Journal of Statistical Software

April 2007, Volume 20, Issue 2. http://www.jstatsoft.org/

Estimating the Multilevel Rasch Model: With the lme4 Package

Harold Doran
American Institutes for Research

Douglas Bates
University of Wisconsin – Madison

Paul Bliese
Walter Reed Army Institute of Research

Maritza Dowling
University of Wisconsin – Madison

Abstract

Traditional Rasch estimation of the item and student parameters via marginal maximum likelihood, joint maximum likelihood or conditional maximum likelihood, assume individuals in clustered settings are uncorrelated and items within a test that share a grouping structure are also uncorrelated. These assumptions are often violated, particularly in educational testing situations, in which students are grouped into classrooms and many test items share a common grouping structure, such as a content strand or a reading passage. Consequently, one possible approach is to explicitly recognize the clustered nature of the data and directly incorporate random effects to account for the various dependencies. This article demonstrates how the multilevel Rasch model can be estimated using the functions in R for mixed-effects models with crossed or partially crossed random effects. We demonstrate how to model the following hierarchical data structures: a) individuals clustered in similar settings (e.g., classrooms, schools), b) items nested within a particular group (such as a content strand or a reading passage), and c) how to estimate a teacher × content strand interaction.

Keywords: generalized linear mixed models, item response theory, sparse matrix techniques.

1. Introduction

The analysis of response data to test items or survey questions often requires psychometric methods of analysis to investigate properties of the items or characteristics of individuals taking those items. Item response theory (IRT) is the prominent application for behavioral scientists involved in such analyses (Lord 1980). In most cases, specialized software programs

are used to calibrate items such as **WINSTEPS** (Linacre 2006), **BILOG-MG** (Zimowski, Muraki, Mislevy, and Bock 2005), **PARSCALE** (Muraki and Bock 2005), or **ltm** (Rizopoloulos 2006) in R (R Development Core Team 2007), with each program using a different estimating algorithm such as joint maximum likelihood (JML) or marginal maximum likelihood (MML).

Using traditional likelihood estimation, such as JML or MML, a strong assumption is posited referred to as local item independence. That is, items sharing a particular grouping structure are independent from one another. When tenable, the joint density is the product of the individual densities and maximum likelihood estimation proceeds in a straightforward manner. However, this assumption is often violated, especially in educational settings where test items share a common stem, such as a multiple items in a reading passage or multiple math items based on the same table or graphic. A second form of dependence is also found when individuals participating in the test share a common grouping structure, such as a classroom.

These important dependencies are commonly ignored, resulting in estimates that may be inconsistent and standard errors that do not adequately characterize the true variance. One option to consider in the presence of a non-zero design effect (Kish 1965) is to regard the point estimates as retaining some utility, but construct robust standard errors in recognition of the fact that correlated observations provide less information than an equivalent number from a simple random sample (Binder 1983; Cohen, Jiang, and Seburn 2005).

A second option is to incorporate random effects in the structural model to aptly model the various dependencies Johnson and Raudenbush (2006); Kamata (2001). In this paper we describe how to use the `lmer` function in the **lme4** package (Bates and Sarkar 2007) to fit the Rasch model and to fit extensions to the Rasch model that take into account correlation of the scores for groups of students or groups of items. We begin by explicitly linking the Rasch measurement model with generalized linear models for the reader to better understand why software for fitting generalized linear models can also serve as a tool for item response applications.

1.1. The Rasch model

The analysis of item response data often begins from the classical linear measurement model $x_{is} = \theta_i + \epsilon_{is}$ where x_{is} is the observed score for individual i to item s, θ_i is the true score for individual i, and ϵ_{is} is the error term.

Given a dichotomous response variable, it is only observed that:

$$y_{is} = \begin{cases} 1 \text{ if } x_{is} > b_s, \\ 0 \text{ otherwise} \end{cases}$$

where b_s is a threshold for item s. The probability that $y_{is} = 1$ conditional on the student's true score and the item threshold can be expressed as:

$$\begin{aligned} \text{Prob}(y_{is} = 1|\theta_i, b_s) &= \text{Prob}(\theta_i + \epsilon_{is} > b_s) \\ &= \text{Prob}(\epsilon_{is} > b_s - \theta_i) \end{aligned} \tag{1}$$

As noted by Greene (2000), if the distribution of the disturbance term is symmetric, such as the standard logistic, then

$$\begin{aligned} \text{Prob}(y_{is} = 1|\theta_i, b_s) &= \text{Prob}(\epsilon_{is} < \theta_i - b_s) \\ &= F(\theta_i - b_s) \end{aligned} \tag{2}$$

Under the assumption that the disturbances are logistic, we can form the following logit model:

$$\text{logit}(\text{Prob}(y_{is} = 1|\theta_i, b_s)) = \log\left(\frac{\text{Prob}(y_{is} = 1|\theta_i, b_s)}{1 - \text{Prob}(y_{is} = 1|\theta_i, b_s)}\right) \tag{3}$$

which gives rise to the familiar Rasch model:

$$\text{Prob}(Y_{is} = 1|\theta_i, b_s) = P_{is} = \frac{1}{1 + \exp(b_s - \theta_i)} \qquad i = 1, \ldots, m;\ s = 1, \ldots, n \tag{4}$$

The parameters b_s, $s = 1, \ldots, m$ are called the *item difficulties* and the θ_i, $i = 1, \ldots, n$ are the *subject abilities*.

This development permits for us to view the Rasch model as connected to the classicial measurement model given certain assumptions regarding the distribution of the error term. Subsequently, traditional item response theory (IRT) further assumes that responses to test items are conditionally independent, which gives rise to the following likelihood:

$$\mathcal{L} = \prod P_{is}^{y_{is}}(1 - P_{is})^{1-y_{is}} \tag{5}$$

First order conditions necessary for maximization of Equation (5) are simple to derive thus making traditional estimation methods, such as marginal maximum likelihood, joint maximum likelihood, or conditional maximum likelihood, for the parameters in the Rasch model quite simple to evaluate when every subject is scored on every item (i.e. the subject and item factors are *completely crossed*) and we can assume that the scores for different subjects are independent and, for a given subject, the scores on different items are independent.

1.2. Generalized linear models

As described in McCullagh and Nelder (1989), a generalized linear model is a statistical model in which the *linear predictor* for the ith response, $\eta_i = \boldsymbol{x}_i\boldsymbol{\beta}$ where $\boldsymbol{x}_i$ is the ith row of the $n \times p$ model matrix $\boldsymbol{X}$ derived from the form of the model and the values of any covariates, is related to the *expected value of the response*, μ_i, through an invertible *link function*, g. That is

$$\boldsymbol{x}_i\boldsymbol{\beta} = \eta_i = g(\mu_i) \quad i = 1, \ldots, n \tag{6}$$

and

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\boldsymbol{x}_i\boldsymbol{\beta}) \quad i = 1, \ldots, n \tag{7}$$

The *natural* link (McCullagh and Nelder 1989) for a binomial response is the *logit* link defined as

$$\eta_i = g(\mu_i) = \log\left(\frac{\mu_i}{1 - \mu_i}\right) \quad i = 1, \ldots, n \tag{8}$$

with inverse link

$$\mu_i = g^{-1}(\eta_i) = \frac{1}{1 + \exp(-\eta_i)} \quad i = 1, \ldots, n \tag{9}$$

from which we can see the relationship to the Rasch model as developed in (4). Because μ_i is the probability of the ith observation being a "success", η_i is the log of the odds ratio.

The parameters $\boldsymbol{\beta}$ in a generalized linear model are generally estimated by *iteratively reweighted least squares* (IRLS). At each iteration in this algorithm the current parameter estimates are replaced by the parameter estimates of a weighted least squares fit with model matrix $\boldsymbol{X}$ to an adjusted dependent variable. The weights and the adjusted dependent variable are calculated from the link function and the current parameter values.

1.3. Extension to clustered settings

The issue at hand, however, is that when individuals are grouped into similar settings (e.g., students in classrooms) the assumption of independence may no longer remain tenable. For example, we would expect the scores of students in the same classroom to be correlated and we would expect scores on items in a topic group to be correlated. As such, traditional methods for obtaining parameter estimates may return biased parameter estimates or incorrect standard errors (Müller 2004; McCullagh and Nelder 1989). In other words, it is no longer the case that $\text{cov}(\epsilon_{j(i)}, \epsilon_{j(i')}) = 0$ for all $i \neq i'$, where the notation $j(i)$ denotes the nesting of unit i in the group j. In fact, the clustering of units into similar groups typically results in a clustering of the measurement error term, $\epsilon_{j(i)} = \nu_j + \epsilon_{ij}$, where $\nu_j \sim \mathcal{N}(0, \sigma^2_\nu)$ and $\epsilon_{ij} \sim \mathcal{L}(0, \sigma^2_\epsilon)$ where $\mathcal{L}$ denotes that the disturbances follow a standard logistic distribution.

This covariance among units within a group motivates us to consider incorporating random effects into the linear predictor in order to more aptly handle the dependencies among units in similar groups. In other words, the error term can now be viewed as having *multiple levels* of random variation, thereby making a *multilevel* statistical model an appropriate approach

1.4. Generalized linear mixed models

In a generalized linear mixed model (GLMM) the n-dimensional vector of linear predictors, $\boldsymbol{\eta}$, incorporates both fixed effects, $\boldsymbol{\beta}$, and random effects, $\boldsymbol{b}$, as

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} \tag{10}$$

where $\boldsymbol{X}$ is an $n \times p$ model matrix and $\boldsymbol{Z}$ is an $n \times q$ model matrix.

Each component of the random effects vector $\boldsymbol{b}$ is associated with a level of a *grouping factor* such as "student" or "class" or "item". Because the number of levels of a factor such as "student" can be very large, the dimension, q, of the random effects vector, $\boldsymbol{b}$, can be very large. In one of the examples in Section 3 q is over 8000.

We model the distribution of the random effects as a multivariate normal (Gaussian) distribution with mean $\boldsymbol{0}$ and $q \times q$ variance-covariance matrix $\boldsymbol{\Sigma}$. That is,

$$\boldsymbol{b} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})\right). \tag{11}$$

Although $\boldsymbol{\Sigma}$ is a very large matrix, it is determined by a parameter vector, $\boldsymbol{\theta}$, whose dimension is typically very small. In the example from Section 3 where q is over 8000, the dimension of $\boldsymbol{\theta}$ is only 5.

The maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ maximize the likelihood of the parameters, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, given the observed data, $\boldsymbol{y}$. This likelihood is numerically equivalent to the marginal

density of $\boldsymbol{y}$ given $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, which is

$$f(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\theta}) = \int_{\boldsymbol{b}} p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{b}) f(\boldsymbol{b}|\boldsymbol{\Sigma}(\boldsymbol{\theta}))\, d\boldsymbol{b} \tag{12}$$

where $p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{b})$ is the probability mass function of $\boldsymbol{y}$, given $\boldsymbol{\beta}$ and $\boldsymbol{b}$, and $f(\boldsymbol{b}|\boldsymbol{\Sigma})$ is the (Gaussian) probability density at $\boldsymbol{b}$.

Unfortunately the integral in (12) does not have a closed-form solution when $p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{b})$ is binomial. However, we can approximate this integral quite accurately using a *Laplace* approximation. For given values of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ we determine the *conditional modes* of the random effects

$$\tilde{\boldsymbol{b}}(\boldsymbol{\beta},\boldsymbol{\theta}) = \arg\max_{\boldsymbol{b}} p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{b}) f(\boldsymbol{b}|\boldsymbol{\Sigma}(\boldsymbol{\theta})), \tag{13}$$

which are the values of the random effects that maximize the conditional density of the random effects given the data and the model parameters. The conditional modes can be determined by a penalized iteratively reweighted least squares algorithm (PIRLS, see Section 2.1) where the contribution of the fixed effects parameters, $\boldsymbol{\beta}$, is incorporated as an offset, $\boldsymbol{X}\boldsymbol{\beta}$, and the contribution of the variance components, $\boldsymbol{\theta}$, is incorporated as a penalty term in the weighted least squares fit.

At the conditional modes, $\tilde{\boldsymbol{b}}$, we evaluate the second order Taylor series approximation to the log of the integrand (i.e. the log of the conditional density of $\boldsymbol{b}$) and use its integral as an approximation to the likelihood.

It is the Laplace approximation to the likelihood that is optimized to obtain approximate values of the mle's for the parameters and the corresponding conditional modes of the random effects vector $\boldsymbol{b}$.

1.5. Fixed-effects parameters versus random-effects parameters

In addition to having potentially a very large number of levels, a grouping factor for a random effect typically has levels that are not *repeatable* in the sense that, if the experiment were to be repeated it would be with different levels of this grouping factor. For example, if we were to repeat a test for a given set of students we would generally use a different set of items. Conversely if we were to administer the same test to a new group of students then the levels of the student factor would be different.

Frequently the levels of such a factor represent a sample from a population. For example, the items used on a particular test can be considered as a sample from the set of all possible items on the subject matter. Even if we administer a test to every subject in a population we can consider this to be an exhaustive sample from the population. Hence we model both the subjects' abilities and the item difficulties as random-effects terms.

Factors with repeatable levels, such as item types, can be modeled as fixed-effects terms.

An interaction between a factor that is modeled as a random effect and another factor that is modeled as a fixed effect is modeled as a random effect.

1.6. Nested versus non-nested grouping factors

In addition to distinguishing between grouping factors that are modeled as fixed effects terms and those that are modeled as random effects, it is helpful to distinguish between nested and

non-nested structures for grouping factors. Grouping factor $\mathcal{A}$ is said to be *nested within* grouping factor $\mathcal{B}$, written $\mathcal{A} \preceq \mathcal{B}$, if each level of $\mathcal{A}$ occurs in conjunction with one and only one level of $\mathcal{B}$. For example, if each student is observed in only one class then the student grouping factor is nested within the class grouping factor.

Obviously, if $\mathcal{A} \preceq \mathcal{B}$ then the number of levels in $\mathcal{A}$ cannot be less than the number of levels in $\mathcal{B}$, with equality occuring only in the case that $\mathcal{A}$ and $\mathcal{B}$ are identical up to changes in the names of the levels (in which case both $\mathcal{A} \preceq \mathcal{B}$ and $\mathcal{B} \preceq \mathcal{A}$ hold). A collection of grouping factors is said to be *strictly nested* if, when ordered according to non-decreasing numbers of levels, each factor is nested within its successor. Otherwise the collection is said to be *non-nested*.

Grouping factors $\mathcal{A}$ and $\mathcal{B}$ are said to be *completely crossed* if every level of $\mathcal{A}$ occurs in conjunction with every level of $\mathcal{B}$. For example if each student takes the same test then the student and item grouping factors are completely crossed. Completely crossed factors are an example of non-nested factors. In fact, they are an extreme example of non-nested factors.

Many computational methods for mixed models with multiple grouping factors are designed for hierarchical models (also called multilevel models) in which the grouping factors form a strictly nested sequence. In such a case the model matrix $\boldsymbol{Z}$ for the random effects, defined in Section 2.1 below, has a special structure that induces a simple, separable structure on $\boldsymbol{Z}^\top\boldsymbol{Z}$ and matrices derived from it, such as $\boldsymbol{Z}^\top\boldsymbol{W}^{(r)}\boldsymbol{Z} + \boldsymbol{\Sigma}^{-1}$ in Equation 18. Computational methods that assume and exploit such a structure are quite effective for hierarchical models but, although they can be "tricked" into fitting a model with crossed or partially crossed grouping factors, they generally are slow and memory-inefficient on such models.

In contrast, the computational methods used in the `lmer` function from the **lme4** package (described in the next section) do not assume nested grouping factors. They are effective and efficient for models with nested or with partially crossed or with completely crossed grouping factors. Thus they allow for fitting IRT models and generalization of IRT models with random effects for subject and for item.

2. Design of `lmer`

2.1. Details of the PIRLS algorithm

Recall from (13) that the conditional modes of the random effects $\tilde{\boldsymbol{b}}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y})$ maximize the conditional density of $\boldsymbol{b}$ given the data and values of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. The penalized iteratively reweighted least squares (PIRLS) algorithm for determining these conditional modes combines characteristic of the iteratively reweighted least squares (IRLS) algorithm for generalized linear models (McCullagh and Nelder 1989, Section 2.5) and the penalized least squares representation of a linear mixed model (Bates and DebRoy 2004).

At the rth iteration of the IRLS algorithm the current value of the vector of random effects. $\boldsymbol{b}^{(r)}$ (we use parenthesized superscripts to denote the iteration) produces a linear predictor

$$\boldsymbol{\eta}^{(r)} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} \tag{14}$$

with corresponding mean vector $\boldsymbol{\mu}^{(r)} = \boldsymbol{g}^{-1}\boldsymbol{\eta}^{(r)}$. (The vector-valued link and inverse link functions, $\boldsymbol{g}$ and $\boldsymbol{g}^{-1}$, apply the scalar link and inverse link, g and g^{-1}, componentwise.)

A vector of weights and a vector of derivatives of the form $d\eta/d\mu$ are also evaluated. For convenience of notation we express these as diagonal matrices, $\boldsymbol{W}^{(r)}$ and $\boldsymbol{G}^{(r)}$, although calculations involving these quantities are performed component-wise and not as matrices.

The adjusted dependent variate at iteration r is

$$\boldsymbol{z}^{(r)} = \boldsymbol{\eta}^{(r)} + \boldsymbol{G}^{(r)}\left(\boldsymbol{y} - \boldsymbol{\mu}^{(r)}\right) \tag{15}$$

from which the updated parameter, $\boldsymbol{b}^{(r+1)}$, is determined as the solution to

$$\boldsymbol{Z}^\top \boldsymbol{W}^{(r)} \boldsymbol{Z} \boldsymbol{b}^{(r+1)} = \boldsymbol{Z}^\top \boldsymbol{z}^{(r)}. \tag{16}$$

McCullagh and Nelder (1989, Section 2.5) show that the IRLS algorithm is equivalent to the Fisher scoring algorithm for any link function and also equivalent to the Newton-Raphson algorithm when the link function is the natural link for a probability distribution in the exponential family. That is, IRLS will minimize $-\log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{b})$ for fixed $\boldsymbol{\beta}$. However, we wish to determine

$$\begin{aligned} \tilde{\boldsymbol{b}}(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \arg\max_{\boldsymbol{b}} p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{b}) f(\boldsymbol{b}|\boldsymbol{\Sigma}(\boldsymbol{\theta})) \\ &= \arg\min_{\boldsymbol{b}} \left[-\log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{b}) + \frac{\boldsymbol{b}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \boldsymbol{b}}{2} \right]. \end{aligned} \tag{17}$$

As shown in Bates and DebRoy (2004) we can incorporate the contribution of the Gaussian distribution by adding q "pseudo-observations" with constant unit weights, observed values of 0 and predicted values of $\boldsymbol{\Delta}(\boldsymbol{\theta})\boldsymbol{b}$ where $\boldsymbol{\Delta}$ is any $q \times q$ matrix such that $\boldsymbol{\Delta}^\top \boldsymbol{\Delta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})$.

Thus the update in the penalized iteratively reweighted least squares (PIRLS) algorithm for determining the conditional modes, $\tilde{\boldsymbol{b}}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y})$, expresses $\boldsymbol{b}^{(r+1)}$ as the solution to the penalized weighted least squares problem

$$\left(\boldsymbol{Z}^\top \boldsymbol{W}^{(r)} \boldsymbol{Z} + \boldsymbol{\Sigma}^{-1}\right) \boldsymbol{b}^{(r+1)} = \boldsymbol{Z}^\top \boldsymbol{z}^{(r)}. \tag{18}$$

The sequence of iterates $\boldsymbol{b}^{(0)}, \boldsymbol{b}^{(1)}, \ldots$ is considered to have converged to the conditional modes $\tilde{\boldsymbol{b}}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y})$ when the relative change in the linear predictors $\|\boldsymbol{\eta}^{(r+1)} - \boldsymbol{\eta}^{(r)}\|/\|\boldsymbol{\eta}^{(r)}\|$ falls below a threshold. The variance-covariance matrix of $\boldsymbol{b}$, conditional on $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, is approximated as

$$\mathrm{Var}\left(\boldsymbol{b}|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y}\right) \approx \boldsymbol{D} \equiv \left(\boldsymbol{Z}^\top \boldsymbol{W}^{(r)} \boldsymbol{Z} + \boldsymbol{\Sigma}^{-1}\right)^{-1}. \tag{19}$$

This approximation is analogous to using the inverse of Fisher's information matrix as the approximate variance-covariance matrix for maximum likelihood estimates.

2.2. Details of the Laplace approximation

The Laplace approximation to the likelihood $L(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{y})$ is obtained by replacing the logarithm of the integrand in (12) by its second-order Taylor series at the conditional maximum, $\tilde{\boldsymbol{b}}(\boldsymbol{\beta}, \boldsymbol{\theta})$. On the scale of the deviance (negative twice the log-likelihood) the approximation is

$$\begin{aligned} -2\ell(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{y}) &= -2\log\left\{\int_{\boldsymbol{b}} p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{b}) f(\boldsymbol{b}|\boldsymbol{\Sigma}(\boldsymbol{\theta}))\, d\boldsymbol{b}\right\} \\ &\approx 2\log\left\{\int_{\boldsymbol{b}} \exp\left\{-\frac{1}{2}\left[d(\boldsymbol{\beta}, \tilde{\boldsymbol{b}}, \boldsymbol{y}) + \tilde{\boldsymbol{b}}^\top \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{b}} + \log|\boldsymbol{\Sigma}| + \boldsymbol{b}^\top \boldsymbol{D}^{-1} \boldsymbol{b}\right]\right\} d\boldsymbol{b}\right\} \\ &= d(\boldsymbol{\beta}, \tilde{\boldsymbol{b}}, \boldsymbol{y}) + \tilde{\boldsymbol{b}}^\top \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{b}} + + \log|\boldsymbol{\Sigma}| + \log|\boldsymbol{D}| \end{aligned} \tag{20}$$

where $d(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{y})$ is the deviance function from the linear predictor only. That is, $d(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{y}) = -2 \log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{b})$. This quantity can be evalated as the sum of the deviance residuals (McCullagh and Nelder 1989, Section 2.4.3).

2.3. Sparse matrix methods

The PIRLS algorithm for determining the conditional modes of the random effects and the use of the Laplace approximation (20) to the deviance require the solution of the positive-definite system of linear equations (18) and evaluation of $\log |\boldsymbol{D}| = -\log |\boldsymbol{Z}^\top \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{\Sigma}^{-1}|$. One way to accomplish both of these tasks is to obtain the Cholesky decomposition of $\boldsymbol{Z}^\top \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{\Sigma}^{-1}$. One way of writing the Cholesky decomposition is as a lower triangular matrix $\boldsymbol{L}$ such that

$$\boldsymbol{L}\boldsymbol{L}^\top = \boldsymbol{Z}^\top \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{\Sigma}^{-1} \tag{21}$$

from which we obtain $\log |\boldsymbol{D}| = -\log |\boldsymbol{Z}^\top \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{\Sigma}^{-1}| = -2 \log |\boldsymbol{L}|$. Because $\boldsymbol{L}$ is triangular its determinant is easily evaluated as the product of its diagonal elements. The triangularity of $\boldsymbol{L}$ also simplifies solution of a linear system like (18).

Even when using the Cholesky decomposition we must be careful when working with $q \times q$ matrices for large q, which can be the case for the models we are considering. Fortunately the matrix $\boldsymbol{Z}^\top \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{\Sigma}^{-1}$ is sparse and the Cholesky decomposition of sparse, positive definite, symmetric matrices has been studied extensively. Davis (2006) gives a general introduction to sparse matrix methods and describes his **Csparse** library of C functions that implement these methods. Another library of C functions called **CHOLMOD**, also written by Tim Davis, implements more sophisticated algorithms for the sparse Cholesky decomposition, including the supernodal Cholesky decomposition used in the `lmer` function.

The **Csparse** and **CHOLMOD** libraries of C functions for sparse matrices are incorporated in the **Matrix** package for R.

Operations with sparse matrices are often performed in two stages: a symbolic stage in which the number and positions of the non-zero elements in the result are determined and a numeric stage in which the numerical values of these elements are calculated. For a Cholesky decomposition the symbolic phase can be particularly important because the number of nonzeros in $\boldsymbol{L}$ can be changed dramatically by permuting the rows and columns of the original positive definite matrix. Determining a fill-reducing permutation can be time consuming but doing so can save considerable time and storage in the subsequent numerical phase.

Although the numeric values of the nonzeros in $\boldsymbol{Z}^\top \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{\Sigma}^{-1}$ change with each iteration of the PIRLS algorithm and with every change in $\boldsymbol{\theta}$ during the optimization of the Laplace approximation, the number and positions of these nonzeros are constant. Thus we only need to perform the symbolic computation once. The numeric computation is performed many times. Both the **Csparse** and the **CHOLMOD** libraries allow the symbolic computation to be performed separately from the numeric computation.

3. Code and examples

In this section we show how `lmer` can be used to fit multilevel Rasch models and generalizations of these models to a dichotomized version of the responses in the `lq2002` data in the **multilevel** package.

3.1. Preliminary data manipulation

The `lq2002` data contain the responses of 2042 soldiers to a total of 19 items, 11 of which are related to leadership, 3 of which are related to task significance and 5 of which measure the hostility felt by the soldier. The soldiers are grouped into 49 companies. Thus both the subjects and the items are grouped.

The responses to the leadership and task significance questions are on a 1 to 5 scale where 5 indicates strong positive feelings. The responses to the hostility questions are on a 0 to 4 scale where 0 indicates no hostility and 4 indicates strong feelings of hostility. We therefore dichotomized the responses so that a positive response (1) was a 4 or a 5 on the leadership and task significance questions and a 0 or a 1 on the hostility questions.

The `lmer` function requires the data to be in the "long" or "subject-item" form where each row corresponds contains the response of one subject to one item. The `reshape` function can be used to convert a data set in the "wide" format to the "long" format and to add appropriate indicators of the item and subject. The indicator of item type must be added separately.

```
R> data("lq2002", package = "multilevel")
R> wrk <- lq2002
R> for (i in 3:16) wrk[[i]] <- ordered(wrk[[i]])
R> for (i in 17:21) wrk[[i]] <- ordered(5 - wrk[[i]])
R> lql <- reshape(wrk, varying = list(names(lq2002)[3:21]),
+     v.names = "fivelev", idvar = "subj", timevar = "item",
+     drop = names(lq2002)[c(2, 22:27)], direction = "long")
R> lql$itype <- with(lql, factor(ifelse(item < 12, "Leadership",
+     ifelse(item < 15, "Task Sig.", "Hostility"))))
R> for (i in c(1, 2, 4, 5)) lql[[i]] <- factor(lql[[i]])
R> lql$dichot <- factor(ifelse(lql$fivelev < 4, 0, 1))

R> str(lql)

'data.frame':        38798 obs. of  6 variables:
 $ COMPID : Factor w/ 49 levels "2","3","4","5",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ item   : Factor w/ 19 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ fivelev: Ord.factor w/ 5 levels "1"<"2"<"3"<"4"<..: 2 4 4 1 1 2 3 3 3 4 ...
 $ subj   : Factor w/ 2042 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ itype  : Factor w/ 3 levels "Hostility","Leadership",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ dichot : Factor w/ 2 levels "0","1": 1 2 2 1 1 1 1 1 1 2 ...

R> summary(lql)

     COMPID           item       fivelev        subj
 13     : 1881   1      : 2042   1: 5208   1      :   19
 18     : 1786   2      : 2042   2: 5357   2      :   19
 46     : 1710   3      : 2042   3: 8879   3      :   19
 15     : 1691   4      : 2042   4:11353   4      :   19
 29     : 1615   5      : 2042   5: 8001   5      :   19
 34     : 1482   6      : 2042             6      :   19
```

```
 (Other):28633   (Other):26546             (Other):38684
        itype       dichot
 Hostility :10210   0:19444
 Leadership:22462   1:19354
 Task Sig. : 6126
```

3.2. Fitting an initial multilevel Rasch model

Because the item type (`itype`) is a repeatable factor and of interest itself, we model it as a fixed-effects term. The subject (`subj`), company (`COMPID`) and item (`item`) factors are modeled as random effects. An initial model fit with fixed effects for `itype` and random effects for `subj`, `COMPID` and `item` is

```
R> (fm1 <- lmer(dichot ~ 0 + itype + (1 | subj) + (1 | COMPID) +
+     (1 | item), lql, binomial))

Generalized linear mixed model fit using Laplace
Formula: dichot ~ 0 + itype + (1 | subj) + (1 | COMPID) + (1 | item)
   Data: lql
 Family: binomial(logit link)
   AIC   BIC logLik deviance
 40722 40773 -20355    40710
Random effects:
 Groups Name        Variance Std.Dev.
 subj   (Intercept) 2.30528  1.51831
 COMPID (Intercept) 0.25449  0.50447
 item   (Intercept) 0.37700  0.61400
number of obs: 38798, groups: subj, 2042; COMPID, 49; item, 19

Estimated scale (compare to  1 )  0.9386558

Fixed effects:
                Estimate Std. Error z value Pr(>|z|)
itypeHostility    1.6721     0.2883   5.801  6.6e-09
itypeLeadership  -0.4921     0.2036  -2.417   0.0157
itypeTask Sig.   -0.1308     0.3654  -0.358   0.7203

Correlation of Fixed Effects:
            itypHs itypLd
itypeLdrshp 0.117
itypeTskSg. 0.066  0.093
```

The coefficients in a generalized linear model for a binomial response with the logit link generate the log-odds for a positive response. For this model we are considering three different types of items and, rather than using a coding in which one of the item types is taken as the reference value, we suppress the intercept in the model and estimate a marginal log-odds for

each item type. Thus the log-odds of a positive response on a hostility question (recall that the coding of the response is such that a positive value indicates a relative lack of hostility) is 1.672, corresponding to a probability of 84.19%. Similarly, the marginal log-odds for a leadership question eliciting a positive response is -0.4921 corresponding to a probability of 37.94%.

We extract and save the random effects for this model including the "posterior variances" (or conditional variances given the data) of these random effects.

```
R> rr <- ranef(fm1, postVar = TRUE)
R> str(rr$COMPID)
```

```
'data.frame':        49 obs. of  1 variable:
 $ (Intercept): num   0.0510 -0.0221 -0.1845  0.0502  0.0978 ...
 - attr(*, "postVar")= num [1, 1, 1:49] 0.0812 0.0596 0.0523 0.0437 0.1193 ...
```

```
R> head(rr$COMPID)
```

```
  (Intercept)
2  0.05095204
3 -0.02208041
4 -0.18445551
5  0.05016374
6  0.09781214
7  0.34845152
```

The value of `ranef()` applied to this model is a list with three components named `subj`, `COMPID` and `item`. Each of these components is a data frame with, in this case, a single column and one row for each level of the grouping factor. The contents are the conditional modes of the random effects evaluated at the parameter estimates.

In Figure 1 we present normal probability plots of the conditional modes of the random effects for the each of the three grouping factors.

```
R> qq <- qqmath(rr)
R> print(qq$subj)
```

To provide a measure of the precision of the conditional distribution of these random effects we add lines extending ± 1.96 conditional standard deviations in each direction from the plotted point. We can see that many of the intervals created in this way overlap with the zero line but for all three of the grouping factors there are several levels that are clearly greater than zero or clearly less than zero.

As indicated by the estimates of the variances of the random effects, the subject factor accounts for the greatest level of variability.

Plots like those in Figure 1 with vertical lines to indicate the precision of the conditional distribution of the random effects are sometimes called "caterpillar plots" because of their appearance.

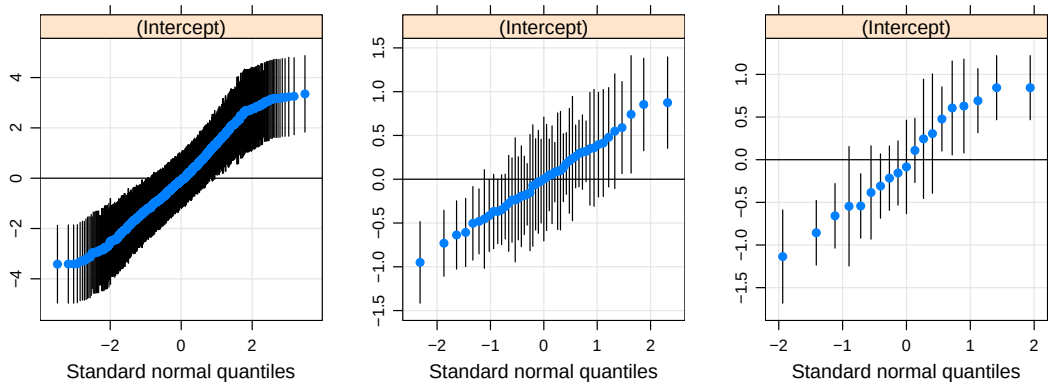

Figure 1: Normal probability plots of the conditional modes of the random effects from model `fm1` for the subject (left panel), company (middle panel) and item (right panel) grouping factors. The precision of the conditional distribution of the random effects is indicated by a line that extends ±1.96 conditional standard deviations in each direction

We could analyze these results in greater detail but first we should check on the possible presence of interactions.

3.3. Allowing for interactions of company and item type

It is of interest to determine if there are significant differences between companies in the probabilities for the different item types. One way to allow for this is to include a random effect for the `COMPID:itype` interaction.

This can be modeled in two different ways: as a random effect for the `COMPID:itype` interaction or by extending the random effect for `COMPID` to be three dimensional with a general variance-covariance matrix. Let us fit the more general model first, using the indicators coding for the random effects for `COMPID`.

```
R> fm2 <- lmer(dichot ~ 0 + itype + (1 | subj) + (0 + itype |
+     COMPID) + (1 | item), lql, binomial)
```

The summary output for this model includes

```
   AIC   BIC logLik deviance
 40334 40428 -20156    40312
Random effects:
 Groups Name             Variance Std.Dev. Corr
 subj   (Intercept)      2.38522  1.54442
 COMPID itypeHostility   0.39297  0.62687
        itypeLeadership  0.36761  0.60631  0.651
        itypeTask Sig.   0.45356  0.67347  0.612 0.168
 item   (Intercept)      0.39092  0.62523
```

from which we can see that the variances of the random effects at the company level for the different item types are similar.

There is some correlation within company between the random effects for the different item types but it could still be of interest to check if a model with independent random effects for the `itype:COMPID` interaction provides an adequate fit.

```
R> fm3 <- lmer(dichot ~ 0 + itype + (1 | subj) + (1 | COMPID:itype) +
+     (1 | item), lql, binomial)
R> fm3a <- lmer(dichot ~ 0 + itype + (1 | subj) + (1 | COMPID:itype) +
+     (1 | COMPID) + (1 | item), lql, binomial)
```

Model `fm3` allows for a random effect for each combination of item type (`itype`) and company (`COMPID`) (in addition to the random effects for subject and item). Model `fm3a` extends model `fm3` by allowing for an overall effect for each company in addition to the effects for the combinations of item type and company.

It is interesting to compare these model fits according to various criteria.

```
R> anova(fm3, fm3a, fm2)
```

```
Data: lql
Models:
fm3: dichot ~ 0 + itype + (1 | subj) + (1 | COMPID:itype) + (1 | item)
fm3a: dichot ~ 0 + itype + (1 | subj) + (1 | COMPID:itype) + (1 | COMPID) +
fm2:     (1 | item)
fm3: dichot ~ 0 + itype + (1 | subj) + (0 + itype | COMPID) + (1 |
fm3a:     item)
     Df    AIC    BIC logLik  Chisq Chi Df Pr(>Chisq)
fm3   6  40352  40403 -20170
fm3a  7  40340  40400 -20163 14.239      1  0.0001610
fm2  11  40334  40428 -20156 14.204      4  0.0066714
```

According to the likelihood ratio tests model `fm3a`, with one more parameter than model `fm3`, is clearly superior to `fm3` and model `fm2`, with four more parameters than model `fm3a`, is clearly superior to `fm3a`. The values of Akaike's Information Criterion (AIC) also favor model `fm2` (AIC and BIC are both on the scale where "smaller is better"). However, Schwartz's Bayesian criterion (BIC) prefers model `fm3a` with model `fm3` close behind. Both these models are clearly superior to model `fm2` according to BIC.

Thus we have a "split decision" in model comparisons according to the information criteria and the hypothesis tests, a not uncommon situation.

Even with this ambiguity it appears that model `fm2` is worthy of further investigation. A scatterplot matrix (Figure 2) of the conditional modes of the trivariate random effects for the `COMPID` factor provides visual verification of the correlation pattern. The company-level random effects for the Task Significance and Leadership questions are essentially uncorrelated but both are positively correlated with the random effect for the Hostility questions. (Recall that the dichotomization of the answers to the Hostility questions was performed in such a way that positive responses indicate a lack of hostility. Thus a more positive attitude regarding leadership and task significance at the company level is associated with lower incidence of hostility.)

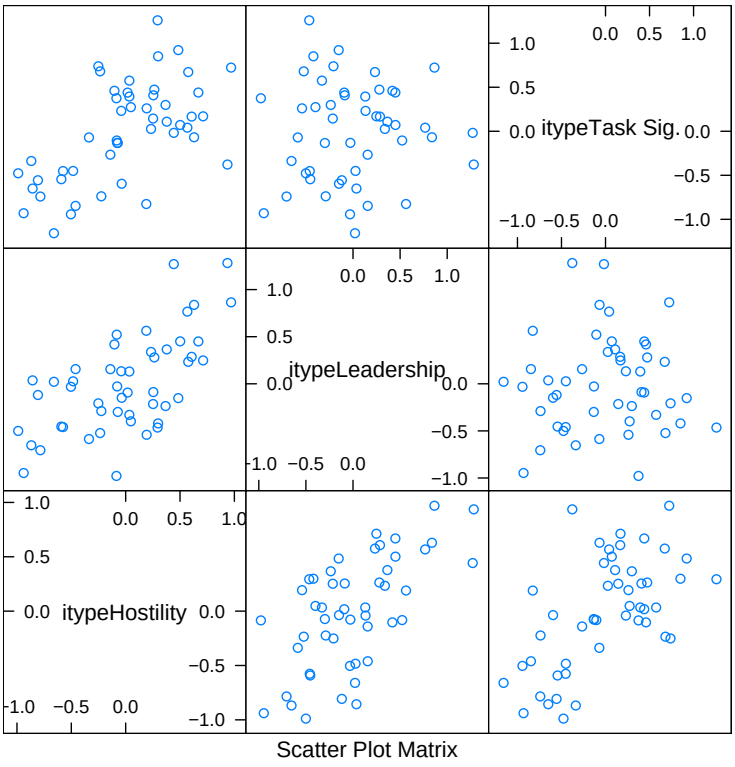

Figure 2: Scatterplot matrix of the conditional modes of the trivariate random effects for `COMPID` in model `fm2`.

The caterpillar plots of the components of the random effects for the `COMPID` factor are shown in Figure 3.

3.4. Extracting item parameters and subject ability estimates

Once a suitable model has been fit, the psychometrician is typically interested in the item parameters. Assuming items were modeled as random effects, the estimated "easiness" parameters (i.e. the negative of the item difficulties $b_s, s = 1, \ldots, n$) are obtained from the estimates of the fixed effects and the conditional modes of the random effects.

For most `lmer` models we could obtain these with the `coef` extractor. In this case we need to do a bit more work because the items are nested in the item types. Because the first item is a leadership question we add the conditional mode for the first level of the `item` factor to the estimate of the leadership fixed effect.

One way of getting the required mapping is to check for the unique combinations of `item` and `itype` and use the resulting table for indexing.

```
R> str(imap <- unique(lql[, c("itype", "item")]))
```

```
'data.frame':        19 obs. of  2 variables:
 $ itype: Factor w/ 3 levels "Hostility","Leadership",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ item : Factor w/ 19 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
```

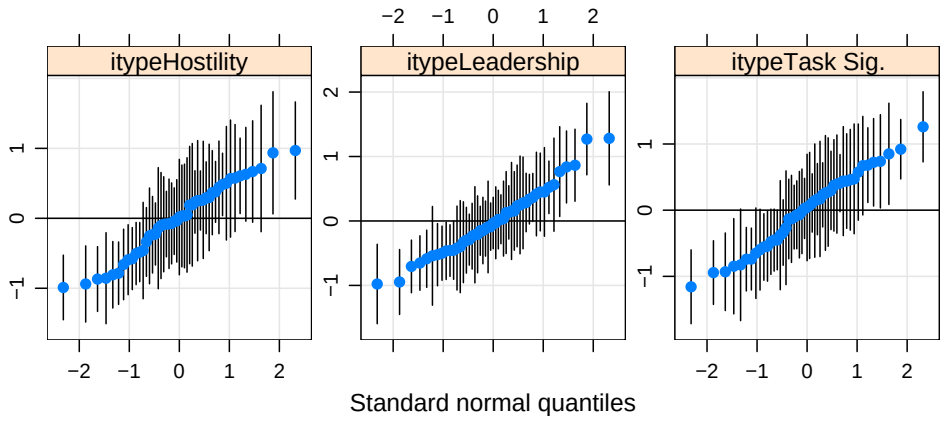

Figure 3: Conditional modes of the random effects for the `COMPID` grouping factor in model `fm2`

```
R> (easiness <- ranef(fm2)$item[[1]] + fixef(fm2)[imap$itype])

itypeLeadership itypeLeadership itypeLeadership itypeLeadership
    -0.39323469      0.35185609     -1.37283998     -0.66116204
itypeLeadership itypeLeadership itypeLeadership itypeLeadership
    -1.05402762      0.19748985     -0.81672112      0.35185609
itypeLeadership itypeLeadership itypeLeadership  itypeTask Sig.
    -1.17151585     -0.01963886     -0.72340696     -0.70154810
 itypeTask Sig.  itypeTask Sig.  itypeHostility  itypeHostility
     0.11001996      0.17329887      0.57992270      2.35226128
 itypeHostility  itypeHostility  itypeHostility
     1.34383628      1.64932424      2.37481906
```

We obtain estimates of the log-odds for a positive response for each company on each item type as

```
R> compPar <- t(fixef(fm2) + t(ranef(fm2)$COMPID))
R> head(compPar)

  itypeHostility itypeLeadership itypeTask Sig.
2       1.494369     -0.71301488     0.59026774
3       2.112532     -0.74182615     0.15087181
4       1.510899     -1.02848250     0.53257639
5       2.039211     -0.97002974     1.11183371
6       1.935374      0.05723364    -0.97401360
7       2.247076     -0.05507709    -0.07656463
```

or, on the probability scale,

```
R> head(binomial()$linkinv(compPar))
```

```
  itypeHostility itypeLeadership itypeTask Sig.
2      0.8167331       0.3289330      0.6434266
3      0.8921152       0.3226049      0.5376466
4      0.8191944       0.2633784      0.6300838
5      0.8848530       0.2748746      0.7524708
6      0.8738431       0.5143045      0.2740812
7      0.9043980       0.4862342      0.4808682
```

These represent typical probabilities for the company/item-type combinations. To obtain a probability for a specific item in a particular company we would need to create the log-odds by adding the random effect for the item to the appropriate company/item-type log-odds then convert the result to the probability scale.

The random effects for the `subj` factor measure the change in the log-odds for a given soldier providing a positive response after accomodating for item and the company/item-type combination.

4. Conclusion

In this paper, we demonstrate how the `lmer` function in R can be a useful tool for psychometric applications. Even though this function is commonly viewed as a tool for generalized and linear mixed models, we demonstrate how the general statistical problem is equivalent with item response theory applications, hence making it tranparent as to why `lmer` can be used as a psychometric tool.

However, `lmer` is significantly more flexible than conventional IRT packages. In particular, the methods demonstrated in this paper do not rely on the untenable assumption that either items or students are independent. Consequently, we are able to freely estimate and account for the covariance structure among items and students that is most commonly ignored. In addition, the multilevel functions sit within a powerful programming environment, making subsequent analyses of the data very convenient.

This provides multiple benefits to the behavorial scientist. For instance, the standard errors associated with the the item and student parameters are more realistic, and more than likely larger than those obtained from conventional methods. One immediate practical benefit is with respect to studies of differential item functioning (DIF). In some cases, DIF is detected when an item behaves significantly different between a focal and reference group. However, with estimation techniques that ignore dependencies in the data, the item standard errors would be too small and one may make claims of DIF when it does not exist.

A second practical benefit is that the sparse matrix methods used by `lmer` are extremely fast, thus making estimation feasible for partially or fully crossed data sets with a large number of items and students. To our knowledge, `lmer` is the only software that can proceed with estimation for large data problems with crossed random effects without reverting to simulation methods such as markov chain monte carlo (MCMC).

While `lmer` is useful for the Rasch model, other IRT models, such as the two- and three-parameter logistic models are currently not available.

References

Bates D, Sarkar D (2007). **lme4***: Linear Mixed-Effects Models Using S4 Classes*. R package version 0.9975-12, URL http://CRAN.R-project.org/.

Bates DM, DebRoy S (2004). "Linear Mixed Models and Penalized Least Squares." *Journal of Multivariate Analysis*, **91**(1), 1–17.

Binder DA (1983). "On the Variances of Asymptotically Normal Estimators from Complex Surveys." *International Statistical Review*, **51**, 279–292.

Cohen J, Jiang T, Seburn M (2005). "Consistent Estimation of Rasch Item Parameters and Their Standard Errors Under Complex Sample Designs." *Technical report*, American Institutes for Research, Washington, DC.

Davis T (2006). *Direct Methods for Sparse Linear Systems*. SIAM, Philadelphia, PA.

Greene WH (2000). *Econometric Analysis*. Prentice-Hall, Saddle River, New Jersey, fourth edition.

Johnson C, Raudenbush S (2006). "A Repeated Measures, Multilevel Rasch Model with Application to Self-reported Criminal Behavior." In CS Bergeman, SM Boker (eds.), "Methodological Issues in Aging Research," pp. 131–64. Lawrence Erlbaum, Hillsdale, New Jersey.

Kamata A (2001). "Item Analysis by the Hierarchical Generalized Linear Model." *Journal of Educational Measurement*, **38**, 79–93.

Kish L (1965). *Survey Sampling*. Wiley, New York.

Linacre JM (2006). *A User's Guide to* **WINSTEPS** *and* **MINISTEP** *– Rasch-Model Computer Programs*. Chicago, IL. ISBN 0-941938-03-4, URL http://www.winsteps.com/.

Lord FM (1980). *Applications of Item Response Theory to Practical Testing Problems*. Erlbaum, Hillsdale, New Jersey.

McCullagh P, Nelder J (1989). *Generalized Linear Models*. Chapman and Hall, 2nd edition.

Müller M (2004). "Generalized Linear Models." In JE Gentle, W Härdle, Y Mori (eds.), "Handbook of Computational Statistics: Concepts and Methods," pp. 592–619. Springer-Verlag, New York.

Muraki E, Bock D (2005). **PARSCALE** *4*. Scientific Software International, Inc., Lincolnwood, IL. URL http://www.ssicentral.com/.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rizopoloulos D (2006). "**ltm**: An R Package for Latent Variable Modeling and Item Response Theory." *Journal of Statistical Software*, **17**(5). URL http://www.jstatsoft.org/v17/i05/.

Zimowski M, Muraki E, Mislevy R, Bock D (2005). **BILOG-MG** *3 – Multiple-Group IRT Analysis and Test Maintenance for Binary Items*. Scientific Software International, Inc., Lincolnwood, IL. URL http://www.ssicentral.com/.

Affiliation:

Harold Doran
American Institutes for Research
Washington, DC 20007, United States of America
E-mail: hdoran@air.org

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 20, Issue 2 *Submitted:* 2006-10-01
April 2007 *Accepted:* 2007-02-22