The estimation of item response models with the lmer function from the lme4 package in R

de Boeck, P.; Bakker, M.; Zwitser, R.; Nivard, M.; Hofman, A.; Tuerlinckx, F.; Partchev, I.

Published in:
Journal of Statistical Software

Journal of Statistical Software

March 2011, Volume 39, Issue 12. *http://www.jstatsoft.org/*

The Estimation of Item Response Models with the `lmer` Function from the lme4 Package in R

Paul De Boeck
University of Amsterdam

Marjan Bakker
University of Amsterdam

Robert Zwitser
Cito Arnhem

Michel Nivard
University of Amsterdam

Abe Hofman
University of Amsterdam

Francis Tuerlinckx
K.U. Leuven

Ivailo Partchev
K.U. Leuven

Abstract

In this paper we elaborate on the potential of the `lmer` function from the **lme4** package in R for item response (IRT) modeling. In line with the package, an IRT framework is described based on generalized linear mixed modeling. The aspects of the framework refer to (a) the kind of covariates – their mode (person, item, person-by-item), and their being external vs. internal to responses, and (b) the kind of effects the covariates have – fixed vs. random, and if random, the mode across which the effects are random (persons, items). Based on this framework, three broad categories of models are described: Item covariate models, person covariate models, and person-by-item covariate models, and within each category three types of more specific models are discussed. The models in question are explained and the associated `lmer` code is given. Examples of models are the linear logistic test model with an error term, differential item functioning models, and local item dependency models. Because the **lme4** package is for univariate generalized linear mixed models, neither the two-parameter, and three-parameter models, nor the item response models for polytomous response data, can be estimated with the `lmer` function.

Keywords: generalized linear mixed models, item response models, multidimensional IRT, item covariates, person covariates.

1. Introduction

The number of software packages for IRT models is clearly on the rise, and an interesting new development is the tendency to migrate onto general-interest platforms such as R (R

Development Core Team 2010). For example, in 2007 the *Journal of Statistical Software* published a special issue on psychometrics in R (de Leeuw and Mair 2007). Software packages for IRT can be categorized in many ways, among others using the following three major categories: Model-oriented packages, extended packages, and general statistical packages.

1. Model-oriented packages concentrate on sets of related item response models, such as the one-parameter (1PM), two-parameter (2PM), and three-parameter (3PM) models, of the logistic or normal-ogive type, and models for ordered-category data, such as the partial credit model (PCM) and the graded response model (GRM), and a variety of other models. Some packages concentrate on a broader family, such as the Rasch family. Examples of packages in R of this first type are **eRm** (Mair and Hatzinger 2007), **mlirt** (Fox 2007), **ltm** (Rizopoulos 2006). They differ in the estimation approach used, such as conditional maximum likelihood (**eRm**), marginal maximum likelihood with Gauss-Hermite quadrature (**ltm**), and Markov chain Monte Carlo (MCMC, **mlirt**).

2. Extended packages stem from a broad category of other than IRT models, such as structural equation models (SEM), multilevel models, and mixture models, and are extended so that they can be used also for IRT models. Examples are **LISCOMP** (Muthén 1987) and its successor **Mplus** (Muthén and Muthén 1998), **HLM** (Raudenbush *et al.* 2004), and **Latent GOLD** (Vermunt and Magidson 2005). Also apart from IRT, they have been developed into very general and flexible model estimation tools.

3. General statistical packages have their origin in generalized linear and nonlinear mixed models. Examples are **lme4** for generalized linear mixed models (Bates *et al.* 2011), SAS `PROC NLMIXED` (SAS Institute Inc. 2008) for nonlinear (and generalized linear) mixed models, and **gllamm** (Rabe-Hesketh *et al.* 2004) for the same kind of framework, but extended, among others elements, with SEM possibilities.

For the third category, it is not always clear what the full potential is for item response modeling, because of the broad purpose of the approach. It is therefore worth specifying the potential explicitly. For SAS `PROC NLMIXED`, suchlike descriptions can be found in De Boeck and Wilson (2004) and in Sheu *et al.* (2005).

For **lme4**, Doran *et al.* (2007) have published an article on the multilevel Rasch model, and a special issue of the *Journal of Memory and Language* (Forster and Masson 2008) contains several articles with useful information, although not in an explicit IRT context. The aim of the present paper is to lay out in an explicit way indeed the possibilities for item response modeling with the `lmer` function, because the generalized linear mixed model approach (as well as the nonlinear mixed model approach) extends the possibilities of IRT modeling, and because this framework links psychometrics to broader domains of statistical modeling.

The models are *generalized* linear models because they allow for a transformation of the expected values of the data in order to rely on a linear formulation, and they are *mixed* because one or more weights in the linear component are random variables (McCulloch and Searle 2001). The linear mixed model (LMM) is a special case of the broader category of generalized linear mixed models (GLMM).

In Section 2, a brief description will be given of GLMM for the IRT context, how the simplest item response model fits into the category of GLMM, and how the `lmer` function to estimate this IRT model reflects the GLMM structure. In the following sections an example dataset

will be described (Section 3.1), a framework for item response models of the GLMM type will be given (Section 4), three broad categories of such models will be presented and their estimation with `lmer` will be explained (Sections 5 to 7), model comparison and testing will be discussed (Section 8), and a comparison with other R packages for item response models will be made (Section 9), followed by a discussion and conclusions (Section 10).

2. Generalized linear mixed models

The models will be described for an item response context, with persons as clusters, items for the repeated observations, and binary responses. The data are denoted as $Y_{pi} = 0, 1$, with $p = 1, \ldots, P$ as an index for persons, and $i = 1, \ldots, I$ as an index for items. The use of `lmer` for IRT is limited to binary data and ordered-category data that can be decomposed into binary data, such as for the continuation ratio model (Tutz 1990). For a more complete discussion of GLMM for IRT, see (Rijmen *et al.* 2003).

2.1. GLMM components

Following a GLMM, data can be understood as generated in a sequence of three steps:

Linear component: For each pair of a person p and an item i, (p, i), a linear combination of the predictors determines the linear component value. This value is denoted here as η_{pi}.

Linking component: The resulting η_{pi} is mapped into the interval $[0, 1]$ based on a link function, yielding a probability π_{pi}, the expected value of Y_{pi}.

Random component: Probability π_{pi} is the parameter of the Bernoulli / binomial distribution on the basis of which a binary observation is generated for the pair (p, i), denoted as $Y_{pi} \in 0, 1$.

Each of these three will now be explained in more detail, and in the reverse order.

The *random component* is the Bernoulli distribution. Probability π_{pi}, is the parameter of the Bernoulli distribution: $Y_{pi} \sim \mathit{Bernoulli}(\pi_{pi})$. The Bernoulli distribution is the binomial distribution with one observation ($n = 1$). It is typical for IRT data to have only one observation per pair of a person and an item, so that

$$Y_{pi} \sim \mathit{binomial}(1, \pi_{pi}). \tag{1}$$

The *linking component* maps the expected value of Y_{pi}, which is π_{pi} for the Bernoulli distribution, on the real line from $-\infty$ to $+\infty$ through a link function:

$$\eta_{pi} = f_{\text{link}}(\pi_{pi}). \tag{2}$$

The link function is commonly chosen to be the logit function, or $\eta_{pi} = \ln(\pi_{pi}/(1 - \pi_{pi}))$. The logit link is the natural link for the Bernoulli / binomial distribution. It leads to logistic IRT models. An alternative for the logit link is the probit link, which is based on the cumulative standard normal distribution, also called the normal ogive. The inverse of the cumulative

probability function yields the value of the probit link function. The two links lead to different scales. The parameters in a probit model are scaled relative to the standard normal distribution, while in a logistic model they are scaled relative to the standard logistic distribution. Because the logistic distribution is approached quite well by a normal distribution with a standard deviation of 1.7, the probit scale can be transformed to the logit scale by multiplication with the well-known $D = 1.7$.

The *linear component* is a linear combination of predictors. In the statistical literature, often the term "covariate" is used instead of "predictor", and the weights are called "effects". Expanding and developing the linear component in different ways is an important part of this paper, so we will explain in detail its structure for the simplest IRT model, the Rasch model.

2.2. Linear component of the Rasch model

The Rasch model uses $I+1$ item covariates: I item indicators plus a constant 1 for all items. As a result, the $I \times (I+1)$ matrix $\mathbf{X}$ of item covariates is the concatenation of a 1-vector and an $I \times I$ identity matrix. The 1-vector has a random effect, which is often called the ability or latent trait in an IRT model, while the effects of the covariates from the identity matrix are fixed, one per item, corresponding to the so-called difficulty parameters. The logistic version of the model is known as the Rasch model or one-parameter logistic (1PL) model. In terms of item predictors and their effects, the model can be formulated as follows:

$$\eta_{pi} = \theta_p X_{i0} + \sum_{k=1}^{K} \beta_i X_{ik}, \tag{3}$$

with $X_{i0} = 1$ for all items; $X_{ik} = 1$ if $i = k (k = 1, \ldots, K$; index k has the same range as index $i)$, and 0 otherwise; and $\theta_p \sim N(0, \sigma_\theta^2)$.

The model in (3) can also be written in a simpler form: $\eta_{pi} = \theta_p + \beta_i$. It follows from (3) that the 1PL and its normal-ogive equivalent are random intercept models. Note that the plus sign in (3) implies that the β_i should be interpreted as item easiness instead of item difficulty. This is in conformity with the `lmer` notation, but not with common psychometric practice, where a minus sign is used and the β_i is interpreted as the item difficulty.

As a general principle, when viewing IRT models from a GLMM perspective, all forms of measurement rely on covariate effects. For example, in most models, the latent traits are random effects, and the item difficulties are fixed effects.

2.3. The `lmer` function

In the `lmer` function, the *random component* for the case of the binomial distribution is specified as `ir ~ ..., family = binomial`, as illustrated in Figure 1. The first part, `ir ~ ...`, tells us that the binary variable containing item responses, called `ir`, is distributed with an expected value determined by the linear component (symbolized here by the dots). The last part, `family = binomial`, indicates the binomial nature of the distribution. This specification is independent of the number of observations per pair (p, i), but, of course, in the common IRT case, there is just one observation.

The *linking component* in the case of binomial data can be specified either as the `"logit"` or `"probit"` argument of `family = binomial()` (Figure 1). Because the logistic link is the default, the specification may be omitted for logistic models.

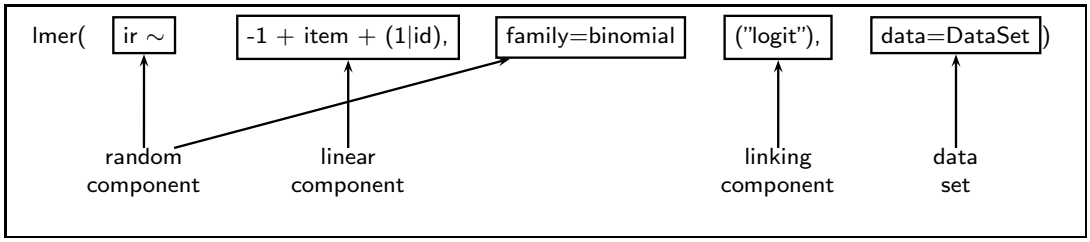

Figure 1: The three GLMM components for a logit model and the dataset as arguments of the `lmer` function.

The *linear component* is specified after `ir ~`, in the form of an linear expression that represents the covariates and the kind of effect (fixed or random) they have (Figure 1). The linear component in Figure 1 is for the Rasch model. It contains fixed effects for the items, indicated with the term `item`, and a random effect across persons, indicated with the term `(1 | id)`. The value of the linear component is the expected value of Y_{pi}.

Combining the three components and the dataset label then leads to the following `lmer` code for the 1PL model: `lmer(ir ~ -1 + item + (1 | id), data = DataSet, family = "binomial")` (see also Figure 1). The linear component expression `-1 + item + (1 | id)` will be explained in Section 5. Alternative ways to specify the 1PM model are possible, as will be explained in the remainder.

3. Data set and format

3.1. Example data set

An example dataset, **VerbAgg** from De Boeck and Wilson (2004), is included in the **lme4** package. It consists of item responses to a self-report questionnaire with a design. The topic is verbal aggression. There are 24 items, based on four frustrating situations, two of which where someone else is to be blamed (e.g., "A bus fails to stop for me"), and two of which where one is self to be blamed (e.g., "I am entering a grocery store when it is about to close"). Each of these situations is combined with each of three behaviors, cursing, scolding, and shouting, leading to 4×3 combinations. These 12 combinations are formulated in two modes, a wanting mode and a doing mode, so that in total there are 24 items. An example is "A bus fails to stop for me. I would want to curse". This is an other-to-blame item with a cursing reaction and a wanting mode. The corresponding doing mode reads as follows: "A bus fails to stop for me. I would curse". The response scale has three categories, "yes", "perhaps", and "no". The first two are coded as 1, the third as 0. Of the 316 respondents 243 are females and 73 are males.

The labels are `r2` for the binary response, `item` for the items, `id` for the persons, `btype` for the behavior type (with levels `"curse"`, `"scold"`, `"shout"`), `situ` for other-to-blame and self-to-blame (with levels `"other"`, `"self"`), `mode` for the behavioral mode (with levels `"want"`, `"do"`), and `Gender` for the person's gender (with levels `"F"` for men, `"M"` for women).

Commonly an IRT dataset with item responses has the form of an array with P rows and

I columns. However, `lmer` needs a "long form" for the data to be modeled, with one row per response and a column length equal to the number of persons times the number of items ($P \times I$), or 316×24 in the example. These responses constitute one of the columns (vectors) in the data frame. At least two other columns must be present, one to identify the person and the other to identify the item. For richer models, there will be additional covariates whose values must be repeated on each row associated with the carrier of the covariate value – for instance, the gender indicator for person p must appear on each row containing a response from person p. See Section 3.2 for how one can proceed if the data format is wide.

Because there is a row for each single response, the row corresponding to a missing response should either be removed, or the missing responses must be coded such that a selection for the analysis is possible. Consequently, missing data will treated as missing at random (MAR). Neither deletion of cases nor data imputation is needed.

The columns can be of two kinds: Quantitative variables (numeric vectors), and nominal or qualitative variables ("factors"). Both types can be used for categorical variables. For example, one can either use the factor `btype` (with levels `"curse"`, `"scold"`, and `"shout"`), or one can define two quantitative binary variables, using either dummy coding, contrast coding, reference coding, or any other type of coding. The factor format will be used for the item and person indicators, for the item design factors, and for gender. An extra predictor variable, `Anger`, in the same dataset, refers to a trait measure of the person's anger and is of the quantitative type.

The default coding of the factors is dummy coding with the first level of each factor as the reference level. In an IRT context this means that the first item functions as the reference item, and that all other item parameters are estimated as deviations from the first. All effects are expressed as deviations from an intercept, unless it is specified that no intercept is used, as will be explained in Section 5.1. If the intercept is removed, also the effect of the first level of the first mentioned factor will be estimated, which is the first item parameter estimate for the factor `item` if `item` is mentioned as the first element of the linear component. Alternative codings for the factors can be chosen.

3.2. Long format

"Wide form" data can be translated to the long format in various ways. We will show how it can be done with function `melt()` from the additional package **reshape** (Wickham 2007), which we have found a bit easier than the in-built function `reshape()`.

Function `melt` distinguishes between measured variables and ID variables. One can think of the former as within-person variables (typically, responses), and of the latter as between-person variables (such as subject's ID, sex, age...) The variables in the data frame can be declared as either measured or ID variables by their names or by their column number. When variables of both types are explicitly declared, the remaining variables in the data frame will be omitted from the operation. When only one type (measured or ID) is specified, all remaining variables are assumed to be of the other type. If nothing is specified, all variables are treated as measured.

As an example, let us use `melt()` to transform the data set `LSAT` included in package `ltm` from wide to long form. `LSAT` contains the responses of 1000 persons to 5 dichotomous items. These are all measured variables that will be "melted" into a single variable, called by default `value`, plus an additional factor, `variable` – the latter is produced from the variable names

of the five original variables to indicate the item to which each response corresponds. `LSAT` does not contain an ID variable for the persons, which is needed by `lmer`, so we shall have to construct one from the row names of the data frame before melting. We will specify explicitly only the new ID variable because all other variables are measured and none of them will be dropped. The necessary R code is then

```
R> library("ltm")
R> data("LSAT")
R> LSAT$person <- rownames(LSAT)
R> library("reshape")
R> LSATlong <- melt(LSAT, id = "person")
```

Data set `LSATlong` is in the shape needed by `lmer`. Each of its rows contains a person ID variable, an item ID variable, and the response given by the person to that particular item:

```
  person variable value
1      1   Item 1     0
2      2   Item 1     0
3      3   Item 1     0
. . .
```

4. Common and less common IRT models

An important type of item response models (IRT) belongs to the GLMM category, because these IRT models are linear models for the logit or probit transformation of the expected values of a binary variable (i.e., probabilities), and because one or more weights in the resulting linear model are random variables. Commonly, these random variables are latent traits. Three important types of models do not belong to the GLMM category. First, models for ordered categories or nominal categories, with a number of categories larger than two, rely on multivariate extensions of the GLMM framework (Fahrmeir and Tutz 2001), also called vector GLMMs (Yee and Wild 1996; Yee 2010). An exception is the continuation ratio model, which can be formulated as a GLMM (Tutz 1990). Second, the 2PM is not a GLMM because it relies on products of parameters through the introduction of a discrimination parameter. Finally, also the 3PM is not a GLMM, not only because it is an extension of the 2PM, but also because it is a mixture model for the item responses. One mixture component refers to the guessing responses, and the other to the responses governed by the 2PM.

On the other hand, not just the 1PM or Rasch model remains as a possible model. Instead, a large variety of models, some of which may be less familiar, do belong to the GLMM category and can therefore be estimated using `lmer`. In the following, this variety of models will be explicated and it will also be explained how to use the `lmer` function for these models.

In order to explicate these models, a short taxonomy is presented here, based on the kind of covariates involved and the kind of effects they have. The taxonomy has four dimensions, two referring to the kind of covariates, and two referring to the kind of effects, so that there are four dimensions in total:

Mode of the covariates: The covariates can refer to items, to persons, or to pairs of persons and items.

External versus internal covariates: An external covariate is external to the item responses to be modeled. An internal covariate does stem from the responses to be modeled. Models with internal covariates are sometimes called conditional models in the statistical literature (Fahrmeir & Tutz, 2001). For example, the number of previous successful responses in a test is an internal (person × covariate) covariate for a model that is meant for the item responses, whereas gender is an external (and person) covariate.

Fixed versus random effects: Fixed effects are unknown constants and do not vary as a function of observational units, whereas random effects do vary across observational units and are drawn from a distribution, which is commonly the normal distribution.

Mode of randomness: A random effect follows a distribution associated with the population of the observational units one wants to consider. An effect can be random across persons, across items, across persons within a person group, across items within an item type, across groups of persons, across groups (types) of items, and even more complicated cases can be constructed. It is clear from the nested elements in the list (e.g, persons and person groups) that also multilevel models are a possibility.

Different combinations of choices from the four taxonomic dimensions lead to different IRT models. For example, in the Rasch model, item indicators are used as item covariates with a fixed effect, called item difficulties, while a constant 1–covariate is considered to have a random effect across persons, called the ability.

Three model categories will be considered, based on the first dimension from the taxonomy:

1. Item covariate models, which rely primarily on item covariates;
2. Person covariate models, which rely in addition on person covariates;
3. Person-by-item covariate models, which rely also on person-by-item covariates.

The other three taxonomic dimensions refer to particular elements of the models within these three categories. All models will be formulated as logistic models. If their normal-ogive equivalent is wanted, the probit link should be used.

For the explanation of the covariates and the models, we use the quantitative variable representation. The `lmer` code will mostly be given in terms of factors. For the distribution of the random effects the program assumes a normal distribution with a mean of zero, and random effects are seen as deviations from this mean and thus from the intercept (or zero if the intercept is removed).

5. Item covariate models

Let us collect the item covariates in an item-by-covariate matrix, with dimensions $I \times (K+1)$, and with $k(k = 0, 1, \ldots, K)$ as an index for the item covariates. Because of the long form format of all variables, the item covariate matrix needs to be repeated for each person, so that a concatenated matrix $\mathbf{X}$ of size $(P \times I) \times (K + 1)$ is obtained, with entries $X_{(p,i)k}$, instead of X_{ik} as in (3). See the first three columns of Table 1 for the simple case of three persons and two items.

	Item 1	Item 2	X_0
Person 1 item 1	1	0	1
Person 1 item 2	0	1	1
Person 2 item 1	1	0	1
Person 2 item 2	0	1	1
Person 3 item 1	1	0	1
Person 3 item 2	0	1	1

Table 1: Item covariate matrix of the identity type and a 1-covariate.

The item covariates can be of different kinds:

1. The 1-covariate: $X_{(p,i)0} = 1$ for all values of i, as in the last column of Table 1;
2. The indicator covariate: $X_{(p,i)k} = 1$ if $k = i$, 0 otherwise, as in the 2nd and 3rd column of Table 1;
3. The item property covariate: $X_{(p,i)k} = 1$ if item i has property k, 0 otherwise;
4. The item partition covariate: $X_{(p,i)k} = 1$ if item i belongs to element k of a partition, and $X_{(p,i)k} = 0$ otherwise.

Each factor corresponds to one partition and generates as many partition covariates as there are levels of the factor. The list of four item covariate types is not exhaustive. For example, the covariates do not need to be binary, but can be integer-valued or real-valued instead. However, for all item covariate models to be presented here, binary covariates are used, of the four types just described.

In general, the covariate matrix $\mathbf{X}$ consists of two possibly overlapping submatrices, one for covariates with a fixed effect and one for covariates with a random effect. Both submatrices need to be of full rank in order for the model to be identified.

Depending on which of the three models to be presented in this section, either indicator covariates, or property covariates, or partition covariates are used. The 1-covariate is used when to define a general latent trait, such as the person ability.

5.1. The one-parameter logistic model (item indicator covariates)

The collection of item indicator covariates constitutes an identity matrix for each person. The long-form item covariate matrix for the one-parameter logistic (1PL) model in (3) consists of a vertical concatenation of P $I \times I$ identity matrices, as in the 2nd and 3rd column of Table 1. The effect of the $k = i$-th indicator covariate is the easiness of item i (or difficulty if a minus sign is used in $\mathbf{X}$): β_i. It is a fixed effect. In order to represent the overall individual differences, a random effect based on the 1-covariate in $\mathbf{X}$ (last column of Table 1) is included. It corresponds to the random intercept θ_p and is often called the ability in an IRT context.

The linear component for the 1PL is specified as `-1 + item + (1 | id)`. The `-1` (or `0`) avoids that the first item is used as the reference item and the basis for the intercept. The term `+ item` defines the fixed effects of the items, while the term `(1 | id)` defines a 1-vector `(1 | ...)` with an effect that is random over persons `(.. | id)`. If an effect is random, it

is put within parenthesis followed by a vertical bar, and after the vertical bar, the units are mentioned across which the effect is random.

In order to obtain the so-called person parameter estimates, the function `ranef()` with the model name as an argument can be used. It returns conditional modes of the random effect, taking into account the observed data of the person and the estimated parameter values, including the variance of the unobserved normally distributed random effect. In an IRT context, this type of estimation is called the maximum a posteriori method (MAP) for person parameter estimation. The function `mcmcsamp()` with the object name of the estimation result as an argument, can be used to obtain standard errors. It generates the posterior distribution using Markov chain Monte Carlo. However, the function is presently being updated and is therefore inactive (**lme4** version 0.999375-32 of October 2009). The function `ranef()` is for all kinds of random effects, and the function `mcmcsamp()` works for fixed effects as well as for random effects.

5.2. The LLTM (item property covariates)

Item properties are covariates which are not just item indicators. Only binary properties are considered here, but that is not a necessity. Again, the covariate matrix is repeated in a vertical concatenation, yielding a $(P \times I) \times (K+1)$ long-form matrix $\mathbf{X}$. An evident example for the verbal aggression data is the item design. For example, for want versus do, $X_{(p,i)k} = 1$ if item i is a want-item, and 0 if it is a do-item. The item property matrix is needed for the linear logistic test model (LLTM) (Fischer 1973; Scheiblechner 1972) which explains the item easiness (difficulty) in terms of item properties. Also for this model, a random person effect, θ_p, is required, to define the latent trait:

$$\eta_{pi} = \theta_p + \sum_{k=1}^{K} \beta_k X_{(p,i)k}, \tag{4}$$

with θ_p as in (3), omitting the 1-covariate $X_{(p,i)0}$, and β_k as the fixed effect of item property covariate $X_{(p,i)k}$.

An interesting extension of this model is the LLTM plus error (Janssen *et al.* 2004; De Boeck 2008), which means that an error term is added in (4):

$$\eta_{pi} = \theta_p + \sum_{k=1}^{K} \beta_k X_{(p,i)k} + \varepsilon_i, \tag{5}$$

with $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, just as in the regular regression model.

The addition of an error term to the model is very useful. The original LLTM is like a regression model that explains all variance, and it is therefore almost always rejected. The error terms allows for an imperfect prediction. Doran *et al.* (2007) describe a similar application.

Note that the model in (5) implies homoscedasticity of the error variance. This assumption can be relaxed. For example, the error variance may be different for the *do* items and the *want* items. A larger error variance means that the item properties have less explanatory power to explain the item difficulties.

The linear component for the regular LLTM and the LLTM with error, either homoscedastic or heteroscedastic, can be specified as follows:

- For the regular LLTM:
  `-1 + btype + mode + situ + (1 | id)`
- For the LLTM plus homoscedastic error:
  `-1 + btype + mode + situ + (1 | id) + (1 | item)`
- For the LLTM plus heteroscedastic error for want and do:
  `-1 + btype + mode + situ +  (1 | id) + (-1 + mode | item)`

In the following, the estimation of the LLTM plus heteroscedastic error for want and do is illustrated. The full `lmer` code one can use is:

```
R> library("lme4")
R> lltmhe <- lmer(
+    r2 ~ -1 + btype + mode + situ + (1 | id) + (-1 + mode | item),
+    data = VerbAgg, family = binomial)
R> print(lltmhe)

Generalized linear mixed model fit by the Laplace approximation
Formula: r2 ~ -1 + btype + mode + situ + (1 | id) + (-1 + mode | item)
   Data: VerbAgg

  AIC  BIC logLik deviance
 8165 8227  -4073     8147

Random effects:
 Groups Name        Variance Std.Dev. Corr
 id     (Intercept) 1.881943 1.37184
 item   modewant    0.044218 0.21028
        modedo      0.212138 0.46058  0.000
Number of obs: 7584, groups: id, 316; item, 24

Fixed effects:
           Estimate Std. Error z value Pr(>|z|)
btypecurse   1.7202     0.1559  11.036  < 2e-16 ***
btypescold   0.6434     0.1530   4.206 2.60e-05 ***
btypeshout  -0.1885     0.1527  -1.234    0.217
modedo      -0.7117     0.1570  -4.533 5.81e-06 ***
situself    -1.0579     0.1286  -8.225  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
           btypcr btypsc btypsh modedo
btypescold  0.482
btypeshout  0.470  0.475
modedo     -0.226 -0.220 -0.215
situself   -0.431 -0.422 -0.406  0.008
```

The output first shows the model formulation, and the name of the dataset. Next, Akaike's information criterion (AIC, Akaike 1974), Schwartz's information criterion (BIC, Schwartz 1978), the loglikelihood, and the deviance are given. This information is followed by the result for the random effects. The estimated variance and standard deviation of the person random effect (latent trait) are 1.88 and 1.37, respectively (on line `id (Intercept)`), the estimates for the unexplained item variance are 0.044 for the *want* items (line `item modewant`), and 0.212 for the *do* items (line `item modedo`).

In the next output section, the estimates of the fixed effects are given. The very first -1 in the model specification has suppressed the overall intercept, so the fixed effect of the behavior type is expressed as three means: 1.72 for cursing (line `btypecurse`), 0.64 for scolding (line `btypescold`), and -0.19 for shouting (line `btypeshout`), which apply to the case when the mode is "want" and the situation is "other to blame". Obviously, cursing is more popular than scolding, which is in turn more popular than shouting. The effects are deviations from the probability of 0.50 on the logistic scale, and no direct comparison between the three behaviors is made. This is different for the factors `mode` and `situ`. The effects on the lines `modedo` and `situself` show the effects of the "do" mode in comparison with the "want" mode (-0.71) and of the "self" situations as compared with the "other" situations (-1.06). Because the model is a main effects model, the same fixed effects hold for the "do" mode and the "other to blame" situations (no interaction). For each fixed effect, the standard error is also given along with the corresponding z-value and the p-value under the null hypothesis, such that the effect can be tested. The effects of "do" and "self" are both highly significant, which means that people report to be less verbally aggressive in what they would actually do than in what they want to do, and that they are also less verbally aggressive when they are self to be blamed in comparison with situations in which other people are to be blamed.

An alternative formulation of the same model would be to remove the initial `-1`, so that the cell which combines cursing with "want" and "other" situations becomes the reference basis. This formulation leads to the same estimates except for a reparameterization of the `btype` effect:

```
btypescold   -1.0767     0.1572  -6.848 7.49e-12 ***
btypeshout   -1.9087     0.1589 -12.016  < 2e-16 ***
```

The reparameterized effects are the differences of the previously estimated effects for `btypescold` and `btypeshout` from the previous effect of `btypecurse`.

Going back to the original output, the final section shows the error covariances or correlations for the estimates of the five effects. The output for the error correlations may be excessively bulky when the number of fixed effects in the model is large (as in the case of the 1PL model). To suppress it, print the output of the `lmer` by calling explicitly the `print` function like this: `print(lltmhe, cor = FALSE)`, given that `lltmhe` is the label that is assigned to the model output.

5.3. The multidimensional 1PL model (item partition covariates)

The design factors can be used to define an item partition matrix, so that all levels of all factors are each represented with a binary covariate. One factor defines one partition. If the partitions are hierarchically ordered, a nested structure is obtained. If the partitions are not hierarchically ordered, a crossed structure is obtained. For items, a nested structure is

rather unusual, while a partially or completely crossed structure is rather common. For the example data, the structure is completely crossed, with three design factors: Do vs. want, self vs. other, and curse vs. scold vs. shout, yielding two partitions of two and one partition of three. The item partition matrix is an interesting tool to define a multidimensional 1PL (Rijmen and Briggs 2004). It plays the same role as a confirmatory factor loading matrix.

In order to include also the item parameters, the item partition matrix must be extended with the item identity matrix, so that the corresponding fixed effects are contained in the model. As a result, the $\mathbf{X}$ matrix is a long-form $(P \times I) \times K$ matrix, with $K = I + K^*$, and K^* is the number of binary item partition covariates, which equals the sum of partition elements over the partitions, or, in other words, the total number of levels for all item design factors. The model can be written as

$$\eta_{pi} = \beta_i + \sum_{k^*=1}^{K^*} \beta_{k^*p} X_{(p,i)k^*}, \tag{6}$$

with β_i as defined in (3), but omitting the item indicator covariate $X_{(p,i)k=i}$; with β_{k^*p} as the random effect of item partition covariate $X_{(p,i)k^*}$; with $\boldsymbol{\beta}_p \sim \text{MVN}(0, \boldsymbol{\Sigma}_\beta)$, and $\boldsymbol{\Sigma}_\beta$ as the covariance matrix of random effects.

Whether this model is identified depends on the structure of the K^* covariates. If only one partition is involved, the model is a between-item multidimensional model (as opposed to a within-item multidimensional model, see Adams *et al.* (1997)), and it is identified indeed. If more than one item partition is involved, the item design can be hierarchical or (partially) crossed. We will not treat the hierarchical case here (nested between-item multidimensional model), but the crossed case instead. A multidimensional model for a crossed item design is not identified unless restrictions are imposed on the model. For example, the model is again identified if the correlations between the dimensions referring to different partitions are fixed to zero.

For a model with F fully crossed design factors (F partitions) (e.g., three in the example data), and m_f levels for factor $f (f = 1, \ldots, F)$ (e.g., $m_1 = 3$, $m_2 = 2$, and $m_3 = 2$ in the example data), the following models can be formulated without identification problems:

Model 1: $\sum_{f=1}^{F}(m_f - 1)$ levels have a random effect (random slope) plus a random intercept (a random intercept, and random effects for scold, shout, self, and do, but not for curse, other, and want);

Model 2: $m_{f=1} + \sum_{f=2}^{F}(m_f - 1)$ levels have a random effect (random slope) (random effects for curse, scold, shout, self, and do, but not for other and want);

Model 3: $\sum_{f=1}^{F} m_f$ levels have a random effect (random slope), but with the constraint of a zero correlation between the levels of the different design factors (random effects for curse, scold, shout, other, self, want, and do, but with a zero correlation between the random effects belonging to different factors, for example, between the curse and want random effects);

Model 4: The $\prod_{f=1}^{F} m_f$ cells of the design have a random effect (12 random effects, one per cell in the $3 \times 2 \times 2$ design).

	Gender	Native
Person 1 item 1	1	1
Person 1 item 2	1	1
Person 2 item 1	0	0
Person 2 item 2	0	0
Person 3 item 1	0	1
Person 3 item 2	0	1

Table 2: Example of a person covariate matrix including a 1-covariate.

Note that the dimensionality is different: $\sum_{f=1}^{F}(m_f - 1) + 1$ in models 1 and 2, $\sum_{f=1}^{F} m_f$ in model 3, and $\prod_{f=1}^{F} m_f$ in model 4. If in models 2 to 4 a random intercept is added, the model is still estimated, but without any improvement of the goodness of fit, because the model is not identified.

As an alternative, instead of item parameters, fixed item property effects can be included in (6), in a similar way as in (4), but with the general latent trait replaced with a multidimensional part as in (6). With this replacement, one can see the model as a second kind of random effect extension of the original LLTM. It is called the random-weight LLTM (Rijmen and De Boeck 2002).

The linear component for the four models can be specified as follows:

```
-1 + item + (1 + btype + mode +  situ | id)
-1 + item + (-1 + btype + mode +  situ | id)
-1 + item + (-1 + btype | id) + (-1 + mode | id) + (-1 + situ | id)
-1 + item + (btype:mode:situ | id)
```

For the random-weight LLTM, replace `item` with `btype + mode + situ`.

6. Person covariate models

Let us collect the person covariates in a person-by-covariate matrix, with dimensions $P \times J$, with $j(j = 1, \ldots, J)$ as the subscript for person covariates. Because of the long form format, the row of a person p needs to be repeated for all I responses per person, so that a $(P \times I) \times J$ matrix $\mathbf{Z}$ is obtained, with entries $Z_{(p,i)j}$. See the second and third column of Table 2, assuming that person 1 is female and persons 2 and 3 are males, while persons 1 and 3 are native speakers, and person 2 is not.

Again, three models will be presented to show the flexibility. As for the item covariate models, the models are based on indicator covariates, property covariates, and partition covariates, but now they are person covariates. The 1-covariate is redundant because it is equivalent with the 1-covariate from $\mathbf{X}$. Just as the item covariates, also the person covariates can be integer-valued or real-valued.

6.1. The JML version of the 1PL model (person indicator covariates)

The collection of person indicator covariates constitutes an identity matrix. In the corresponding long-form covariate matrix, $Z_{(p,i)j} = 1$ if $j = p$, as in the second and third column

of Table 1, and $J = P$. The effect of the p-th covariate is person parameter p, which is a fixed effect. This covariate matrix is needed for the joint maximum likelihood (JML) version of the 1PL. The JML model also needs an item identity matrix $\mathbf{X}$ to define the item difficulties as in (3):

$$\eta_{pi} = \sum_{j=1}^{J} \theta_p Z_{(p,i)j} + \sum_{k=1}^{K} \beta_i X_{(p,i)k}, \tag{7}$$

with θ_p as a fixed effect of the person indicator covariate $Z_{(p,i)j} = 1$ if $p = j$, and 0 otherwise.

The JML version of the 1PL model is the fixed person effect alternative for the model in (3), which is the marginal maximum likelihood (MML) version. The labels JML and MML refer to the estimation of the item parameters, either jointly with the person parameters (JML) or integrating over the random person effect (MML). Apart from the estimation method, also the models differ – the person effects are either fixed (JML) or random (MML), respectively. It is known that the JML model does not lead to consistent estimates which need to be corrected (Andersen 1980; Ghosh 1995). In fact, there are four possible 1PL models, one of which is the MML model, and one other of which is the JML model. They are obtained by crossing random vs. fixed for persons and items: Random persons and random items, random persons and fixed items (MML), fixed persons and random items, and fixed persons and fixed items (JML) (De Boeck 2008). The linear component for the four versions is formulated as follows:

- For random persons and random items `1 + (1 | id) + (1 | item)`
- For random persons and fixed items `-1 + item + (1 | id)`
- For fixed persons and random items `-1 + id + (1 | item)`
- For fixed persons and fixed items `-1 + id + item + (1 | item)`

The `id + item` part in the latter defines the person and item effects as fixed. The term `(1 | item)` is added to meet the `lmer` condition that the model contains a random effect (in order to be a mixed model), but, given that there is already a fixed effect requested for the items, it does not add to the model, other than making estimation possible.

6.2. The latent regression 1PL (person property covariates)

Person properties are covariates which are not just person indicators. Not only binary properties are considered here, but also an integer valued quantitative property. The rows of this $P \times J$ matrix are repeated for all I items, so that a $(P \times I) \times J$ long-form matrix $\mathbf{Z}$ is obtained. Two person properties are available in the example dataset, the factor `Gender`, and the integer-valued quantitative variable `Anger`.

The latent regression model (Zwinderman, 1991) is a model with fixed effects of the person covariates, and it can be understood as a latent regression model for the θ_p from the 1PL model in (3). It can be combined with item property effects, as in (4), or with item parameters, so that the corresponding $\mathbf{X}$ matrix is also needed. The version with item parameters is as

follows:

$$\theta_p = \sum_{j=1}^{J} \zeta_j Z_{(p,i)j} + \varepsilon_p + \beta_i, \tag{8}$$

with ζ_j as the fixed effect of person property covariate $Z_{(p,i)j}$; with ε_p as the unexplained part of θ_p, assuming that $\varepsilon_p \sim N(0, \sigma^2_\varepsilon)$, and β_i as in (3).

The latent regression model is especially helpful if subpopulations are represented in the sample of persons, because the assumption of a global normal distribution for θ_p would not hold if the subpopulations have different means. Note that the model in (8) implies homoscedasticity of the unexplained variance. This assumption can be relaxed using random effects within the levels of the factor allowing for heteroscedasticity.

The linear component can be formulated as follows:

```
-1 + item + Anger + Gender +  (1 | id),
```

and with heteroscedasticity depending on gender:

```
-1 + item + Anger + Gender +  (-1 + Gender | id)
```

When the model is multidimensional, and the effect of a person covariate is assumed to be same for all dimensions, then a simple fixed effect must be included in the model, in the same way as shown for `Anger` and `Gender`. If the effect is assumed to be dimension specific or different depending on the dimension, then the linear component should contain a fixed interaction effect of the person covariate and the item covariate that defines the dimension in question.

The linear component for a model with a differential effect of gender on the want-dimension and the do-dimension can be formulated as follows:

```
-1 + btype + mode + situ + Gender:mode + Gender + (-1 + mode | id)
```

Note that the model cannot have at the same time item parameters and an interaction between `Gender` and `mode`.

6.3. Multilevel models (person partition covariates)

Let us denote the elements of person partitions as person groups. In total there are as many person partition covariates as there are person groups. The structure of the partitions can be nested or (partly or fully) crossed. It is common to use the term *multilevel* for nested person partitions. For persons, the nested structure is the more common one, but Raudenbush (1993) describes also crossed person groups. Like for all person covariate matrices, the rows must be repeated for all I items in order to obtain the long form. There is only one person partition in the example dataset, with only two groups which are also fixed (men and women). Often, data do have a structure which lends itself to a multilevel analysis, for example, for educational data the levels are persons, classes and/or schools, and possibly there is also a higher level such as states.

	Cov. 1	Cov. 2	Cov. 3
Person 1 item 1	1	1	0
Person 1 item 2	1	0	0
Person 2 item 1	0	0	0
Person 2 item 2	1	0	1
Person 3 item 1	1	0	1
Person 3 item 2	0	1	1

Table 3: Example of a person-by-item covariate.

A simple multilevel version of the 1PL is the following:

$$\eta_{pi(g)} = \theta_p + \beta_i + \theta_g X_{(p,i)j=0}, \tag{9}$$

with $\eta_{pi(g)}$ as the logit for item i and person p belonging to group g; with θ_p and β_i as in (3); with θ_g as a random group effect, $\theta_g \sim N(0, \sigma_g^2)$; and $X_{(p,i)j=0} = 1$ for all p and i.

The simple multilevel model in (9) can be extended, among others, with components from the previous models. Another extension is that a third level is added, beyond the persons (first level) and the first-order person groups (second level).

Suppose there were person groups defined in the example data, then the linear component for the multilevel model can be formulated as follows:

```
-1 + item + (1 | id) + (1 | group)
```

with `group` as the label for the group factor.

Heteroscedasticity can be included in the model in the same way as for other models. The use of `lmer` for the multilevel 1PL model is described in a more elaborated way by Doran *et al.* (2007), including random effects for items.

7. Person-by-item covariates

Let us collect the person-by-item covariates in a long-form person-by-covariate matrix $\mathbf{W}$ with dimensions $(P \times I) \times H$, with $h(h = 1, \ldots, H)$ as an index for the person-by-item covariates. The covariates now refer to the pairs of persons and items (p, i). An example of a person-by-item covariate matrix with three covariates is given in Table 3.

Among the three kinds of covariates (indicators, properties, partition subsets), we will concentrate on property covariates. Indicator covariates do not make sense, because one would need an indicator per response. The partition covariates do make sense, for example, to define a different dimension depending on the person-by-item block, but they would lead us too far. Here, three models with property covariates will be presented. Because the external versus internal dimension of the taxonomy becomes meaningful for person-by-item covariates, both types will be illustrated. One model has external covariates, and two models have internal covariates – one model with binary covariates, and another with an integer-valued covariate.

7.1. DIF models (external person-by-item property covariates)

Differential item functioning (DIF) means that for the item in question the response probabilities of persons with the same ability differ depending on the group the persons belong to. Items showing DIF are a problem for measurement equivalence and may lead to bias in the measurement of a latent trait (Millsap and Everson 1993). Holland and Wainer (1993) give an overview of the topic.

Commonly, DIF is studied in a focal group in comparison with a reference group. Therefore, a DIF model requires a **Z** matrix. Furthermore, it requires also the **X** matrix of the regular 1PL model. The **W** matrix of a DIF model consists of covariates which are the product of two other covariates: The focal group covariate $Z_{(p,i)\text{focal}}$ and either an item indicator covariate $X_{(p,i)k=i}$ (item specific DIF) or an item property covariate $X_{(p,i)k}$ (item subset DIF). For item specific DIF, a pair (p, i) has a value of 1 on covariate h if person p belongs to the focal group and item i is an hypothesized DIF item. The fixed effect of this covariate is the DIF parameter of item i. It refers to the deviance of the item i easiness (difficulty) in the focal group from its easiness (difficulty) in the reference group. For item subset DIF, a pair (p, i) has a value of 1 on covariate h if person p belongs to the focal group and item i belongs to the subset in question (has the corresponding property), and a value if 0 otherwise. Both types of DIF can be combined in one model.

A DIF model can be formulated as follows:

$$\eta_{pi} = \theta_p + \beta_i + \zeta_{\text{focal}} Z_{(p,i)\text{focal}} + \sum_{h=1}^{H} \omega_h W_{(p,i)h}, \tag{10}$$

with θ_p and β_i as in (3), with ζ_{focal} as the global effect of the focal group in comparison with the reference group; with $Z_{(p,i),\text{focal}} = 1$ for the focal group, and 0 for the reference group; with $W_{(p,i)h}$ as the person-by-item covariate h, defined in such a way that $W_{(p,i)h} = 1$ if both $Z_{(p,i),\text{focal}} = 1$ and either $X_{(p,i)k=i} = 1$ (item specific DIF), or $X_{(p,i)k} = 1$ (item subset DIF), and $W_{(p,i)h} = 0$ otherwise; and ω_h as the corresponding DIF parameter.

Notice that the dummy coding of the DIF covariate affects the group effect and the corresponding fixed item or covariate effect. Alternatives are contrast coding and effect coding.

Using item specific DIF modeling, one can test items one by one, and compare the likelihood of the one-item DIF models with the likelihood of the regular 1PL model, based on a likelihood ratio test. In a next stage the model can be reformulated with item subset DIF if several items show approximately the same amount of DIF.

For example, in the verbal aggression dataset, the do-items referring to cursing and scolding, eight items in total, seem to show about the same gender DIF, which can be captured with one common DIF parameter (Meulders and Xie 2004). The DIF parameter is the fixed effect of a person-by-item covariate which is the product of an item property covariate (the subset of eight items) and the gender covariate. From the sign of the DIF effect, it must be concluded that men (say they) curse and scold more easily than women, independent of their latent verbal aggression trait. Using DIF modeling, one can also test items one by one, and compare the likelihood of the one-item DIF models with the likelihood of the regular 1PL model, based on a likelihood ratio test.

Note that only DIF of the uniform type is studied in this way and that the approach is model based. Uniform DIF means that DIF does not depend on the value of the latent trait.

The approach is model-based because an IRT model is used. Other methods exist to study non-uniform DIF, and to study DIF without modeling (Millsap and Everson 1993). DIF is commonly studied as a fixed effect, but if individual differences in DIF are expected, the DIF can be defined as random by adding a subscript p to ω (Van den Noortgate and De Boeck 2005) and defining a multivariate normal distribution for all random person effects accordingly.

In order to estimate the model in (10), first a new covariate must be defined, which we call `dif`, and which is not part of the example dataset. As an illustration, the covariate is defined for item subset gender DIF of the eight items from the verbal aggression data:

```
R> dif <- with(VerbAgg,
+    factor(0 + (Gender == "F" & mode == "do" & btype != "shout")))
```

(note that `Gender` is coded with `"F"` for men). The linear component for the model can now be formulated as follows:

```
-1 + item + dif + Gender + (1 | id)
```

More than one DIF covariate can be defined, also for individual items, depending on what one wants to investigate.

For DIF which is random across persons, the following formulation can be used:

```
-1 + item +  dif + Gender + (1+ dif | id)
```

In this way, a random intercept (for the latent trait) as well as a random DIF effect is obtained, and the correlation between both is estimated.

7.2. Local dependency models (internal binary person-by-item covariates)

Local independence is a basic assumption of IRT models. However, local independence does not always apply. One solution is to increase the dimensionality of the model but this is sometimes an overkill for this kind of problem. In such a case, or if dependency between certain items is the topic of interest, the model may be extended with a local item dependency (LID) component based on an internal item covariate (Meulders and Xie 2004).

A LID model can be formulated making use of a matrix $\mathbf{W}$ that is constructed as follows. First, define an $I \times I$ dependency matrix $\mathbf{D}$, so that $D_{ii'} = 1$ if the response to item i' is expected to depend on the response to item i, and $D_{ii'} = 0$ otherwise. Second, multiply $\mathbf{Y}$ (the $P \times I$ data matrix) with $\mathbf{D}$. Finally, transform the wide form of the product $\mathbf{YD}$ to its long form. The result is one person-by-item covariate $w_{(h=1)}$ for the case all dependencies contained in $\mathbf{D}$ have the same effect. If they are not expected to have the same effect, then $\mathbf{D}$ needs to be decomposed into more elementary matrices $\mathbf{D}_h$, one per LID parameter ω_h, so that $\mathbf{D}= \sum_{h=1}^{H} \mathbf{D}_h$. The corresponding long form is the $(P \times I) \times H$ matrix $\mathbf{W}$, with one column per dependency effect.

The resulting model is a recursive dependency model and differs from non-recursive variants as described by Kelderman (1984) and Wilson and Adams (1995). See (Tuerlinckx *et al.* 2004) for a discussion of the two types. The model is as follows:

$$\eta_{pi} = \theta_p + \beta_i + \sum_{h=1}^{H} \omega_h W_{(p,i)h}, \tag{11}$$

with θ_p and β_i as in (3), with $W_{(p,i)h}$ as the value of the item dependency covariate h, and ω_h as the corresponding LID parameter.

As explained by Hoskens and De Boeck (1997), the dependency may depend on the person. If that is the case, the LID effect should be defined as a random effect, adding a subscript p: ω_{hp} and defining a multivariate normal distribution for all random person effects accordingly.

In order to estimate the model in (11), first, a new covariate must be defined, which we call `dep`, and which is not part of the dataset. As an illustration, the covariate is defined for the dependency of do responses on want responses, assuming that the effect is the same for all 12 pairs:

```
R> dep <- with(VerbAgg, factor((mode == "do") * (r2[mode == "want"] == "Y")))
```

The linear component for the model can now be formulated as follows:

```
-1 + item + dep + (1 | id)
```

In order to include individual differences in the dependency, the formulation is as follows:

```
-1 + item + dep + (1 + dep | id)
```

More than one such covariate can be defined, also for individual pairs of items, depending on the dependency one wants to investigate.

7.3. Dynamic 1PL model (internal integer-valued person-by-item covariates)

An example of an internal integer-valued covariate is the person's progressive sum of correct responses preceding the item in question. For example, for a person with a response pattern for six items 011101, the corresponding covariate values for the six items are 001233. The covariate values clearly depend on the person and on the item. The effect of the progressive sum covariate is a learning effect induced by the amount of previous successes. As one can see, the covariate is of an internal nature because it is based on the responses to be modeled.

The dynamic Rasch model, as formulated by Verhelst and Glas (1993), makes use of this covariate:

$$\eta_{pi} = \theta_p + \beta_i + \omega_{\text{sum}} W_{(p,i)\text{sum}}, \tag{12}$$

with θ_p and β_i as in (3), and $W_{(p,i)\text{sum}}$ as the progressive sum, and ω_{sum} as its fixed effect.

If one wants the model to include individual differences in learning, as has been suggested by Verguts and De Boeck (2000), the learning effect should be defined as a random effect, adding a subscript p : $\omega_{\text{sum},p}$ and defining a multivariate normal distribution for all random person effects accordingly.

The learning model does not make much sense for the example dataset, but for illustrative reasons, we will nevertheless use this dataset. In order to estimate the model in (12), first a new covariate must be defined, which we shall call `prosum`, and which refers to the number of previous 1-responses. This involves some R syntax that we leave unexplained:

```
R> long <- data.frame(id = VerbAgg$id, item = VerbAgg$item, r2 = VerbAgg$r2)
R> wide <- reshape(long, timevar = "item", idvar = "id",
+    dir = "wide")[, -1] == "Y"
R> prosum <- as.vector(t(apply(wide, 1, cumsum)))
```

The linear component for the dynamic 1PL model can now be formulated as follows:

```
-1 + item + prosum + (1 | id)
```

To include individual differences in learning, the formulation is as follows:

```
-1 + item + prosum + (1+ prosum | id)
```

8. Model comparison and testing

A familiar method to compare two models is the likelihood ratio (LR) test. The test requires that the compared models are nested, which means that one or more parameters in a more general model (M_1) are constrained, most often to zero, in the more restricted model (M_0). It is well-known that asymptotically $-2\ln(L0/L1) \sim \chi^2(df)$, with df equal to the difference in the number of free parameters, and with L_0 and L_1 as the likelihood of M_0 and M_1, respectively. For two models which differ only in a single effect, $df = 1$.

In the present context, there are two problems with the LR test, a major one and a minor one. The first and major problem is that when the null hypothesis implies a zero variance, there cannot be a random fluctuation on both sides of the hypothesized value. Zero variance is a boundary value in the parameter space. As a consequence, the regular LR statistic with a χ^2-distribution does no longer apply. Stram and Lee (1994, 1995) have shown that comparing a model, M_1, with $r+1$ random effects, and a model M_0, with r random effects, the resulting LR statistic under the null hypothesis follows a mixture of $\chi^2(r)$ and $\chi^2(r+1)$ with a mixing probability of $\frac{1}{2}$. Further studies can be found in Giampaoli and Singer (2009) and Vu and Zhou (1997). The recommendation of Baayen *et al.* (2008) to use the LR test for variances as a conservative test (p-value too large) is in line with that result.

The score test and the Wald test are problematic for the same reason as the LR test, so that also for those two one would have to rely on a mixture of chi-squares. Molenberghs and Verbeke (2003) advise the LR test as the default because it is computationally easier. Although computationally more laborious, based on their application, the score test seems more robust than the other two. The computational burden can be overcome by using a GAUSS program provided by the same authors.

The second and minor problem is that the estimation method is based on a Laplace approximation of the likelihood, as explained in Doran *et al.* (2007). Strictly speaking, it is not the likelihood that is maximized, but an approximation of it, so that under the null hypothesis, the LR statistic cannot assumed to be asymptotically χ^2-distributed, but only approximately so. Because the Laplace approximation is reasonably accurate, this is not a major problem. From a small simulation study we made, comparing the Laplace approximation of the integrand with a Gauss-Hermite approximation of the integral, it turns out that the LR test works well for fixed effects, but less so to test the null hypothesis of a zero variance (absence of a random effect), using the mixture approach. It would require a much larger simulation study to draw more definite conclusions.

The LR test for nested models can be performed by using the `anova` function. For example, the earlier DIF model of (10) for gender DIF of do-items referring to cursing and scolding (model denoted as `m1`) can be tested against a model without such DIF (model denoted as

m0). The latter has one degree of freedom less than the former because all items in question are assumed to show the same DIF effect. In the following the code as well as extracts of the output are shown.

```
R> library("lme4")
R> m0 <- lmer(r2 ~ -1 + item + Gender + (1 | id), data = VerbAgg,
+    family = binomial)
R> dif <- with(VerbAgg,
+    factor(0 + (Gender == "F" & mode == "do" & btype != "shout")))
R> m1 <- lmer(r2 ~ -1 + item + dif + Gender + (1 | id), data = VerbAgg,
+    family = binomial)
```

Among the various output for model M_1 there is an important line with the estimate for the fixed effect associated with DIF:

```
dif1           1.00435   0.14375   6.987  2.81e-12***
```

For a formal test of M_1 against M_0 use function anova():

```
R> anova(m0, m1)

   Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
m0 26 8128.0 8308.3 -4038.0
m1 27 8081.2 8268.4 -4013.6 48.818      1  2.808e-12 ***
```

It is clear from the $\chi^2(1)$-statistic for the LR test that model M_0 must be rejected.

Another familiar method to compare models is to use Akaike's information criterion, AIC, and Schwartz' s information criterion, BIC. These criteria can also be used for non-nested models. Because they are based on the loglikelihood, strictly speaking, similar problems stemming from the approximate method apply. The lmer and the anova output shows the AIC and the BIC. It is clear from a comparison of the models with and without DIF that the model with DIF has a smaller AIC and BIC, and is therefore the preferred model.

The output of the lmer function shows also a z-value for the fixed parameters based on the estimated standard error. This z-statistic can be used to test the effect. The z-test is asymptotically equivalent with the LR test, as is illustrated by the almost perfect correspondence between the p-values of both tests for the above investigation of DIF. The z-test must not be used for variances because the distribution of the z-statistic cannot be normal given that the null hypothesis is located on the boundary of the parameter space.

Finally, one can inspect the posterior of the parameters using the mcmcsamp function and plot the estimated posterior densities with densityplot(). Baayen *et al.* (2008) explain how to proceed from there to derive Bayesian confidence intervals with an ancillary R function.

Note on some lmer options The manual offers an option for the number of nodes for the estimation (nAGQ = x). When using the 1PL for the example data, the effect is negligible, except for the likelihood, which decreases slightly with the number of nodes. The manual also offers the option of restricted maximum likelihood as a default. For the example data, the effect of REML = TRUE versus REML = FALSE is null.

9. A comparison with other IRT packages in R

For the 1PL applied to the dataset, the results of the `lmer` function are compared with the six programs described by Tuerlinckx *et al.* (2004) for the same dataset. The `lmer` variance is 1.90, which is lower than the estimate obtained with the three Gauss-Hermite quadrature based algorithms among the six (estimates of 1.98), but higher than PQL and PQL2 based estimates from the other three (estimates of 1.70 and 1.87, respectively). It is known that the variance estimates are lower for estimation methods based on an approximation of the integrand, such as the Laplace approximation, compared with methods based on Gauss-Hermite integration. However, the item parameter estimates are nearly identical. The absolute deviation is at most 0.01. Also the standard error estimates are highly similar.

These findings are confirmed with an analysis based on **ltm** (Rizopoulos 2006). The general discrimination parameter estimate is 1.455, which corresponds to the 1.98 variance estimate obtained with other Gauss-Hermite based programs.

The **ltm** package can be used to estimate the 1PL, 2PL, and 3PL for binary items and the graded-response model for polytomous items with a logit link. Several other programs in R (R Development Core Team 2010) are available for item response modeling. In contrast with the **lme4** package and its `lmer` function, **ltm** and the other packages are rather model-oriented, and therefore of the first type as mentioned in the introduction. One way to categorize the R programs for IRT further is as follows:

1. Two packages for Rasch families of models

   - The **eRm** package (Mair and Hatzinger 2007) using conditional maximum likelihood estimation for the Rasch model, the partial credit and rating scale models, and the corresponding models with property covariates;
   - The **plRasch** package (Anderson *et al.* 2007) using maximum likelihood and pseudolikelihood estimation for loglinear formulations of the Rasch family of models for response patterns of binary and polytomous items and a single or multiple latent traits.

2. Two packages with Bayesian estimation for a family of models and one specific model

   - The **mlirt** package (Fox 2007) for a Bayesian estimation of the 2P normal-ogive model for binary and polytomous items, and multilevel extensions;
   - The **gpcm** package (Johnson 2007) for a Bayesian estimation of the generalized partial credit model.

3. Two programs from a political science perspective (IRT models are called ideal-point models in political sciences). The packages are for one-dimensional and multidimensional models, and a probit link is used:

   - The **MCMCpack** package (Martin *et al.* 2011) for the Bayesian estimation of the two-parameter one-dimensional and multidimensional normal-ogive models for binary responses. Also four-parameter variants can be estimated, which are seen as robust models because the manifest response may deviate from the covert response, by allowing for an upper and lower asymptote of the item response function;

- The **pscl** package (Jackman 2010) with a Bayesian estimation of the two-parameter one-dimensional and multidimensional normal-ogive models for binary items.

The packages from the above category 1 do not yield variance estimates. The packages from the above categories 2 and 3 are based on MCMC and therefore may be expected to yield variance estimates which are about the same as the Gauss-Hermite based programs. When **WinBUGS** was used to estimate the 1PL for the example data, it was found indeed that the variance estimate was very similar (Tuerlinckx *et al.* 2004).

10. Discussion and conclusion

The taxonomy of models and the large number of model variants presented here are meant to illustrate the rich potential of the GLMM and explanatory perspective on item response models and beyond. We have followed the tradition of identifying separate models but, within the more general framework, they differ only in the covariates and effects that are included.

Given the variety of the item response models that can be estimated with the `lmer` function, it is a highly flexible tool and not just for item response model estimation, but also as an alternative for analysis of variance with binary (and Gaussian) repeated observations as illustrated in the special issue of *Journal of Memory and Language* (Forster and Masson 2008). Apart from its flexibility, a major asset is the GLMM background, which is conceptually interesting and facilitates the links with other domains of modeling and with the statistical findings in those domains.

On the other hand, the `lmer` function cannot be used for popular IRT models such as the two-parameter and three-parameter models, and the partial-credit and graded-response models. The two former models are not GLMMs, and the latter two require multivariate logits or probits. In comparison with other R (R Development Core Team 2010) programs, such as **ltm** and **eRm**, `lmer` needs more runtime. For the Rasch model with fixed item effects and the example dataset, the runtimes in seconds are 1.56 (**ltm**), 4.70 (**eRm**), and 26.73 (`lmer`) on an Intel T4200 processor (2 GHz). However, in comparison with another general package such as SAS `PROC NLMIXED`, the runtime of the `lmer` function is rather small.

When `lmer` possibilities for IRT are compared with those of SAS `PROC NLMIXED`, which is based on a similar statistical background, several major differences can be noticed: SAS `PROC NLMIXED` can estimate nonlinear mixed models, such as the two-parameter and three-parameter models, and models for categorical data, such as the partial-credit and graded-response models, latent variables can be made a function of one another (as in SEM), and its estimation method relies on the Gauss-Hermite approximation of the integral. On the other hand, SAS `PROC NLMIXED` cannot be used for multilevel models and for crossed random effect modeling, it cannot handle more than a few dimensions, and is rather slow. In contrast, because `lmer` is based on an approximation of the integrand, it is really fast and seems to have no problems with higher dimensionalities, although further study is required to test its qualities in that respect. The ideal would be to combine the qualities of the two kinds of programs. However, it may be difficult to compete with a Bayesian package when it comes to flexibility. It is interesting that both Baayen *et al.* (2008) and Doran *et al.* (2007) recommend the Bayesian approach with the `mcmcsamp` function to inspect the posterior for a statistical evaluation of the effects.

In sum, `lmer` is a highly interesting tool added to the toolkit of item response modeling.

We have highlighted some of its potentialities, but more is possible. One example is random item models, with random item variables linked to subsets of persons, just as random person variables (latent person dimensions) are linked to subsets of items (De Boeck 2008).

Acknowledgments

The article is based on a course, "Explanatory Item Response Models", held by the first author at the University of Amsterdam in the Fall of 2009, with the assistance of the second author. Robert Zwitser, Michel Nivard, and Abe Hofman have contributed to the manuscript through extra work related to the course. We hereby thank also the other students who have participated in the course, and Peter Halpin and Sun-Joo Cho for their helpful comments on earlier drafts of this manuscript. Finally, we are grateful to the anonymous reviewers of the manuscript for their suggestions.

References

Adams RJ, Wilson MR, Wang W (1997). "The Multidimensional Random Coefficients Multinomial Logit Model." *Applied Psychological Measurement*, **21**, 1–23.

Akaike M (1974). "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control*, **19**, 716–723.

Andersen EB (1980). *Discrete Statistical Models with Social Science Applications*. North-Holland, Amsterdam.

Anderson CJ, Li Z, Vermunt JK (2007). "Estimation of Models in a Rasch Family for Polytomous Items and Multiple Latent Variables." *Journal of Statistical Software*, **20**(6), 1–36. URL http://www.jstatsoft.org/v20/i06/.

Baayen RH, Davidson DJ, Bates DM (2008). "Mixed-Effects Modelling with Crossed Random Effects for Subjects and Items." *Journal of Language and Memory*, **59**, 390–412.

Bates D, Maechler M, Bolker B (2011). ***lme4:*** *Linear Mixed-Effects Models Using S4 Classes*. R package version 0.999375-38, URL http://CRAN.R-project.org/package=lme4.

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*. Springer-Verlag, New York.

de Leeuw J, Mair P (2007). "An Introduction to the Special Volume of "Psychometrics in R"." *Journal of Statistical Software*, **20**(1), 1–5. URL http://www.jstatsoft.org/v20/i01/.

Doran H, Bates D, Bliese P, Dowling M (2007). "Estimating the Multilevel Rasch Model: With the **lme4** Package." *Journal of Statistical Software*, **20**(2), 1–18. URL http://www.jstatsoft.org/v20/i02/.

Fahrmeir L, Tutz G (2001). *Multivariate Statistical Modeling Based on Generalized Linear Mixed Models*. 2nd edition. Springer-Verlag, New York.

Fischer GH (1973). "The Linear Logistic Test Model as an Instrument in Educational Research." *Acta Psychologica*, **3**, 359–374.

Forster KI, Masson MEJ (2008). "Special Issue: Emerging Data Analysis." *Journal of Memory and Language*, **59**, 387–556.

Fox JP (2007). "Multilevel IRT Modeling in Practice with the Package **mlirt**." *Journal of Statistical Software*, **20**(5), 1–16. URL http://www.jstatsoft.org/v20/i05/.

Ghosh M (1995). "Inconsistent Maximum Likelihood for the Rasch Model." *Statistics & Probability Letters*, **23**, 165–170.

Giampaoli V, Singer JM (2009). "Likelihood Ratio Tests for Variance Components in Linear Mixed Models." *Journal of Statistical Planning and Inference*, **139**, 1435–1448.

Holland PW, Wainer H (eds.) (1993). *Differential Item Functioning*. Erlbaum, Hillsdale, NJ.

Hoskens M, De Boeck P (1997). "A Parametric Model for Local Item Dependencies Among Test Items." *Psychological Methods*, pp. 261–277.

Jackman S (2010). ***pscl**: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University*. Department of Political Science, Stanford University, Stanford, California. R package version 1.03.6, URL http://CRAN.R-project.org/package=pscl.

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. Springer-Verlag, New York.

Johnson MS (2007). "Marginal Maximum Likelihood Estimation of Item Response Models in R." *Journal of Statistical Software*, **20**(10), 1–14. URL http://www.jstatsoft.org/v20/i10/.

Kelderman H (1984). "Loglinear Rasch Model Tests." *Psychometrika*, **49**, 223–245.

Mair P, Hatzinger R (2007). "Extended Rasch Modeling: The **eRm** Package for the Application of IRT Models in R." *Journal of Statistical Software*, **20**(9), 1–20. URL http://www.jstatsoft.org/v20/i09/.

Martin AD, Quinn KM, Park JH (2011). ***MCMCpack:** Markov Chain Monte Carlo in R*. Forthcoming.

McCulloch CE, Searle SR (2001). *Generalized, Linear, and Mixed Models*. John Wiley & Sons, New York.

Meulders M, Xie Y (2004). "Person-by-Item Predictors." In P De Boeck, M Wilson (eds.), *Explanatory Item Response Models*, pp. 213–240. Springer-Verlag, New York.

Millsap RE, Everson HT (1993). "Methodology Review: Statistical Approaches for Assessing Measurement Bias." *Applied Psychological Measurement*, **17**, 297–334.

Molenberghs G, Verbeke G (2003). "Likelihood Ratio, Score, and Wald Tests in a Constrained Parameter Space." *The American Statistician*, **61**, 1–6.

Muthén BO (1987). ***LISCOMP:*** *Analysis of Linear Structural Equations with a Comprehensive Measurement Model*. Scientific Software, Mooresville, IN.

Muthén LK, Muthén BO (1998). ***Mplus*** *User's Guide*. Muthén & Muthén, Los Angeles, CA. URL http://www.statmodel.com/.

Rabe-Hesketh S, Skrondal A, Pickles A (2004). "**gllamm** Manual." *Working paper 160*, U.C. Berkeley Division of Biostatistics Working Paper Series. URL http://www.bepress.com/ucbbiostat/paper160/.

Raudenbush SW (1993). "A Crossed Random Effects Model for Unbalanced Data with Applications in Cross-Sectional and Longitudinal Research." *Journal of Educational & Behavioral Statistics*, **18**, 321–349.

Raudenbush SW, Bryk AS, Cheong YF, Congdon R, du Toit M (2004). ***HLM*** *6: Hierarchical Linear and Nonlinear Modeling*. Scientific Software International, Lincolnwood, IL. URL http://www.ssicentral.com/hlm/.

R Development Core Team (2010). *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/.

Rijmen F, Briggs D (2004). "Multiple Person Dimensions and Latent Item Predictors." In P De Boeck, M Wilson (eds.), *Explanatory Item Response Models*, pp. 247–265. Springer-Verlag, New York.

Rijmen F, De Boeck P (2002). "The Random Weights Linear Logistic Test Model." *Applied Psychological Measurement*, **26**, 269–283.

Rijmen F, Tuerlinckx F, De Boeck P, Kuppens P (2003). "A Nonlinear Mixed Model Framework for Item Response Theory." *Psychological Methods*, **8**, 185–205.

Rizopoulos D (2006). "**ltm**: An R Package for Latent Variable Modeling and Item Response Theory Analyses." *Journal of Statistical Software*, **17**(5), 1–25. URL http://www.jstatsoft.org/v17/i05/.

SAS Institute Inc (2008). *SAS/STAT Software, Version 9.2*. Cary, NC. URL http://www.sas.com/.

Scheiblechner H (1972). "Das Lernen und Lösen komplexer Denkaufgaben." *Zeitschrift für experimentelle und angewandte Psychologie*, **19**, 746–506.

Schwartz G (1978). "Estimating the Dimension of a Model." *The Annals of Statistics*, **6**, 461–464.

Sheu CF, Chen CT, Su YH, Wang WC (2005). "Using SAS PROC NLMIXED to Fit Item Response Theory Models." *Behavior Research Methods*, **37**, 202–218.

Stram DO, Lee JW (1994). "Variance Components Testing in the Longitudinal Mixed-Effects Model." *Biometrics*, **50**, 1171–1177.

Stram DO, Lee JW (1995). "Correction to: Variance Components Testing in the Longitudinal Mixed-Effects Model." *Biometrics*, **51**, 1196.

Tuerlinckx F, Rijmen F, Molenberghs G, Verbeke G, Briggs D, Van den Noortgate W, Meulders M, De Boeck P (2004). "Estimation and Software." In P De Boeck, M Wilson (eds.), *Explanatory Item Response Models*. Springer-Verlag, New York.

Tutz G (1990). "Sequential Item Response Models with an Ordered Response." *British Journal of Mathematical and Statistical Psychology*, **43**, 39–55.

Van den Noortgate W, De Boeck P (2005). "Assessing and Explaining Differential Item Functioning Using Logistic Mixed Models." *Journal of Educational & Behavioral Statistics*, **30**, 443–464.

Verguts T, De Boeck P (2000). "A Rasch Model for Learning while Solving an Intelligence Test." *Applied Psychological Measurement*, **24**, 151–162.

Verhelst N, Glas K (1993). "A Dynamic Generalization of the Rasch Model." *Psychometrika*, **58**, 395–415.

Vermunt JK, Magidson J (2005). ***Latent Gold*** *4.0 User's Guide*. Statistical Innovations, Inc., Belmont, MA. URL http://www.statisticalinnovations.com/.

Vu HTV, Zhou S (1997). "Generalization of Likelihood Ratio Tests under Nonstandard Conditions." *The Annals of Statistics*, **25**, 847–916.

Wickham H (2007). "Reshaping Data with the **reshape** Package." *Journal of Statistical Software*, **21**(12), 1–20. URL http://www.jstatsoft.org/v21/i12/.

Wilson M, Adams RJ (1995). "Rasch Models for Item Bundles." *Psychometrika*, **60**, 181–198.

Yee TW (2010). "The **VGAM** Package for Categorical Data Analysis." *Journal of Statistical Software*, **32**(10), 1–34. URL http://www.jstatsoft.org/v32/i10/.

Yee TW, Wild CJ (1996). "Vector Generalized Additive Models." *Journal of the Royal Statistical Society B*, **58**, 481–493.

Affiliation:

Paul De Boeck
Department of Psychology
University of Amsterdam
Roeterstraat 15
1018 WB Amsterdam, The Netherlands
E-mail: paul.deboeck@uva.nl
URL: http://www.fmg.uva.nl/psychologicalmethods/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 39, Issue 12 *Submitted:* 2010-07-01
March 2011 *Accepted:* 2010-12-13