \frac{t-\mu}{\sqrt{2}\sigma}$. The transformed nodes $\mu + \sigma\sqrt{2}t_b$ and weights $v_b/\sqrt{\pi}$ constitute a new set of nodes and weights that are said to be based on the normal kernel. Because it is useful below, we will explicitly define a set of nodes and weights based on the standard normal kernel ($\mu = 0, \sigma = 1$): $\mathrm{d}_b = \sqrt{2}t_b$ and $w_b = v_b/\sqrt{\pi}$.

Integration over a multivariate normal distribution of dimension Q is, in effect, an extension of the unidimensional case. The vector of nodes based on the multivariate standard normal kernel $\mathbf{d}_{b_1,\ldots,b_Q} = (d_{b_1}, \ldots, d_{b_Q})$ has to be pre-multiplied by $\boldsymbol{\Sigma}^{1/2}$, the

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

Cholesky decomposition of the variance–covariance matrix (the matrix equivalent of a square root). Applied to the likelihood for person n this gives:

$$L_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{y}_n) \approx \sum_{b_1=1}^{m} \cdots \sum_{b_Q=1}^{m} f(\mathbf{y}_n|\boldsymbol{\beta}, \boldsymbol{\Sigma}^{1/2}\mathbf{d}_{b_1,\ldots,b_Q}) w_{b_1} \ldots w_{b_Q}. \tag{10}$$

If both m and N in this equation tend to infinity, standard asymptotical results about maximum likelihood estimation hold for GLMMs (i.e. consistency and asymptotic normality: Fahrmeir & Tutz, 2001). This holds under a set of regularity conditions (Lehman & Casella, 1998), which are so general that they apply to all models considered here.

The accuracy of the approximation depends in the first place on the number of nodes m and on the proximity of $f(\mathbf{y}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n)$ to a polynomial of degree $2m - 1$. For none of the GLMMs is $f(\mathbf{y}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n)$ a polynomial, and moreover it remains unclear what degree a well-approximating polynomial should have. Therefore, for the choice of the number of nodes, a researcher has to rely on the literature and experience. But to make things even more complicated, there is no agreement in the literature concerning the number of nodes. For unidimensional models, some authors judge 10 nodes to be satisfactory (Bock, Gibbons, & Muraki, 1988), while others claim that at least 20 are needed to achieve sufficient accuracy (Crouch & Spiegelman, 1990). Lesaffre and Spiessens (2001) discuss a relatively simple example where the final answer from the analysis strongly depends on the number of nodes. For multidimensional models, Bock *et al.* suggest that in two dimensions, 5 nodes per dimension are sufficient, and if there are more than two dimensions, 3 nodes per dimension are enough. The accuracy of the approximation is also related to the size of the variance of the random effects distribution and to the size of clusters (StataCorp., 2001). Larger cluster sizes and large variances generally have an accumulated negative impact on the accuracy of the GH quadrature. The latter observation is also made by Lesaffre and Spiessens. A heuristic strategy to determine whether the quadrature has enough nodes is to perform several analyses with an increasing number of nodes. If the estimates from m to $m + 1$ nodes do not change very much, then one might opt for a quadrature with m nodes. The influence of cluster size and random effects variance will be taken up again in detail below.

As the number of dimensions Q increases, the computational burden rapidly grows because the total number of nodes (m^Q) increases at an exponential rate. Even if we decrease the number of nodes per dimension as suggested by Bock *et al.* (1988), the total number of terms in the quadrature still increases rapidly. For high-dimensional problems, stochastic GH quadrature (see below) may be a resolution.

In equation (10), a GH quadrature with the same centring (around the zero vector, the population mean for random effects) and with the same scaling (derived from the estimated variance–covariance matrix of the random effects at population level) is used for all clusters. Common centring and scaling of the GH quadrature is not very efficient when the variance components are large because then the nodes may not be located in the integrand's most interesting region. This is illustrated in Figure 1. Our explanation in the following paragraphs is based on Rabe-Hesketh, Pickles, and Skrondal (2001).

Starting from the Rasch model, we simulated a data set with 100 clusters each having 50 measurements (this is a logistic regression model with a random intercept, θ_n, and only indicator variables as predictors). Next, we selected the response pattern for the

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

clusters with the highest and lowest θ_n and plotted the integrand $a(\theta_n|\mathbf{y}_n, \boldsymbol{\beta}, \sigma^2) = f(\mathbf{y}_n|\boldsymbol{\beta}, \theta_n)\phi(\theta_n|0, \sigma^2)$ as a function of θ_n (for $\boldsymbol{\beta}$ and σ we used the true population values). In fact, the integrand $a(\theta_n|\mathbf{y}_n, \boldsymbol{\beta}, \sigma^2) = f(\mathbf{y}_n|\boldsymbol{\beta}, \theta_n)\phi(\theta_n|0, \sigma^2)$ can be considered as the unnormalized posterior density for θ_n given $\mathbf{y}_n$, $\boldsymbol{\beta}$, and σ. Therefore, the mode of the integrand is located at the most likely a posteriori value of θ_n.

In the upper panel, this was done for $\sigma = 0.0001$, which meant that there are almost no between-cluster differences. One can see that for both clusters the integrand is unimodal and symmetric around zero. In the lower panel of Figure 1, the θ_n-values are simulated from a normal distribution with $\sigma = 2$. The integrands of the two clusters have moved in opposite directions on the θ-continuum compared with the upper panel and become more peaked. This makes sense because in the simulation step more extreme values for θ_n lead to more extreme data patterns, which lead to more extreme a posteriori likely values for θ_n. However, because of the increased peakedness of the integrands, the GH quadrature points do not cover the integrands well. Therefore, if the variance of the random effects distribution is large, the number of GH quadrature points has to be increased. However, this may not work because the new points are mainly added at the extremes of the continuum rather than resulting in an increased density of grid points (moreover, the weights of the extreme nodes are very small, and therefore could be ignored). A larger cluster size will also increase the peakedness and consequently it also affects the accuracy of the approximation.

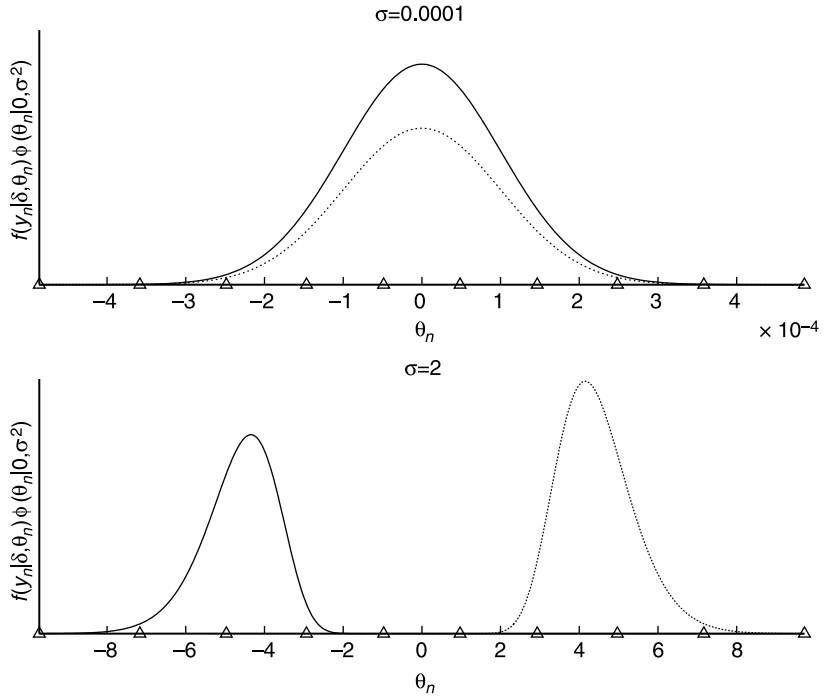

Figure 1. Plot of the integrand $f(\mathbf{y}_n|\boldsymbol{\beta}, \theta_n)\phi(\theta_n|\mathbf{0}, \sigma^2)$ for a person with a large true θ_n (dotted line) and one with a small true θ_n (solid line). In the upper panel, the θ-values are drawn from a normal distribution with mean 0 and standard deviation $\sigma = 0.0001$. In the lower panel the θ-values come from a normal distribution with mean 0 and standard deviation $\sigma = 2$. The nodes of the Gauss–Hermite quadrature are denoted by triangles.

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

An improved version of the regular GH quadrature (Pinheiro & Bates, 1995) is obtained by centring and scaling the GH quadrature nodes differently for each person according to the most likely value for the random effects vector given the observed data and (current) estimates of the fixed effects and variance components. This technique is called adaptive Gauss–Hermite (AGH) quadrature.

The AGH quadrature starts by maximizing the integrand $a(\boldsymbol{\theta}_n|\mathbf{y}_n, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = f(\mathbf{y}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n)\phi(\boldsymbol{\theta}_n|\mathbf{0}, \boldsymbol{\Sigma})$ in equation (7) for each person with respect to the random effects $\boldsymbol{\theta}_n$; the resulting estimates are the joint posterior modes for the random effects since the function $a(\boldsymbol{\theta}_n|\mathbf{y}_n, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ is equal to the unnormalized posterior density of $\boldsymbol{\theta}_n$ (i.e. it is proportional to the posterior distribution of the random effects). It is easier to find the maximum of $\log(a(\boldsymbol{\theta}_n|\mathbf{y}_n, \boldsymbol{\beta}, \boldsymbol{\Sigma}))$ than of $a(\boldsymbol{\theta}_n|\mathbf{y}_n, \boldsymbol{\beta}, \boldsymbol{\Sigma})$. Because $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are unknown, they are replaced by the current parameter estimates of iteration c, $\hat{\boldsymbol{\beta}}^{(c)}$ and $\hat{\boldsymbol{\Sigma}}^{(c)}$. The matrix $\hat{\mathbf{H}}_n$ is the Hessian matrix which contains the second-order partial derivatives of $\log(a(\boldsymbol{\theta}_n|\mathbf{y}_n, \hat{\boldsymbol{\beta}}^{(c)}, \hat{\boldsymbol{\Sigma}}^{(c)}))$, evaluated at $\hat{\boldsymbol{\theta}}_n$. The matrix $-\hat{\mathbf{H}}_n$ is then the observed information matrix, and consequently, $\hat{\boldsymbol{\Omega}}_n = -\hat{\mathbf{H}}_n$ is the estimated asymptotic variance–covariance matrix for the random effects posterior modes. Under a GLMM, an explicit expression for $\hat{\boldsymbol{\Omega}}_n$ can be derived (McCulloch & Searle, 2001), but that will not be done here. The contribution of cluster n to the likelihood can now be written as:

$$\begin{aligned} L_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{y}_n) &= \int f(\mathbf{y}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n)\phi(\boldsymbol{\theta}_n|\mathbf{0}, \boldsymbol{\Sigma})\mathrm{d}\boldsymbol{\theta}_n \\ &= \int \frac{f(\mathbf{y}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n)\phi(\boldsymbol{\theta}_n|\mathbf{0}, \boldsymbol{\Sigma})}{\phi(\boldsymbol{\theta}_n|\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\Omega}}_n)}\phi(\boldsymbol{\theta}_n|\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\Omega}}_n)\mathrm{d}\boldsymbol{\theta}_n. \end{aligned} \tag{11}$$

The GH quadrature approximation can be applied again, but now considering $\phi(\boldsymbol{\theta}_n|\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\Omega}}_n)$ as the mixing normal density. The node vector $\mathbf{d}_{b_1,\ldots,b_Q}$ from the standard normal kernel has to be transformed to $\hat{\boldsymbol{\theta}}_n + \hat{\boldsymbol{\Omega}}_n^{1/2}\mathbf{d}_{b_1,\ldots,b_Q}$ to accommodate the mean vector and variances and covariances of the new multivariate normal mixing distribution. The approximation to equation (11) then gives:

$$\begin{aligned} &L_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{y}_n) \\ &\approx \sum_{b_1=1}^{m}\cdots\sum_{b_Q=1}^{m}\frac{f(\mathbf{y}_n|\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}_n + \hat{\boldsymbol{\Omega}}_n^{1/2}\mathbf{d}_{b_1,\ldots,b_Q})\phi(\hat{\boldsymbol{\theta}}_n + \hat{\boldsymbol{\Omega}}_n^{1/2}\mathbf{d}_{b_1,\ldots,b_Q}|\mathbf{0}, \boldsymbol{\Sigma})}{\phi(\hat{\boldsymbol{\theta}}_n + \hat{\boldsymbol{\Omega}}_n^{1/2}\mathbf{d}_{b_1,\ldots,b_Q}|\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\Omega}}_n)}w_{b_1}\ldots w_{b_Q} \\ &= (2\pi)^{Q/2}|\hat{\boldsymbol{\Omega}}_n|^{1/2}\sum_{b_1=1}^{m}\cdots\sum_{b_Q=1}^{m}f(\mathbf{y}_n|\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}_n + \hat{\boldsymbol{\Omega}}_n^{1/2}\mathbf{d}_{b_1,\ldots,b_Q})\phi(\hat{\boldsymbol{\theta}}_n + \hat{\boldsymbol{\Omega}}_n^{1/2}\mathbf{d}_{b_1,\ldots,b_Q}|\mathbf{0}, \boldsymbol{\Sigma}) \\ &\quad \times \exp\left(\frac{1}{2}\mathbf{d}_{b_1,\ldots,b_Q}^{T}\mathbf{d}_{b_1,\ldots,b_Q}\right)w_{b_1}\ldots w_{b_Q}, \end{aligned} \tag{12}$$

where the last equation follows by inserting the equation for the normal density.

The principle of the AGH quadrature is illustrated in Figure 2. The upper panel shows the integrands for a large and a small value of θ_n under a Rasch model with $\sigma = 0.0001$. The triangles pointing upwards and downwards are the recentred and rescaled quadrature nodes corresponding to the integrand represented by the solid

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

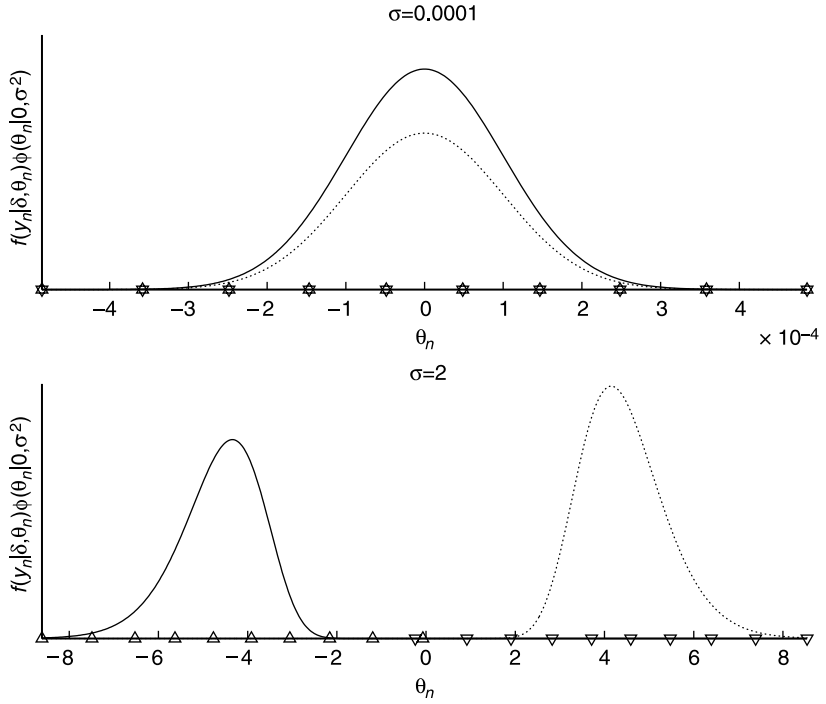

Figure 2. Similar plot to Figure 1. The nodes of the AGH quadrature for the solid curve are triangles pointing upwards and for the dotted curve are triangles pointing downwards.

and the dotted line, respectively. The adaptive procedure does not produce a lot of difference and, in fact, the nodes are placed at the same locations as the nodes in the upper part of Figure 1. In the lower panel, however, we see that the AGH quadrature aptly recentres and rescales the nodes so that they sample the region of interest for both integrands.

When applying the AGH quadrature, fewer nodes are necessary to achieve equal accuracy as for the regular GH quadrature. However, the gain in speed by having fewer terms in the approximating sum is offset by the time-consuming procedure of estimating the mode and Hessian matrix of the log of the integrand for each cluster. As an anonymous reviewer pointed out, one could design a version of the method where the centring and scaling of the nodes is only updated after a few iterations.

In principle, the AGH quadrature could be made even more accurate. As proposed here, the new centre $\hat{\boldsymbol{\theta}}_n$ and scale $\hat{\boldsymbol{\Omega}}_n$ of the quadrature are based on the current parameter estimates for $\hat{\boldsymbol{\beta}}^{(c)}$ and $\hat{\boldsymbol{\Sigma}}^{(c)}$. One could also view the centre and scale of the quadrature as a function of the fixed effects parameters and variance components: $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ and $\hat{\boldsymbol{\Omega}}_n = \hat{\boldsymbol{\Omega}}_n(\boldsymbol{\beta}, \boldsymbol{\Sigma})$. When we substitute these expressions into equation (12), the expression can be maximized directly as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. The dependence of $\hat{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\Omega}}_n$ on $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ may be complicated and this will slow down the AGH quadrature method further. Therefore, this approach is not considered for the AGH quadrature.

Monte Carlo integration. An alternative to the non-stochastic Gaussian quadrature rules is Monte Carlo integration. All the methods we discuss here simulate the likelihood rather

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

than really computing it (therefore these methods are sometimes called simulated maximum likelihood methods). The starting-point is to view the integral as an expectation of the function $f(\mathbf{y}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n)$ of a normally distributed random variable $\boldsymbol{\theta}_n$:

$$L_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{y}_n) = \int f(\mathbf{y}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n)\phi(\boldsymbol{\theta}_n|\mathbf{0}, \boldsymbol{\Sigma})\mathrm{d}\boldsymbol{\theta}_n = E[f(\mathbf{y}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n)]. \tag{13}$$

A straightforward finite-sample approximation to this expectation is calculated by sampling m independent realizations, $\boldsymbol{\theta}_n^1, \ldots, \boldsymbol{\theta}_n^m$, from $N(\boldsymbol{\theta}_n|\mathbf{0}, \boldsymbol{\Sigma})$ and then computing the sample average:

$$L_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{y}_n) \approx \frac{1}{m}\sum_{b=1}^{m} f(\mathbf{y}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n^b). \tag{14}$$

As m approaches infinity, the sample average converges to the true likelihood. If both N and m go to infinity, the maximum likelihood estimators converge to their true value under suitable regularity conditions. Note that a Monte Carlo approximation to the integral introduces another source of error: the error due to sampling variance. The sampling variance can be estimated by the sample variance of the calculated values of $f(\mathbf{y}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n^b)$. As m goes to infinity, error due to sampling variability disappears. One of the important factors influencing the error due to sampling is whether $N(\boldsymbol{\theta}_n|\mathbf{0}, \boldsymbol{\Sigma})$ has enough mass in the region of interest of $f(\mathbf{y}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n)$.

The method described above is the stochastic counterpart of the GH quadrature: the only difference is that instead of a fixed set, the nodes in Monte Carlo integration are drawn at random from the currently estimated normal random effects population distribution. For all clusters in the sample, the same distribution is used to sample the nodes from, but the nodes may also be sampled for each cluster from the posterior distribution of random effects as in the AGH. The result is an importance sampling approximation. As indicated by the right-hand side of equation (11), we can draw m independently and identically distributed and samples from $N(\boldsymbol{\theta}_n|\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\Omega}}_n)$ and then compute the following sample average:

$$L_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{y}_n) \approx \frac{1}{m}\sum_{b=1}^{m} \frac{f(\mathbf{y}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n^b)\phi(\boldsymbol{\theta}_n^b|\mathbf{0}, \boldsymbol{\Sigma})}{\phi(\boldsymbol{\theta}_n^b|\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\Omega}}_n)}. \tag{15}$$

The added value of Monte Carlo integration is especially noticeable with multidimensional integrals. For the GH quadrature, the total number of nodes increases exponentially with an increasing number of dimensions, while this is not the case with Monte Carlo integration methods.

Other types of computer-intensive Monte Carlo methods within a classical statistical framework are rather specialized and are not discussed here. For instance, a popular method in econometrics is the method of simulated moments (see Jiang, 1998; McFadden & Train, 2000; Train, 2003).

Optimization methods and software. Once the problem of the intractable integral is solved, the actual maximization of the likelihood can commence. Usually, the log-likelihood is easier to maximize than the likelihood. Therefore, the logarithms of the individual contributions to likelihood are summed to yield the total log-likelihood $l(\boldsymbol{\beta}, \boldsymbol{\Sigma})$, which is maximized as a function of the fixed effects parameters $\boldsymbol{\beta}$ and the variance components

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

$\boldsymbol{\Sigma}$. Algorithms for solving this optimization problem differ in the amount of information they extract from the log-likelihood surface to find the maximum.

The first class of algorithms uses only function values to locate the maximum and nothing else. The simplex algorithm (Nelder & Mead, 1965) is probably the best-known exemplar of this class. On the other side of the continuum, algorithms such as Newton–Raphson or the related Fisher scoring algorithm (see Everitt, 1987; Tanner, 1996) also use first- and second-order partial derivatives of the log-likelihood function, which provide information about the steepness and curvature, respectively, of the log-likelihood surface. An intermediate class of algorithms relies only on function values and first-order derivatives. The first-order derivatives can be used to create at each iteration step of the algorithm an updated approximation to the Hessian matrix of second derivatives (as in the quasi-Newton methods). Other algorithms of this class do not make approximate the Hessian matrix (e.g. the steepest descent method).

There exists a trade-off between the computational effort to execute a single step of the algorithm and the total number of steps it takes to locate the maximum of the log-likelihood. On the one hand, taking into account more information about the surface of the log-likelihood function will generally lead to fewer iterations and faster convergence than when that information is not used. On the other hand, extracting information about the surface of the log-likelihood through the first- and second-order derivatives may be computationally expensive, especially when many parameters are estimated. A balance should be achieved between computational and algorithmic speed. The point of equilibrium depends on many factors such as the type of the model and the number of parameters.

Currently there are three popular software packages that aim at a direct maximization of the marginal GLMM likelihood: PROC NLMIXED (SAS Institute, 1999), GLLAMM (Rabe-Hesketh *et al.*, 2001), MIXOR (Hedeker & Gibbons, 1996) and MIXNO (Hedeker, 1999). With PROC NLMIXED and GLLAMM, one can choose between a non-adaptive and an adaptive Gaussian quadrature. Moreover, in PROC NLMIXED both deterministic and Monte Carlo procedures are available (GLLAMM has only the deterministic methods). MIXOR and MIXNO use a non-adaptive Gaussian quadrature and the optimization is carried out through the Fisher scoring algorithm.

3.1.2. Indirect maximization

The major indirect maximization algorithm used in quantitative research is the expectation–maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977).[3] In this approach, the random effects from the N clusters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)$ are collectively considered as missing data and, together with the observed data $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$, they form the complete data. If the random effects were known, one could construct the complete-data log-likelihood, $l_C(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{y}, \boldsymbol{\theta})$, that is, the joint log-likelihood of the observed and missing data, given the parameter estimates. This is not possible because the random effects $\boldsymbol{\theta}$ are not known. But one can compute the expected value of the complete log-likelihood given the current estimates of the fixed effects $\hat{\boldsymbol{\beta}}^{(c)}$, the variances and covariances $\hat{\boldsymbol{\Sigma}}^{(c)}$ and the observed data $\mathbf{y}$. This function, denoted as $Q(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\hat{\boldsymbol{\beta}}^{(c)}, \hat{\boldsymbol{\Sigma}}^{(c)})$, is defined as follows:

[3] *Strictly speaking, the EM algorithm is an optimization tool and not a method for numerical integration. However, we discuss it here because it may lead to a major simplification of the estimation process.*

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

$$
\begin{aligned}
Q(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\hat{\boldsymbol{\beta}}^{(c)}, \hat{\boldsymbol{\Sigma}}^{(c)}) &= E(l_C(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{y}, \boldsymbol{\theta})) \\
&= E\left(\log \prod_{n=1}^{N} (f(\mathbf{y}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n)\phi(\boldsymbol{\theta}_n|\mathbf{0}, \boldsymbol{\Sigma}))\right) \\
&= \sum_{n=1}^{N} E(\log (f(\mathbf{y}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n)\phi(\boldsymbol{\theta}_n|\mathbf{0}, \boldsymbol{\Sigma}))) \qquad (16) \\
&= \sum_{n=1}^{N} \int \{\log (f(\mathbf{y}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n)) \\
&\quad + (\phi(\boldsymbol{\theta}_n|\mathbf{0}, \boldsymbol{\Sigma}))\} q(\boldsymbol{\theta}_n|\hat{\boldsymbol{\beta}}^{(c)}, \hat{\boldsymbol{\Sigma}}^{(c)}, \mathbf{y}) \mathrm{d}\boldsymbol{\theta}_n,
\end{aligned}
$$

where $q(\boldsymbol{\theta}_n|\hat{\boldsymbol{\beta}}^{(c)}, \hat{\boldsymbol{\Sigma}}^{(c)}, \mathbf{y})$ is the conditional density of the random effects given the observed data and current estimates of the fixed parameters and variance components derived using Bayes' theorem:

$$
q(\boldsymbol{\theta}_n|\hat{\boldsymbol{\beta}}^{(c)}, \hat{\boldsymbol{\Sigma}}^{(c)}, \mathbf{y}) = \frac{f(\mathbf{y}_n|\hat{\boldsymbol{\beta}}^{(c)}, \boldsymbol{\theta}_n)\phi(\boldsymbol{\theta}_n|\mathbf{0}, \hat{\boldsymbol{\Sigma}}^{(c)})}{\int f(\mathbf{y}_n|\hat{\boldsymbol{\beta}}^{(c)}, \boldsymbol{\theta}_n)\phi(\boldsymbol{\theta}_n|\mathbf{0}, \hat{\boldsymbol{\Sigma}}^{(c)})}. \qquad (17)
$$

The denominator in (17) does not depend on the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, nor on $\boldsymbol{\theta}_n$ and therefore it plays no further role.

Once the expected complete-data log-likelihood $Q(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\hat{\boldsymbol{\beta}}^{(c)}, \hat{\boldsymbol{\Sigma}}^{(c)})$ for iteration c is computed, it is maximized as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. This results in parameter estimates in iteration $c+1$ that will be used to compute the updated value of $Q(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\hat{\boldsymbol{\beta}}^{(c+1)}, \hat{\boldsymbol{\Sigma}}^{(c+1)})$. Cycling between the E-step (computing the expected complete-data log-likelihood given the observed data and the current estimates) and the M-step (optimizing the function calculated in the E-step) continues until convergence (the observed-data log-likelihood value does not increase by more than a very small amount, or the parameters do not change by more than a small amount).

Unfortunately, the intractable integral encountered before remains present (see equation (16)) and therefore we have to approximate it. The most common choices in this context are a GH quadrature or a Monte Carlo approximation (leading to the MCEM algorithm: see Booth & Hobert, 1999; McCulloch & Searle, 2001).

The EM algorithm has three advantages. First, the marginal log-likelihood (the log of equation (7)) increases at every iteration of the EM algorithm, although the algorithm does not directly maximize it (Lange, 2004; McLachlan & Krishnan, 1997). This is called the *ascent property* of the EM algorithm, and it renders the algorithm its renowned numerical stability. Maximizing the marginal likelihood directly with a technique such as Newton–Raphson does not have this numerical stability because it may converge to a minimum instead of a maximum, and if not combined with a line search procedure it may undershoot the maximum (because the algorithm takes excessively large steps). Second, the function $Q(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\hat{\boldsymbol{\beta}}^{(c)}, \hat{\boldsymbol{\Sigma}}^{(c)})$ can be written as the sum of a part pertaining to the regression parameters and a part pertaining to the variance–covariance parameters. This means that the estimation of both sets of parameters can be done separately in the M-step, thereby reducing the dimensionality of the optimization problem somewhat (this is also the case in a direct maximization approach with Fisher scoring; see Hedeker & Gibbons, 1994). Third, the M-step in the EM algorithm is sometimes particularly

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

simple to carry out and closed-form solutions are available for some of the parameters. This is the case, for instance, for the variance components under a normal mixing distribution. For other parameters, one has to rely in the M-step again on an iterative optimization method such as Newton–Raphson.

A disadvantage of the EM algorithm is that the convergence is usually very slow, especially in the neighbourhood of the maximum of the marginal likelihood. The rate of convergence depends on the ratio of missing information to the complete information (see Tanner, 1996). Thus, one expects that for the same data the convergence is slower for multidimensional models than for unidimensional models. Different kinds of modifications of the original EM algorithm have been presented to accelerate convergence or to facilitate the computation of the maximization step; we will not discuss those extensions here, but instead refer the interested reader to McLachlan and Krishnan (1997), Meng and van Dyk (1997, 1998) and Tanner (1996).

As noted by Bock and Aitkin (1981), in the context of IRT models with only indicator predictors coding for the different measurements (i.e. items in the context of IRT models), the application of the EM algorithm has a major advantage. In that specific case, the vector with item parameters $\boldsymbol{\beta}$ can be subdivided into I disjoint subsets of parameters, $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_I$, each pertaining to a single measurement (or item). Given the random effects, there is independence between the items, and consequently the expected log-likelihood can be expressed as a sum of independent terms, one for each item, that can be maximized separately. Hence, data sets with a large number of measurements per cluster (50 or more) and a separate parameter for each measurement can be analysed easily. It would be a much more difficult task to optimize directly a 50-dimensional log-likelihood with the standard methods.

Software. The EM algorithm for estimating IRT-type GLMMs is implemented in IRT software packages such as MULTILOG (Thissen, 1991) and CONQUEST (Wu, Adams, & Wilson, 2005).

3.2. Approximation to the integrand

Instead of numerically approximating the integral, the integrand itself may be approximated. The goal is then to find an approximation that leads to a tractable integral, so that the closed-form expression that follows from it can be maximized. In this section, we will discuss two classes of such techniques. We start with the Laplace approximation and a related extension. Then, we present some popular methods that are known as *quasi-likelihood methods*.

3.2.1. Laplace's method

Laplace's method (Tierny & Kadane, 1986) is used for approximating an integral of the form $\int e^{l(\mathbf{t})} d\mathbf{t}$, where it is assumed that $l(\mathbf{t})$ is a smooth, bounded and unimodal function ($\mathbf{t}$ is a Q-dimensional variable). The basic idea is to approximate $l(\mathbf{t})$ by a quadratic Taylor series expansion about the maximum $\hat{\mathbf{t}}$ of $l(\mathbf{t})$:

$$l(\mathbf{t}) \approx l(\hat{\mathbf{t}}) + \left.\frac{\partial l(\mathbf{t})}{\partial \mathbf{t}}^T\right|_{\mathbf{t}=\hat{\mathbf{t}}} (\hat{\mathbf{t}} - \hat{\mathbf{t}}) + \frac{1}{2}(\mathbf{t} - \hat{\mathbf{t}})^T l''(\hat{\mathbf{t}})(\mathbf{t} - \hat{\mathbf{t}}), \tag{18}$$

where $l''(\hat{\mathbf{t}})$ is the Hessian matrix evaluated at the maximum:

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

$$l''(\hat{\mathbf{t}}) = \left.\frac{\partial^2 l(\mathbf{t})}{\partial \mathbf{t} \partial \mathbf{t}^T}\right|_{\mathbf{t}=\hat{\mathbf{t}}}.$$

The second term in equation (18) cancels because at the maximum of $l(\mathbf{t})$, the first derivative is zero. Replacing $l(\mathbf{t})$ by the truncated Taylor series yields:

$$\int e^{l(\mathbf{t})} d\mathbf{t} \approx \int e^{l(\hat{\mathbf{t}}) + \frac{1}{2}(\mathbf{t}-\hat{\mathbf{t}})^T l''(\hat{\mathbf{t}})(\mathbf{t}-\hat{\mathbf{t}})} d\mathbf{t} = e^{l(\hat{\mathbf{t}})} \int e^{-\frac{1}{2}(\mathbf{t}-\hat{\mathbf{t}})^T(-l''(\hat{\mathbf{t}}))(\mathbf{t}-\hat{\mathbf{t}})} d\mathbf{t}.$$

In the function after the integral sign in the right-hand expression, the kernel of a multivariate normal density with mean vector $\hat{\mathbf{t}}$ and inverse covariance matrix $-l''(\hat{\mathbf{t}})$ can be recognized. This integral is equal to $(2\pi)^{Q/2}|-l''(\hat{\mathbf{t}})|^{-1/2}$, and so the Laplace approximation to the integral is:

$$\int e^{l(\mathbf{t})} d\mathbf{t} \approx (2\pi)^{Q/2}|-l''(\hat{\mathbf{t}})|^{-1/2} e^{l(\hat{\mathbf{t}})}.$$

Let us now turn to the contribution of cluster n to the likelihood, $L_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{y}_n)$. Laplace's approximation can be used after expressing the integral as:

$$\int f(\mathbf{y}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n)\phi(\boldsymbol{\theta}_n|\mathbf{0}, \boldsymbol{\Sigma}) d\boldsymbol{\theta}_n = \int a(\boldsymbol{\theta}_n|\mathbf{y}_n, \boldsymbol{\beta}, \boldsymbol{\Sigma}) d\boldsymbol{\theta}_n = \int e^{\log(a(\boldsymbol{\theta}_n|\mathbf{y}_n, \boldsymbol{\beta}, \boldsymbol{\Sigma}))} d\boldsymbol{\theta}_n.$$

We approximate $\log(a(\boldsymbol{\theta}_n|\mathbf{y}_n, \boldsymbol{\beta}, \boldsymbol{\Sigma}))$ with a quadratic Taylor series about its mode. The mode is found by solving the equation:

$$\frac{\partial \log a(\boldsymbol{\theta}_n|\mathbf{y}_n, \boldsymbol{\beta}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\theta}_n} = \mathbf{0}.$$

The expansion about the mode then becomes:

$$\begin{aligned} \log a(\boldsymbol{\theta}_n|\mathbf{y}_n, \boldsymbol{\beta}, \boldsymbol{\Sigma}) &\approx \log\left(f(\mathbf{y}_n|\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}_n)\phi(\hat{\boldsymbol{\theta}}_n|\mathbf{0}, \boldsymbol{\Sigma})\right) + \tfrac{1}{2}(\boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_n)^T \hat{\mathbf{H}}_n (\boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_n) \\ &= \log\left(f(\mathbf{y}_n|\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}_n)\phi(\hat{\boldsymbol{\theta}}_n|\mathbf{0}, \boldsymbol{\Sigma})\right) - \tfrac{1}{2}(\boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_n)^T \hat{\boldsymbol{\Omega}}_n^{-1} (\boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_n), \end{aligned} \tag{19}$$

where $\hat{\mathbf{H}}_n$ is the Hessian matrix of $\log(a(\boldsymbol{\theta}_n|\mathbf{y}_n, \boldsymbol{\beta}, \boldsymbol{\Sigma}))$ evaluated at the mode $\hat{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\Omega}}_n = (-\hat{\mathbf{H}}_n)^{-1}$ is the asymptotic covariance matrix. The Laplace method yields

$$L_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{y}_n) \approx (2\pi)^{Q/2}|\hat{\boldsymbol{\Omega}}_n|^{1/2} f(\mathbf{y}_n|\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}_n)\phi(\hat{\boldsymbol{\theta}}_n|\mathbf{0}, \boldsymbol{\Sigma}). \tag{20}$$

Laplace's method is precise when $f(\mathbf{y}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n)\phi(\boldsymbol{\theta}_n|\mathbf{0}, \boldsymbol{\Sigma})$ is exactly equal to the kernel of a normal distribution. That will be the case if the unnormalized posterior distribution of the random effects $a(\boldsymbol{\theta}_n|\mathbf{y}_n, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ is proportional to a normal density (e.g. as in the linear mixed models). For other types of models, the posterior distribution of $\boldsymbol{\theta}_n$ is only asymptotically normal (Gelman, Carlin, Stern, & Rubin, 2003). Consequently, a quadratic approximation of the log of the integrand at its mode will work well for non-normal data provided that the cluster size is large enough. Stated more precisely, the leading error term of the Laplace approximation is of order $O(I^{-1})$. Note that the asymptotics refer here to the cluster size, I, and not to the number of clusters, N.

This all looks straightforward, but in the application of Laplace's method to $L_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{y}_n)$ we ignored the fact that both the maximum $\hat{\boldsymbol{\theta}}_n$ and the observed information matrix $\hat{\boldsymbol{\Omega}}_n$ evaluated at the maximum depend on the (unknown) fixed effects

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

parameters $\boldsymbol{\beta}$ and variance components $\boldsymbol{\Sigma}$; thus $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ and $\hat{\boldsymbol{\Omega}}_n = \hat{\boldsymbol{\Omega}}_n(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ (see also the discussion on the AGH quadrature in section 3.1.1). There are two solutions to this problem.

The first solution is to take the dependency of $\hat{\boldsymbol{\theta}}_n(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ and $\hat{\boldsymbol{\Omega}}_n(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ (or only of $\hat{\boldsymbol{\theta}}_n(\boldsymbol{\beta}, \boldsymbol{\Sigma})$) on the unknown common parameters explicitly into account and find at the same time the posterior mode and the values that maximize the Laplace approximation.

The second solution circumvents the problem of the dependence of the random effects on the unknown fixed effects parameters and variance components, by using the current estimates of the fixed effects parameters, $\hat{\boldsymbol{\beta}}^{(c)}$, and variance components, $\hat{\boldsymbol{\Sigma}}^{(c)}$, as in AGH quadrature. The relation between this form of the Laplace method and the AGH quadrature is even more fundamental. The AGH quadrature from equation (12) simplifies to the Laplace approximation described here if only one quadrature point is used ($m = 1$: Liu & Pierce, 1994; Pinheiro & Bates, 1995). This can be seen as follows: with a single node, $\mathbf{d}_{b_1,\ldots,b_Q}$ is a vector of zeros and all weights $w_{b_1}, \ldots, w_{b_Q}$ are equal to one; substituting these values into equation (12) gives the Laplace approximation of equation (20).

In the Laplace method, the log of the integrand is expanded as a quadratic Taylor series about its mode. This suggests that we may increase its accuracy by also including higher-order terms in the Taylor series. Raudenbush, Yang, and Yosef (2000; but see also Breslow & Lin, 1995) proposed to use a sixth-order approximation (called 'Laplace6'). Let us denote such higher-order terms in a Taylor expansion up to order k as $T_3, \ldots, T_k$. Equation (19) now has to be completed with terms $T_3, \ldots, T_k$ to obtain the kth-order approximation:

$$\begin{aligned}&\log\left(f(\mathbf{y}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n)\phi(\boldsymbol{\theta}_n|\mathbf{0}, \boldsymbol{\Sigma})\right)\\ &\approx \log\left(f(\mathbf{y}_n|\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}_n)\phi(\hat{\boldsymbol{\theta}}_n|\mathbf{0}, \boldsymbol{\Sigma})\right) - \tfrac{1}{2}(\boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_n)^T \hat{\boldsymbol{\Omega}}_n^{-1}(\boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_n) + T_3 + \ldots + T_k.\end{aligned} \tag{21}$$

After exponentiating and integrating over $\boldsymbol{\theta}_n$, we can extract the first term, which is a constant, outside the integral:

$$\begin{aligned}&f(\mathbf{y}_n|\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}_n)\phi(\hat{\boldsymbol{\theta}}_n|\mathbf{0}, \boldsymbol{\Sigma}) \int \exp\left(-1/2(\boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_n)^T \hat{\boldsymbol{\Omega}}_n^{-1}(\boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_n) + T_3 + \ldots + T_k\right)\mathrm{d}\boldsymbol{\theta}_n\\ &= f(\mathbf{y}_n|\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}_n)\phi(\hat{\boldsymbol{\theta}}_n|\mathbf{0}, \boldsymbol{\Sigma}) \int \exp\left(T_3 + \ldots + T_k\right)\\ &\quad \exp\left(-1/2(\boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_n)^T \hat{\boldsymbol{\Omega}}_n^{-1}(\boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_n)\right)\mathrm{d}\boldsymbol{\theta}_n\\ &= (2\pi)^{Q/2}|\hat{\boldsymbol{\Omega}}_n|^{1/2} f(\mathbf{y}_n|\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}_n)\phi(\hat{\boldsymbol{\theta}}_n|\mathbf{0}, \boldsymbol{\Sigma}) E(\exp\left(T_3 + \ldots + T_k\right))\\ &= (2\pi)^{Q/2}|\hat{\boldsymbol{\Omega}}_n|^{1/2} f(\mathbf{y}_n|\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}_n)\phi(\hat{\boldsymbol{\theta}}_n|\mathbf{0}, \boldsymbol{\Sigma})\\ &\quad \left[1 + E(T_3 + \ldots + T_k) + \frac{1}{2!}E((T_3 + \ldots + T_k)^2) + \ldots\right];\end{aligned}$$

in the last equality, we used the series expansion for $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$.

Raudenbush *et al.* (2000) show that the expectations of odd terms and cross-products with one even and one odd term are equal to zero. Intuitively this makes sense, as can be illustrated easily for the unidimensional case. If there is only a single random effect, all

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

terms in the Taylor expansion are of the form $(\theta_n - \hat{\theta}_n)^p$. The expectation of such an odd term with respect to the normal distribution $\phi(\theta_n|\hat{\theta}_n, \hat{\sigma}^2)$ indeed vanishes. For all even terms and cross-products, Raudenbush *et al.* (2000) derive simplified expressions.

The terms that contribute most to improving the accuracy of approximation are T_4, $T_3^2/2$, T_6, $T_3T_5/2$ and $T_4^2/2$, but Raudenbush *et al.* (2000) include terms up to the sixth order (i.e. T_4, $T_3^2/2$ and T_6), leading to an approximation error of order $O(I^{-2})$. Higher-order approximations, while possible, result only in a slight additional improvement.

Raudenbush *et al.* (2000) maximize the sixth-order Laplace approximation by an approximate Fisher scoring algorithm that avoids the computation of second derivatives. They solve the issue of the dependence of the mode $\hat{\boldsymbol{\theta}}_n$ and curvature $\hat{\boldsymbol{\Omega}}_n$ on the unknown fixed effects parameters and variance components using implicit differentiation. In an example and a simulation study, Raudenbush *et al.* (2000) found that the performance of the 'Laplace6' method is comparable with a regular Gaussian quadrature with at least 20 nodes and an adaptive Gaussian quadrature with 7 nodes. The method clearly outperforms PQL, another integrand approximation method, to be discussed next.

3.2.2. Quasi-likelihood approaches

Estimation methods for linear mixed models are well established and this motivates approximations of the GLMM by a linear mixed model, so that the estimation methods for the latter class of models can be applied. Two such approaches are the penalized quasi-likelihood (PQL: Breslow & Clayton, 1993; Schall, 1991; Stiratelli, Laird, & Ware, 1984) and marginal quasi-likelihood (MQL: Goldstein, 1991) and their extensions (MQL2, PQL2 and corrected PQL).

The names PQL and MQL were introduced by Breslow and Clayton (1993). The reason for baptizing these methods as quasi-likelihood is that they require only the specification of the mean and variance for the observations and there is no need to spell out explicitly the distribution of the data. In a GLM framework, the same terminology is used for fitting models in which only means and variances are specified. The adjectives *penalized* and *marginal* will be clarified after the presentation of the methods.

Although PQL and MQL are classified as methods that approximate the integrand, we will explain them primarily as approximations to the data themselves because of conceptual simplicity. After this discussion, we present briefly how the same method can be derived directly from an approximation to the integrand.

To explain the linearized approximation methods, we start by decomposing the observations into their mean and an error term, conditional upon the random effect:

$$y_{ni} = \mu_{ni} + \varepsilon_{ni} = h(\mathbf{x}_{ni}^T\boldsymbol{\beta} + \mathbf{z}_{ni}^T\boldsymbol{\theta}_n) + \varepsilon_{ni}, \tag{22}$$

where $h(\cdot)$ denotes the response function (i.e. the inverse of the link function).

To illustrate equation (22), an observation y_{ni} in a mixed logistic regression model has mean

$$\boldsymbol{\mu}_{ni} = h(\mathbf{x}_{ni}^T\boldsymbol{\beta} + \mathbf{z}_{ni}^T\boldsymbol{\theta}_n) = \Pr(y_{ni} = 1) = \pi_{ni} = \frac{\exp(\mathbf{x}_{ni}^T\boldsymbol{\beta} + \mathbf{z}_{ni}^T\boldsymbol{\theta}_n)}{1 + \exp(\mathbf{x}_{ni}^T\boldsymbol{\beta} + \mathbf{z}_{ni}^T\boldsymbol{\theta}_n)}$$

(see equations (4) and (5)). The error term then equals $-\pi_{ni}$ with probability $1-\pi_{ni}$ (for the observation $y_{ni} = 0$) and $1 - \pi_{ni}$ with probability π_{ni} (for the observation $y_{ni} = 1$). The error has expectation zero and variance $\pi_{ni}(1 - \pi_{ni}) = \mu_{ni}(1 - \mu_{ni}) = v(\mu_{ni})$.

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

PQL starts with a linear Taylor approximation of the response function about the current estimate of the fixed effect, $\hat{\boldsymbol{\beta}}$, and about the posterior mode of the random effect. The mean evaluated at $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}_n$ is denoted as μ_{ni*}. The approximate variance of the error term is denoted as $v(\mu_{ni})$; it is only approximate because the true variance depends on the unknown $\boldsymbol{\beta}$ and $\boldsymbol{\theta}_n$. The observation y_{ni} can now be written as:

$$\begin{aligned} y_{ni} &\approx h(\mathbf{x}_{ni}^T\hat{\boldsymbol{\beta}} + \mathbf{z}_{ni}^T\hat{\boldsymbol{\theta}}_n) + h'(\mathbf{x}_{ni}^T\hat{\boldsymbol{\beta}} + \mathbf{z}_{ni}^T\hat{\boldsymbol{\theta}}_n)\mathbf{x}_{ni}^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \\ &\quad + h'(\mathbf{x}_{ni}^T\hat{\boldsymbol{\beta}} + \mathbf{z}_{ni}^T\hat{\boldsymbol{\theta}}_n)\mathbf{z}_{ni}^T(\boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_n) + \varepsilon_{ni} \\ &= \mu_{ni*} + v(\mu_{ni*})\mathbf{x}_{ni}^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + v(\mu_{ni*})\mathbf{z}_{ni}^T(\boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_n) + \varepsilon_{ni}. \end{aligned} \tag{23}$$

We can stack all observations y_{ni} and errors ε_{ni} in the column vectors $\mathbf{y}$ and $\boldsymbol{\varepsilon}$, respectively, and the design vectors $\mathbf{x}_{ni}^T$ and $\mathbf{z}_{ni}^T$ likewise in the respective design matrices $\mathbf{X}$ and $\mathbf{Z}$. The means μ_{ni*} are collected in a vector $\boldsymbol{\mu}_*$ and the approximate variances $v(\mu_{ni*})$ in a diagonal matrix $\mathbf{V}_*$. Equation (23) can now be written for all observations as follows:

$$\mathbf{y} \approx \boldsymbol{\mu}_* + \mathbf{V}_*\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \mathbf{V}_*\mathbf{Z}(\boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_n) + \boldsymbol{\varepsilon}.$$

Reorganizing this equation by bringing $\boldsymbol{\mu}_*$ to the left-hand side, pre-multiplying by V_*^{-1} and then also bringing $\boldsymbol{X\hat{\beta}}$ and $\boldsymbol{X\hat{\theta}}_n$ to the left-hand side yields:

$$\mathbf{y}_* \equiv \mathbf{V}_*^{-1}(\mathbf{y} - \boldsymbol{\mu}_*) + \mathbf{X}\hat{\beta} + Z\hat{\boldsymbol{\theta}}_n \approx \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\theta}_n + \mathbf{V}_*^{-1}\boldsymbol{\varepsilon}. \tag{24}$$

The error term $\mathbf{V}_*^{-1}\boldsymbol{\varepsilon}$ on the right-hand side will be denoted as $\boldsymbol{\varepsilon}_*$. The marginal mean of $\mathbf{y}_*$ equals $\mathbf{X}\boldsymbol{\beta}$. To derive the marginal variance–covariance matrix, we first find the variance of $\boldsymbol{\varepsilon}_*$:

$$\mathrm{Var}\,(\boldsymbol{\varepsilon}_*) = \mathrm{Var}\,(\mathbf{V}_*^{-1}\boldsymbol{\varepsilon}) = \mathbf{V}_*^{-1}\mathrm{Var}\,(\boldsymbol{\varepsilon})\mathbf{V}_*^{-1} \approx \mathbf{V}_*^{-1}.$$

Consequently, the variance of $\mathbf{y}_*$ is approximately:

$$\begin{aligned} \mathrm{Var}\,(\mathbf{y}_*) &\approx \mathrm{Var}\,(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\theta}_n + \boldsymbol{\varepsilon}_*) \approx \mathbf{Z}\,\mathrm{Var}\,(\boldsymbol{\theta}_n)\mathbf{Z}^T + \mathbf{V}_*^{-1} \\ &= \mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T + \mathbf{V}_*^{-1}. \end{aligned} \tag{25}$$

In the GLM literature, the vector $\mathbf{y}_*$ is also called the *adjusted dependent variable* (McCullagh & Nelder, 1989). The same result follows by defining the adjusted dependent variable directly; it is a linearized version of the link function $g(\cdot)$ applied to the data: $y_{ni*} \equiv g(\mu_*) + g'(\mu_*)(\mu_{ni} - \mu_*)$.

Equations (24) and (25) appear also in the context of linear mixed models with $\mathbf{y}_*$ as the data vector, $\mathbf{X}\boldsymbol{\beta}$ as the mean vector and $\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T + \mathbf{V}_*^{-1}$ as the marginal variance–covariance matrix. Therefore, to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, we may apply an estimation procedure for a linear mixed model.[4] Once the estimates are updated, linearization is applied again (see equation (23)), leading to a further update of the adjusted dependent variable $\mathbf{y}_*$. The algorithm cycles between these two steps until convergence.

The MQL method is very similar to the PQL approach. It starts with a linear Taylor approximation of the mean, but the expansion is now about the current estimates for $\boldsymbol{\beta}$

[4] *Note that the variance components in a linear mixed model are often estimated using restricted maximum likelihood (McCulloch & Searle, 2001; Verbeke & Molenberghs, 2000).*

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

and the zero vector for the random effects (their population mean):

$$\begin{aligned} y_{ni} &\approx h(\mathbf{x}_{ni}^T\hat{\boldsymbol{\beta}}) + h'(\mathbf{x}_{ni}^T\hat{\boldsymbol{\beta}})\mathbf{x}_{ni}^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + h'(\mathbf{x}_{ni}^T\hat{\boldsymbol{\beta}})\mathbf{z}_{ni}^T\boldsymbol{\theta}_n + \varepsilon_{ni} \\ &= \mu_{ni0} + v(\mu_{ni0})\mathbf{x}_{ni}^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + v(\mu_{ni0})\mathbf{z}_{ni}^T\boldsymbol{\theta}_n + \varepsilon_{ni}, \end{aligned} \tag{26}$$

where μ_{ni0} stands for the mean evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ and $\boldsymbol{\theta}_n = \mathbf{0}$ and $v(\mu_{ni0})$ for the approximate variance of the error. We now form the vector $\mathbf{y}$ and the matrices $\mathbf{X}$ and $\mathbf{Z}$ by analogy with the definitions for the PQL method and the mean vector $\boldsymbol{\mu}_0$ and diagonal variance matrix $\mathbf{V}_0$. This implies the approximation:

$$\mathbf{V}_0^{-1}(\mathbf{y} - \boldsymbol{\mu}_0) + \mathbf{X}\hat{\boldsymbol{\beta}} \approx \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\theta}_n + \mathbf{V}_0^{-1}\boldsymbol{\varepsilon}. \tag{27}$$

The left-hand side, denoted by $\mathbf{y}_0$, is the adjusted dependent variable. Parameter estimation can now proceed for MQL in the same way as for PQL, cycling between a linear mixed model estimation step and an adjusted dependent variable updating step until convergence. The method is somewhat confusingly called *marginal* because the linear approximation to the response function is carried out with all random effects set equal to zero, $\boldsymbol{\theta}_n = \mathbf{0}$.

Justifying PQL and MQL by a linear approximation to the response function is simple and leads to the revelation of the subtle but important differences between them. However, the same result can be obtained by directly considering an approximation to the integrand. For PQL, one has to start from the Laplace approximation method in equation (20) and apply one more approximation. (For simplicity, let us assume here that we are interested in the fixed effect parameters in the first place. The estimation of the variance components needs to be done in a secondary second step at each iteration, for example, using restricted maximum likelihood.) Although the variance–covariance matrix $\hat{\boldsymbol{\Omega}}_n = \hat{\boldsymbol{\Omega}}_n(\boldsymbol{\beta})$ depends on the unknown fixed effects parameters, we assume that $\hat{\boldsymbol{\Omega}}_n$ is almost constant with respect to the fixed effects parameters (i.e. slow varying), and therefore that it can be ignored in the optimization of the Laplace approximation. Next, we may opt for the first solution outlined in Section 3.2.1 on the Laplace approximation. By ignoring the dependence of $\hat{\boldsymbol{\Omega}}_n$ on the fixed effects parameters, finding the random effects posterior modes and the fixed effects parameter estimates jointly, reduces to maximizing:

$$\sum_{n=1}^{N}\left(\log\left(f(\mathbf{y}_n|\boldsymbol{\beta}, \boldsymbol{\theta}_n)\right) - \frac{1}{2}\boldsymbol{\theta}_n^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}_n\right), \tag{28}$$

with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}_n$. This result can be obtained starting from equation (20), ignoring $\hat{\boldsymbol{\Omega}}_n$ and all terms not dependent upon $\boldsymbol{\beta}$ and $\boldsymbol{\theta}_n$. In equation (28), the term $\frac{1}{2}\boldsymbol{\theta}_n^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}_n$ acts as a penalty to the quasi-likelihood (which is the first term) for the estimation of the random effects (the random effects are not allowed to deviate too much from the zero vector) and hence the method's name – penalized quasi-likelihood.

MQL can be justified similarly, based not on the Laplace approximation but on a quadratic expansion of $\log\left(f(\mathbf{y}_n|\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}_n)\right)$ about $\boldsymbol{\theta}_n = \mathbf{0}$. This method was proposed by Longford (1994) and Rodriguez and Goldman (1995) show that it leads to the same approximation as in equation (27).

Following the development of PQL and MQL, some attempts have been made to improve both methods. This resulted in PQL2, MQL2 and corrected PQL. Both PQL2 and MQL2 make use of second-order Taylor expansions, so that the terms pertaining to the

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

random effects appearing in the approximation are:

$$T_{\mathrm{PQL2}} = h'(\mathbf{x}_{ni}^T\boldsymbol{\beta}_0 + \mathbf{z}_{ni}^T\hat{\boldsymbol{\theta}}_n)\mathbf{z}_{ni}^T(\boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_n) + \frac{1}{2}(\boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_n)^T\mathbf{z}_{ni}h''(\mathbf{x}_{ni}^T\boldsymbol{\beta}_0 + \mathbf{z}_{ni}^T\hat{\boldsymbol{\theta}}_n)\mathbf{z}_{ni}^T(\boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_n) \tag{29}$$

$$T_{\mathrm{MQL2}} = h'(\mathbf{x}_{ni}^T\boldsymbol{\beta}_0)\mathbf{z}_{ni}^T\boldsymbol{\theta}_n + \frac{1}{2}\boldsymbol{\theta}_n^T\mathbf{z}_{ni}h''(\mathbf{x}_{ni}^T\boldsymbol{\beta}_0)\mathbf{z}_{ni}^T\boldsymbol{\theta}_n.$$

For PQL2, the last term in T_{PQL2} is replaced by its expected value (see Goldstein & Rasbash, 1996). For MQL2, the quadratic term is considered to be another set of random effects, with mean and variances calculated from the current estimates of the random effects variances (see Rodriguez & Goldman, 1995). Then, one cycles between an algorithm for linear mixed models to obtain estimates of the fixed effects and variance components and updating the adjusted dependent variable until convergence. Breslow and Lin (1995) and Lin and Breslow (1996) present other bias correction methods for PQL.

Because both PQL and MQL are based on approximations for which the accuracy is difficult to assess, it comes as no surprise that these methods do not work very well in some cases. We will discuss briefly in what situations the methods fail and why. Both for PQL and MQL, the original data are transformed linearly and then regarded as normally distributed, so that a linear mixed model estimation routine can be applied. Because the performance of the two methods depends on the validity of the normal approximation, they tend to perform poorly when the original data are far from normal, as is the case with binary data, for instance. This has been confirmed in a simulation study by Breslow and Clayton (1993) for PQL and MQL and by Rodriguez and Goldman (1995) for MQL. The fixed effects and/or variance component estimates are biased towards zero. Furthermore, Breslow and Lin (1995) demonstrate that the PQL estimates for the regression coefficients also have an appreciable downward asymptotic bias (and hence are inconsistent).

Comparing different quasi-likelihood methods, Rodriguez and Goldman (1995) show that MQL2 performs only slightly better than MQL in terms of bias, but Goldstein and Rasbash (1996) demonstrate that PQL2 leads to a substantial improvement over PQL. In a recent simulation study by Browne and Draper (2004), it appears that PQL2 leads to much better results than MQL (the authors did not use PQL).

Normality of the transformed data in the quasi-likelihood methods also depends on the number of elementary units in a cluster. In a linear mixed model, the individual approximated observations affect the fixed effects parameter estimates through their sufficient statistics, which are linear combinations of the approximated data. Since linear combinations of normal random variables are also normally distributed, PQL and MQL are expected to perform less well if the number of observations per cluster is small. (For the PQL method, this hypothesis can also be derived starting from the Laplace approximation to the log-likelihood. As stated before, the order of the leading error term in the approximation is $O(I^{-1})$, showing that for small I the approximation will be worse. Thus, if the posterior distribution of random effects can be approximated by a normal density, PQL will work fine.) Thus, while the regular GH quadrature tends to work less well for large cluster sizes, the opposite holds for PQL and MQL.

Moreover, MQL uses a linear Taylor approximation around the current fixed effects and zeros for all random effects, and therefore we expect this method to work not as

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

well for data with large random effects variances. This is also confirmed by the simulation study of Rodriguez and Goldman (1995).

Finally, we note that the deviance measure produced by using PQL or MQL methods cannot be used in subsequent model testing. Making only assumptions about the first and second moment of the data (mean and variance) is acceptable for the estimation of the parameters. However, for testing the adequacy of the model, reference distributions for the test statistics have to be derived and the distribution of the data has to be specified.

Software The MQL and PQL methods are implemented in MLwiN (Rasbash, Steele, Browne, & Prosser, 2005) to fit GLMMs for binary data (currently it is not possible to fit models for polytomous data). Besides the first-order versions of PQL and MQL, PQL2 and MQL2 are also implemented. The PQL method is the standard tool for fitting models in HLM 5 (Raudenbush, Bryk, & Congdon, 2001), although binary response models can be estimated with the Laplace6 method too. Besides MLwiN, the SAS macro GLMMIX (Wolfinger, 1993), programmed to estimate GLMMs, uses a variant of the PQL method. Because the adaptive Gaussian quadrature with a single node reduces to a Laplace approximation, the PROC NLMIXED procedure of SAS (SAS Institute, 1999) can be used as a Laplace approximation by restricting the adaptive quadrature to have a single node only.

4. Testing: Hypothesis testing and goodness of fit

In hypothesis testing, one assesses whether one or more functions of the parameters are equal to some constant value(s). Typically, and the only case considered here, a linear hypothesis is tested, with as an important special case the assumption that one or more parameters are equal to zero. The model restricted according to the hypothesis is the restricted or null model, and the more general model to which the restriction(s) of the null hypothesis does not apply is the unrestricted or alternative model.

In goodness-of-fit testing, the model to be evaluated with respect to its fit to the data is the restricted model, and the unrestricted model is the so-called saturated model. The latter is the most general model, with the maximum likelihood achievable. Hypothesis testing and testing for the goodness of fit are also referred to in the literature as relative and absolute model testing, respectively.

As was pointed out by an anonymous reviewer, there is some logical inconsistency in the way model selection is commonly carried out in that failure to reject a hypothesis (the restricted model is true) is regarded as evidence in favour of the hypothesis. Furthermore, one should keep in mind that usually many candidate models are assessed which may lead to considerable discrepancies between 'true' and 'observed' *p*-values (Draper, 1995).

4.1. Testing of linear hypotheses

We distinguish between testing linear hypotheses with respect to the fixed effects, and testing hypotheses with respect to the variance–covariance structure of the random effects or the variance components (testing for the need for random effects).

4.1.1. Fixed effects

In testing the set of s linear hypotheses

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \boldsymbol{\xi} \text{ against } H_1 : \mathbf{C}\boldsymbol{\beta} \neq \boldsymbol{\xi},$$

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

where the matrix of coefficients $\mathbf{C}$ has full row rank $S \leq P$, one can rely on three types of tests: a likelihood-ratio (LR) test, a Wald test and a score test (Fahrmeir & Tutz, 2001). Under the restricted model, the three tests are asymptotically chi-squared distributed with s degrees of freedom.

As a simple case, consider a logistic regression model with a random intercept. Assume β_j is the regression coefficient of the jth predictor and suppose we want to test the hypothesis $H_0 : \beta_j = 0$. In this case, $\mathbf{C}$ is a row vector with all zeros except for the jth entry, which is equal to one.

The LR statistic contrasts the log-likelihood $l(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\Sigma}}|\mathbf{y})$ of the MLE $\tilde{\boldsymbol{\beta}}$ obtained under the restricted model with the log-likelihood of the MLE $l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}}|\mathbf{y})$ obtained under the unrestricted model:

$$LR = -2[l(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\Sigma}}|\mathbf{y}) - l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}}|\mathbf{y})].$$

The Wald statistic,

$$w = [\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\xi}]^T [\mathbf{C}\mathbf{F}_{\hat{\boldsymbol{\beta}}}^{-1}\mathbf{C}^T]^{-1} [\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\xi}],$$

is the weighted distance between the MLE of $\mathbf{C}\boldsymbol{\beta}$ under the unrestricted model, $\mathbf{C}\hat{\boldsymbol{\beta}}$, and its hypothetical value under the null model, $\boldsymbol{\xi}$. The weight $[\mathbf{C}\mathbf{F}_{\hat{\boldsymbol{\beta}}}^{-1}\mathbf{C}^T]^{-1}$ is the inverse of the asymptotic covariance matrix of $\mathbf{C}\hat{\boldsymbol{\beta}}$. The matrix $\mathbf{F}_{\hat{\boldsymbol{\beta}}}$ can be either the observed or the expected information matrix, evaluated at $\hat{\boldsymbol{\beta}}$:

$$\mathbf{F}_{\hat{\boldsymbol{\beta}}} = -\frac{\partial^2 l(\boldsymbol{\beta}, \hat{\boldsymbol{\Sigma}}|\mathbf{y})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \quad \text{or} \quad \mathbf{F}_{\hat{\boldsymbol{\beta}}} = E\left(-\frac{\partial^2 l(\boldsymbol{\beta}, \hat{\boldsymbol{\Sigma}}|\mathbf{y})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}\right).$$

Finally, the score statistic (also called the *Lagrange multiplier Test*) is defined as:

$$u = \mathbf{s}(\tilde{\boldsymbol{\beta}})^T \mathbf{F}_{\tilde{\boldsymbol{\beta}}}^{-1} s(\tilde{\boldsymbol{\beta}}),$$

where $\mathbf{s}(\tilde{\boldsymbol{\beta}})$ is the score function under the restricted model defined as the first derivative of the log-likelihood evaluated at the MLE $\tilde{\boldsymbol{\beta}}$:

$$\mathbf{s}(\tilde{\boldsymbol{\beta}}) = \frac{\partial l(\boldsymbol{\beta}, \tilde{\boldsymbol{\Sigma}}|\boldsymbol{y})}{\partial\boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}.$$

The matrix $\mathbf{F}_{\tilde{\boldsymbol{\beta}}}$ is the expected or observed information matrix, but now evaluated at the MLE $\tilde{\boldsymbol{\beta}}$ under the restricted model. The score statistic is the weighted distance between $\mathbf{s}(\tilde{\boldsymbol{\beta}})$ and $\mathbf{s}(\tilde{\boldsymbol{\beta}}) = 0$.

The Wald and score tests are both based on a quadratic approximation to the log-likelihood function by a second-order Taylor expansion and thus are approximations to the LR statistic. Hence, the p-values obtained for the LR test are more exact than the p-values obtained for the Wald and score tests. However, the disadvantage of the LR test is that both the restricted and the unrestricted model have to be fitted, whereas for the Wald test and the score test only the unrestricted or restricted MLE is needed, respectively. Hence, in a forward predictor selection procedure, one can rely on the score test but not on the Wald test, and vice versa in a backward selection procedure. With increasing sample size, the log-likelihood becomes approximately quadratic, so that all three tests are asymptotically equivalent.

In Figure 3, the three tests are graphically illustrated for testing a hypothesis about a single regression parameter $\beta = \xi$ and with known $\boldsymbol{\Sigma}$ (hence, reference to random

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

effects variances and covariances is dropped from the log-likelihood). The LR test is twice the distance between $l(\hat{\beta}|\mathbf{y})$ and $l(\tilde{\beta} = \xi|\mathbf{y})$. The Wald test is the distance between the quadratic approximations of $l(\hat{\beta}|\mathbf{y})$ and $l(\tilde{\beta} = \xi|\mathbf{y})$. The score test is based on the slope of the log-likelihood of the unrestricted model at ξ.

4.1.2. Covariance structure of the random effects

Testing whether a random effect for a particular covariate should be or should not be included in the model corresponds to testing the hypothesis that the variance of the population distribution is zero. Suppose we want to test a hypothesis about the single variance parameter σ^2 for a GLMM with a single random effect. Because variances cannot be negative, zero is at the boundary of the parameter space, and one actually tests a one-sided hypothesis (Verbeke & Molenberghs, 2003):

$$H_0 : \sigma^2 = 0 \quad \text{against} \quad H_1 : \sigma^2 > 0.$$

The consequence is that the asymptotic distribution of the LR, Wald and score tests differs from chi-squared.

For the linear mixed model, Stram and Lee (1994, 1995) investigated testing for random effects using results of Self and Liang (1987). Stram and Lee showed that the appropriate asymptotic distribution under the restricted model is often a mixture of chi-squared distributions. For example, when there are no random effects in the restricted model, and a single random effect in the full model, the asymptotic distribution of the LR test is a mixture of a chi-squared distribution with, respectively zero and one degree of freedom, both with weight of 1/2 (the chi-squared distribution with zero degrees of freedom is a distribution with point mass one at zero). Hence, the appropriate *p*-value is obtained by halving the 'classical' *p*-value that is obtained when testing a hypothesis for a single parameter. When the restricted model contains q random effects, and the alternative model $q + 1$, the asymptotic distribution of the LR statistic under the restricted model is a mixture with equal component weights of 1/2 of chi-squared distributions with q and $q + 1$ degrees of freedom. For the general case of testing q

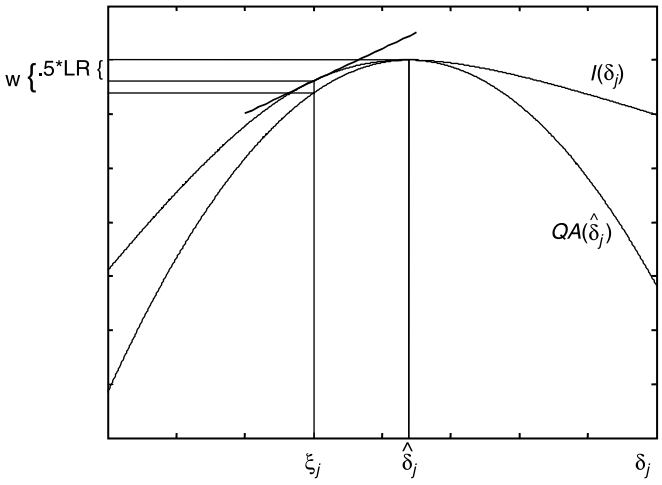

Figure 3. The log-likelihood $l(\delta_j)$, the quadratic approximation to the log-likelihood in $\hat{\delta}_j$, $QA(\hat{\delta}_j)$ and the slope of the log-likelihood at ε_j.

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

random effects against $q + k$ random effects, the mixture involves components other than random variables with a chi-squared distribution, and one has to rely on simulation methods to approximate the distribution of the LR statistic under the restricted model. Under regularity conditions, these results carry over to the generalized linear and non-linear mixed model (Wolak, 1989, 1991).

Since the asymptotic equivalence of the score and the LR tests also holds when testing for one-sided hypotheses (or a combination of one- and two-sided hypotheses: Silvapulle & Silvapulle, 1995; Verbeke & Molenberghs, 2003), the same results apply to the score test. For discussions of the score test applied to testing random effects variances, see Berkhof and Snijders (2001), Hall and Præstgaard (2001) and Lin (1997).

4.2. Goodness-of-fit tests

The global goodness of fit of a particular model can be assessed by a likelihood ratio test for which the alternative model is saturated; that is, the most complex model that is identified by the data. In the saturated model, $E(y_{ni}|\boldsymbol{\beta}, \boldsymbol{\theta}_n) = y_{ni}$ for all n and i.

For categorical dependent variables, another frequently used goodness-of-fit statistic is the Pearson statistic:

$$\chi^2 = \sum_{g=1}^{G} n_g \sum_{j=0}^{J} \frac{(y_{gj} - \hat{\pi}_{gj})^2}{\hat{\pi}_{gj}},$$

where n_g is the number of participants with the same values on the covariates, and y_{gj} and $\hat{\pi}_{gj}$ are respectively the observed and expected proportions of the participants in group g responding in category j. Note that for repeated measurement data, each possible response pattern serves as a distinct category of a multinomial distribution.

The LR test statistic with the saturated model as the alternative and the Pearson statistic are asymptotically equivalent and asymptotically chi-squared distributed. The number of degrees of freedom equals the difference between the numbers of estimated parameters in the saturated and the tested model.

The first problematic aspect of both test statistics is that the asymptotic results are obtained under the assumption that the number of groups, G, is fixed with increasing sample size. The latter assumption does not hold when some covariates are continuous. Second, when a categorical dependent variable is measured on several occasions, the number of possible response patterns increases exponentially, so that the sample is always too small to warrant relying on the asymptotic results. For example, with 10 dichotomous responses, the number of possible response patterns J already equals $2^{10} = 1{,}024$.

A solution to both problems is to simulate the distribution of the test statistic under the restricted model, either by using bootstrap techniques in a maximum likelihood framework (Efron & Tibshirani, 1993).

Instead of testing the global goodness of fit, one can test for specific aspects of the model. The underlying idea is that a model is never true, but acceptable when it succeeds in explaining the aspects of the data that are most relevant. The latter can be achieved by computing statistics that are sensitive to the aspects of interest only. For example, in the context of IRT, Glas and Verhelst (1988) constructed a family of statistics that are based on the Pearson statistic to test for specific model assumptions. The proposed statistics asymptotically follow a chi-squared distribution under the restricted model, with the appropriate number of degrees of freedom. Alternatively, one can simulate the distribution of the statistics under the restricted model.

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

5. Alternative modelling approaches

Besides the GLMM approach to clustered data as presented here, there are alternative modelling approaches not discussed in this paper. Three possibilities will be briefly highlighted in this section.

First, the central class of models in this paper was the GLMMs class. However, many models do not belong to this class. For instance, the common two-parameter logistic item response model (Birnbaum, 1968) is not a GLMM because some of the parameters appear in a multiplicative relationship in the predictor. This broader class of models is called *non-linear mixed models* (NLMMs) and the GLMMs are a subset of theirs. Some of the methods discussed extend or can be adapted to NLMMs. For instance, non-adaptive and adaptive Gaussian quadrature can be used to estimate the parameters of an NLMM. Wolfinger and Lin (1997) have shown that the integrand approximation methods PQL and MQL can be transferred to NLMMs for which the observed data, conditionally on the random effects, have a normal distribution. But it is not clear whether these methods also work for the general class of NLMMs.

Another related class of models we did not discuss in this paper are structural equation models (Muthén & Muthén, 2004; Skrondal & Rabe-Hesketh, 2004). In the structural equation modelling approach, models consist of two parts: a measurement part and a structural part. The measurement part links the observed random variables with predictors and the latent variables or random effects and is a model of the NLMM type. The structural part describes the relations between the latent variables of the measurement part and a (possibly different) set of predictors. It offers the user a very general modelling framework that allows the construction of a broad spectrum of models. As for the NLMMs, some of the approaches discussed in this paper generalize to the structural equation model framework, but some are more specific to GLMMs (such as PQL and MQL).

Second, the main focus in this paper was on methods within the classical statistical or frequentist approach. However, in a Bayesian framework, one could rely on computer-intensive techniques, known as Markov chain Monte Carlo methods (Gelman *et al.*, 2003; Tanner, 1996), to sample from the posterior distribution of the parameters, and hence facilitate the parameter estimation of complex statistical models tailored to specific research questions. However, a full treatment of these methods would almost require a separate paper. Moreover, the Markov chain Monte Carlo methods are not yet implemented in the majority of the GLMM software packages.

Third, there are also methods that take into account the clustered structure of the data, but without introducing cluster-specific effects. The best known of these marginal methods is the generalized estimating equation (GEE) approach (Liang & Zeger, 1986; see also Johnson & Kim, 2004). In the GEE method, a model is proposed for the expected values of the measurements (the level 1 units; see equation (3)) and a working correlation matrix that captures the dependencies among the measurements within a cluster. The final estimates are found by solving the so-called generalized estimating equations. Moreover, robust estimates of the standard errors of the regression parameters can be found. The main advantage of the GEE approach is that it guarantees consistent estimates for the regression parameters, even if the working correlation matrix is misspecified.

A disadvantage of the GEE is the fact that it is based on the specification of the first and second moments. Therefore, there is no likelihood and thus no possibility of assessing the fit of the model or doing model selection. Recently, however, there have been various attempts to overcome the latter disadvantage (e.g. Hardin & Hilbe, 2003). A more serious disadvantage for applications in psychology and other social sciences is

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

that one is often interested in cluster-specific effects (e.g. when measuring persons) and they are not available in the GEE method. Heagerty and Zeger (2000) propose a new method that combines the advantages of a marginal model with the advantages of a mixed model. However, this 'marginalized multi-level' is relatively complex and is not yet implemented in standard software packages.

Acknowledgements

Francis Tuerlinckx and Frank Rijmen are partially supported by the Fund for Scientific Research – Flanders. Francis Tuerlinckx was also partially supported by grant SES00-84368 and Frank Rijmen by a postdoctoral grant from the Research Council and grant GOA/2000/02 from the University of Leuven.

References

Abramowitz, M., & Stegun, I. (1974). *Handbook of mathematical functions*. New York: Dover.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society, Series B*, *32*, 283–301.

Andersen, E. B. (1980). *Discrete statistical models with social science applications*. Amsterdam: North Holland.

Berkhof, J., & Snijders, T. A. B. (2001). Variance component testing in multilevel models. *Journal of Educational and Behavioral Statistics*, *26*, 132–152.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds), *Statistical theories of mental test scores* (pp. 397–424). Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor-analysis. *Applied Psychological Measurement*, *12*, 261–280.

Booth, J. G., & Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B*, *61*, 265–285.

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

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

Browne, W. J., & Draper, D. (2004). *A comparison of Bayesian and likelihood-based methods for fitting multilevel models*. Nottingham Statistics Research Report 04-01.

Bunday, B. D. (1984). *Basic optimisation methods*. London: Edward Arnold.

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

Conaway, M. (1990). A random effects model for binary data. *Biometrics*, *46*, 317–328.

Crouch, A. C., & Spiegelman, E. (1990). The evaluation of integrals of the form $\int f(t)\exp(-t^2)\mathrm{d}t$: Application to logistic-normal models. *Journal of the American Statistical Society*, *85*, 464–469.

Crowder, M. J. (1978). Beta-binomial anova for proportions. *Applied Statistics*, *27*, 34–37.

Daniels, M. J., & Zhao, Y. D. (2003). Modelling the random effects covariance matrix in longitudinal data. *Statistics in Medicine*, *22*, 1631–1647.

Davidian, M., & Giltinan, D. M. (1995). *Nonlinear models for repeated measurement data*. London: Chapman & Hall.

De Leeuw, J., & Verhelst, N. D. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics*, *11*, 183–196.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

Diggle, P. J., Heagerty, P. J., Liang, K.-Y., & Zeger, S. L. (2002). *Analysis of longitudinal data* (2nd ed.). Oxford: Oxford University Press.

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

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. London: Chapman & Hall.

Everitt, B. S. (1987). *Introduction to optimization methods and their application in statistics*. London: Chapman & Hall.

Fahrmeir, L., & Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models* (2nd ed.). New York: Springer.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall.

Gill, P. E., Murray, W., & Wright, M. H. (1981). *Practical optimization*. New York: Academic Press.

Glas, C. A. W., & Verhelst, N. D. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika*, *53*, 525–546.

Goldstein, H. (1991). Nonlinear multilevel models with an application to discrete response data. *Biometrika*, *78*, 45–51.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Edward Arnold.

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

Golub, G. H., & Welsch, J. H. (1969). Calculation of Gauss quadrature rules. *Mathematics of Computation*, *23*, 221–230.

Hall, D. B., & Præstgaard, J. T. (2001). Order-restricted score tests for homogeneity in generalised linear and nonlinear mixed models. *Biometrika*, *88*, 739–751.

Hardin, J. W., & Hilbe, J. M. (2003). *Generalized estimating equations*. Boca Rotan, FL: Chapman & Hall/CRC.

Heagerty, P. J., & Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference. *Statistical Science*, *15*, 1–26.

Hedeker, D (1999). *MIXNO: A computer program for mixed-effects nominal logistic regression.* Unpublished manuscript. http://tigger.uic.edu/%7Ehedeker/manuals.html.

Hedeker, D., & Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, *50*, 933–944.

Hedeker, D., & Gibbons, R. D. (1996). MIXOR: A computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine*, *49*, 157–176.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences*. Thousand Oaks, CA: Sage.

Jansen, M. G. H. (1994). Parameters of the latent distribution in Rasch's Poisson counts model. In G. H. Fischer & D. Laming (Eds.), *Contributions to Mathematical Psychology, Psychometrics, and Methodology*, (pp. 315–322). New York: Springer, .

Jiang, J. (1998). Consistent estimators in generalized linear mixed models. *Journal of the American Statistical Association*, *93*, 720–729.

Johnson, T. R., & Kim, J. S. (2004). A generalised estimating equations approach to mixed-effects ordinal probit models. *British Journal of Mathematical and Statistical Psychology*, *57*, 295–310.

Kleinman, J. (1973). Proportions with extraneous variance: Single and independent samples. *Journal of the American Statistical Association*, *68*, 46–54.

Kreft, I., & De Leeuw, J. (1998). *Introducing multilevel modeling*. London: Sage.

Laird, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, *73*, 805–811.

Lange, K. (2004). *Optimization*. New York: Springer.

Legler, J. M., & Ryan, L. M. (1997). Latent variable models for teratogenesis using multiple binary outcomes. *Journal of the American Statistical Association*, *92*, 13–20.

Lehman, E. L., & Casella, G. (1998). *Theory of point estimation* (2nd ed.). New York: Springer.

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

Lesaffre, E., & Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random-effects model: An example. *Applied Statistics*, *50*, 325–335.

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

Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika*, *84*, 309–326.

Lin, X., & Breslow, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, *91*, 1007–1016.

Lindsay, B. G., Clogg, C. C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models. *Journal of the American Statistical Association*, *86*, 96–107.

Liu, Q., & Pierce, D. A. (1994). A note on Gauss–Hermite quadrature. *Biometrics*, *81*, 624–629.

Longford, N. T. (1993). *Random coefficient models*. New York: Oxford University Press.

Longford, N. T. (1994). Logistic regression with random coefficients. *Computational Statistics and Data Analysis*, *17*, 1–15.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York: Wiley.

McFadden, D., & Train, K. (2000). Mixed MNL models for discrete response. *Journal of Applied Econometrics*, *15*, 447–470.

McLachlan, G. J., & Krishnan, T. (1997). The *EM algorithm and extensions*. New York: Wiley.

Meng, X. L., & van Dyk, D. A. (1997). The EM algorithm – an old folk song sung to a fast new tune. *Journal of the Royal Statistical Society, Series B*, *59*, 511–567.

Meng, X. L., & van Dyk, D. A. (1998). Fast EM implementations for mixed-effects models. *Journal of the Royal Statistical Society, Series B*, *60*, 559–578.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus User's guide* (3rd ed.). Los Angeles, CA: Muthén and Muthén.

Naylor, J. C., & Smith, A. F. M. (1982). Applications of a method for the efficient computation of posterior distributions. *Applied Statistics*, *31*, 214–225.

Nelder, J. A., & Mead, R. (1965). A simplex method for function optimization. *Computer Journal*, *7*, 308–313.

Neuhaus, J. M., Hauck, W. W., & Kalbfleisch, J. D. (1992). The effects of mixture distribution misspecification when fitting mixed effects logistic models. *Biometrika*, *79*, 755–762.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*, 1–32.

Pinheiro, P. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, *4*, 12–35.

Rabe-Hesketh, S., Pickles, A., & Skrondal, A (2001). *GLLAMM manual.* Technical report 2001/01. London: Department of Biostatistics and Computing, University of London.

Rasbash, J., Steele, F., Browne, W., & Prosser, B. (2005). *A user's guide to MLwiN (Version 2.0)*. University of Bristol: Centre for Multilevel Modelling.

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

Raudenbush, S. W., Bryk, A. S., & Congdon, R. T. (2001). *HLM 5*. Lincolnwood, IL: Scientific Software.

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.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, *8*, 185–205.

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

Copyright © The British Psychological Society
Reproduction in any form (including the internet) is prohibited without prior permission from the Society

SAS Institute. (1999). *SAS On-lineDoc* (Version 8) [software manual on CD-ROM]. Cary, NC: SAS Institute, Inc.

Schall, R. (1991). Estimation in generalised linear models with random effects. *Biometrika*, *78*, 719–727.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: Wiley.

Self, S. G., & Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, *82*, 605–610.

Silvapulle, M. J., & Silvapulle, P. (1995). A score test against one-sided alternatives. *Journal of the American Statistical Association*, *90*, 342–349.

Skellam, J. G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society, Series B*, *10*, 257–261.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling*. Boca Raton, FL: Chapman & Hall/CRC.

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

StataCorp. (2001). *Stata statistical software: Release 7*. College Station, TX: Stata Press.

Stiratelli, R., Laird, N., & Ware, J. H. (1984). Random-effects model for serial observations with binary response. *Biometrics*, *40*, 961–971.

Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed-effects model. *Biometrics*, *50*, 1171–1177.

Stram, D. O., & Lee, J. W. (1995). Correction to 'Variance components testing in the longitudinal mixed-effects model'. *Biometrics*, *51*, 1196.

Tanner, M. A. (1996). *Tools for statistical inference* (3rd ed.). New York: Springer.

Tierny, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*, 82–86.

Thissen, D (1991). *MULTILOG: User's guide*. Lincolnwood, IL: Scientific Software Inc.

Train, K. (2003). *Discrete choice methods with simulation*. Cambridge: Cambridge University Press.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer.

Verbeke, G., & Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics*, *60*, 254–262.

Verbeke, G., Spiessens, B., & Lesaffre, E. (2001). Conditional linear mixed models. *American Statistician*, *55*, 25–34.

Williams, D. A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics*, *31*, 949–952.

Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics*, *31*, 144–148.

Wolak, F. A. (1989). Local and global testing of linear and nonlinear inequality constraints in nonlinear econometric models. *Econometric Theory*, *5*, 1–35.

Wolak, F. A. (1991). The local nature of hypothesis tests involving inequality constraints in nonlinear models. *Econometrica*, *59*, 981–995.

Wolfinger, R. D. (1993). The *GLIMMIX SAS macro*. Cary, NC: SAS Institute, Inc.

Wolfinger, R. D., & Lin, X. (1997). Two Taylor-series approximation methods for nonlinear mixed models. *Computational Statistics and Data Analysis*, *25*, 465–490.

Wu, M. L., Adams, R. J., & Wilson, M. (2005). *Acer Conquest: Generalized Item Response Modelling Software* [computer software]. Melbourne, Victoria: Australian Council for Educational Research Ltd.

Received 17 June 2003; revised version received 31 August 2005