JOURNAL OF APPLIED MEASUREMENT, *6*(3), 273–288
Copyright© 2005

Mixed Model Estimation Methods for the Rasch Model

Frank Rijmen

Francis Tuerlinckx

Michel Meulders

Dirk J.M. Smits

Katalin Balázs

K.U. Leuven, Belgium

Mixed models take the dependency between observations based on the same person into account by introducing one or more random effects. After introducing the mixed model framework, it is explained, by taking the Rasch model as a generic example, how item response models can be conceptualized as generalized linear and nonlinear mixed models. Common estimation methods for generalized linear and nonlinear models are discussed. In a simulation study, the performance of four estimation methods is assessed for the Rasch model under different conditions regarding the number of items and persons, and the degree of interindividual differences. The estimation methods included in the study are: an approximation of the integral over the random effect by means of Gaussian quadrature; direct maximization with a sixth-order Laplace approximation to the integrand; a linearized approximation of the nonlinear model employing PQL2; and finally a Bayesian MCMC method. It is concluded that the estimation methods perform almost equally well, except for a slightly worse recovery of the variance parameter for PQL2 and MCMC.

Requests for reprints should be sent to Frank Rijmen, Department of Psychology, HCIV, Tiensestraat, 102, 3000 Leuven, Belgium, e-mail: frank.rijmen@psy.kuleuven.ac.be.

Over the last years, several authors have explained how the Rasch model (Rasch, 1960) and other common item response (IR) models can be conceptualized as generalized linear or nonlinear mixed models, see for example Agresti, Booth, Hobert, and Caffo (2000); De Boeck and Wilson (2004); Kamata (2001); Legler and Ryan (1997); and Rijmen, Tuerlinckx, De Boeck, and Kuppens (2003). Much less attention has been devoted to how well the estimation methods that have been developed for generalized linear and nonlinear mixed models perform when used to estimate IR models. Estimation performance is not a trivial matter, since most estimation methods for generalized linear and nonlinear mixed models rely on approximations to the likelihood. Moreover, some of these approximative methods are known to behave quite poorly under particular conditions (see e.g., Breslow and Clayton, 1993; Breslow and Lin, 1995; Rodriguez and Goldman, 1995). Therefore, it is worthwhile to assess the performance of the estimation methods for generalized linear and nonlinear mixed models in the specific context of IR models.

In the first part, we will shortly recapitulate the mixed model framework, including how IR models can be formulated as generalized linear or nonlinear mixed models. Next, we will turn to a description of the main estimation methods for generalized linear and nonlinear mixed models. In a simulation study, we assess the performance of four common estimation methods under different conditions regarding the number of items and persons, and the degree of interindividual differences. Throughout the paper, we focus on the Rasch model (Rasch, 1960), which takes a prominent place in educational measurement.

Mixed model framework

Mixed models (see e.g., Diggle, Heagerty, Liang, and Zeger, 2002; Verbeke and Molenberghs, 2000) are a collection of statistical tools that are well suited for analyzing clustered data, such as, for example, data from students nested within schools or repeated measurement data (measurements nested within participants). Modeling clustered data under a model that assumes independent observations is inappropriate because the observations on the subunits (students, measurements) within the same unit (school, participant) tend to be more homogeneous than the observations on subunits of different units. The heterogeneity between units can be taken into account by assuming that (some of) the parameters of the model are unit-specific. If it is assumed that those unit-specific parameters are random variables with a distribution defined over the population of units, the model is called a *random-effects* or *mixed model*. The parameters that are not unit-specific are called *fixed effects*.

Mixed models and related methods were first developed in the context of analysis of variance and regression analysis, leading to the linear mixed model. Other commonly used terms are multilevel models (Goldstein, 2003), hierarchical models (Raudenbush and Bryk, 2002), and random coefficient models (Longford, 1993). More recently, also generalized linear mixed models (Clayton, 1996; McCullogh and Searle, 2001; Zeger and Karim, 1991) and nonlinear mixed models (Davidian and Giltinan, 1995) were developed.

The mixed logistic regression model (Hedeker and Gibbons, 1994) belongs to the class of generalized linear mixed models. According to the model, conditional on the random effects, the observations are assumed to be independent Bernoulli observations with probability of success

$$\Pr\left(Y_{ni}=1|\beta_n\right)=\frac{\exp\left(\boldsymbol{x}_{ni}'\boldsymbol{\delta}+\boldsymbol{z}_{ni}'\boldsymbol{\beta}_n\right)}{1+\exp\left(\boldsymbol{x}_{ni}'\boldsymbol{\delta}+\boldsymbol{z}_{ni}'\boldsymbol{\beta}_n\right)}, \quad (1)$$

where

- Y_{ni} is the binary response variable for subunit i of unit n, $i = 1, \ldots, I_n$; $n = 1, \ldots, N$,
- $\boldsymbol{x}_{ni}$ is a known P-dimensional covariate vector for the P fixed effects,
- $\boldsymbol{\delta}$ is the P-dimensional parameter vector of fixed effects,
- $\boldsymbol{z}_{ni}$ is a known Q-dimensional covariate vector for the Q random effects, and

$\boldsymbol{\beta}_n$ is the Q-dimensional parameter vector of random effects for unit n.

The distribution of the random effects $\boldsymbol{\beta}_n$ is commonly assumed to be multivariate normal with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. The random effects have a mean of zero because they represent the deviations from the mean effects of the covariates, which are treated as fixed effects.

Now, according to the Rasch model, the probability that person n gives a correct answer on an item i is, conditional on the person's latent variable β_n:

$$\Pr(Y_{ni}=1|\beta_n)=\frac{\exp(\beta_n-\delta_i)}{1+\exp(\beta_n-\delta_i)}. \quad (2)$$

Assuming that β_n is not a fixed parameter but a random variable with a distribution that is defined over the population of the persons, the Rasch model is a mixed logistic regression model where the fixed effects correspond to the item parameters, and the random intercept to the latent variable of the Rasch model. The fixed-effects covariates are dummy variables coding to identify the items ($x_{ni}=-1$ for item i, $x_{ni}=0$ elsewhere).

Analoguous, other familiar IR models can be framed in a generalized linear (or nonlinear) mixed model framework (Rijmen et al., 2003).

Estimation

The Rasch model will be the central example for which various estimation methods will be explained. The generalization of the presented methods to more dimensions and other models is straightforward, however. In the introduction, we defined the person parameter as a random effect, but in a frequentist statistical framework, we may also view the β_ns as fixed effects. In the latter case, methods like maximum joint and maximum conditional likelihood (Wright and Stone, 1979) could be applied to obtain inferences about the δ_is (and about the β_ns as well for joint maximum likelihood). We will not discuss these two methods further here, and instead concentrate on maximum marginal likelihood (MML).

In the remainder, we consider the person-specific effects as independent random draws from a unidimensional normal population density with mean zero and variance σ^2, $g(\beta_n)=N(0, \sigma^2)$. The parameter σ^2 has to be estimated together with fixed-effects parameters (the δ_is for the Rasch model). The marginal likelihood is then formed as follows:

$$\begin{aligned}&\prod_{n=1}^{N}\Pr(\boldsymbol{y}_n)=\\&\prod_{n=1}^{N}\int\Pr(\boldsymbol{y}_n|\beta_n)g(\beta_n)d\beta_n=\\&\prod_{n=1}^{N}\int\left\{\prod_{i=1}^{I}\frac{\exp(\beta_n-\delta_i)}{1+\exp(\beta_n-\delta_i)}\right\}g(\beta_n)\,d\beta_n,\end{aligned} \quad (3)$$

where $\Pr(\boldsymbol{y}_n)$ is the joint probability of the response pattern obtained for person n. The likelihood in Equation 3 is maximized with respect to the δ_is and σ^2. This method of finding parameter estimates is known as the MML method.

The assumption of a parametric normal population density can be relaxed. The unobserved population distribution can be left completely unspecified (nonparametric maximum likelihood estimation; Laird, 1978) or it can be considered as a discrete probability distribution with masses located at fixed points (semi-parametric maximum likelihood estimation; Heinen, 1996). These two methods will not be treated in this paper. Instead, we will focus on the normal population density.

In the following, we will discuss some of the classical ways (i.e., within a frequentist framework) of finding parameter estimates for the item parameters. In a final subsection, Bayesian estimation will be covered as well. Note that we focus particularly on those techniques that are used later on in the simulation study.

Maximum Likelihood Estimation

In this subsection we give an overview of the most common maximum likelihood estimation methods that are used with an underlying normal random-effects distribution. For the Rasch model, the integral appearing in the marginal like-

lihood is intractable, meaning that it has no analytical or closed-form solution.

There are two general types of solutions to the problem of the intractable integral. The first one is to approximate the integral with numerical integration techniques. The second solution is to approximate the integrand, so that the integral of the approximation does have a closed-form solution. After either approximation is carried out, the problem becomes a standard optimization problem for which many introductory texts are available (e.g., Bunday, 1984; Everitt, 1987; Gill, Murray, and Wright, 1981; Lange, 1999).

Approximation of the Integral

In this method, the integral is approximated by means of a numerical integration rule and then this numerical approximation is maximized. In the unidimensional case, the integral is replaced by a single finite sum of rectangles that approximate the area under the integrand. Because the random effects are assumed to be normally distributed, usually the Gauss-Hermite (GH) quadrature (Abramowitz and Stegun, 1974) is chosen.

In a standard Gaussian quadrature approximation the nodes are rescaled (and recentered, but the recentering does not have an effect because of the zero population mean) such that they cover the range of the normal population distribution. However, this rescaling is identical for every person n, which is not always the most accurate thing to do because if the data for person n are extreme (e.g., almost all ones or zeros), the posterior distribution of β_n will deviate largely from the population distribution, which puts more mass in the region where the moderate β_n values are located. Consequently, it might be more appropriate to do a rescaling and recentering for each individual separately and that is the basic idea behind the adaptive GH quadrature. Details on this method can be found in Pinheiro and Bates (1995).

The adaptive Gaussian quadrature needs fewer quadrature nodes, because it is more concentrated in the informative region of the continuum. The price to pay is that in a single step of the optimization algorithm the individual recentering and rescaling has to be computed, which may be very time-consuming. That is the main reason why in many cases (as is done in this paper) the non-adaptive or regular Gaussian quadrature rule is used.

Several algorithms are available for the maximization of the approximated likelihood function obtained from the adaptive or non-adaptive Gaussian quadrature. Some well-known techniques are the Newton-Raphson and the Fisher scoring algorithms. Usually, the logarithm of the marginal likelihood is maximized to facilitate the computations.

Instead of maximizing the approximated (log)likelihood function directly, one can also use an indirect maximization technique. This means that the maximization problem is transferred to another function for which it can be shown that as a by-product of its maximization the (log)likelihood function is maximized too. A prominent example of an indirect maximization technique is the EM algorithm (McLachlan and Krishnan, 1997).

Approximation of the Integrand

The goal of approximating the integrand is to obtain an expression so that the integral of the approximated integrand has a closed-form solution. Two types of techniques will be discussed: Laplace's method and a class of methods called the quasi-likelihood methods.

Laplace's method. In Laplace's method (Tierny and Kadane, 1986), we take the integrand of the contribution of person n to the marginal likelihood, $\Pr(\boldsymbol{y}_n|\beta_n)g(\beta_n)$ which is then approximated by a normal distribution (using a quadratic Taylor series expansion of the logarithm of $\Pr(\boldsymbol{y}_n|\beta_n)g(\beta_n)$ about its maximum).

Raudenbush, Yang, and Yosef (2000) proposed an extension of Laplace's method by including also higher order terms in the Taylor approximation to the logarithm of $\Pr(\boldsymbol{y}_n|\beta_n)g(\beta_n)$. Raudenbush et al. (2000) argue that six terms are sufficient for an accurate approximation and therefore they call their method Laplace6.

Quasi-likelihood approaches. Because estimation methods for linear mixed models are well-established, several researchers have tried to approximate the models for categorical data by a linear mixed model so that the estimation methods for the latter class of models can be applied. The two most popular approaches here are the Penalized Quasi-Likelihood (PQL; Breslow and Clayton, 1993; Schall, 1991; Stiratelli, Laird, and Ware, 1984) and Marginal Quasi-Likelihood (MQL; Goldstein, 1991). There are also several extensions proposed for both methods (MQL2, PQL2, and corrected PQL).

The term quasi-likelihood methods is used because the methods require only the specification of a 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 the estimation of models where only means and variances are specified. Although PQL and MQL are classified here as methods that approximate the integrand, they can also be explained as linear approximations to the data. That is also the perspective we take for didactic reasons.

To start with, an observation is written as the sum of its mean and some random error. In a GLMM, the mean equals the response function which is equal to the inverse link function (see e.g., Fahrmeir and Tutz, 2001; McCullagh and Nelder, 1989). In both PQL and MQL, the response function is approximated by a linear Taylor series expansion. For PQL, the expansion is about the current estimates of the fixed-effect regression coefficients and about the empirical Bayes estimates for the random effects. For MQL, the expansion is also about the current estimates of the fixed-effect regression coefficients, but about 0 for the random effects. It can be shown that applying this linear approximation of the response function leads to a linear mixed model for a linear transformation of the original data. The parameters of this model can be estimated and the estimates can in turn be used to update the linear approximation. The algorithm cycles between these two steps until convergence.

Because the PQL and MQL methods rely on a number of approximations, for which it is hard to assess their accuracy, one should not be surprised to learn that the methods often do not work very well. The original data are transformed linearly in PQL and MQL, and then they are assumed to be normally distributed such that a linear mixed model estimation routine can be applied. Consequently, the methods will perform poorly when the original data are already far from normal, which is the case for binary data. This was confirmed by Breslow and Clayton (1993) in a simulation study for PQL and MQL. Rodriguez and Goldman (1995) came to the same conclusion for MQL. A bias towards zero was found for the fixed effects and/or variance component estimates. Breslow and Lin (1995) demonstrated an asymptotic downward bias also for the regression coefficients estimated with PQL.

Another factor affecting the quality of the quasi-likelihood methods is the number of items. The reason is that in a linear mixed model, the estimates of the fixed-effects parameters are only affected by the observations through the sufficient statistics of the fixed-effect regression parameters, which are linear combinations of the data. Hence, it is the distribution of the sufficient statistics that should be close to normal. For a large number of items or sufficiently continuous data, this is not problematic but for fewer items or with less sufficiently continuous data, the normality approximation can be poor. Thus, it is expected that PQL and MQL perform less well if the number of items is small, for binary data, or both.

Moreover, MQL uses a linear Taylor approximation about the current fixed effects and zeros for the random effects. As a consequence, if the random-effect variances are large, this method will lead to less accurate results. Rodriguez and Goldman (1995) have confirmed this hypothesis in a simulation study.

To overcome some of the problems of PQL and MQL, there have been attempts to improve both methods. PQL2 and MQL2 make use of second-order (as opposed to first-order) Taylor ex-

pansions of the response function. Breslow and Lin (1995), and Lin and Breslow (1996) present a bias-corrected version of PQL. Comparing the 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).

As a final disadvantage to the quasi-likelihood methods, note that the deviance measure produced by using PQL or MQL methods cannot be used for model testing. These methods are based only on assumptions about the first and second moment of the data (mean and variance), which is acceptable when estimating the parameters but not for testing. Testing the model requires reference distributions for the test statistics and for this purpose one must specify the full distribution of the data, not only their first and second moments.

Bayesian Estimation

As an alternative to the classical approach in which the marginal likelihood is optimized, one may also use a Bayesian approach (Gelman, Carlin, Stern, and Rubin, 2004) for estimating the parameters of item response models. Classical and Bayesian approaches differ in several respects: First, classical approaches clearly distinguish between fixed and random effects (parameters of interest and nuisance variables) whereas Bayesian approaches consider all effects to be essentially random. Bayesian approaches use sampling-based MCMC algorithms such as the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990) and the Metropolis algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller, 1953; Metropolis and Ulam, 1949) to obtain a sample of the entire posterior distribution of all model parameters in one stage. To guarantee the propriety of the posterior distribution it is necessary to specify a prior distribution for all model parameters. For person parameters and location parameters a normal distribution can be used.

Second, classical approaches yield standard errors of parameters that are based on an asymptotic normal approximation to the likelihood (or posterior) whereas Bayesian approaches yield posterior intervals of parameters that are also valid in small samples.

Third, monitoring convergence is more straightforward when using a classical approach than when using a Bayesian approach to approximate the entire posterior. A popular maximization algorithm such as the EM algorithm has the strong property of increasing the likelihood at each iteration and converging to a stationary point of the parameter space. For popular MCMC algorithms such as the Gibbs sampler it is known that, under mild regularity conditions, the simulated sequences converge to the true posterior distribution, but assessing whether convergence has been attained is a difficult problem (for an overview, see Cowles and Carlin, 1996).

Fourth, when using a classical approach for estimation, it is straightforward to check whether a necessary condition for (local) identifiability of the model is satisfied (i.e., evaluate whether the Hessian in the estimated point is negative definite). In contrast, such a check is not readily available when using a Bayesian approach for parameter estimation.

Simulation Study

A simulation study was set up in order to investigate the performance of different estimation methods for estimating the Rasch model. We considered eight conditions by crossing all levels of three factors: the number of persons (100 vs. 500), the number of items (5 vs. 25) and the variance of the latent trait (1 vs. 3). Two sets of true item parameters were chosen, one set for the conditions with five items and one set for the conditions with 25 items. The item parameters of each set ranged between minus two and two. Furthermore, the items were placed symmetrically around zero and were placed more densely the closer they were to zero, see Figure 1 displaying the item parameters and the distributions of the latent trait β_n. In addition, it can be seen that in conditions with $\sigma^2=1$, the item parameters approximately covered the

entire distribution of β_n, whereas in conditions with σ^2=3, the item parameters only covered the central part of the distribution.

For each condition 30 data sets were generated under the Rasch model using the true values of the item parameters and taking a random sample of person parameters from a normal distribution with mean equal to zero and the variance associated with the condition (i.e. 1 vs. 3).

In order to evaluate the performance of different algorithms for parameter estimation we computed a number of goodness-of-recovery (GOR) measures for each parameter in each condition. More specifically, we considered three GOR measures (Maris, 1999): First, the BIAS, defined as the difference between the average estimated parameter value and the true parameter value,

$$\mathrm{BIAS}(\xi_j) = \frac{\sum_{r=1}^{R} \hat{\xi}_{jr}}{R} - \xi_j,$$

where ξ_j denotes the true value of parameter j and $\hat{\xi}_{jr}$ the estimate of parameter j for replicated data set r (r = 1,..,R). Second, the root mean square deviation (RMSD), defined as the square root of the average squared deviations between the estimated parameter values and the true parameter values,

$$\mathrm{RMSD}(\xi_j) = \sqrt{\frac{\sum_{r=1}^{R} (\hat{\xi}_{jr} - \xi_j)^2}{R}}.$$

Finally, the standard error of the parameter estimate was estimated by the Monte Carlo standard error (MCSE), namely,

$$\mathrm{MCSE}(\xi_j) = \sqrt{\frac{\sum_{r=1}^{R} (\hat{\xi}_{jr} - \bar{\xi}_j)^2}{R}},$$

with $\bar{\xi}_j$ being the average estimate of parameter j in R replicated data sets. Note that for each parameter j, the above GOR measures are related in that $\mathrm{RMSD}^2 = \mathrm{MCSE}^2 + \mathrm{BIAS}^2$.

As it would be too much information to report the above GOR measures for each item parameter in each condition, we present average GOR measures for the item parameters. In particular, the following average measures are reported:

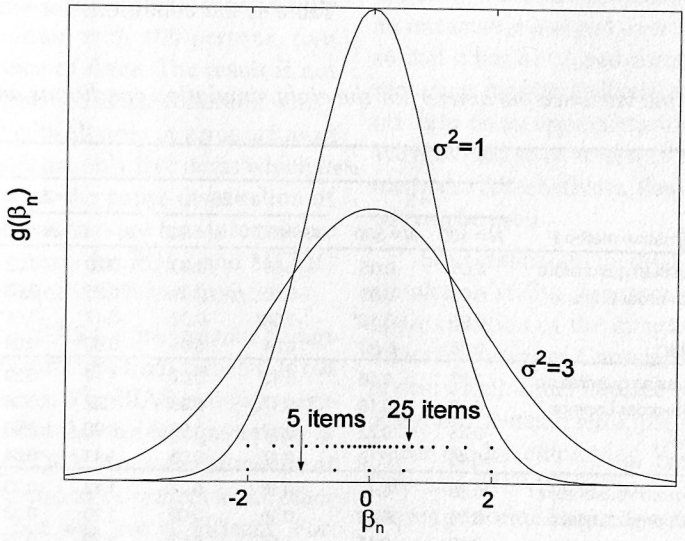


Figure 1. Item parameters (5 *vs.* 25) and distribution of the latent trait (normally distributed with mean zero and variance of 1 *vs.* 3)

(1) $\sum_i \text{BIAS}(\delta_i) / I$,

(2) $\sqrt{\sum_i (\text{RMSD}(\delta_i))^2 / I}$, and

(3) $\sqrt{\sum_i (\text{MCSE}(\delta_i))^2 / I}$,

with I being the number of items in the condition (5 vs. 25).

Four estimation techniques were employed in the study: direct maximization with an approximation of the integral over β_n by means of Gaussian quadrature over 20 nodes; a sixth-order Laplace approximation to the integrand; a linearized approximation of the nonlinear model employing PQL2; and finally a Bayesian MCMC method. The software packages used were the SAS procedure NLMIXED (SAS Institute Inc., 1999; Gaussian quadrature), HLM (version 5.04; Raudenbush, Bryk, and Congdon, 2001; sixth-order Laplace), MLwiN (version 1.10; Rasbash, Browne, Healy, Cameron, and Charlton, 2001; PQL2) and WinBUGS (version 1.2; Spiegelhalter, Thomas, and Best, 1999; MCMC). The Appendix contains a short description of how to estimate the Rasch model for each of the four software packages.

The GOR measures considered in this study arise from a frequentist statistical framework in which parameters are considered fixed but unknown quantities. In a Bayesian context on the other hand, parameters are considered to be random variables, as discussed in the estimation section. Instead of point estimates of the parameters, MCMC techniques supply a sample from the posterior distribution of the parameters. As a consequence, the GOR measures here considered can strictly speaking not be applied to Bayesian estimation methods. However, summary statistics of the posterior sample are often pragmatically treated as if they were point estimates of the parameters. In the same vein, we computed the GOR measures for the Bayesian MCMC method based on the posterior means for the item parameters and the posterior median for the variance parameters (the posterior median of the variance resulted in less BIAS than the posterior mean because the posterior distribution of the variance parameter was skewed to the right).

Results

Table 1 contains the GOR measures for the variance parameter for the different estimation methods discussed, and Table 2 the GOR measures for the item parameters. In the top panel of Table 1, the conditions for which the BIAS of

Table 1

GOR measures for the variance parameter for the eight simulation conditions and four estimation techniques.

		σ^2=1				σ^2=3			
		I = 5		I = 25		I = 5		I = 25	
	Estimation method	N = 100	N = 500	N = 100	N = 500	N = 100	N = 500	N = 100	N = 500
(overall) BIAS	Gaussian quadrature	0.06	-0.05	-0.05	-0.00	0.20	0.02	0.02	-0.07
	Sixth-order Laplace	0.06	-0.05	-0.05	-0.00	0.25	0.05	-0.01	0.03
	PQL2	-0.06	-0.16*	-0.08*	-0.02	-0.47*	-0.51*	-0.13	-0.11*
	MCMC	0.18*	-0.03	0.01	0.01	0.60*	0.09	0.17	0.06
(overall) RMSD	Gaussian quadrature	0.37	0.18	0.17	0.07	1.14	0.33	0.65	0.25
	Sixth-order Laplace	0.37	0.18	0.17	0.08	1.22	0.34	0.65	0.26
	PQL2	0.33	0.22	0.17	0.08	0.92	0.56	0.59	0.26
	MCMC	0.45	0.18	0.17	0.08	1.41	0.34	0.71	0.27
(overall) MCSE	Gaussian quadrature	0.36	0.17	0.16	0.07	1.12	0.33	0.65	0.24
	Sixth-order Laplace	0.36	0.18	0.16	0.09	1.20	0.33	0.65	0.26
	PQL2	0.32	0.15	0.16	0.07	0.79	0.23	0.57	0.23
	MCMC	0.41	0.18	0.17	0.08	1.28	0.33	0.69	0.26

*:BIAS(σ^2) $\neq$ 0, (p < .05)

the variance estimates significantly differed from zero ($p < .05$ in a one-sample t-test) are identified with an asterisk. In Table 2, the number of item parameters for which the BIAS significantly differed from zero ($p < .05$ in a one-sample t-test) is superscripted for each condition.

A first observation is that the item parameters are recovered well under all conditions. Because the BIAS was averaged over the items however, a bias towards zero for the estimates of the item parameters may pass unnoticed. It is possible that the BIAS for the negative and positive item parameters compensated for each other. Therefore, we inspected the box plots of the estimated minus the true item parameters for each item. For each condition, the box plots for all items were displayed in a single figure and sorted by the true item parameter, so that a bias towards zero could be read off easily. Figure 2 displays the boxplots for the condtion with 25 items, 100 persons and a variance of three for β_n. No bias towards zero for the estimates of the item parameters was found for any of the estimation methods. It seems that PQL2, contrary to PQL, does not suffer from a downward bias, even if the number of items is small.

Second, for all estimation techniques the recovery of the variance parameter is substantially worse in the condition with 100 persons, five items and a variance of three. The result is not unexpected, because this is the condition where response patterns with all ones or zeros are most likely to occur: there are only five items which in addition do not cover the entire distribution of β_n. Such response patterns are less informative with respect to β_n. Only for PQL2 and MCMC the BIAS is significantly different from zero.

For PQL2, the BIAS for the variance parameter is significantly different from zero in five out of the eight conditions. The BIAS is always negative which means that the variance parameter is underestimated by PQL2. An underestimation of the variance seems also to be present in the other three conditions, but was not significant. For MCMC on the other hand, a positive BIAS for the variance parameter is found in the two conditions with 100 persons and five items. These are the conditions that are the least informative with respect to β_n. A positive BIAS seems to be present in some other conditions as well, but was not significant. For sixth-order Laplace and Gaussian quadrature, the BIAS for the variance parameter never significantly differed from zero.

RMSD and MCSE show no systematic differences across estimation techniques. As can be expected, both measures decrease as a function of sample size, and, for the variance parameter, as a function of number of items.

In conclusion, all four estimation techniques perform equally well with respect to the item parameters. When interest also lies in the variance parameter, we recommend the Gaussian quadrature and the sixth-order Laplace estimation techniques.

Conclusion

In the first part, we explained how item response models fit into the generalized linear and nonlinear mixed model framework. We focused on the Rasch model, but the rationale is quite similar for other item response models (De Boeck and Wilson, 2004; Rijmen et al., 2003).

For most generalized linear and nonlinear mixed models, the (marginal) likelihood contains an intractable integral over the random effects, so that it has no closed-form solution. As a consequence, maximum likelihood estimation methods rely on an approximation to the likelihood. We provided an overview of such approximation methods. Alternatively, Bayesian MCMC methods can be used.

Four estimation methods were assessed in a simulation study: direct maximization with an approximation of the integral over the random effect by means of Gaussian quadrature over 20 nodes; a sixth-order Laplace approximation to the integrand; a linearized approximation of the nonlinear model employing PQL2; and finally a Bayesian MCMC method. All estimation techniques performed equally well with respect to the item parameters. For the variance parameter on the other hand, PQL2 and MCMC suffered, at least in some conditions of the simulation study, from a downward and upward bias, respectively.

Table 2

Overall GOR measures for the item parameters for the eight simulation conditions and four estimation techniques.

		σ^2=1				σ^2=3			
		I=5		I=25		I=5		I=25	
	Estimation method	N=100	N=500	N=100	N=500	N=100	N=500	N=100	N=500
Overall BIAS	Gaussian quadrature	0.01	0.00[2]	0.02[1]	0.01	0.04[2]	-0.00	0.04	0.01
	Sixth-order Laplace	0.01	0.00[2]	0.02[1]	0.01	0.04[2]	-0.00	0.05	0.02[3]
	PQL2	0.01	0.00[2]	0.02	0.01	0.04	-0.00[2]	0.05[1]	0.02[3]
	MCMC	0.01	0.00	0.01[2]	0.01	0.04[2]	-0.00	0.05	0.02[1]
Overall RMSD	Gaussian quadrature	0.30	0.13	0.25	0.11	0.36	0.15	0.35	0.17
	Sixth-order Laplace	0.30	0.13	0.25	0.11	0.36	0.15	0.31	0.15
	PQL2	0.29	0.13	0.25	0.11	0.31	0.15	0.26	0.14
	MCMC	0.32	0.13	0.25	0.11	0.40	0.15	0.31	0.15
Overall MCSE	Gaussian quadrature	0.30	0.12	0.25	0.11	0.34	0.14	0.35	0.16
	Sixth-order Laplace	0.30	0.12	0.25	0.11	0.34	0.14	0.30	0.14
	PQL2	0.29	0.12	0.24	0.11	0.31	0.13	0.30	0.14
	MCMC	0.31	0.12	0.25	0.11	0.36	0.14	0.31	0.14

[1, 2, ...] : number of item parameters with BIAS(δ_i) $\neq$ 0, (p < .05)

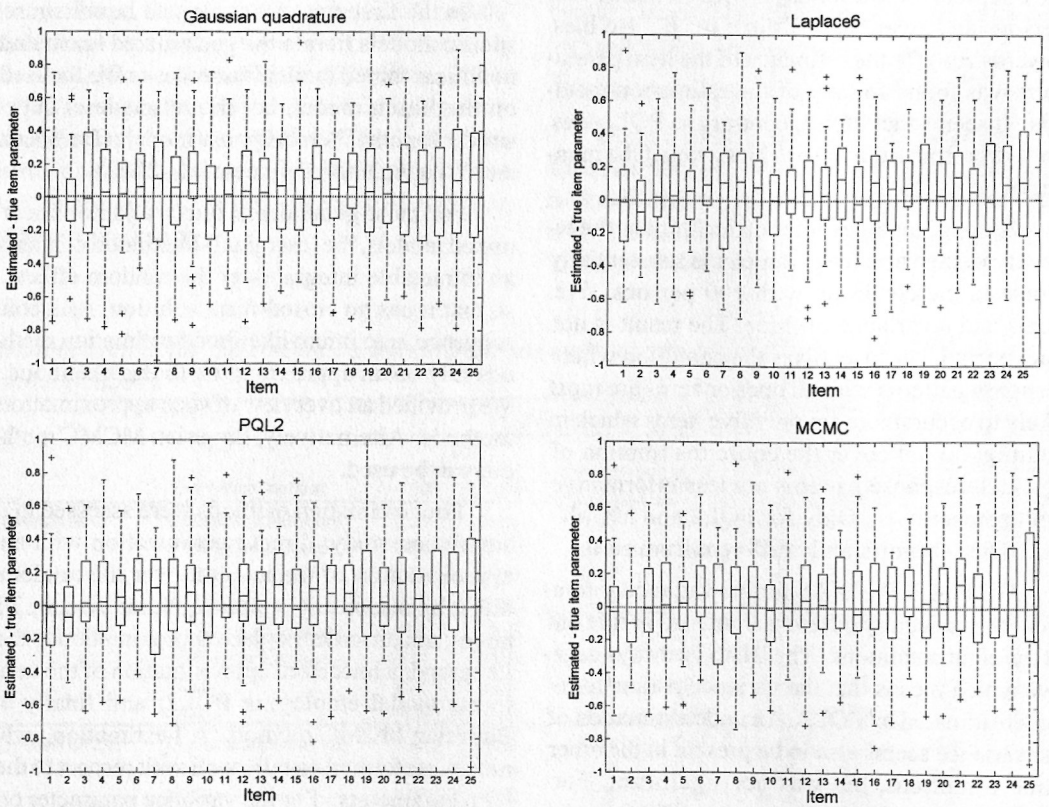


Figure 2. Boxplots of the estimated minus true item parameters. The boxplots are ordered from lowest to highest true item parameter (I = 25, N = 100, σ^2 = 3).

Acknowledgement

The research was supported by the Fund for Scientific Research Flanders (Francis Tuerlinckx) and by the following grants obtained from the K. U. Leuven: PDM/02/067, PDM/02/066, PDM/03/078, and GOA/2000/2 (other authors).

References

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

Agresti, A., Booth, J. G., Hobert, J. P., and Caffo, B. (2000). Random-effects modeling of categorical response data. *Sociological Methodology, 30*, 27-80.

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

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

Browne, W. J., and Draper, D. (2004). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. Manuscript submitted for publication.

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

Clayton, D. (1996). Generalized Linear Mixed Models. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo methods in practice* (pp. 275-302). New York: Chapman and Hall.

Cowles, K., and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association, 91*, 883-904.

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

De Boeck, P., and Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* New York: Springer-Verlag.

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

Everitt, B. S. (1987). *Optimisation methods and their applications in statistics.* London: Chapman and Hall.

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

Gelfand, A. E., and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85*, 398-409.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). London: Chapman and Hall.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*, 721-741.

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

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

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

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

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

Lange, K. (1999). *Numerical methods for statisticians*. New York: Wiley.

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

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*, 79-93.

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

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

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187-212.

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

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

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

Metropolis, N., and Ulam, S. (1949). The Monte Carlo Method. *Journal of the American Statistical Association, 44*, 335-341.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics, 21*, 1087-1092.

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

R Project (2002). The R project for statistical computing. www.r-project.org

Rasbash, J., Browne, W., Healy, M., Cameron, B., and Charlton, C. (2001). MlwiN version 1.10. [Computer program.] London: Institute of Education.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago: University of Chicago Press.)

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

Raudenbush, S. W., Bryk, A. S., and Congdon, R. T. (2001). HLM version 5.04. [Computer program.] Lincolnwood, IL: Scientific Software Inc.

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

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

Rodriguez, G., and 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.

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

Spiegelhalter, D. J., Thomas, A., and Best, N. G. (1999). *WinBUGS version 1.2 user manual*. Cambridge, United Kingdom: MRC Biostatistics Unit.

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

SAS Institute Inc. (1999). The SAS system for Windows Release 8.02. [Computer program.] Cary, NC: SAS Institute.

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

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

Wright , B. D., and Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

Zeger, S. L., and Karim, M. R. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association, 86*, 79-86.