British Journal of Mathematical and Statistical Psychology (2012), 65, 438–466
© 2011 The British Psychological Society



The British Psychological Society

www.wileyonlinelibrary.com

Parameter estimation of multiple item response profile model

Sun-Joo Cho[1]*, Ivailo Partchev[2] and Paul De Boeck[3,4]

[1]Vanderbilt University, Nashville, Tennessee, USA
[2]Cito, Arnhem, The Netherlands
[3]University of Amsterdam, The Netherlands
[4]K. U. Leuven, Belgium

Multiple item response profile (MIRP) models are models with crossed fixed and random effects. At least one between-person factor is crossed with at least one within-person factor, and the persons nested within the levels of the between-person factor are crossed with the items within levels of the within-person factor. Maximum likelihood estimation (MLE) of models for binary data with crossed random effects is challenging. This is because the marginal likelihood does not have a closed form, so that MLE requires numerical or Monte Carlo integration. In addition, the multidimensional structure of MIRPs makes the estimation complex. In this paper, three different estimation methods to meet these challenges are described: the Laplace approximation to the integrand; hierarchical Bayesian analysis, a simulation-based method; and an alternating imputation posterior with adaptive quadrature as the approximation to the integral. In addition, this paper discusses the advantages and disadvantages of these three estimation methods for MIRPs. The three algorithms are compared in a real data application and a simulation study was also done to compare their behaviour.

1. Introduction

The main purpose of this paper is to compare three methods for the estimation of a model for binary data stemming from a design with crossed random effects (persons and items) and crossed fixed effects (between persons and within persons). We will explain in the following why such a design is of interest and why binary data are considered.

Repeated measurement designs and tests have in common that persons are presented with a set of items. The term 'item' is commonly used for tests, but we will use it in a broader sense for all kinds of stimuli, situations, tasks, questions, etc., presented to persons in an experiment or in a test. The perspective we take here on items is to consider them as random. As a consequence, the models are models for *crossed random effects*, since the persons with whom the items are crossed are also treated as random. These

*Correspondence should be addressed to Dr. Sun-Joo Cho, Peabody College of Vanderbilt University, 230 Appleton Place, Nashville, TN 37203-5721, USA (e-mail: sj.cho@vanderbilt.edu).

DOI:10.1111/j.2044-8317.2011.02036.x

models are sometimes called multilevel models, because they share with the typical multilevel models that more than one mode of the data is treated as random: persons and groups in the typical multilevel case, and persons and items in the kind of design we are considering here. Strictly speaking they are not multilevel models because they do not necessarily have nested random effects (two or more levels of random effects), as explained by Baayen, Davidson, and Bates (2008).

Apart from crossed random persons and items, other ingredients of the type of design we are considering are a between-persons factor, stemming from a manipulation (in an experiment) or based on natural groups (e.g., gender), and a within-persons factor, stemming from a manipulation (in an experiment) or based on different types of items. The full design has two fixed factors, one differentiating between persons and the other differentiating within persons, and two random factors – persons and items.

The essential elements for the present study are that the items are treated as random and that the observations are binary. Treating items as random is not a new statistical approach. The model we are studying is a member of the broader family of generalized linear mixed models (McCulloch & Searle, 2001). While a random item approach is highly useful for psychological data of various kinds, it is still not commonly used.

For the context of *psychological experiments*, the random effects approach has been advocated in the psycholinguistic and perception literature. A special issue of the *Journal of Memory and Language* in 2008 is devoted to this approach (Forster & Masson, 2008). Treating items as fixed when one is interested in the items as a sample and not so much in the specific items leads to inflated Type I errors. One solution to this problem is the $F_{\min}$ statistic (Clark, 1973; Raaijmarkers, 2003; Raaijmarkers, Schrijnemakers, & Gremmen, 1999). A more elegant and model-based solution is to use mixed models with random person and random item effects (Baayen *et al.*, 2008; Jaeger, 2008; Quené, 2008). Covariates with fixed effects can be easily included in the model, as is illustrated for recognition data by Freeman, Heathcote, Chalmers, and Hockley (2010). The fixed factors of the design we are interested in can be seen as covariates.

The special issue of the *Journal of Memory and Language* gives special attention to binary data. Binary data are rather common in psychology: in the fields of memory (recognition, recall), psychophysics, perception, lexical and categorical decision. Whenever accuracy is important, binary data play a role, sometimes in combination with response time data, such as for the Stroop task, and in priming paradigms. When the data are binary, the link function of a generalized linear mixed model can be used to avoid artefacts stemming from a discrete and bounded scale. Item-level data can be analysed instead of aggregated data, such as proportions, as is rather common.

In a *test and item response theory* (IRT) context, treating the items as random makes sense for reasons explained, among others, by Cho and Rabe-Hesketh (2011) and De Boeck (2008). Sometimes items are literally drawn from a population, such as an item bank, or they are generated at random (Doran, Bates, Bliese, & Dowling, 2007), such as in an item generation paradigm (Glas & van der Linden, 2003; Johnson & Sinharay, 2005; Geerlings, Glas, & van der Linden, 2011). When an explanatory approach is used with item covariates, then it is not always realistic to limit the model to fixed effects on the item side, because it implies that a perfect explanation can be given. Including an error term, and hence considering items as random, solves this problem and is analogous to the common regression model (Janssen, Tuerlinckx, Meulders, & De Boeck, 2000; Janssen, Schepers, & Peres, 2004). Also from a hierarchical Bayesian approach, it is natural to treat items as random variables (Fox, 2010; Soares, Gonçalves, & Gamerman, 2009). Finally,

random item effects are a parsimonious way of modelling. Instead of many fixed item effects, only the item distribution parameters are needed.

The simplest case with all the ingredients we are interested in is a two-by-two design, with a between-persons factor and a within-persons factor. It can be considered as representative of more complex designs. There is no reason why the estimation method should be different or have a different quality depending on the number of fixed effect factors and the number of levels of the factors. The estimation quality would rather depend on aspects of the random effects (persons and items): the number of units and the variance–covariance structure. Therefore we concentrate on a variation of the latter for the simulation study.

The random effects models for the design we are interested in can be formulated with different kinds of parameterization. For the random items effects, an important choice is between (a) random intercept plus random slope, and (b) level-specific random effects. The choice applies to both the persons and the items, and is explained for the items as follows: (a) a random item intercept defines the general item effect, whereas the random item slope (random difference) defines the item-dependent difference between the levels of the between-person factor, and (b) level-specific random item effects are item effects per level of the between-person factor. Let us assume that the between-person factor is gender. The random item intercept defines the gender-independent item effect and the random slope defines the item-dependent difference between the two gender groups. The alternatives are gender-specific item effects, which are the effects per gender group. For example, if the data are test data, the two random item variables would be the item difficulties for men and for women. In this paper we focus on level-specific random effects, the equivalent of simple effects in an analysis of variance (ANOVA) context. First, level-specific random item effects reflect in a direct way within-person effects, and not group differences, so that a within-person interpretation of the phenomena and the processes involved is evident. Second, level-specific effects, for persons and for items, are the equivalent of a simple structure in factor analysis, with one random variable per level of the within-person factor for the individual differences, and one per level of the between-person factor for the item differences. In that way, each random variable corresponds to just one level. We use the term *item response profile* for the random item effect of one group. Because the design we are interested in has more than one group, we call the model (with the parameterization in question) the *multiple item response profiles* (MIRP) model (De Boeck & Wilson, 2010). For example, when a test is presented to men and women, then each group has its random item profile, and in an IRT context, these profiles would be the (random) item difficulties.

1.1. Model formulation

Figure 1 gives a presentation of the design with G person groups (the between-person factor, $g = 1, \ldots, G$) and K item types (the within-person factor, $k = 1, \ldots, K$). The random person effect is defined as $\theta_{g[p]k}$, with p ($p = 1, \ldots, P$) as a person index nested within g (denoted by $g[p]$) for the kth item type. $\boldsymbol{\theta}_{g[p]}$ follows a multivariate normal (MNormal) distribution with mean vector $\mathbf{0}$ for $k = 1, \ldots, K$, that is, $\boldsymbol{\theta}_{g[p]} = [\theta_{g[p]1}, \ldots, \theta_{g[p]k}, \ldots, \theta_{g[p]K}]' \sim \text{MNormal}(\mathbf{0}, \Sigma_\theta)$. The random item effect is defined as $\zeta_{k[i]g}$, with i ($i = 1, \ldots, I$) as an item index nested within k (denoted by $k[i]$) for the gth person group. $\boldsymbol{\zeta}_{k[i]}$ follows a MNormal with mean vector $\mathbf{0}$ for $g = 1, \ldots, G$, that is, $\boldsymbol{\zeta}_{k[i]} = [\zeta_{k[i]1}, \ldots, \zeta_{k[i]g}, \ldots, \zeta_{k[i]G}]' \sim \text{MNormal}(\mathbf{0}, \Sigma_\zeta)$.

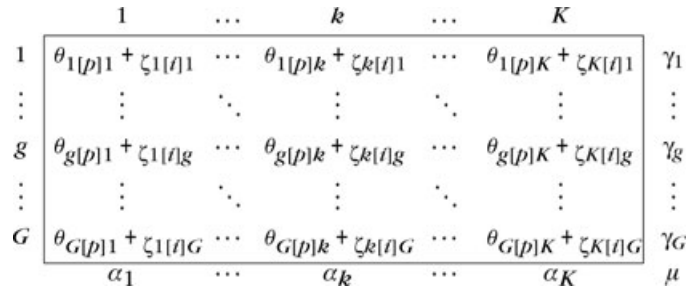


Figure 1. MIRP model configuration.

The MIRP model, taking the person group and item type fixed effects and their interactions into account, is specified as:

$$\text{logit}\left[\Pr_{g[p]k[i]}(y_{pi} = 1)\right] = \mu + \alpha_k + \gamma_g + \lambda_{kg} + \theta_{g[p]k} + \zeta_{k[i]g}, \tag{1}$$

where $\Pr_{g[p]k[i]}(y_{pi} = 1)$ is the probability of having $y_{pi} = 1$ for a person p nested within a person group g and an item i nested within an item type k, μ is an overall mean, α_k is an item type effect, γ_g is a person group effect, and λ_{kg} is an interaction effect between the item type and the person group. The following model identification constraints are imposed on parameters: (1) $\sum_{k=1}^{K} \alpha_k = 0$, (2) $\sum_{g=1}^{G} \gamma_g = 0$, (3) $\sum_{k=1}^{K} \lambda_{kg} = \sum_{g=1}^{G} \lambda_{kg} = 0$, and (4) means of random effects are $\mathbf{0}$.

1.2. Some possible applications

Applications of the MIRP model include the following:

Explanation. It can be of great help for the understanding of test data when an explanatory model is set up to explain the item difficulties (De Boeck & Wilson, 2004). This can be realized through the inclusion of item covariates (the α_k in equation (1)). The fixed effects of these covariates almost never provide a perfect explanation. It is therefore useful to have a residual random term ($\zeta_{k[i]g}$ in equation (1)) to reflect the unexplained part. This term makes the explanatory model formally equivalent to a random item model. Because the explanatory effect may differ depending on the person group, it helps to have a separate random item variable per person group. The covariance structure of the random terms (Σ_ζ) is a helpful way to check whether there are shared item covariates with explanatory power which are not included in the model. Covariates are also meaningful in an experimental context – for example, to investigate the effect of word characteristics, such as familiarity, in a psycholinguistic experiment – and also in such a context a random residual term may be needed.

Qualitative differences. The MIRP model is an interesting way to capture qualitative differences between groups, either natural groups or differentially treated groups, and also between levels of a within-person factor, such as pre-test and post-test. The model nicely separates quantitative effects as reflected in the fixed effects (α_k, γ_g, and λ_{kg} in equation (1)) from qualitative effects as reflected in the imperfect correlations between the random effects ($\theta_{g[p]k}$ and $\zeta_{k[i]g}$ in equation (1)). In terms of ANOVA for repeated measures, the compound symmetry assumptions are relaxed.

Measurement invariance. The variance–covariance matrix, Σ_ζ, allows for lack of measurement invariance. Measurement invariance would mean that one random item

profile suffices for all person groups. Therefore, the MIRP model can be used to test measurement invariance and to deal with lack of it. Given that the items belong to one test, the MIRP model is a well-suited approach to differential test functions. It is a covariance alternative to the variance approach for differential test functioning (Camilli & Penfield, 1997; Longford, Holland, & Thayer, 1993; Penfield & Algina, 2006).

Even though the use of crossed random effects models is increasing, there is a lack of research on the estimation methods. The purpose of this paper is to investigate three distinct estimation methods for the MIRP version of the crossed random effects model:

(1) the Laplace approximation to the integrand;
(2) hierarchical Bayesian analysis;
(3) an alternating imputation posterior (AIP) with adaptive quadrature as the approximation to the integral.

In Section 2, results from the three algorithms are compared for a real data example. The different estimation methods are further described and compared in Section 3. In Section 4, a simulation study is presented to examine their performance. We end with a summary of results and discussion in Section 5.

2. Empirical example: Verbal aggression data

In this section, an empirical example is given to illustrate how the MIRP model can be used for explanation, qualitative differences, and measurement invariance, and to illustrate the three different estimation methods.

2.1. Data description and model

The empirical data set is from Smits, De Boeck, and Vansteelandt (2004). The participants were 316 first-year psychology students from a university in the Dutch-speaking part of Belgium. Participation in the study was a partial fulfilment of the requirement to participate in research. The sample consists of 73 males and 243 females. The average age was 18.4 (standard deviation 1.2). There are 90 items with 3 within-person factors in the design: 15 (situations) $\times$ 3 (reactions) $\times$ 2 (modes). All items are based on a frustrating situation, (e.g., 'A bus fails to stop for me'), and a reaction. There are three kinds of reaction (cursing, scolding, shouting), and two modes of reacting (wanting, doing). For example, the wanting mode for cursing sounds as follows for the example situation: 'A bus fails to stop for me. I would want to curse', while the doing mode is formulated as 'A bus fails to stop for me. I would curse.' The two formulations are used in order to study verbal aggression and its inhibition, the discrepancy between wanting and doing. Two behaviour reaction item covariates are created, 'blaming' (cursing and scolding) and 'expressing' (cursing and shouting).

The following model will be estimated:

$$\text{logit}\left[\Pr_{g[p]k[i]}(y_{pi}=1)\right] = \mu + \alpha_k + \gamma_g + \lambda_{kg} + \beta_1 + \beta_2 + \theta_{g[p]k} + \zeta_{k[i]g}, \qquad (2)$$

where $\Pr_{g[p]k[i]}(y_{pi}=1)$ is a probability of having $y_{pi}=1$ for person p nested within person group g ($g=1$ and $g=2$) and item i nested within item type k ($k=1$ and $k=2$), μ is an overall mean, α_k is an item type effect ('want' item group and 'do' item group), γ_g is a person group effect ('female' person group and 'male' person group),

λ_{kg} is an interaction effect between the item type and the person group, β_1 is a fixed item effect for 'blaming', β_2 is a fixed item effect for 'expressing', $\theta_{g[p]k}$ is a random person effect, and $\zeta_{k[i]g}$ is a random item effect. The fixed effects express how much more or how much less verbally aggressive the respondents report being as a function of the item covariates and their gender. The person random effects take the individual differences within gender groups into account and the random item effects do the same for item differences within item types. The covariance structure of the person random effects inform us about the qualitative difference between wanting and doing, and the item random effects (item profiles) inform us about the qualitative differences between men and women with respect to verbal agression.

The covariates for parameters are coded as follows: an item type coded as 'want' $= -1$ and 'do' $= 1$ for α_k; a person group effect coded as 'female' $= -1$ and 'male' $= 1$ for γ_g; 'blaming' coded as cursing and scolding $= 1/2$ and shouting $= -1$ for β_1; and 'expressing' coded as cursing and shouting $= 1/2$ and scolding $= -1$ for β_2.

The following research questions are of interest for the verbal aggression data set using the MIRP model:

(1) Is there any interaction between 'want' and 'do' items and between 'female' and 'male'? (Is λ_{kg} significant?)
(2) Is there any quantitative difference between 'want' and 'do' items? (Is α_k significant?)
(3) Is there any quantitative difference between 'female' and 'male'? (Is γ_g significant?)
(4) Do 'blaming' and 'expressing' as features of a verbally aggressive reaction (cursing, scoling, shouting) make a quantitative difference? (Are β_1 and β_2 significant?)
(5) Is there any qualitative difference between the persons' 'want' mode and 'do' mode? (Is the correlation between $\theta_{g[p]1}$ and $\theta_{g[p]2}$ high?)
(6) Is there any qualitative difference in item difficulty profiles between 'female' and 'male'? (Is the correlation between $\zeta_{k[i]1}$ and $\zeta_{k[i]2}$ high?)

The effects of 'want' and 'do' (question 2) and of 'blaming' and 'expressing' (question 4) illustrate the explanatory value of the model. The correlations between 'want' and 'do' (question 5) and between 'female' and 'male' (question 6) illustrate how qualitative differences can be investigated. Finally, question 6 is also relevant for the issue of measurement invariance.

2.2. *Analyses*

The R *lmer* function (Bates & Maechler, 2009) code for the Laplace estimation of the model is shown in Appendix A, and the WinBUGS (Spiegelhalter, Thomas, & Best, 2003) code for Markov chain Monte Carlo (MCMC) is presented in Appendix B. The Stata do file for AIP with adaptive quadrature is available from the first author upon request. Four chains were implemented with 10,000 iterations in the MCMC analysis (with a burn-in of 4,000). Convergence checking was performed in WinBUGS using the Gelman and Rubin statistic (Gelman & Rubin, 1992) along with the condition that $\sqrt{\hat{R}}$ is less than 1.001. In AIP with adaptive quadrature analysis, 10 quadrature points were used for the estimation. A burn-in of 5 was set, and 15 used to calculate the posterior moments for AIP with adaptive quadrature.

2.3. Results

Results from the three methods are reported in Table 1. Unlike the Laplace approximation and AIP, MCMC provides a sample of the entire posterior distribution of all model parameters. Posterior mean, median, standard deviation (SD), and 95% credibility interval are shown in the table. Estimates of the fixed effects are similar across the three estimation methods. All fixed effects are statistically significant except the gender group effect, γ_g, based on z-test and 95% interval testing for Laplace and AIP and credibility interval testing for MCMC. The estimates of the variances of the random effects and their standard errors are similar for MCMC and AIP with adaptive quadrature. The Laplace approximation variance estimates are slightly smaller. Since standard errors for the estimates of variance components are not provided in the *lmer* function for the Laplace approximation, it is hard to tell how reliable the difference between the methods is. To test the significance of the distribution parameters, one can rely on the 95% credibility interval based on the posterior as derived from the iterations when using MCMC. AIP does provide standard errors for the distribution parameters of the random effects, unlike the Laplace approximation using the *lmer* function. However, it is not recommended to create a t- or z-statistic to test the significance of variance estimates since the distribution of variance estimates of random effects is not symmetric.

From the estimated fixed effects, it can be inferred that the effect of gender is not statistically significant (γ estimate), that people say they do less than they want with respect to verbal aggression (α estimate), and that this is especially true for women (λ estimate). Furthermore, the blaming and the expressive nature of the responses both have a positive effect on the probability of the corresponding behaviours (β_1 and β_2 estimates, respectively), but blaming is more popular than expressing as a reaction to frustrating situations. The fixed effects express a quantitative difference in the tendency

Table 1. Estimates for verbal aggression data

	MCMC					Laplace		AIP	
	Mean	Median	SD	2.50%	97.50%	Est	SE	Est	SE
Fixed part									
μ [Intercept]	−0.445*	−0.445*	0.127	−0.692	−0.182	−0.435*	0.124	−0.447*	0.130
α [Want versus Do]	−0.282*	−0.278*	0.092	−0.468	−0.117	−0.296*	0.093	−0.280*	0.091
γ [Female versus Male]	−0.028	−0.026	0.089	−0.200	0.141	0.026	0.093	−0.030	0.090
λ [Interaction]	0.199*	0.198*	0.045	0.111	0.290	0.201*	0.042	0.200*	0.047
β_1 [Blaming]	1.214*	1.244*	0.146	0.879	1.468	1.211*	0.136	1.210*	0.150
β_2 [Expressing]	0.487*	0.488	0.146	0.188	0.749	0.437*	0.135	0.485*	0.151
Random part									
$\sigma^2_{\theta_1}$ [Want]	2.064*	2.054*	0.189	1.722	2.466	2.030		2.060	0.187
$\sigma^2_{\theta_2}$ [Do]	2.107*	2.095*	0.197	1.757	2.527	2.076		2.100	0.201
$\rho(P)$ [Person correlation]	0.742*	0.743*	0.030	0.678	0.797	0.745		0.740	0.035
$\sigma^2_{\zeta_1}$ [Female]	0.846*	0.832*	0.141	0.610	1.167	0.801		0.841	0.152
$\sigma^2_{\zeta_2}$ [Male]	0.612*	0.602*	0.111	0.426	0.863	0.537		0.610	0.120
$\rho(I)$ [Item correlation]	0.888*	0.891*	0.030	0.822	0.937	0.935		0.880	0.031

*Statistically significant.

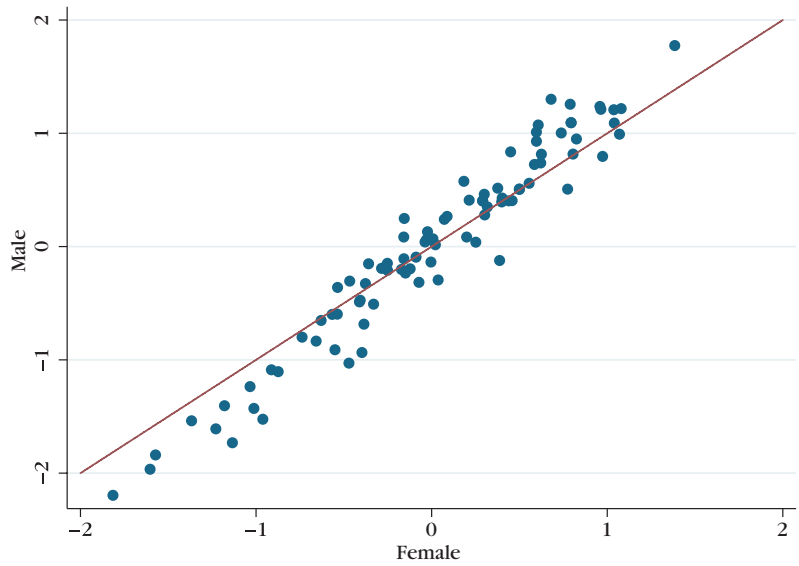


Figure 2. Measurement invariance across gender.

to react in a certain way, and they correspondingly also provide an explanation of the tendency to react with verbal aggression.

The correlations of the random item effects, on the other hand, indicate qualitative differences. It can be concluded from the correlation of .745 (Laplace estimate) between the random person effects for wanting and doing, that the difference between the two is not just quantitative (fixed effect Laplace estimate of −0.296). Although the random item profiles of men and women are highly correlated (.935 based on the Laplace estimation), the Akaike information criterion (AIC: Akaike, 1974) and the Schwarz Bayesian information criterion (BIC: Schwarz, 1978) values for the Laplace estimation with only one random item profile are higher than for the Laplace estimation with two random item profiles. This implies that there is a lack of measurement invariance and a qualitative difference in verbal aggression between men and women, as shown in Figure 2 for the maximum a posteriori estimates of the item random effects. Not all estimates are on the straight line, indicating some lack of invariance, and therefore some small qualitative differences.

3. MIRP model estimation

In this section, the different estimation methods for the MIRP model are reviewed and compared.

3.1. Estimation methods

Two random effects (i.e., $\theta_{g[p]k}$ and $\zeta_{k[i]g}$) in equation (1) are crossed, not nested. Maximum likelihood estimation (MLE) of models for categorical data with crossed random effects is challenging. This is because the marginal likelihood does not have a closed form, so that MLE requires numerical or Monte Carlo integration. If the random

effects are nested, the integrals are also nested (e.g., Rabe-Hesketh, Skrondal, & Pickles, 2005), keeping the computational burden low. Models with crossed random effects can be reformulated as models with nested random effects (Goldstein, 1987; Rasbash & Goldstein, 1994), but this approach requires evaluation of high-dimensional integrals and is therefore computationally demanding.

Approximations to MLE have been proposed to avoid high-dimensional numerical integration in generalized linear mixed models, including marginal quasi-likelihood (MQL: Goldstein, 1991), penalized quasi-likelihood (PQL: Breslow & Clayton, 1993) and its second-order improvement (PQL-2: Goldstein & Rasbash, 1996), bias-corrected PQL (Breslow & Lin, 1995; Lin & Breslow, 1996), Laplace approximations (Tierny & Kadane, 1986; Pinheiro & Bates, 1995; Raudenbush, Yang, & Yosef, 2000), and the hierarchical likelihood method (Lee & Nelder, 1996, 2006). However, both MQL and PQL perform poorly for binary responses with small cluster sizes, with a downward bias in the estimated variance components (Goldstein & Rasbash, 1996; Raudenbush, *et al.*, 2000; Rodríguez & Goldman, 1995). Although PQL-2 performs considerably better than PQL, this downward bias often remains a problem (Breslow, 2004; Browne & Draper, 2006; Rodríguez & Goldman, 2001). Joe (2008) found similar results for the Laplace approximation for binary responses with small cluster sizes. Diaz (2007) found that the higher-order Laplace approximation proposed by Raudenbush *et al.* (2000) reduces the bias of PQL but increases the mean squared error.

It is easy to incorporate crossed random effects by imposing hyperpriors on variances of parameters in hierarchical Bayesian analysis using MCMC (Karim & Zeger, 1992; Rasbash & Browne, 2007). However, MCMC is computationally expensive, and it may be difficult to specify vague hyperpriors for the variance parameters in hierarchical models that result in a posterior mean (or mode) close to the MLE (Browne & Draper, 2006; Natarajan & Kass, 2000). When the number of higher-level units is small, choice of prior distribution becomes more important (Lambert, 2006).

Similar to MCMC, the AIP algorithm (Clayton & Rasbash, 1999) makes use of data augmentation. The aim is to obtain *approximate* MLEs. The algorithm alternates between an item wing in which item difficulties are sampled for given person abilities and a person wing in which person abilities are sampled for given item difficulties, by sampling from the respective conditional posterior distributions. However, instead of drawing individual model parameters from their posterior distributions as in MCMC, parameters are estimated by maximum likelihood in the person wing (for given item parameters), and then sampled from their estimated sampling distribution. Similarly, parameters are estimated by maximum likelihood in the item wing (for given person parameters). Each MLE involves only one random effect and can therefore be accomplished relatively easily.

Unlike MCMC, the AIP algorithm does not require specification of prior distributions for the model parameters. Furthermore, the algorithm typically converges much more rapidly because several model parameters are updated simultaneously. Clayton and Rasbash (1999) used MQL and PQL within the AIP algorithm, as implemented in MLwiN (Goldstein *et al.*, 1998). As discussed above, MQL and PQL sometimes underestimate the variance components. Clayton and Rasbash (1999) found that this problem can also occur when MQL and PQL are used within their AIP algorithm. For a simulated data set, AIP with PQL-2 produced a variance estimate that was just over half the true value of the MCMC estimate. Cho and Rabe-Hesketh (2011) therefore developed an AIP algorithm that uses MLE with adaptive quadrature (Bock & Schilling, 1997; Pinheiro & Bates, 1995; Rabe-Hesketh, Skrondal, & Pickles, 2005). Adaptive quadrature is an improvement of

Gauss–Hermite quadrature (Bock & Liberman, 1970; Butler & Moffitt, 1982; Hedeker & Gibbons, 1994) that performs well in a wide range of situations (Rabe-Hesketh *et al.*, 2005).

The purpose of the Laplace approximation and AIP with adaptive quadrature estimation methods is to obtain the approximate MLE. However, the approximate MLE is from the approximation to the integrand with Laplace approximation, while it is from the approximation to the integral with AIP with adaptive quadrature. Thus, the approximate MLE from AIP with adaptive quadrature is likely to produce more precise estimates unless the higher-order approximation is used in Laplace approximation. The main goal of Bayesian inference is to obtain the posterior moments by sampling parameters from the posterior distribution based on both the prior distributions of parameters and the likelihood. The MLE approaches are based on asymptotic (large-sample) theory, while sampling-based Bayesian approaches are not based on asymptotic theory. Unlike the Laplace approximation, AIP with adaptive quadrature shares the feature of the Bayesian inference that parameters are sampled using MLEs based on the likelihood. An important difference between the hierarchical Bayesian analysis and the AIP algorithm with adaptive quadrature is that the former makes use of a prior for sampling parameters, while the latter does not.

The comparison of the three methods can be summarized as follows. First, the Laplace approximation is an approximation to MLE. Second, the hierarchical Bayesian analysis is based on sampling. And third, the AIP method is based on sampling and MLE with adaptive quadrature. These main differences lead to further differences regarding: (1) memory problems due to high-dimensionality, (2) computational efficiency, (3) standard error calculation and significance testing, (4) the estimates of random effects, (5) 'uncertainty' indications of the parameters, and (6) model comparison. These six differences are described below.

(1) Since the MIRP model involves high-dimensional integration, memory usage can be greatly increased. The Laplace approximation avoids integration by approximating the integrand. In hierarchial Bayesian analysis, it is possible to sample complex and high-dimensional posterior densities by MCMC methods such as the Gibbs sampler (Geman & Geman, 1984) through sampling from the conditional distributions of parameters, so that memory usage is rather modest. In the AIP with adaptive quadrature, the MIRP model is divided into sub-models depending on the kind of random effect. Since each sub-model has only one kind of random effect (persons or items), memory problems can be avoided.

(2) The Laplace approximation was computationally very efficient since numerical integration is avoided. However, both MCMC and AIP with adaptive quadrature may exhibit computational inefficiency since the MCMC and the part of AIP with adaptive quadrature are sampling-based methods. AIP with adaptive quadrature typically converges much more rapidly than MCMC since several model parameters are updated simultaneously. However, AIP with adaptive quadrature may suffer from computational inefficiency since adaptive multidimensional quadrature points are evaluated.

(3) Since the the distribution of the MLE estimate is normal based on asymptotic theory, the standard error and confidence interval of the MLE estimate can be constructed based on the symmetric distribution. However, Bayesian approaches provide the whole posterior distribution of a parameter, which is not assumed to be a normal distribution. Standard errors come out automatically from the MCMC simulations. Bayesian confidence intervals (credibility intervals) are based on the percentiles of

the posterior, and this allows for a strongly skewed distribution. Often the confidence interval based on MLE and Bayesian credibility intervals have essentially the same value, but the interpretational difference remains. The key point is that the Bayesian prior allows us to make direct probability statements about a parameter, while under classical statistics we can only make statements about the behaviour of the statistic if we repeat an experiment a large number of times. In Laplace approximation as implemented in the R *lmer* function, a widely used software package, standard errors of variance components are not provided. In AIP, both within variance (squared standard errors of estimates) and between variance (variance across iterations after burn-in) are available to calculate the standard errors as well as for variance estimates.

(4) Technically speaking, 'estimates' of the random effects are not provided because they are not parameters in the Laplace approximation and the AIP with adaptive quadrature. Point estimates of random effects are obtained for both these methods via empirical Bayes prediction after estimates of the fixed effects and population parameter estimates of the random effects are obtained. In contrast, a sample of the entire posterior distribution of all model parameters can be obtained in one stage in MCMC.

(5) The Laplace approximation does not provide the user with inferential 'uncertainty' indications on the variance estimates for the distribution of the random effects. MCMC and AIP with adaptive quadrature yield simulations (i.e., different estimates over iterations) representing inferential uncertainty about all parameters in the model.

(6) When MLE is used, the likelihood ratio test (LRT) or information-based statistics, such as AIC and BIC, are appropriate ways to compare models. Comparisons of models estimated via quasi-log-likelihood, such as the Laplace approximation, can be problematic, since quasi-likelihood is an approximation to the integrand. The problem is likely to become less of an issue as the approximations employed improve. For a (hierarchical) Bayesian analysis uniquely developed methods are available. They include the pseudo-Bayes factor (Geisser & Eddy, 1979; Gelfand & Dey, 1994), the deviance information criterion (Spiegelhalter, Best, & Carlin, 1998), and posterior predictive model checks (Gelman, Carlin, Stern, & Rubin, 2004). The AIC and BIC are appropriate when maximum likelihood estimates of model parameters are obtained, and they are also used in Bayesian analysis; see Congdon (2003) for examples of such usage in Bayesian analysis. In AIP with adaptive quadrature, the marginalized log-likelihood can be obtained by a combination of adaptive quadrature integration and Monte Carlo integration. Thus, the marginalized log-likelihood in AIP with adaptive quadrature is a more 'exact' log-likelihood than the quasi-likelihood in Laplace approximation. The 'asymptotic' reference distribution of a χ^2 does not apply for the variance parameters being tested against null because zero variance is the boundary of the parameter space. The mixture χ^2 distribution can be used in this case (Stram & Lee, 1994).

3.2. MIRP model parameter estimation

In this subsection, three different estimation methods are described to estimate the MIRP model parameters.

3.2.1. Laplace approximation

The *lmer* function was used for the Laplace approximation of the integrand. It is known that approximation routines of the *lmer* function tend to work well when the sample size and number of groups in multilevel modelling are moderate to large (Gelman & Hill, 2007).

In this section, the implementation of the *lmer* function is described for the MIRP model. The likelihood can be expressed as follows:

$$\begin{aligned} &L(Y|\mu, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \Sigma_\theta, \Sigma_\zeta) \\ &\quad = \prod_p \prod_i \int_{\mathbb{R}^P} \int_{\mathbb{R}^I} f(\boldsymbol{\theta}) f(\boldsymbol{\zeta}) \cdot \Pr_{g[p]k[i]}(y_{pi} = 1)^{y_{pi}} \\ &\qquad \cdot (1 - \Pr_{g[p]k[i]}(y_{pi} = 1))^{1-y_{pi}} d\boldsymbol{\zeta}\, d\boldsymbol{\theta}. \end{aligned} \tag{3}$$

Unfortunately the integral in equation (3) does not have a closed-form solution with the logit link. The *lmer* function uses the second-order Taylor series approximation to the log of the integrand at the conditional modes of the random effects from the penalized iteratively reweighted least squares (PIRLS) algorithm.

The PIRLS algorithm combines the characteristics of the iteratively reweighted least squares (IRLS) algorithm for generalized linear models (McCullagh & Nelder, 1989) and the penalized least squares representation of a linear mixed model (Bates & DebRoy, 2004).

If the data are sorted by item responses within persons, this allows us to split the covariance matrix of item responses into two components: a component for persons and a component for items. The crossed random effects are constructed in a $(P + I) \times (P + I)$ matrix as follows: $\mathbf{b}_{(P+I)\times(P+I)} = \text{diag}([\boldsymbol{\theta}_{P\times P}\ \boldsymbol{\zeta}_{I\times I}]')$. The variance–covariance matrix, $\Sigma_\mathbf{b}$, is a block-diagonal matrix in two blocks, Σ_θ and Σ_ζ such that $\Sigma_\mathbf{b}(\Theta) = \text{diag}([\Sigma_\theta\ \Sigma_\zeta]')$. As an example, the block-diagonal structure of the variance–covariance matrix for 400 persons and 80 items is shown in Figure 3.

The likelihood is rewritten as follows with crossed random effects implemented in the *lmer* function:

$$\begin{aligned} &L(Y|\mu, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \Theta) \\ &= \prod_p \prod_i \int_{\mathbb{R}^{P+I}} f(\mathbf{b}) \cdot \Pr_{g[p]k[i]}(y_{pi} = 1)^{y_{pi}} (1 - \Pr_{g[p]k[i]}(y_{pi} = 1))^{1-y_{pi}} d\mathbf{b}. \end{aligned} \tag{4}$$

In the PIRLS algorithm, the contribution of the fixed effects parameters $(\mu, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$ is incorporated as an offset, and the contribution of the variance components (Σ_θ and Σ_ζ) is incorporated as a penalty term in the weighted least squares fit. The approximate inference from the PIRLS algorithm does not fully account for 'uncertainty' in the estimated variance parameters and is therefore not so reliable.

The Laplace approximation to the likelihood is obtained by replacing the logarithm of the integrand in equation (4) by its second-order Taylor series at the conditional modes

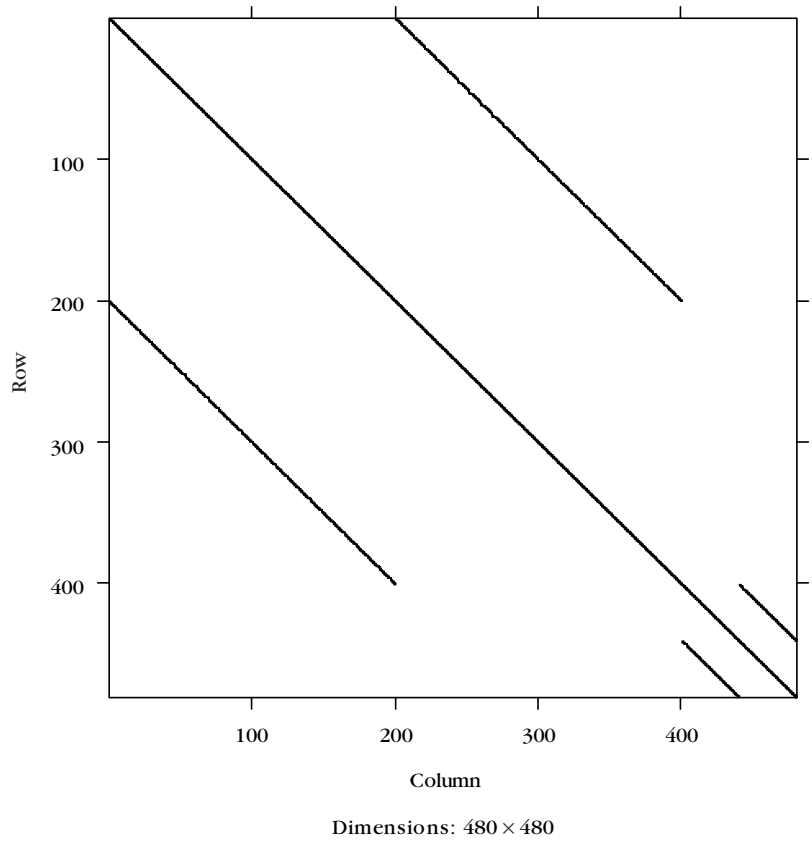


Figure 3. Block diagonal structure of the variance–covariance matrix for crossed random effects in *lmer*.

of the random effects, $\widetilde{\mathbf{b}}(\mu, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \Theta)$. The approximation is

$$
\begin{aligned}
&-2l(\mu, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \Theta|Y) = \\
&-2\log\left\{\prod_p \prod_i \int_{\mathbb{R}^{P+I}} f(\mathbf{b}) \cdot \Pr_{g[p]k[i]}(y_{pi}=1)^{y_{pi}}(1-\Pr_{g[p]k[i]}(y_{pi}=1))^{1-y_{pi}} d\mathbf{b}\right\} \\
&\quad \approx 2\log\left\{\prod_p \prod_i \int_{\mathbb{R}^{P+I}} \exp\left\{-\frac{1}{2}[d(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \widetilde{\mathbf{b}}, Y) + \widetilde{\mathbf{b}}^T \Sigma_{\mathbf{b}} \widetilde{\mathbf{b}} + \right.\right. \\
&\qquad \left.\left. \log|\Sigma_{\mathbf{b}}| + \widetilde{\mathbf{b}}^T D^{-1} \widetilde{\mathbf{b}}]\right\} d\mathbf{b}\right\} \\
&\quad = d(\mu, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \widetilde{\mathbf{b}}, Y) + \widetilde{\mathbf{b}}^T \Sigma_{\mathbf{b}} \widetilde{\mathbf{b}} + \log|\Sigma_{\mathbf{b}}| + \log|D|,
\end{aligned}
\tag{5}
$$

where $d(\mu, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \widetilde{\mathbf{b}}, Y)$ is the function for the deviance from the linear predictor only and D is the approximation to the variance–covariance matrix of $\mathbf{b}$ conditional on $\mu, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\lambda}$ and Θ.

The 'estimates' of the random effects are not model parameters. The conditional *modes* of the random effects can be extracted using the R extractor function *ranef*().

The *lme4* R package also has a function called *mcmcsamp* to evaluate samples from a fitted model using the MCMC.

There are no standard errors for the estimates of variance components (e.g., Σ_θ and Σ_ζ) from the *lmer* function. Since the distributions of estimators of variances are not symmetric, the regular standard errors do not apply. One solution is to obtain quantities for the variance estimates using the *mcsamp* function from the *arm* library in R (the wrapper for the *mcmcsamp* function that goes with the *lmer* function). This function generates a sample from the posterior distribution of the parameters of a fitted model using MCMC methods. The *mcsamp* function automatically simulates multiple sequences and allows convergence to be monitored. The function relies on *mcmcsamp* in the *lme4* package. The simulations with the *mcsamp* function summarize uncertainty about coefficients, predictions, and other quantities of interest.

3.2.2. Hierarchical Bayesian analysis

The MCMC estimation algorithm with a Gibbs sampler can be used to estimate the model parameters. This algorithm was implemented in the WinBUGS software and used to simulate a Markov chain in which values representing parameters of the model are repeatedly sampled from their full conditional posterior distributions over a large number of iterations.

Prior distributions and posterior distribution. The following priors were used to estimate the parameters of the MIRP model in this study:

$$\begin{aligned}
\mu &\sim \text{Normal}(0, 10),\\
\alpha_k|k = 1 &\sim \text{Normal}(0, 10),\\
\gamma_g|g = 1 &\sim \text{Normal}(0, 10),\\
\lambda_{kg}|k = 1, g = 1 &\sim \text{Normal}(0, 10),\\
\boldsymbol{\theta} &\sim \text{MNormal}(\mathbf{0}, \Sigma_\theta), \text{ and}\\
\boldsymbol{\zeta} &\sim \text{MNormal}(\mathbf{0}, \Sigma_\zeta).
\end{aligned}$$

Since μ, α_k, γ_g, and λ_{kg} are fixed effects, hyperpriors are not imposed.

Different kinds of hyperpriors for multivariate normal (MNormal) distributions have been used (Gelman & Hill, 2007): (1) inverse-Wishart$_K$ (where K is the number of variances as the degrees of freedom), (2) inverse-Wishart$_{K+K'}$ (where K' is the number of correlation coefficients), and (3) scaled inverse-Wishart. In this study, inverse-Wishart$_K$ for Σ_θ and Σ_ζ is investigated as widely used prior:

$$\Sigma_\theta \sim \text{Inverse-Wishart}_K(R_\theta, K),\ R_\theta = I_{K\times K}$$

and

$$\Sigma_\zeta \sim \text{Inverse-Wishart}_G(R_\zeta, G),\ R_\zeta = I_{G\times G},$$

where I denotes an identity matrix.

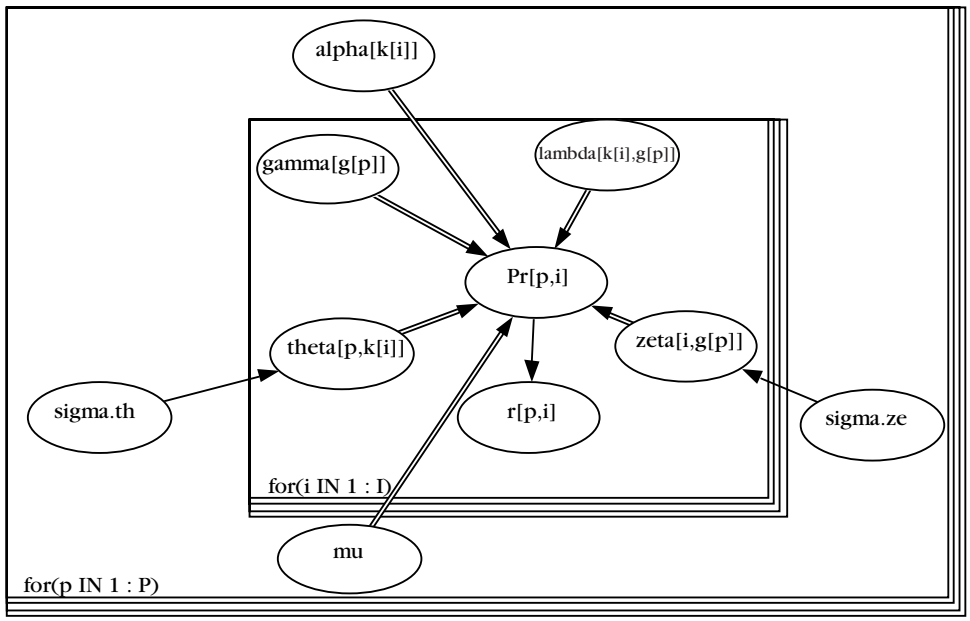


Figure 4. Graphical representation of MIRP model with priors.

These priors result in posterior distributions for each of the parameters that are sampled. The joint posterior distribution of $S = \{\mu, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \Sigma_\theta, \Sigma_\zeta\}$ can be written as

$$\begin{aligned} &\Pr(S|\mathbf{y}) \\ &\propto \mathrm{L}(S) \cdot \Pr(\mu)\Pr(\alpha_k)\Pr(\gamma_g)\Pr(\lambda_{kg})\Pr(\boldsymbol{\theta}|\mathbf{0}, \Sigma_\theta)\Pr(\Sigma_\theta)\Pr(\boldsymbol{\zeta}|\mathbf{0}, \Sigma_\zeta)\Pr(\Sigma_\zeta). \end{aligned} \tag{6}$$

Sampling in WinBUGS. WinBUGS provides the DoodleBUGS graphical package used here to describe how the sampling proceeds. Figure 4 presents the graphical model for the MIRP model obtained from DoodleBUGS. The priors are represented as oval nodes with either solid arrows leading to another node or with hollow arrows leading to Pr[p,i].

The processing in WinBUGS proceeds by sampling all nodes starting at the outer edge of the diagram and working inwards in the diagram to Pr[p,i]. In the program code r[p,i] is an item response, mu is the variable name used for μ, alpha[k[i]] for α_k, gamma[g[p]] for γ_g, lambda[k[i],g[p]] for λ_{kg}, theta[p,k[i]] for $\theta_{g[p]k}$, zeta[i,g[p]] for $\zeta_{k[i]g}$, sigma.th for Σ_θ, and sigma.ze for Σ_ζ.

A solid arrow indicates a stochastic dependence and a hollow arrow indicates a logical function. From the diagram, it can be seen that theta[p,k[i]] depends on sigma.th and zeta[i,g[p]]depends on sigma.ze. Pr[p,i] (which is the program code for 'logit $\left[\Pr_{g[p]k[i]}(y_{pi} = 1)\right] = \mu + \alpha_k + \gamma_g + \lambda_{kg} + \theta_{g[p]k} + \zeta_{k[i]g}$') is a logical function of mu, alpha[k[i]], gamma[g[p]], lambda[k[i],g[p]], theta[p,k[i]], and zeta[i,g[p]]. WinBUGS can also develop code to estimate the model given the specification of a DoodleBUGS diagram. The diagram in Figure 4, for example, would lead to the same solution as the code used in this study. (The variable names in the diagram have no meaning other than that they were used for programming purposes.)

Once the model is fully specified using the distributions given above, WinBUGS then determines the necessary sampling methods directly from the structure in the

diagram. The form of the full conditional distribution of μ, α_k, γ_g, λ_{kg}, $\boldsymbol{\theta}$, $\boldsymbol{\zeta}$, Σ_θ, and Σ_ζ is a conjugate distribution of the parameters (i.e., normal and inverse-Wishart distributions), so that in this study, direct sampling was conducted using standard algorithms in WinBUGS.

Starting values are needed for each parameter being sampled in order to define the first state of the Markov chain. The starting values for the remaining model parameters were randomly generated in the WinBUGS software.

Convergence checking. Some information from the initial iterations is discarded because initial sampled values tend to be dependent on the starting values. These are called burn-in iterations. The remaining iterations are based on a chain that is assumed to have converged to its stationary distribution. Estimates of sampled parameters are then calculated from these post-burn-in iterations. The Gelman and Rubin (1992) method is used for checking convergence of the MCMC algorithm.

3.2.3. AIP with adaptive quadrature

AIP algorithm. Clayton and Rasbash (1999) suggested a special kind of MCMC algorithm for generalized linear mixed models with crossed random effects based on the imputation posterior (IP) algorithm of Tanner and Wong (1987) which can be outlined as follows (Tanner, 1996):

I-step (data augmentation). Impute missing data (random effects) by sampling from the distribution of the missing data conditional on the observed data. This requires that first sampling the parameters are sampled from the current approximation of their posterior distribution.

P-step. Update the approximation of the posterior distribution.

For the MIRP model, the algorithm consists of two wings: a person wing and an item wing. In the person wing, the item effects are treated as known and in the item wing, the person effects are treated as known. In the person wing, the parameters μ, α_k, γ_g, λ_{kg} and $\log \Sigma_\theta$ are estimated (P-step) and the θ_{pg} are sampled (I-step) after first sampling parameters from their approximate posterior distribution (treating the item effects as known). In the item wing, the parameters μ, α_k, γ_g, λ_{kg}, and $\log \Sigma_\zeta$ are estimated (P-step) and the ζ_{ik} are sampled (I-step), after first sampling the parameters from their approximate posterior distribution (treating the person abilities as known).

Specifically, after setting initial values $\boldsymbol{\zeta}^0$ for the item effects, the person wing and item wing outlined below are alternated until convergence. In iteration n:

Person wing. Treat the item effects $\boldsymbol{\zeta}^{n-1} = (\zeta_{1[1]1}^{n-1}, \ldots, \zeta_{K(I)G}^{n-1})'$ from the previous iteration as known:

$$\text{logit}\left[\Pr_{g[p]k[i]}\left(y_{pi} = 1 | \theta_{g[p]k}, \zeta_{k[i]g}^{n-1}\right)\right] = \mu + \alpha_k + \gamma_g + \lambda_{kg} + \theta_{g[p]k} + \zeta_{k[i]g}^{n-1}. \qquad (7)$$

Let the parameters be denoted by $\boldsymbol{\vartheta}_1 = \{\mu, \alpha_k, \gamma_g, \lambda_{kg}, \log \Sigma_\theta\}$.

1. Obtain maximum likelihood estimates $\widehat{\boldsymbol{\vartheta}}_1^n$ with the estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\vartheta}_1}^n$.
2. Sample parameters $\boldsymbol{\vartheta}_1^n$ from their approximate sampling distribution

$$\boldsymbol{\vartheta}_1^n | \boldsymbol{\zeta}^{n-1} \sim N\left(\widehat{\boldsymbol{\vartheta}}_1^n, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\vartheta}_1}^n\right). \tag{8}$$

3. Sample $\boldsymbol{\theta}^n$ from its estimated posterior distribution with parameters $\boldsymbol{\vartheta}_1^n$.

Item wing. Treat the person abilities $\boldsymbol{\theta}^n = (\theta_{1(1)1}^n, \ldots, \theta_{G(P)K}^n)'$ from the person wing as known:

$$\text{logit}\left[\Pr_{g[p]k[i]}\left(y_{pi} = 1 | \theta_{g[p]k}^n, \zeta_{k[i]g}\right)\right] = \mu + \alpha_k + \gamma_g + \lambda_{kg} + \theta_{g[p]k}^n + \zeta_{k[i]g}. \tag{9}$$

Let the parameters be denoted by $\boldsymbol{\vartheta}_2 = \{\mu, \alpha_k, \gamma_g, \lambda_{kg}, \log \Sigma_\zeta\}$.

1. Obtain maximum likelihood estimates $\widehat{\boldsymbol{\vartheta}}_2^n$ with the estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\vartheta}_2}^n$.
2. Sample parameters $\boldsymbol{\vartheta}_2^n$ from their approximate sampling distribution

$$\boldsymbol{\vartheta}_2^n | \boldsymbol{\theta}^n \sim N\left(\widehat{\boldsymbol{\vartheta}}_2^n, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\vartheta}_2}^n\right). \tag{10}$$

3. Sample $\boldsymbol{\zeta}^n$ from its estimated posterior distribution with parameters $\boldsymbol{\vartheta}_2^k$.

After convergence is achieved, the algorithm is continued for a fixed number of iterations and the parameter estimates are obtained by averaging the estimates obtained after burn-in.

In the rest of this section, we discuss the implementation of steps 1–3.

Step 1: MLE using adaptive quadrature. In the AIP algorithm, the parameters are estimated using Stata's *gllamm* command (Rabe-Hesketh, Skrondal, & Pickles, 2004), which employs a Newton–Raphson algorithm with analytical first and second derivatives to maximize the likelihood. The number of quadrature points required is determined by fitting the person-wing model with item effects set to zero and the item-wing model with person effects set to zero. The number of quadrature points is increased from 5 in five-point increments. If the change in maximized log-likelihood associated with an increment is less than 1×10^{-10}, the smaller number of adaptive quadrature points is used.

Step 2: Sampling the model parameters. The parameters are sampled from a multivariate normal distribution with mean given by maximum likelihood estimates $\widehat{\boldsymbol{\vartheta}}$ and covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\vartheta}}}$ by the inverse of the estimated information matrix obtained in step 1. This distribution approximates the hierarchical Bayesian posterior if uniform priors are specified for all parameters. In this case, the posterior distribution is just the normalized likelihood which is approximated by a multivariate normal distribution. A log transformation of the variance parameters is used to improve the normal approximation. In step 2, we can see clear differences between AIP and Gibbs sampling. First, AIP does not need a specification of the prior distributions. Second, a whole vector of nodes is sampled in AIP while a scalar is sampled in Gibbs sampling.

Step 3: Sampling the random effects from their posterior distribution. In the imputation step, the person effects and item effects are sampled from their respective conditional posterior distributions. According to the Bayesian central limit theorem, posterior distributions approach normality as the sample size (here the number of items or number of persons) tends to infinity, as noted by Chang and Stout (1993) for binary data. We therefore approximate the posterior by a normal density with person-specific posterior means, $\boldsymbol{\mu}_p^n$, and posterior variance–covariance matrix, Σ_p^n, for parameters $\boldsymbol{\vartheta}_1^n$:

$$\Pr\left(\boldsymbol{\theta}_{g[p]}|Y_{g[p]};\boldsymbol{\zeta}^{n-1},\mu^{n,1},\alpha_1^{n,1},\gamma_1^{n,1},\lambda_1^{n,1}\right)\simeq g\left(\boldsymbol{\theta}_{g[p]};\boldsymbol{\mu}_{g[p]}^n,\Sigma_{g[p]}^n\right). \tag{11}$$

We compute the posterior mean and variance–covariance matrix using *gllamm* and the corresponding prediction command (Rabe-Hesketh & Skrondal, 2008). This normal approximation ignores any skewness for clusters with large or small cluster totals (Thomas & Gan, 1997). A discrete approximation is available in AIP with adaptive quadrature (Cho & Rabe-Hesketh, 2011), but only a normal approximation was implemented for the MIRP model.

The correlated random effects, $\boldsymbol{\theta}_{g[p]}=[\theta_{1(p)1},\ldots,\theta_{g[p]k},\ldots,\theta_{G(p)K}]'$, are represented by a linear combination of independent standard normal random effects $\mathbf{z}$ using the Cholesky decomposition of the covariance matrix Q with $QQ'=\Sigma_\theta$, so that $\boldsymbol{\theta}=Q\mathbf{z}$. Refer to Cho and Rabe-Hesketh (2011) for the detail of convergence checking and posterior moment calculation in AIP with adaptive quadrature.

4. Simulation study

In this section, a simulation study is presented to assess the performance of three algorithms for the MIRP model.

4.1. Conditions

Two item types and two person groups ($K=2$ and $G=2$) are generated. The following four factors are considered to investigate the recovery of the model parameters in various conditions: two different numbers of items (20 and 50) for each type; two different numbers of persons (100 and 250) for each group; two different correlation coefficients of person random effects ($\rho(P)=0,.5$); and two different correlation coefficients of item random effects ($\rho(I)=0,.5$). Conditions for the following are fixed: variances of person and item random effects are 1, $\mu=0$, $\alpha_1=-0.5$, $\gamma_1=-0.5$, and $\lambda_{11}=0.5$. The number of conditions is thus $2\times2\times2\times2=16$. Random effects were generated following a bivariate normal distribution given the variance and covariance described above. Each condition was replicated 10 times and is fitted with three different estimation algorithms using the same data sets. It will be shown that this rather low number does not invalidate the study.

4.2. Analyses

Convergence checking was performed in WinBUGS using similar graphical checks as described in convergence checking along with the condition that $\sqrt{\widehat{R}}$ is less than 1.001. In addition, autocorrelation plots from WinBUGS were examined. A conservative burn-in of 1,800–3,600 iterations was used in this study followed by 4,800–6,000 post-burn-in iterations depending on conditions. Thinning was set at 3, meaning that 1,600–2,000

iterations were used to obtain a posterior mean and median across conditions. In AIP with adaptive quadrature, 5–15 quadrature points were used. One simulated data set for each condition was used for convergence checking and the same burn-in was set across replications. An additional 10 iterations were obtained to estimate the posterior moments with the small sample size correction.

4.3. Results

Tables 2–4 show the average values for the root mean square error (RMSE, given by $\sqrt{\sum_{r=1}^{10}(\bar{\mu}_r - \mu)^2/10}$, where r denotes a replication for μ as an example), the average bias across replications (Bias, given by $\sum_{r=1}^{10} \bar{\mu}/10 - \mu$, where r again denotes a replication for μ as an example), the standard deviation of the estimates across replications (SD), the mean of the standard errors of estimates across replications (M(SE)), and parameter coverage in percentage (C(%)), all of these for each of the two levels of each of the four factors of the simulation design. As reviewed in Section 3, it is known that the approximation methods such as the Laplace approximation have the problem of downward bias in variance estimation. Therefore, bias is a special point of interest.

For the fixed effects, parameter coverage is calculated as the percentage of simulation runs for which the 95% confidence intervals for Laplace and AIP and credibility intervals for MCMC contain the corresponding true value. For variance estimates of the random item and person effects, the parameter coverage was not reported for Laplace and AIP since the distribution of the estimator of the random effect variance is not symmetric. However, MCMC provides the whole posterior distribution of the parameters, so that an appropriate credibility interval can be determined. The posterior median was used to calculate the RMSE, bias, and coverage for MCMC since some posterior distributions are not symmetric. For the purpose of comparing the three estimation methods, the proportion of times that the true value has a cumulative proportion less than or equal to .10 or greater than or equal to .90 (from the 10 replication in each cell) is calculated for the variance estimates of the random effects. This means that the true value is within the range of the 10 replications.

An important finding is that in general the uncertainty of the sampling error (the SD of estimates across replications) is close to or smaller than the uncertainty of the estimates (the SE of estimates across replications) for all parameters. The absolute difference between two uncertainty quantities is less than 0.3 for all parameters from the three methods (except population parameters of random effects in Laplace where M(SE) was not obtained). The implication is that the 10 replications per cell give a realistic idea of parameter estimation. The largest discrepancies and rather irregular patterns can be found for the item random effects when MCMC is used. Much smaller but systematic differences are found for the correlation between random item effects and for the fixed effect α when AIP is used. Since standard errors of variances were not calculated in the *lmer* function, M(SE)s were not reported for variance and correlation for Laplace and were not compared to the standard deviation of estimates across replications.

Figure 5 shows the boxplot with the median, the two quartiles, the range, and the outliers for the fixed effect estimates (4 latent correlation conditions $\times$ 10 replications for all fixed effects). It is evident that the medians are close to the true parameters ("mu" for $\mu = 0$, "gamma" for $\gamma_1 = -0.5$, "alpha" for $\alpha_1 = -0.5$, and "lambda" for $\lambda_{11} = 0.5$). They are similar across the three methods, and there are only a few outliers.

Table 2. Simulation results of three estimation methods: fixed effects

			Laplace					MCMC					AIP				
Parameter	Con.	Level	RMSE	Bias	SD	M(SE)	C(%)	RMSE	Bias	SD	M(SE)	C(%)	RMSE	Bias	SD	M(SE)	C(%)
μ	I	40	0.136	0.008	0.128	0.133	95.0	0.143	0.008	0.136	0.138	92.4	0.147	0.019	0.138	0.133	90.0
		100	0.080	−0.005	0.075	0.092	95.0	0.081	−0.006	0.076	0.091	93.6	0.102	0.002	0.097	0.097	92.5
	P	200	0.116	0.009	0.111	0.119	93.8	0.118	0.008	0.113	0.121	92.4	0.127	0.017	0.123	0.119	92.5
		500	0.107	−0.006	0.098	0.109	96.2	0.114	−0.007	0.106	0.113	93.6	0.126	0.004	0.115	0.112	90.0
	$\rho(P)$	0.0	0.099	0.017	0.093	0.112	97.5	0.103	0.012	0.096	0.115	94.7	0.114	0.027	0.102	0.115	91.2
		0.5	0.122	−0.014	0.116	0.117	92.5	0.128	−0.010	0.122	0.119	91.2	0.138	−0.006	0.134	0.116	91.2
	$\rho(I)$	0.0	0.105	0.007	0.097	0.105	95.0	0.108	0.009	0.101	0.107	92.5	0.125	0.021	0.115	0.106	90.0
		0.5	0.118	−0.003	0.112	0.123	95.0	0.124	−0.007	0.118	0.126	93.5	0.128	0.000	0.123	0.124	92.5
γ	I	40	0.113	0.000	0.104	0.109	93.8	0.141	0.014	0.135	0.156	93.6	0.124	0.009	0.114	0.123	93.8
		100	0.083	0.005	0.079	0.078	93.8	0.093	−0.003	0.090	0.089	91.1	0.087	0.010	0.085	0.105	98.8
	P	200	0.110	0.004	0.100	0.100	92.5	0.119	0.005	0.114	0.116	92.5	0.115	0.011	0.104	0.120	97.5
		500	0.087	0.002	0.084	0.089	95.0	0.119	0.006	0.115	0.138	92.2	0.099	0.008	0.097	0.108	95.0
	$\rho(P)$	0.0	0.101	−0.004	0.097	0.092	92.5	0.124	0.007	0.119	0.141	92.2	0.099	0.002	0.096	0.120	98.8
		0.5	0.098	0.009	0.088	0.097	95.0	0.114	0.004	0.110	0.112	92.5	0.114	0.016	0.105	0.109	93.8
	$\rho(I)$	0.0	0.111	0.015	0.102	0.106	96.2	0.107	0.018	0.101	0.104	93.6	0.117	0.025	0.107	0.106	93.8
		0.5	0.086	−0.009	0.083	0.081	91.2	0.131	−0.007	0.127	0.147	91.1	0.096	−0.006	0.093	0.122	98.8
α	I	40	0.138	0.015	0.133	0.130	95.0	0.115	−0.001	0.105	0.114	95.0	0.155	0.021	0.146	0.105	81.2
		100	0.093	−0.003	0.090	0.087	91.2	0.084	0.005	0.081	0.078	93.6	0.108	0.002	0.101	0.082	85.0
	P	200	0.121	0.004	0.115	0.114	92.5	0.112	0.005	0.102	0.103	93.8	0.136	0.013	0.128	0.099	85.0
		500	0.115	0.008	0.112	0.107	93.8	0.087	−0.001	0.084	0.093	94.9	0.132	0.010	0.123	0.088	81.2
	$\rho(P)$	0.0	0.122	0.007	0.117	0.112	92.5	0.101	−0.009	0.098	0.096	93.6	0.135	0.016	0.126	0.094	85.0
		0.5	0.114	0.005	0.110	0.109	93.8	0.100	0.012	0.089	0.100	95.0	0.133	0.007	0.125	0.093	81.2
	$\rho(I)$	0.0	0.106	0.019	0.101	0.100	93.8	0.114	0.014	0.103	0.110	94.9	0.127	0.027	0.115	0.106	91.2
		0.5	0.129	−0.007	0.126	0.120	92.5	0.085	−0.011	0.082	0.084	93.8	0.140	−0.004	0.135	0.082	75.0
λ	I	40	0.100	0.000	0.095	0.104	96.2	0.097	0.001	0.093	0.113	97.5	0.114	0.015	0.109	0.103	87.5
		100	0.078	0.000	0.069	0.072	92.5	0.079	−0.002	0.070	0.075	96.2	0.091	0.008	0.082	0.082	86.2
	P	200	0.097	−0.015	0.090	0.094	92.5	0.095	−0.015	0.087	0.097	96.2	0.109	−0.005	0.103	0.097	86.2
		500	0.082	0.014	0.076	0.086	96.2	0.082	0.015	0.077	0.095	97.5	0.096	0.027	0.089	0.088	87.5
	$\rho(P)$	0.0	0.088	−0.011	0.080	0.092	95.0	0.086	−0.012	0.078	0.101	100.0	0.102	0.006	0.093	0.097	86.2
		0.5	0.092	0.010	0.087	0.087	93.8	0.091	0.011	0.086	0.091	93.8	0.103	0.016	0.099	0.088	87.5
	$\rho(I)$	0.0	0.100	−0.001	0.092	0.102	92.5	0.097	−0.001	0.090	0.107	96.2	0.108	0.012	0.102	0.104	91.2
		0.5	0.079	0.001	0.074	0.076	96.2	0.080	0.001	0.073	0.084	97.5	0.097	0.011	0.090	0.081	82.5

Table 3. Simulation results of three estimation methods: item random effects

			Laplace			MCMC					AIP			
Parameter	Con.	Level	RMSE	Bias	SD	RMSE	Bias	SD	M(SE)	C(%)	RMSE	Bias	SD	M(SE)
$\sigma^2_{\zeta_1}$	I	40	0.210	−0.076	0.189	0.202	−0.005	0.196	0.281	97.5	0.200	0.038	0.191	0.145
		100	0.140	−0.025	0.125	0.142	0.003	0.128	0.157	93.8	0.139	0.025	0.123	0.127
	P	200	0.192	−0.059	0.168	0.187	−0.014	0.174	0.218	93.8	0.180	0.024	0.167	0.145
		500	0.163	−0.041	0.152	0.161	0.012	0.157	0.237	97.5	0.164	0.039	0.154	0.127
	$\rho(P)$	0.0	0.182	−0.061	0.159	0.175	−0.009	0.163	0.238	95.0	0.146	0.027	0.135	0.140
		0.5	0.174	−0.040	0.161	0.175	0.007	0.168	0.218	96.2	0.196	0.036	0.183	0.132
	$\rho(I)$	0.0	0.187	−0.045	0.170	0.190	0.002	0.176	0.220	95.0	0.178	0.032	0.161	0.138
		0.5	0.169	−0.056	0.150	0.158	−0.005	0.154	0.236	96.2	0.166	0.031	0.160	0.134
$\sigma^2_{\zeta_2}$	I	40	0.260	−0.002	0.245	0.283	0.073	0.259	0.472	92.5	0.270	0.100	0.238	0.145
		100	0.142	−0.019	0.139	0.143	0.006	0.140	0.163	96.1	0.123	0.022	0.116	0.127
	P	200	0.233	0.005	0.225	0.252	0.053	0.234	0.247	93.8	0.213	0.069	0.183	0.145
		500	0.184	−0.026	0.169	0.192	0.025	0.179	0.434	94.9	0.207	0.054	0.192	0.127
	$\rho(P)$	0.0	0.219	−0.013	0.209	0.233	0.039	0.220	0.439	93.6	0.191	0.051	0.175	0.140
		0.5	0.201	−0.008	0.189	0.215	0.04	0.195	0.238	95.0	0.227	0.071	0.199	0.132
	$\rho(I)$	0.0	0.232	−0.008	0.225	0.251	0.037	0.238	0.241	94.9	0.185	0.035	0.172	0.134
		0.5	0.186	−0.012	0.168	0.193	0.041	0.173	0.437	93.8	0.232	0.087	0.202	0.138
$\rho(I)$	I	40	0.143	0.004	0.140	0.142	−0.001	0.139	0.145	98.8	0.144	0.003	0.141	0.145
		100	0.096	−0.014	0.092	0.097	−0.014	0.093	0.092	97.5	0.102	−0.004	0.096	0.127
	P	200	0.126	−0.007	0.123	0.124	−0.01	0.121	0.124	98.8	0.122	0.000	0.119	0.145
		500	0.118	−0.003	0.113	0.119	−0.006	0.114	0.119	97.5	0.127	−0.001	0.122	0.127
	$\rho(P)$	0.0	0.125	−0.007	0.122	0.125	−0.01	0.123	0.122	98.8	0.128	0.003	0.125	0.140
		0.5	0.119	−0.003	0.114	0.117	−0.006	0.113	0.121	97.5	0.121	−0.004	0.116	0.132
	$\rho(I)$	0.0	0.147	−0.020	0.143	0.147	−0.02	0.143	0.135	96.2	0.160	−0.008	0.156	0.138
		0.5	0.091	0.011	0.087	0.089	0.004	0.087	0.106	100.0	0.073	0.007	0.069	0.134

Table 2 reports results for all fixed effects. The following trends are apparent. First, the RMSEs decrease with increasing number of items and number of persons for all three estimation methods. Second, the RMSEs of μ decrease with decreasing correlations between random effects. For the other fixed effects, the patterns of the effects are less clear or the differences are only very small. Third, the bias is relatively small and does not show systematic patterns. Fourth, the M(SE) is similar for the three methods and decreases with the number of persons and items in all three methods. The SDs for α and λ are slightly larger when AIP is used. Finally, the parameter coverage is adequate: between 91.1% and 97.5% for Laplace and MCMC. The coverage for α and λ is somewhat smaller for AIP than for the other two methods. This may be because the variability in estimates across replications (represented by RMSE and SD) is somewhat higher for AIP than for the other two methods.

Table 3 presents results for the population parameters of the random item effects, $\sigma^2_{\zeta_1}$, $\sigma^2_{\zeta_2}$, and $\rho(I)$. We observe the following trends. First, the RMSEs of all three parameters decrease with increasing number of items and persons for all three estimation methods, except for $\rho(I)$ and the number of persons in AIP. Second, the RMSEs of all three parameters decrease with increasing correlation between the random effects, albeit less clearly for AIP. Third, as expected, a negative bias of the $\sigma^2_{\zeta_1}$ and $\sigma^2_{\zeta_2}$ estimates is found with the Laplace estimation. For AIP the bias is positive. The bias of the correlation estimates is negligible. Fourth, the SD is similar for the three methods and improves with the number of persons and items, except for two cases in AIP. M(SE) for variances is smaller in AIP than in MCMC. Finally, MCMC yields accurate coverage of the variance parameters, ranging from 92.5% to 97.5%. We have not derived credibility intervals for

Table 4. Simulation results of three estimation methods: person random effects

			Laplace			MCMC					AIP			
Parameter	Con.	Level	RMSE	Bias	SD	RMSE	Bias	SD	M(SE)	C(%)	RMSE	Bias	SD	M(SE)
$\sigma^2_{\theta_1}$	I	40	0.128	−0.046	0.107	0.126	−0.028	0.110	0.117	96.1	0.132	−0.00	0.124	0.145
		100	0.097	−0.008	0.092	0.098	0.000	0.093	0.099	96.1	0.154	0.024	0.144	0.127
	P	200	0.139	−0.043	0.120	0.138	−0.028	0.122	0.130	95.0	0.177	0.005	0.165	0.145
		500	0.079	−0.011	0.075	0.079	0.000	0.076	0.083	97.2	0.099	0.015	0.096	0.127
	$\rho(P)$	0.0	0.110	−0.032	0.091	0.109	−0.020	0.094	0.108	97.2	0.159	0.018	0.147	0.140
		0.5	0.116	−0.022	0.108	0.116	−0.008	0.109	0.110	95.0	0.126	0.003	0.121	0.132
	$\rho(I)$	0.0	0.119	−0.033	0.100	0.117	−0.021	0.102	0.107	96.1	0.121	0.000	0.113	0.134
		0.5	0.107	−0.021	0.099	0.108	−0.007	0.102	0.110	96.1	0.163	0.020	0.154	0.138
$\sigma^2_{\theta_2}$	P	40	0.131	0.002	0.128	0.136	0.023	0.130	0.129	96.2	0.145	0.043	0.133	0.145
		100	0.103	−0.033	0.090	0.101	−0.023	0.090	0.100	92.4	0.110	−0.01	0.102	0.127
	I	200	0.143	−0.013	0.138	0.146	0.007	0.141	0.139	92.4	0.157	0.024	0.149	0.145
		500	0.086	−0.018	0.073	0.086	−0.007	0.072	0.085	96.2	0.093	0.008	0.078	0.127
	$\rho(P)$	0.0	0.120	−0.015	0.113	0.121	0.000	0.114	0.116	94.9	0.130	0.022	0.121	0.140
		0.5	0.116	−0.016	0.107	0.118	−0.001	0.109	0.115	93.8	0.128	0.010	0.117	0.132
	$\rho(I)$	0.0	0.114	0.005	0.110	0.118	0.020	0.112	0.117	96.2	0.131	0.038	0.119	0.138
		0.5	0.122	−0.037	0.111	0.120	−0.020	0.112	0.114	92.4	0.126	−0.00	0.118	0.134
$\rho(P)$	P	40	0.079	0.014	0.071	0.078	0.010	0.071	0.071	93.8	0.086	0.021	0.076	0.145
		100	0.055	−0.005	0.053	0.056	−0.006	0.053	0.061	97.4	0.065	0.008	0.061	0.127
	I	200	0.082	0.003	0.075	0.081	0.000	0.075	0.079	95.0	0.085	0.013	0.076	0.145
		500	0.050	0.006	0.048	0.051	0.005	0.048	0.050	96.1	0.067	0.016	0.061	0.127
	$\rho(P)$	0.0	0.072	0.010	0.065	0.072	0.011	0.065	0.073	97.4	0.093	0.025	0.083	0.140
		0.5	0.064	−0.002	0.061	0.063	−0.006	0.061	0.059	93.8	0.055	0.004	0.052	0.132
	$\rho(I)$	0.0	0.074	0.011	0.067	0.074	0.010	0.067	0.066	92.4	0.085	0.027	0.073	0.138
		0.5	0.061	−0.003	0.059	0.061	−0.005	0.058	0.066	98.8	0.067	0.002	0.065	0.134

Laplace and AIP, given that normality of the estimation error cannot be assumed for the variance (AIP) and that no SEs for the variance are available (Laplace). However, for neither Laplace nor AIP have we found cells where the true value is not within the range of the 10 estimates. For the correlation estimates, the coverage is even better, ranging from 96.2% to 100%, and also for the correlation the true value is always within the range of the 10 estimates per cell.

Table 4 presents results for the population parameters of the random person effects, $\sigma^2_{\theta_1}$, $\sigma^2_{\theta_2}$, and $\rho(P)$. The following trends are evident. First, the RMSEs of all three parameters decrease with increasing number of items and persons for all estimation methods. Second, the RMSEs of all three parameters decrease in most cases with increasing correlation between random effects. There are exceptions, and the pattern is not clear, except that they never appear for $\rho(P)$. Third, as expected, a negative bias of the $\sigma^2_{\theta_1}$ and $\sigma^2_{\theta_2}$ estimates is found with the Laplace estimation. The bias is again slightly positive when AIP is used. The bias of the $\rho(P)$ is negligible. Fourth, the SD is similar for the three methods and improves with the number of persons and items. M(SE) for variances and a correlation is smaller in MCMC than in AIP. Finally, the credibility intervals constructed for MCMC coverage of the variance parameters show that the coverage is satisfactory, ranging from 92.4% to 97.2%. For the other two methods, the true value is within the range of the 10 estimates per cell, except for two cases. The MCMC coverage of $\rho(P)$ is as high as for the variances. The true value of $\rho(P)$ is always within the range of the 10 estimates per cell.

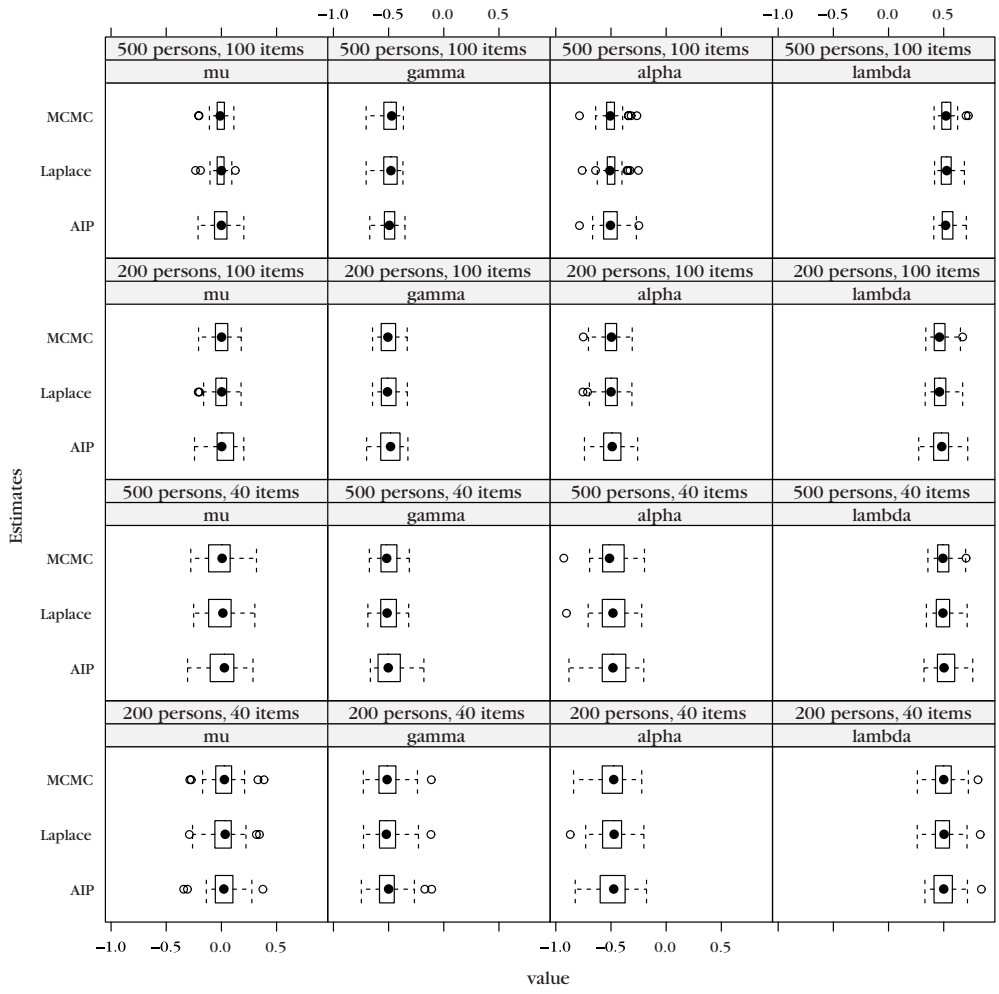


Figure 5. Boxplots for fixed effect estimates.

5. Discussion

The MIRP instantiation of a crossed random effects model for binary data is a member of the broader category of generalized linear mixed models. It is an instantiation with great potential for application in psychology, in test psychology as well as in experimental psychology, as shown in the introduction. In this paper, we compare three different estimation methods to overcome the computational challenges. The model is illustrated with a real data example. A theoretical comparison is presented, followed by a simulation study. The simulation results are of course limited to the simulation conditions included, but also the real data example shows that the three methods give similar results. Only 10 replications have been implemented per simulation condition due the computational burden of MCMC and AIP with adaptive quadrature, and consequently the time taken to run these algorithms. However, the uncertainty of the sampling error (the standard deviation of estimates across replications) is close to the uncertainty of the estimates (the means of the standard errors of estimates across replications) in most cases for

all conditions. This is an indication that even with 10 replications the methodology described is satisfactory.

When practitioners choose an estimation method for the MIRP model, the two important criteria would be computational efficiency (computation time) and estimate precision. Based on our theoretical analysis and the simulation results, we conclude that Laplace approximation is the most efficient in estimating the MIRP model parameters while AIP is computationally very expensive. For example, in the empirical study, there was no a memory problem due to the high dimensionality of the model on a computer equipped with a 1.20 GHz processor with 2.93 GB of RAM. The Laplace approximation with the *lmer* function required 5 minutes, MCMC with WinBUGS required 913 minutes with four chains of 10,000 iterations including 5,000 burn-in for each chain, and AIP with adaptive quadrature required 2250 minutes for 40 iterations, including 10 iterations as burn-in.

The precision of all *fixed effect* estimates was very similar with respect to bias and M(SE) for the three methods when the SD of the estimates is considered and the precision increases with the number of persons and items. Coverage with the AIP method is less accurate than with Laplace and with MCMC for two fixed effects (α and λ), because of the larger variability in estimates across replications in AIP than in the other two methods. The conclusions for the precision of the population parameter estimates of the *random effect* distributions are that the performance is similar for MCMC and AIP based on RMSE. However, when the sample size is small (in our case, the number of items), the item variance estimates have larger bias and smaller mean SE in AIP, while they have smaller bias and larger mean SE in MCMC. AIP with adaptive quadrature does not require specification of prior distributions for model parameters. This may be an advantage since the choice of hyperprior for the variance components can affect the parameter estimates, especially in small sample sizes. For MCMC only one prior on the variance was considered in this study. Further study is required to establish the sensitivity of the estimates to the kind of prior. The most important finding regarding Laplace approximation is that downward bias of the Laplace approximation was found. Such bias has previously been found for Laplace (Joe, 2008) and related methods (MQL and PQL) (Browne & Draper, 2006) for binary response with small cluster sizes and high intraclass correlations.

Model selection with respect to the number of random effects to be included is of interest for the MIRP approach. The AIC and BIC can be used to select the number of random effects, but further study of this issue is important, for example to evaluate the mixture χ^2-test used in the application.

Acknowledgement

The first author thanks Dr. Sophia Rabe-Hesketh for her valuable comments on AIP with adaptive quadrature.

References

Akaike, A. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Baayen, R. H., Davidson, R. H., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Bates, D. M., & DebRoy, S. (2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis*, *1*, 1–17.

Bates, D., & Maechler, M. (2009). lme4: Linear mixed-effects models using s4 classes. Retrieved from http://cran.R-project.org/lme4

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *33*, 179–197.

Bock, R. D., & Schilling, S. G. (1997). High-dimensional full-information item factor analysis. In M. Berkane (Ed.), *Latent variable modelling and applications to causality* (pp. 164–176). New York: Springer.

Breslow, N. E. (2004). Whither PQL?. In D. Y. Lin & P. J. Heagerty (Eds.), *Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of correlated data* (pp. 1–22). New York: Springer.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *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. (2006). A comparison of Bayesian and likelihood methods for fitting multilevel models. *Bayesian Analysis*, *3*, 473–514.

Butler, J. S., & Moffitt, R. (1982). A computationally efficient quadrature procedure for the one-factor multinomial probit model. *Econometrica*, *50*, 761–764.

Camilli, G., & Penfield, D. A. (1997). Variance estimation for differential test functioning based on Maentel–Haenszel statistics. *Journal of Educational Measurement*, *34*, 123–139.

Chang, H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, *58*, 37–52.

Cho, S.-J., & Rabe-Hesketh, S. (2011). Alternating imputation posterior estimation of models with crossed random effects. *Computational Statistics and Data Analysis*, *55*, 12–25.

Clark, H. H. (1973). The language-as-fixed effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–359.

Clayton, D. G., & Rasbash, J. (1999). Estimation in large crossed random-effect models by data augmentation. *Journal of the Royal Statistical Society, Series A*, *162*, 425–436.

Congdon, P. (2003). *Applied Bayesian modelling*. Chichester: Wiley.

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533–559.

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

De Boeck, P., & Wilson, M. (2010). *Multidimensional (multiple) random item profile model*. Denver, CO: National Council Measurement in Education.

Diaz, R. E. (2007). Comparison of PQL and Laplace 6 estimates of hierarchical linear models when comparing groups of small incident rates in cluster randomised trials. *Computational Statistics and Data Analysis*, *51*, 2871–2888.

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: With the lme4 package. *Journal of Statistical Software*, *20*. Retrieved from http://www.jstatsoft.org/v20/i02/

Forster, K. I., & Masson, M. E. J. (2008). Issue: Emerging data analysis. *Journal of Memory and Language*, *59*, 387–556.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York: Springer.

Freeman, E., Heathcote, A., Chalmers, K., & Hockley, W. (2010). Item effects in recognition memory for words. *Journal of Memory and Language*, *62*, 1–18.

Geerlings, H., Glas, C. A. W., & van der Linden, W. J. (2011). Modeling rule-based item generation. *Psychometrika*, *76*(2), 337–359.

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, *74*, 153–160.

Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B*, *56*, 501–514.

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

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/ hierarchical models*. New York: Cambridge University Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Sceience*, *7*, 457–472.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of image. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Glas, C. A. W., & van der Linden, W. J. (2003). Computerized adaptive testing with item clones. *Applied Psychological Measurement*, *27*, 247–261.

Goldstein, H. (1987). Multilevel covariance component models. *Biometrika*, *74*, 430–431.

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

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

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W. J., Yang, M., Woodhouse, G., & Healy, M. (1998). *A user's guide to MLwiN*. London: Multilevel Models Project, Institute of Education, University of London.

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

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck& M. Wilson (Eds.), *Explanatory item response models* (pp. 198–212). New York: Springer.

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics*, *25*, 285–306.

Joe, H. (2008). Accuracy of Laplace approximation for discrete response mixed models. *Computational Statistics and Data Analysis*, *52*, 5066–5074.

Johnson, M. S., & Sinharay, S. (2005). Calibration of polytomous item families using Bayesian hierarchical modeling. *Applied Psychological Measurement*, *29*, 369–400.

Karim, M. R., & Zeger, S. L. (1992). Generalized linear models with random effects: Salamander mating revisited. *Biometrics*, *48*, 631–644.

Lambert, P. C. (2006). Comment on article by Browne and Draper. *Bayesian Analysis*, *1*, 543–546.

Lee, Y., & Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Series B*, *58*, 619–678.

Lee, Y., & Nelder, J. A. (2006). Double-hierarchical generalized linear models (with discussion). *Applied Statistics*, *55*, 1–29.

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.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland& H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 171–196). Hillsdale, NJ: Erlbaum.

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

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

Natarajan, R., & Kass, R. E. (2000). Reference Bayesian methods for generalized linear mixed model. *Journal of the American Statistical Association*, *95*, 227–237.

Penfield, R. D., & Algina, J. (2006). A generalized DIF effect variance estimator for measuring unsigned differential test functioning in mixed format tests. *Journal of Educational Measurement*, *43*, 295–312.

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

Quené, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *Journal of the Acoustical Society of America*, *123*, 1104–1113.

Raaijmarkers, J. G. W. (2003). A further look at the language-as-fixed-effect fallacy. *Canadian Journal of Experimental Psychology*, *57*, 141–151.

Raaijmarkers, J. G. W., Schrijnemakers, J. M. C., & Gremmen, F. (1999). How to deal with "the language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language*, *41*, 416–426.

Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata* (2nded.). College Station, TX: Stata Press.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modelling. *Psychometrika*, *69*, 167–190.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, *128*, 301–323.

Rasbash, J., & Browne, W. J. (2007). Non-hierarchical multilevel models. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 333–336). New York: Springer.

Rasbash, J., & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and crossed random structures using a multilevel model. *Journal of Behavioral Statistics*, 337–350.

Raudenbush, S. W., Yang, M., & 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.

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

Rodríguez, G., & Goldman, N. (2001). Improved estimation procedures for multilevel models with binary response: A case study. *Journal of the Royal Statistical Society, Series A*, *164*, 339–355.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 416–464.

Smits, D. J. M., De Boeck, P., & Vansteelandt, K. (2004). The inhibition of verbally aggressive behavior. *European Journal of Personality*, *18*, 537–555.

Soares, T. M., Gonçalves, F. B., & Gamerman, D. (2009). An integrated Bayesian model for DIF analysis. *Journal of Educational and Behavioral Statistics*, *34*, 348–377.

Spiegelhalter, D., Best, N. G., & Carlin, B. P. (1998). *Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models*. Technical report, MRC Biostatistics Unit, Cambridge.

Spiegelhalter, D., Thomas, A., & Best, N. (2003). *WinBUGS (Version 1.4)* [Computer program]. Cambridge: MRC Biostatistics Unit, Institute of Public Health.

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

Tanner, M. A. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions*. New York: Springer.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528–540.

Thomas, N., & Gan, N. (1997). Generating multiple imputations for matrix sampling data analyzed with item response models. *Journal of Educational and Behavioral Statistics*, *22*, 425–445.

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

Received 30 August 2010; revised version received 9 September 2011

Appendix A

R code for verbal aggression data

```
library(Matrix)
library(lattice)
library(lme4)
library(arm)
VA < − read.table("C:/VA.txt",header=T)
GENDER1 < − factor(VA$gender)
WANTDO1 < − factor(VA$wantdo)
fm1 < − lmer(y ∼ 1 + gender + wantdo + gender*wantdo + expressing + blaming +
(WANTDO1-1|person) + (GENDER1-1|item), VA, binomial)
ranef(fm1)
se.ranef(fm1)
```

Appendix B

WinBUGS code for verbal aggression data

```
# p: Person index
# i: Item index
# g: Person group
# k: Item type
# alpha: Fixed item type effect
# gamma: Fixed person group effect
# lambda: Fixed interaction effect
# theta: Random person effect
# zeta: Random item effect
# beta1: The effect of blamin
# beta2: The effect of expressing
# Pr: Probability
# Part I: Model Specificaiton
model {
for (p in 1:P) { for (i in 1:I) {
logit(Pr[p,i]) < − mu + alpha[k[i]] + gamma[g[p]] + beta1*blaming[i] +
beta2*expressing[i] + lambda[k[i],g[p]] + theta[p,k[i]] + zeta[i,g[p]]
resp[p,i] ∼ dbern(Pr[p,i])
} }
# Priors for fixed effects
mu ∼ dnorm(0, 0.1)
alpha[1] ∼ dnorm(0, 0.1)
alpha[2] < − - alpha[1]
gamma[1] ∼ dnorm(0, 0.1) gamma[2] < − -gamma[1]
# Interaction structure
lambda[1,1] ∼ dnorm(0, 0.1)
lambda[2,2] < − lambda[1,1]
lambda[2,1] < − lambda[1,2]
lambda[1,2] < − -lambda[1,1]
beta1 ∼ dnorm(0,0.1)
```

```
beta2 ~ dnorm(0,0.1)
# Bivariate normal for persons
for (p in 1:P) {
theta[p,1:2] ~ dmnorm(mu.th[1:2],R.th[1:2,1:2])
}
mu.th[1] < − 0; mu.th[2] < − 0
R.th[1:2,1:2] ~ dwish(Omega.th[1:2,1:2], 2)
IR.th[1:2,1:2] < − inverse(R.th[1:2,1:2])
de.th < − sqrt(IR.th[1,1])*sqrt(IR.th[2,2])
corr.th < − IR.th[1,2]/de.th
# Bivariate normal for items
for (i in 1:I) {
zeta[i,1:2] ~ dmnorm(mu.it[1:2],R.it[1:2,1:2])
}
mu.it[1] < − 0; mu.it[2] < − 0
R.it[1:2,1:2] ~ dwish(Omega.it[1:2,1:2], 2)
IR.it[1:2,1:2] < − inverse(R.it[1:2,1:2])
de.it < − sqrt(IR.it[1,1])*sqrt(IR.it[2,2])
corr.it < − IR.it[1,2]/de.it
}
# Part II: Set initial values
list(R.it=structure(.Data=c(0.1,0,0,0.1), .Dim=c(2,2)), R.th=structure(.Data=c(0.1,0,
0,0.1), .Dim=c(2,2)))
list( R.it=structure(.Data=c(0.25,0,0,0.25), .Dim=c(2,2)), R.th=structure(.Data=c
(0.25,0,0,0.25), .Dim=c(2,2)))
list(R.it=structure(.Data=c(1,0,0,1), .Dim=c(2,2)), R.th=structure(.Data=c(1,0,0,1),
.Dim=c(2,2)))
list(R.it=structure(.Data=c(0.16,0,0,0.16), .Dim=c(2,2)), R.th=structure(.Data=c
(0.16,0,0,0.16), .Dim=c(2,2)))
# Part III: Load data
# g(gender): 1-women, 2-men
# k: 1-want, 2-do
list(P=316, I=90,
Omega.th=structure(.Data=c(1,0,0,1), .Dim=c(2,2)),
Omega.it=structure(.Data=c(1,0,0,1), .Dim=c(2,2)),
blaming=c( 0.5,0.5,-1,0.5,0.5,-1,0.5,0.5,-1,0.5,0.5, -1,0.5,0.5,-1,0.5,0.5,-1,0.5,0.5,-1,
0.5, ... -1,0.5,0.5,-1,0.5,0.5,-1,0.5,0.5,-1,0.5, 0.5,-1,0.5,0.5,-1,0.5,0.5,-1,0.5,0.5,-1),
expressing=c( 0.5,-1,0.5,0.5,-1,0.5,0.5,-1,0.5,0.5,-1, 0.5,0.5,-1,0.5,0.5,-1,0.5,0.5,-1,0.5,
0.5, ... 0.5,0.5,-1,0.5,0.5,-1,0.5,0.5,-1,0.5,0.5, -1,0.5,0.5,-1,0.5,0.5,-1,0.5,0.5,-1,0.5),
k=c( 1,1,1,1,1,1,1,1,1,1,1, 1,1,1,1,1,1,1,1,1,1,1, ... 2,2,2,2,2,2,2,2,2,2,2, 2,2,2,2,2,
2,2,2,2,2,2),
g=c( 2,2,1,1,1,1,1,1,1,1,1, 1,2,2,1,1,1,1,1,1,1,1, ... 1,1,1,2,1,1,1,1,2,2,1, 2,1,1,1,1,
2,2,1,1,1,1),
resp=structure(.Data =c( 0,0,0,0,0,0,0,0,0,1,0,0, 0,0,0,0,0,0,0,0,0,0,0,0, ... 1,1,1,0,
0,1,1,0,1,1,0,0, 1,0,1,1,1,1,0,0,0,1,0,1) .Dim= c(316,90)))
```