A Modular Approach for Item Response Theory Modeling with the R Package flirt

Minjeong Jeon[1]

Frank Rijmen[2]

[1] The Ohio State University

[2] CTB/McGraw-Hill

Correspondence:

Minjeong Jeon
Department of Psychology
Faculty of Quantitative Psychology
The Ohio State University
228 Lazenby Hall 1827 Neil Avenue, Columbus, OH 43210
E-mail: jeon.117@osu.edu

Abstract

The new R package **flirt** is introduced for a flexible modular item response theory (IRT) modeling of psychological, educational and behavior assessment data. **flirt** integrates a generalized linear and nonlinear mixed modeling framework with graphical model theory. The graphical model framework allows for efficient maximum likelihood estimation. The key feature of **flirt** is its modular approach to facilitate convenient and flexible model specifications. Researchers can construct customized IRT models by simply selecting various modeling modules that are needed, such as parametric forms, number of dimensions, (number of) item and person covariates, (number of) person groups, link functions (for different response types of response variables), etc. In this paper, we describe major features of **flirt** and provide examples to illustrate how **flirt** works in practice.

Key words:Modular approach; Software; Item response theory; Explanatory models; Multidimensional models; Bifactor models; DIF

1 Introduction

Item response theory (IRT) models are widely used in educational, psychological, and social science research. A number of commercial software packages are available for the estimation of IRT models, such as Bilog-MG (Zimowski et al., 2006), Multilog (Thissen, 1991), ConQuest (Adams et al., 2012), IRTPRO (Cai et al., 2011), and FlexMIRT (Cai, 2012). General purpose statistical software packages and software for structural equation modeling or generalized linear mixed modeling have also been used for the estimation of IRT models: for example, Mplus (Muthén & Muthén, 2012), SAS (e.g., the nlmixed procedure, Wolfinger, 2008), and gllamm (Rabe-Hesketh et al., 2005).

In the past years, many free packages have been developed in the R environment (R Development Core Team, 2013). For instance, **ltm** (Rizopoulos, 2006) for the unidimensional one-, two-, three-parameter logistic models, **eRm** (Mair et al., 2014) for Rasch family models including rating scale and partial credit models, **mlirt** (Fox, 2007) for multilevel IRT models for binary and polytomous responses, **mirt** (Chalmers, 2012) for exploratory and confirmatory multidimensional and bifactor IRT models for binary responses, **sirt** (Robitzsch, 2013) for various IRT models for binary responses, **TAM** (Kiefer et al., 2014) for uni- and multi-dimensional Rasch and two-parameter logistic models. **lme4** (Bates et al., 2014) for generalized linear mixed models has also been used to estimate various Rasch family models for binary responses (De Boeck et al., 2011). For a full list of IRT packages see http://cran.r-project.org/web/views/Psychometrics.html.

Even though a number of free and commercial software packages are currently available as listed above, there is still a need for a new type of software for the following reasons: First, most IRT packages offer only a limited number of models and therefore researchers are forced to choose one among a set of pre-defined models; that is, there is less freedom for researchers in building their own models. In addition, researchers often have to move from one package to another in order to explore other types of models. This is not only inefficient but also could be misleading in that different software packages often employ different estimation methods and model parameterization that may not be directly comparable. Second, most software packages focus on descriptive IRT models without explanations on why some items are more difficult than others and why some people are more able

than others. Such questions can be answered by incorporating a regression model on the item and person sides, which leads to explanatory IRT models. Furthermore, explanatory models can be used to investigate construct validity (Embretson, 1983) and to modify distributional assumptions on the latent variables (Bock & Zimowski, 1997). Third, general commercial statistical software packages, such as Mplus, SAS nlmixed or gllamm, may be not only too general but also not dedicated to IRT analysis; thus, it is often not straightforward on how to specify particular IRT models of interest and interpret outputs with those software packages.

In this paper, we introduce the free R package, **flirt** (flexible item response theory). As the acronym of the package indicates, **flirt** offers flexible modeling of item response data. Flexibility of **flirt** comes from its general statistical framework: By conceptualizing IRT models as generalized linear and nonlinear mixed models, various types of IRT models can be understood and constructed with **flirt** by simply selecting and combining various modules, such as a parametric form, the number of dimensions, the number of item and person covariates, the number of person groups, and a link function for different types of response variables, etc. Furthermore, **flirt** is a dedicated IRT software package and provides IRT-friendly specifications of various models and interpretations of outputs.

Another strength of **flirt** comes from its efficient maximum likelihood (ML) estimation using a modified expectation-maximization (EM) algorithm based on graphical model theory. The modified EM algorithm implements the expectation (E) step in an efficient way such that computations can be carried out in lower-dimensional latent spaces. Additional computational efficiency is achieved by adopting adaptive quadrature for numerical integration. The gain in computational efficiency can be nontrivial, in particular for high-dimensional models. Details on the computational strength will be discussed in Section 3.

The rest of this paper is organized as follows: In Section 2, we first describe the statistical framework of **flirt**, generalized linear and nonlinear mixed models, and explain how a variety of IRT models can be conceptualized in this framework. In Section 3, the estimation framework of **flirt** is described. In Section 4, important features of **flirt** are illustrated using empirical examples. In Section 5, we end with some concluding remarks.

2 Statistical framework

2.1 Generalized linear and nonlinear mixed models

The flexibility of **flirt** comes from its underlying general statistical framework - generalized linear and nonlinear mixed models (GLNMMs). Generalized linear mixed models (GLMMs) are a class of models for analysis of clustered normal and non-normal data, such as repeated measurements (e.g., item responses) within subjects. Correlations within clusters are accounted for by incorporating random cluster effects, i.e., by assuming a cluster specific effect that has a distribution over the populations of clusters. In GLMMs, it is assumed that a within-subject model (or linear predictor) is related to the conditional mean of the response variable given random effects via a link function. GLNMMs are a broader class of GLMMs in which the within-subject model allows for a nonlinear combination of model parameters.

How IRT models can be conceptualized as nonlinear mixed models has been discussed by Rijmen et al. (2003). Here we briefly describe a GLNMM framework and show how a variety of IRT models can be specified in this framework.

For simplicity, let us assume binary responses y_{pi} for person p $(= 1, ..., N)$ to item i $(= 1, ..., I)$. The observation y_{pi} is assumed to have a Bernoulli distribution with conditional probability π_{pi} given latent variables or random effects $\boldsymbol{\theta}_p$. The conditional expectation of the responses, $\mu_{pi} = E(y_{pi}|\boldsymbol{\theta}_p)$ is related to the linear predictor ν_{pi} via a link function $g(\cdot)$

$$g(\mu_{pi}) = \nu_{pi}.$$

With binary responses, the conditional expectation of the responses is equivalent to the conditional probability of a correct response $(y_{pi} = 1)$ given the cluster-specific random effects (or latent variables)

$$g(\mu_{pi}) = g(\pi_{pi}) = g(P(y_{pi} = 1|\boldsymbol{\theta}_p)).$$

A commonly used link function for binary data is the logit link

$$\begin{aligned} g(\pi_{pi}) &= \log \frac{\pi_{pi}}{1-\pi_{pi}} \\ &= \text{logit}\left(\pi_{pi}\right). \end{aligned}$$

Two other link functions for binary data are the probit link and the complementary log-log link (McCullagh & Nelder, 1989; Rijmen et al., 2003).

In GLMMs, we can write the linear predictor with e.g., a single latent variable (or random effect) and Q observed covariates

$$\nu_{pi} = \sum_{q=1}^{Q} \beta_q X_{iq} + \theta_p, \tag{1}$$

where β_q is the regression coefficient for the qth observed covariate X_{iq} $(q = 1, ..., Q)$ and θ_p is the random effect with $\theta_p \sim N(0, \sigma^2)$.

In GLNMMs with the same setting as (1), the linear predictor can be written as

$$\nu_{pi} = \sum_{q=1}^{Q} \beta_q X_{iq} + \alpha_i \theta_p, \tag{2}$$

where α_i is the scaling (or loading) parameter for the random effect θ_p. Model (2) is a nonlinear model because of the scaling parameter α_i that is multiplied by the random effect θ_p. For identification, one of the α_i (typically $i = 1$) is fixed to 1 or the variance of θ_p is fixed to 1 in (2). Note that GLNMMs (2) can be reduced to GLMMs (1) by fixing $\alpha_i = 1$ for all i.

2.2 IRT models as generalized linear and nonlinear mixed models

Now we show how IRT models can be specified as GLNMMs. We begin with a one parameter logistic (1PL) model followed by a two parameter logistic (2PL) model for binary responses. Various other models are described as extensions of the 2PL model.

1. 1PL model

The 1PL model is formulated by setting $\alpha_i = 1$ in model (2)

$$\text{logit}(P(y_{pi} = 1|\theta_p)) = \nu_{pi} = \sum_{d=1}^{I} \beta_d X_{id} + \theta_p, \tag{3}$$

where β_d is the regression coefficient for dummy (or indicator) variables for items, i.e., $X_{id} = 1$ if $i = d$ and 0 otherwise, and θ_p is the latent variable with $\theta_p \sim N(0, \sigma^2)$. Typically β_d is referred to as the item intercept parameter and θ_p represents the ability or proficiency of person p.

2. 2PL model

The 2PL model is formulated by freely estimating α_i in model (2)

$$\text{logit}(P(y_{pi} = 1|\theta_p)) = \nu_{pi} = \sum_{d=1}^{I} \beta_d X_{id} + \alpha_i \theta_p, \tag{4}$$

where α_i is the scaling parameter for the latent variable is θ_p, $\theta_p \sim N(0, \sigma^2)$. To identify the model, σ^2 is fixed to 1 or α_1 is fixed to 1.

Equation (4) can be written in an IRT-familiar form

$$\text{logit}(P(y_{pi} = 1|\theta_p)) = \alpha_i \theta_p + \beta_i, \tag{5}$$

where β_i is the intercept and α_i is the loading or slope parameter.

Equation (5) is referred to as an item-intercept parameterization of the 2PL model. An item-difficulty parameterization can be obtained as

$$\text{logit}(P(y_{pi} = 1|\theta_p)) = \alpha_i(\theta_p + \beta_i'), \tag{6}$$

where α_i is referred to as the discrimination parameter and $-\beta_i'$ as the item difficulty parameter (and β_i' is the item easiness parameter). It can be easily shown that $\beta_i' = \beta_i/\alpha_i$. **flirt** allows for both parameterizations expressed in (5) and (6).

3. Multiple groups

The basic model (4) assumes that all persons are sampled from a common distribution. We can relax this assumption by assuming separate distributions for different groups of people. Then, a multiple group extension of (4) can be specified

$$\text{logit}(P(y_{pi} = 1|\theta_p)) = \sum_{d=1}^{I} \beta_d X_{id} + \alpha_i \theta_{pg}, \tag{7}$$

where the latent variable θ_{pg} depends on group g for person p and $\theta_{pg} \sim N(\mu_g, \sigma_g^2)$. For the reference group, μ_g and σ_g are set to 0 and 1, respectively. For the focal groups, μ_g and σ_g are freely estimated. The 1PL multiple-group model is also available in **flirt**.

4. Multiple dimensions

The basic model (4) assumes a single underlying latent variable θ_p, implying all items are located on a single test scale. In practice, however, the latent trait of interest may be more complex than that. For example, a complex performance can be understood by taking into account knowledge structures, cognitive processes, or interactions between multiple component behaviors (Kelderman & Rijkes, 1994). Assuming that a test consists of K subscales (or dimensions), model (4) can be extended for K latent variables (or dimensions)

$$\text{logit}(P(y_{pi} = 1|\boldsymbol{\theta}_p)) = \sum_{d=1}^{I} \beta_d X_{id} + \sum_{k=1}^{K} \alpha_{i(k)} \theta_{pk}, \tag{8}$$

where $\alpha_{i(k)}$ is the loading (or slope) for item i in dimension k that item i belongs to and θ_{pk} is the kth latent variable ($k = 1, ..., K$). If items belong to no more than one dimension, the model is called a between-item multidimensional model; then the loading matrix $\Lambda = (\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_K)$ where $\boldsymbol{\alpha}_k$ is a I dimensional vector of item loadings for dimension k, can be written, e.g., with $I = 6$ and

$K = 3$

$$\Lambda = \begin{bmatrix} \alpha_{1(1)} & 0 & 0 \\ \alpha_{2(1)} & 0 & 0 \\ 0 & \alpha_{3(2)} & 0 \\ 0 & \alpha_{4(2)} & 0 \\ 0 & 0 & \alpha_{5(3)} \\ 0 & 0 & \alpha_{6(3)} \end{bmatrix},$$

where two non-overlapping items belong to dimension 1, 2, and 3, respectively. On the other hand, items might belong to multiple dimensions at the same time; such models are referred to as a within-item multidimensional model (Adams et al., 1997). The loading matrix Λ can then be written for a within-item multidimensional model

$$\Lambda = \begin{bmatrix} \alpha_{1(1)} & 0 & \alpha_{1(3)} \\ \alpha_{2(1)} & 0 & 0 \\ 0 & \alpha_{3(2)} & \alpha_{3(3)} \\ 0 & \alpha_{4(2)} & 0 \\ \alpha_{5(1)} & 0 & \alpha_{5(3)} \\ 0 & 0 & \alpha_{6(3)} \end{bmatrix}.$$

where item 1 belongs to dimensions 1 and 3, item 3 belongs to dimensions 2 and 3, and item 5 belongs to dimensions 1 and 3, respectively, and item 2 belongs to dimension 1, item 4 to dimension 2, and item 6 to dimension 3. The package **flirt** allows for specification of both between-item and within-item multidimensional models.

In Equation (8), the K latent variables are allowed to be correlated with each other and assumed to follow a multivariate normal distribution, $\boldsymbol{\theta}_p = (\theta_{p1}, ..., \theta_{pK})' \sim N(\mathbf{0}, \Sigma)$. The package **flirt** estimates the elements of a lower triangular Cholesky matrix L for the covariance matrix Σ, where

$\Sigma = L \cdot L^\top$, with $L^\top$ is a transpose of L. For example, with $K = 3$

$$\underbrace{\begin{bmatrix} c1 & 0 & 0 \\ c4 & c2 & 0 \\ c5 & c6 & c3 \end{bmatrix}}_{L} \cdot \underbrace{\begin{bmatrix} c1 & c4 & c5 \\ 0 & c2 & c6 \\ 0 & 0 & c3 \end{bmatrix}}_{L^\top} = \underbrace{\begin{bmatrix} c1^2 & c1 \cdot c4 & c1 \cdot c5 \\ c1 \cdot c4 & c4^2 + c2^2 & c4 \cdot c5 + c2 \cdot c6 \\ c1 \cdot c5 & c4 \cdot c5 + c2 \cdot c6 & c5^2 + c6^2 + c3^2 \end{bmatrix}}_{\Sigma}, \tag{9}$$

where $c1$ to $c6$ are estimated and then estimated covariance matrix Σ is constructed. For identification, the diagonal elements of L, $c1, c2$, and $c3$ are fixed to 1 when all loading parameters are freely estimated in the 2PL model family and for the reference group with a multiple-group models; hence, some diagonal elements (that is, variances) in Σ (i.e., $c4^2 + c2^2$ and $c5^2 + c6^2 + c3^2$) may not be 1. Standardized covariance matrices (or correlation matrices) can be obtained using a built-in function in **flirt**. Details are illustrated in Section 4.

The 1PL multidimensional models can be formulated in **flirt** and $\boldsymbol{\theta}_p \sim N(\mathbf{0}, \Sigma)$, where Σ is freely estimated; that is, $c1$ to $c6$ in L are all freely estimated.

5. Bifactor structure

An item bifactor model is a special case of multidimensional IRT models with a bifactor structure. The bifactor structure is characterized by a general dimension in addition to specific (or group) dimensions. Given the general dimension, specific dimensions are assumed to be independent of each other and of the general dimension. An item bifactor model is useful for tests that consist of item bundles or testlets, such as reading test items within reading passages and science test items within common stimuli. The general dimension is the primary dimension of interest, while items within testlets are likely to be correlated with each other and dependencies within testlets constitute specific dimensions.

Basic model (4) can be extended with a bifactor structure

$$\text{logit}(P(y_{pi} = 1|\theta_p)) = \sum_{d=1}^{I} \beta_d X_{id} + \sum_{k=1}^{K+1} \alpha_{i(k)} \theta_{pk}, \tag{10}$$

where θ_{pk} is the general dimension when $k = 1$ and when $k > 1$, θ_{pk} is the $(k-1)$th specific

dimension ($k = 1, ..., (K+1)$). $\alpha_{i(k)}$ is the loading parameter for the general dimension when $k = 1$ and when $k > 1$, the $(k-1)$th specific dimension. That is, item i has two loadings, one for the general dimension and another for the specific dimension that item i belongs to. For a single-group bifactor model, all $(K+1)$ latent variables are independent of each other and follow a multivariate normal distribution, $\boldsymbol{\theta}_p = (\theta_p^G, \theta_{p1}^K, ..., \theta_{pK}^K)' \sim N(\mathbf{0}, \Sigma^D)$, where Σ^D is a $(K+1) \times (K+1)$ diagonal matrix. For a single-group model, Σ^D becomes the identity matrix. As in the multidimensional model (8), the Cholesky matrix of Σ^D is estimated in **flirt**.

For a multiple-group bifactor model, the assumption of independent latent variables is relaxed to conditional independence for the focal group; that is, K specific dimensions are assumed to be conditionally independent of each other given the general dimension θ_p^G and Σ for the focal group is freely estimated. For the reference group, the independence assumption of the $K+1$ latent variables holds to resolve rotational indeterminacy and Σ becomes the identity matrix. For details on the multiple-group bifactor model, see Jeon et al. (2013).

6. Explanatory models

The basic IRT formulation in (4) is a descriptive model and useful for measurement purposes. With a descriptive model, however, questions on items and persons, such as why some items are more difficult than others and why some people are more able than others, remain unexplained. Those questions can be answered by incorporating explanatory variables in the model.

Here we formulate explanatory extensions of model (4) with two types of covariates: item covariates and person covariates.

1) Item covariates

Model (4) can be extended with item covariates for the item intercept parameters

$$\text{logit}(P(y_{pi} = 1|\theta_p)) = \sum_{q=1}^{Q} \beta_q^* X_{iq} + \alpha_i \theta_p, \tag{11}$$

where β_q^* is the regression coefficient for qth item covariates X_{iq} ($q = 1, ..., Q$, $Q < I$). A well-known IRT model with item covariates is the linear logistic test model (LLTM; Fischer, 1973)

with known loading parameters. **flirt** can estimate model (11) with known as well as unknown loading parameters with any types of item covariates (e.g., categorical, numerical). Similarly, the multidimensional model (8) can be extended with item covariates

$$\text{logit}(P(y_{pi}=1|\boldsymbol{\theta}_p)) = \sum_{q=1}^{Q} \beta_q^* X_{iq} + \sum_{k}^{K} \alpha_{i(k)} \theta_{pk}. \tag{12}$$

Extensions of the bifactor model and the 1PL versions of (11) and (12) are also available in **flirt**. Item covariates can be incorporated for the item discrimination (loading) parameters as a linear constraint for α_i, i.e., $\alpha_i = \sum_{p=1}^{P} \alpha_p^* Z_{ip}$, which was proposed in a 2PL constrained model (Embretson, 1991).

2) Person covariates

Model (4) can be extended with person covariates

$$\begin{aligned} \text{logit}(P(y_{pi}=1|\theta_p)) &= \sum_{d=1}^{I} \beta_d X_{id} + \alpha_i \theta_p, \\ \theta_p &= \sum_{r=1}^{R} \gamma_r W_{pr} + \theta'_p, \end{aligned} \tag{13}$$

where γ_r is the regression coefficient for the rth person covariates W_{pr} $(r = 1, ..., R)$, θ'_p is the residual for θ_p after being explained by the person covariates, and $\theta'_p \sim N(0, \sigma'^2)$. An IRT model with person covariates are referred to as a latent regression (Adams et al., 1997). Equation (13) is parameterized as a two-level multilevel model formulation, where the first line is level-1 for item responses and the second line is level-2 for persons. The person covariates are entered in level-2. With **flirt**, person covariates can also be entered as the main effects in the first line as

$$\text{logit}(P(y_{pi}=1|\theta_p)) = \sum_{d=1}^{I} \beta_d X_{id} + \sum_{r=1}^{R} \gamma_r W_{pr} + \alpha_i \theta'_p. \tag{14}$$

Note that α_i is not affected by $\sum_{r=1}^{R} \gamma_r W_{pr}$ in formulation (14). In addition, the package **flirt** can include any types of person covariates (e.g., categorical, numerical). Multidimensional models can

be extended with person covariates

$$\text{logit}(P(y_{pi}=1|\boldsymbol{\theta}_p)) = \sum_{d=1}^{I}\beta_d X_{id} + \sum_{k}^{K}\alpha_{i(k)}\theta_{pk}, \tag{15}$$
$$\theta_{pk} = \sum_{r_k=1}^{R_k}\gamma_{r_k}W_{pr_k} + \theta'_{pk},$$

where γ_{r_k} is the regression coefficient for the r_kth person covariates W_{pr_k} $(r_k = 1, ..., R_k)$ for the dimension θ_{pk}, and θ'_{pk} is the residual for θ_{pk}. Note that each latent variable θ_{pk} can have a different set of person covariates. The person covariates can also enter as the main effects as in (14). The 1PL versions of models (13) to (15) are also available in **flirt**.

7. Differential item functioning analysis

Suppose test takers come from diverse ethnic, socio-economic, or gender groups. Differential item functioning (DIF) occurs if items in a test are more or less difficult to a group of people than to another group of people with the same ability level. In other words, DIF may indicate that items are biased against one group, or more generally, function differently. DIF is of critical interest for most measurement situations, in particular for high-stake tests.

DIF can be analyzed as an interaction effect between a person group and an item under investigation. It is important to adjust for possible distributional differences between groups to interpret the interaction effect as DIF.

Person-by-item covariates can be created as interaction variables between person groups and item indicators. Adjustment for distributional differences between groups can be done by allowing for (i.e., estimating) different means and variances for different groups as in multiple group analysis.

Suppose we are suspect of item f for DIF between group 1 and 2 with group 1 as the reference group and group 2 as the focal group. Model (4) can be written for DIF analysis

$$\text{logit}(P(y_{pi}=1|\theta_p)) = \sum_{d=1}^{I}\beta_d X_{id} + \gamma_\beta G_p X_{if} + \alpha_i\theta_{pg}, \tag{16}$$

where γ_β is the regression coefficient for the interaction variable $(G_p X_{if})$ between person group

indicator variable G_p and item indicator variable X_{if} where $G_p X_{if} = 1$ if p belongs to group 2 and $i = f$, otherwise 0. The latent variable for group g has a distribution $\theta_{pg} \sim N(\mu_g, \sigma_g^2)$ for group 2 (focal group), and for group 1 (reference group), $\theta_{p(g=1)} \sim N(0, 1)$. For 1pl model families, $\theta_{p(g=1)} \sim N(0, \sigma_{g=1}^2)$

In model (16), γ_β represents uniform DIF for the item intercept parameter β_i. Non-uniform DIF can also be analyzed by incorporating a regression effect of $G_p X_{id}$ for α_i as well as for β_i

$$\text{logit}(P(y_{pi} = 1|\theta_p)) = \sum_{d=1}^{I} \beta_d X_{id} + \gamma_\beta G_p X_{if} + (\alpha_i + \gamma_\alpha G_p X_{if})\theta_{pg}, \tag{17}$$

where γ_β represents the DIF for the intercept parameter β_i for item f and γ_α represents the DIF for the loading parameter α_i for item f. The package **flirt** allows both for uniform and non-uniform DIF analyses for all models (1PL and 2PL versions of the uni- and multi-dimensional models and bifactor models).

8. Polytomous responses

In the generalized linear and nonlinear mixed model framework, polytomous item responses can be handled by forming logits for polytomous data. Given $J_i + 1$ response categories $(0, ..., J_i)$ for item i, J_i nonredundant logits can be formulated using different logit functions. **flirt** provides two link functions:

1) Adjacent-category logits

The adjacent-category logit function (Mellenbergh, 1995) contrasts each category j with adjacent category $(j - 1)$ or $(j + 1)$. Using the $(j - 1)$th category as the adjacent category, the adjacent-category link function becomes $\log\left(\frac{\pi_{pij}}{\pi_{pi(j-1)}}\right)$. With unknown loading parameters, this leads to the generalized partial credit model (Muraki, 1992) that can be expressed as

$$P(y_{pi} = j|\theta_p) = \frac{\exp \sum_{m=0}^{j} (\alpha_i \theta_p + \beta_{ij})}{\sum_{r=0}^{J} \exp \sum_{m=0}^{r} (\alpha_i \theta_p + \beta_{ij})},$$

where β_{ij} is the step parameter for category j of item i, and $\sum_{r=0}^{0}(\alpha_i \theta_p + \beta_{ij}) \equiv 0$. With known

loading parameters, the partial credit model (Masters, 1982) is obtained. The rating scale model (Andrich, 1978) is obtained as a special case of the partial credit model with equal category parameters across items. For example, a 2PL version of the rating scale model can be written as

$$P(y_{pi} = j|\theta_p) = \frac{\exp \sum_{m=0}^{j} (\alpha_i \theta_p + \beta_i + \delta_j)}{\sum_{r=0}^{J} \exp \sum_{m=0}^{r} (\alpha_i \theta_p + \beta_i + \delta_j)},$$

where β_i is the intercept for item i and δ_j is the step parameter for category j. And $\sum_{r=0}^{0}(\alpha_i \theta_p + \beta_i + \delta_j) \equiv 0$.

2) Cumulative logits

For each category j, the cumulative logit is defined as the logit of category j or higher. The cumulative logit function is written as $\log\left(\frac{\pi_{pi(j^+)}}{1-\pi_{pi(j^+)}}\right)$, where $\pi_{pi(j^+)}$ is the probability for responding in category j or higher. This link function leads to the graded response model (Samejima, 1969) with unknown loading parameters

$$P(y_{pi} \geq j|\theta_p) = \frac{\exp(\alpha_i \theta_p + \beta_{ij})}{1 + \exp(\alpha_i \theta_p + \beta_{ij})},$$

where β_{ij} is the step parameter for category j of item i. The category probabilities are obtained by subtracting the conditional probability for responding in category greater than j, i.e., $P(y_{pi} = j|\theta_p) = P(y_{pi} \geq j|\theta_p) - P(y_{pi} > j|\theta_p)$.

There are other link functions such as the multinomial (or baseline category) logit and continuation-ratio logit functions, which can lead to the nominal response model (Bock, 1972) and the sequential response model (Tutz, 1990), respectively. For binary responses, **flirt** uses the multinomial logit link that leads to a regular logit link. The current version of **flirt** uses adjacent and cumulative link functions for polytomous responses, but with the MATLAB code **BNLflirt** (Rijmen & Jeon, 2013), the baseline category and continuation-ratio link functions are available as well as the adjacent and cumulative link functions.

Note that a variety of other IRT models can be constructed by incorporating more than one modeling option that is discussed above. In other words, users can develop different types of

models by simply combining several modeling options within **flirt**. For example, an extension of the generalized partial credit model can be developed with item covariates, person covariates, and multiple groups.

3 Estimation

The R package **flirt** uses a modified EM algorithm (Lauritzen, 1995; Rijmen et al., 2008) for ML estimation. The E-step of the modified algorithm is based on a graphical model framework: First, based on the conditional independent relationship between variables, an initial graphical representation of the statistical model is obtained. A junction tree is then constructed based on the graph, where the junction tree provides a sequence of low-dimensional latent subspaces for efficient computation during the E-step. The M-step proceeds in the same way as the traditional M-step to update parameter estimates. For details on the algorithm, see e.g., Rijmen et al. (2008); Rijmen (2009); Jeon et al. (2013); Rijmen et al. (2014).

The gain of the efficient E-step can be considerable for multidimensional models for which high-dimensional numerical integration is required over the joint space of all latent variables, which can be computationally very demanding. The modified E-step replaces the numerical integration over the joint latent space by a sequence of integrations over smaller subsets of (i.e. low dimensional) latent variables. For example, for a three-dimensional 1PL model with 108 items (36 items per dimension) and 1,069 subjects, the Laplace approximation with the R package **lme4** (Bates et al., 2014) took nearly four hours (14,264 seconds) whereas **flirt** took 809 seconds on a Intel Pentium Dual-Core 2.5-GHz processor computer with 3.2 GB of memory. For additional computational efficiency, **flirt** adopts adaptive quadrature for numerical integration which is more precise and requires fewer quadrature points than ordinary Gaussian quadrature (Pinheiro & Bates, 1995; Rabe-Hesketh et al., 2005).

The package **flirt** uses two methods to obtain standard errors depending on how the observed information matrix is approximated: by the empirical information matrix (Meilijson, 1989) or by a numerical differentiation of the score function (which is routinely obtained in the M-step of the EM

algorithm).

The package **flirt** is written in the R environment (R Development Core Team, 2013) and available in R ($\geq$ 2.13) for Windows 32/64 bit operating systems. Mac and Linux versions currently work in progress. For estimation, **flirt** relies on the Matlab code, **BNLflirt** (Rijmen & Jeon, 2013) that employs sub-functions from the Matlab toolbox BNL (Rijmen, 2006). **flirt** requires the Matlab Compiler Runtime (MCR) for Matlab 2014a, Windows 32/64 bit. The MCR can be freely downloadable on the website http://www.mathworks.com/products/compiler/mcr/ or by contacting the first author.

4 Illustrations

To illustrate the R package **flirt**, we utilized the verbal aggression data (Vansteelandt, 2000; De Boeck & Wilson, 2004). The data consist of responses from 316 first-year psychology students (73 men and 243 women) on 24 items. Each item is a combination of one of four scenarios or situations (e.g., "A bus fails to stop for me"), nested within two situation types (self-to-blame vs. other-to-blame), three behavior types (cursing, scolding, and shouting), and two behavior modes (doing vs. wanting). For each combination, students were asked whether they were likely to exhibit the behavior that was described in each item. The original responses include three categories, No (0), Perhaps (1), and Yes (2). Binary responses were also created by combining Perhaps with Yes categories: No (0) and Perhaps and Yes (1).

We used four item covariates: 1) Do (vs. Want) 2) Other-to-blame (vs. Self-to-blame) 3) Blame (Curse, Scold vs. Shout), and 4) Express (Scold vs. Curse, Shout) and one person covariate: Gender (male(1) vs. female(0)). Gender was also used to define person group membership.

4.1 Data preparation

To specify IRT models using various modeling options in **flirt**, the following data preparations are needed.

1. Item response data matrix Y

$N \times I$ item response data matrix Y in wide form where N persons are placed in rows and I item responses in columns. Both for binary and polytomous responses, the lowest response category should be 0 (users should recode the data otherwise). Missing responses are treated as ignorable (i.e., missing completely at random (MCAR) or missing at random (MAR)). Users can adjust a minimum percentage for category collapsing (default is 0%) such that categories with responses lower than the specified minimum percentage are collapsed within lower adjacent categories.

2. Explanatory models with person design matrix W (if used)

   $N \times R$ person design matrix W for N persons in rows and R person covariates in columns. The persons should be placed in the same order as the persons (rows) in the item response matrix Y. Cases with missing values in W are deleted (listwise deletion) as well as in the item response matrix Y.

3. Explanatory models with item design matrix X (if used)

   $I \times Q$ item design matrix X for I items in rows and Q item covariates in columns. The items should be ordered in the same order as the items (columns) in the item response matrix Y.

4. Multiple-group analysis with person group matrix G (if used)

   N dimensional column vector or $N \times 1$ person group matrix G for N persons in rows and 1 person group variable in column. The person group membership variable should take consecutive integers from 0 and N persons in rows should be ordered by the group membership. It is important to place persons in the item response matrix Y and the person design matrix W (if used) in the same order as the persons in the G matrix. Cases with missing values in G are deleted (listwise deletion), as well as in the item response matrix Y and the person design matrix W (if used).

5. Frequency or sampling weights with person weight matrix V (if used)

   N dimensional column vector or $N \times 1$ person weight matrix V for N persons in rows and 1 person group variable in column. Using frequency weights can be useful for reducing the size

of item response data and reducing computation time. It is important to place persons in the item response matrix Y and the person design matrix W (if used) in the same order as the persons in the V matrix. Cases with missing values in V are deleted (listwise deletion), as well as in the item response matrix Y and the person design matrix W (if used).

All these data can be prepared as a matrix or a data frame.

4.2 Features

The item response, person design, item design, and person group matrices were created for the verbal aggression data and included in the R package **flirt**. We begin by loading **flirt** and the 316×24 binary item response data matrix in the R console.

```
R> library("flirt")
R> data("verb2")
```

flirt includes one major fitting function `flirt` whose arguments consist of the following three components:

1. Data options

   Item response data are specified with the `data` option. A subset of persons (rows) or items (columns) can be chosen with the `subset` and `select` options, respectively. With `subset` being used, the sizes of person design, person group, and weight matrices (if used) should be adjusted accordingly. With `select` being used, the size of the item design matrix (if used) should be adjusted.

2. Modeling options

   Models discussed in Section 2.2 can be specified as the following modeling options of **flirt**.

   - `loading`: 2PL models with two parameterizations 1) item-intercept and 2) item-difficulty parameterizations
   - `mg`: Multiple-group analysis

- `mul`: Multidimensional models
- `bifac`: Bifactor models
- `person_cov`: Explanatory models with person covariates
- `item_cov`: Explanatory models with item covariates
- `dif`: Differential functioning analysis

All modeling options include a list of sub-options with the `on` option. When any one of the modeling options is used, the corresponding `on` option should be `TRUE`.

3. Practical options

flirt provides several useful options such as the prediction of latent variables (or factor scoring) and IRT reliabilities. For polytomous responses, a different type of link function (discussed in Section 2.2) can be chosen in the `control` option. These and other practical options are summarized below.

- `post`: Prediction of latent variables (posterior means (expected a posteriori; EAP) and covariances), expected sum-scores, and IRT reliabilities that are computed based on a formulation of Haberman & Sinharay (2010).
- `weight`: Person-level frequency or sampling weight
- `start`: Starting values
- `constraint`: Constrain model parameters at fixed values
- `evaluate`: Evaluate the log-likelihood at given values of model parameters
- `control`: Several estimation options including
  - `link`: Link functions (1: Multinomial, 2: Cumulative, 3: Adjacent logits)
  - `adapt`: Adaptive quadrature
  - `nq`: Number of quadrature points
  - `conv`: Convergence criterion
  - `max_it`: Number of maximum iterations

- `minpercent`: Minimum category percentage (for polytomous responses)

- `se_num`: Standard errors using the numerical information matrix

- `se_emp`: Standard errors using the empirical information matrix

- `alp_bounds`: Maximum boundary value for the discrimination parameters

- `verbose`: Print estimation process

- `show`: Print messages from **BNLflirt**

4.3 Model specifications

The only required argument for running **flirt** is `data`. For all default options, the 1PL model is estimated for the specified item response data.

```
R> model1 <- flirt(data=verb2)
```

The 2PL model is specified by using the `loading` option with a choice of parameterization. With the item-intercept parameterization (5), the `inside` option should be `FALSE` (default).

```
R> model2 <- flirt(data=verb2, loading=list(on=TRUE, inside=FALSE) )
```

For the item-difficulty parameterization (6), the `inside` option needs to be `TRUE`.

```
R> model3 <- flirt(data=verb2, loading=list(on=TRUE, inside=TRUE) )
```

Person and item covariates can be incorporated using the `person_cov` and `item_cov` options. First, load the person design matrix that includes the person covariate gender.

```
R> data("person_design")
R> head(person_design)
     male
[1,]    0
[2,]    0
[3,]    0
[4,]    0
[5,]    0
[6,]    0
```

The first six people are all females. Then load the item design matrix.

```
R> data("item_design_bin")
R> head(item_design_bin)
  intercept blame express dowant otherself
1         1   0.5     0.5      0         1
2         1   0.5     0.5      0         1
3         1   0.5     0.5      0         0
4         1   0.5     0.5      0         0
5         1   0.5     0.5      1         1
6         1   0.5     0.5      1         1
```

The item design matrix has five columns including the intercept in the first column. Each row represents an item. Now, the 2PL explanatory model with person and item covariates can be specified as

```
R> model4 <- flirt(data=verb2, loading=list(on=TRUE, inside=FALSE),
 +             person_cov=list(on=TRUE, person_matrix=person_design),
 +             item_cov=list(on=TRUE, item_matrix_beta=item_design_bin) )
```

The person covariates (person design matrix, `person_design`) and the item covariates (item design matrix, `item_design_bin` for the item intercept parameters) are specified using the `person_matrix` and `item_matrix_beta` sub-options in the `person_cov` and `item_cov` options, respectively.

Multiple dimensions can be specified using the `mul` option. For example, with a two dimensional model where the first 12 items belong to dimension 1 and the next 12 items belong to dimension 2, the multidimensional 2PL model can be specified as

```
R> model5 <- flirt(data=verb2, loading=list(on=TRUE, inside=FALSE),
 +             mul=list(on=TRUE, dim_info=list(dim1=1:12, dim2=13:24 ) ) )
```

Information on which items belong to which dimensions should be specified using the `dim_info` sub-option. `dim_info` is a list of items for each dimension and is equal in length to the number of dimensions.

A bifactor structure can be specified using the `bifac` option. For example, with two specific dimensions that include 12 items in each of the specific dimensions in order, the bifactor model can be specified

```
R> model6 <- flirt(data=verb2, loading=list(on=TRUE, inside=FALSE),
 +              bifac=list(on=TRUE, dim_info=list(dim1=1:12,dim2=13:24)) )
```

In the `bifac` option, `dim_info` needs information only for specific dimensions.

For DIF analysis, the `dif` option is used to specify items that are suspect of DIF and the `mg` option is used for multiple-group analysis. For example, to specify non-uniform DIF for the first two items between genders, we use the person covariate, gender, as the person group variable.

```
R> person_group <- person_design
R> model7 <- flirt(data=verb2, loading=list(on=TRUE, inside=FALSE),
 +             dif=list(on=TRUE, dif_beta=c(1,2), dif_alpha=c(1,2) ),
 +             mg=list(on=TRUE, group_matrix=person_group) )
```

The `dif` option includes two sub-options `dif_beta` for DIF for intercepts (β_i) and `dif_alpha` for DIF for loadings (α_i). The `mg` option needs to specify a person group matrix within the `group_matrix` sub-option.

For polytomous item responses, all modeling options can be specified in the same way as binary responses. The only difference is to choose one of the link functions that are discussed in Section 2.2. To illustrate, first load polytomous item responses and an item design matrix for polytomous data.

```
R> data("verb3")
R> data("item_design_pol")
R> head(item_design_pol)
  intercept blame express dowant otherself category
1         1   0.5     0.5      0         1        0
2         1   0.5     0.5      0         1        1
3         1   0.5     0.5      0         1        0
4         1   0.5     0.5      0         1        1
5         1   0.5     0.5      0         0        0
6         1   0.5     0.5      0         0        1
```

The item design matrix `item_design_pol` is 48×6 in size. Each item is repeated twice (i.e., two adjacent lines represent one item), one for category 1 and another for category 2. In addition to the fist five item covariates, an additional column is included (column 6) in order to estimate a category effect that is assumed to be equal across all items. For differential category effects, 24 additional columns (one for each item) are needed instead of one column.

Now specify a multidimensional 1PL model with item covariates (for item intercept parameters) for multiple groups with an adjacent logit link.

```
R> model8 <- flirt(data=verb3,
 +            mul=list(on=TRUE, dim_info=list(dim1=1:12, dim2=13:24)),
 +            item_cov=list(on=TRUE, item_matrix_beta=item_design_pol),
 +            mg=list(on=TRUE, group_matrix=person_group),
 +            control=list(link="adjacent") )
```

See how the link function is chosen in the control option. Specification of other options is the same as for binary responses.

4.4 Output options

The `flirt` fitting function returns an object of class `flirtClass`, for which several methods are available to examine parameters estimates, standard errors, and model fit statistic. These methods are summarized as follows:

- `summary`: Print a long summary of estimation results
- `show`: Print a short summary of estimation results
- `coef`: Print a table of parameter estimates and standard errors
- `logLik`: Returns the log-likelihood of the fitted model
- `anova`: Returns likelihood ratio tests between nested models

For example, estimation results of `model5` can be summarized using `summary`.

```
R> summary(model5)
Estimation of Multidimensional 2PL Model Family

Data:
  nobs  nitem maxcat ngroup
   316     24      2      1

Model fit:
  npar    AIC    BIC loglik
    49   8029   8213  -3965

Parameterization:
```

```
"a*th+b"

Type:

"between-item"

Dimension:
ndim dim1 dim2
   2   12   12

Parameter estimates:
                Est     SE
alp1       1.908787 0.3204
alp2       2.049491 0.3268
alp3       1.380444 0.2191
alp4       1.993594 0.3578
...
bet1      -1.465713 0.2379
bet2      -0.745050 0.2056
bet3      -0.106366 0.1540
bet4      -2.117973 0.3001
...
mg_mean11  0.000000     NA
mg_mean21  0.000000     NA
mg_sd11    1.000000     NA
mg_sd21    1.556884     NA
mg_cov11   1.193268     NA

(some results are abbreviated with ellipsis (...))
```

The `summary` function prints the title of the estimation on the top, and gives a summary of data (number of observations, number of items, maximum number of categories, and number of groups), model fit (number of parameters, AIC, BIC, and log-likelihood), parameterization, and dimensions. A matrix of parameter estimates and standard errors are then provided, where `alp1` to `alp24` are the loading parameter estimates, `bet1` to `bet24` are the intercept parameter estimates, `mg_mean11` and `mg_mean12` are the estimated means for dimensions 1 and 2 for group 1, and `mg_sd11`, `mg_sd11`, and `mg_cov11` are the estimated standard deviations and covariance between dimensions 1 and 2, respectively, for group 1. The standard errors for the variance parameters are not provided (NA) because the use of standard errors for Wald-type tests and confidence intervals are not appropriate

for these parameters (e.g., Stram & Lee, 1994). The standard errors for the means are not provided since they are fixed to 0 and not estimated in this model.

Note that since `mg_sd11`, `mg_sd11`, and `mg_cov11` are unstandardized estimates, the loading estimates (`alp1` to `alp24`) are also unstandardized. **flirt** provides two functions `std_coef` and `std_cov` for computing standardized covariance matrix (correlation matrix) and standardized loading estimates.

Here we illustrate how to use the `std_coef` and `std_cov` functions. What is required is the estimated loading parameters and covariance matrix that can be extracted from the `pars` slot of `flitClass`

```
R> est_alp <- model5@pars[1:24,1]
R> est_cov <- model5@pars[51:53,1]
R> cov_matrix <- matrix(c(est_cov[1]^2,est_cov[3],est_cov[3],
    est_cov[2]^2),2,2, byrow=F)
R> cov_matrix
         [,1]     [,2]
[1,] 1.000000 1.193268
[2,] 1.193268 2.423889
```

We then can obtain standardized loading estimates and correlation matrix using `std_coef`

```
R> std_alp <- std_coef(est=est_alp, dim_info=list(dim1=1:12, dim2=13:24) ,
+  cov_matrix= cov_matrix)
```

`std_coef` returns standardized loading estimates in `std_est` and correlation matrix in `cor_mat`.

```
R> std_alp
$std_est
 [1] 1.9087865 2.0494913 1.3804436 1.9935944 2.2181836
 [6] 1.2882939 1.1993109 1.5567419 0.9062966 1.2823367
[11] 1.5971148 1.0024666 2.0500335 2.7648832 1.4576887
[16] 1.9029332 2.3368675 1.6475668 1.2071455 1.4710054
[21] 1.1968724 1.6201829 1.7427217 1.1983337

$cor_mat
          [,1]      [,2]
[1,] 1.0000000 0.7664462
[2,] 0.7664462 1.0000000
```

The `std_cov` function simply returns the correlation matrix.

```
R> corr_mat <- std_cov(dim_info=list(dim1=1:12, dim2=13:24) ,
+  cov_matrix= cov_matrix)
R> corr_mat
          [,1]      [,2]
[1,] 1.0000000 0.7664462
[2,] 0.7664462 1.0000000
```

To obtain standard errors for the standardized loading estimates, the delta method can be used. For example, the delta method can be implemented using the `deltamethod` function in the **msm** package (Jackson, 2011). The estimated covariance matrix can be obtained from the information matrix (evaluated at the parameter estimates) that is stored in the `info_num` and `info_emp` slots of `flirtClass`.

Finally, we illustrate the `post` option of **flirt** that provides EAP estimates with standard errors, expected scores, and IRT reliability. For example, `model1` can be rerun with the `post` option

```
R> model1 <- flirt(data=verb2 , post=TRUE )
```

The `post` slot of the `flirtClass` includes EAP (posterior mean) estimates, posterior variance (squared standard errors), expected scores, and IRT reliability. They can be accessed as follows

```
R> exp_score <- model1@post$exp
R> eap <- model1@post$eap
R> postvar <- model1@post$eap_var
R> irt_rel <- model1@post$rel
R> irt_rel
[1] 0.8823181
```

To illustrate, we compare the observed scores (computed as `rowSums` of the observed data) with expected scores and EAP estimates. Figure 1 shows the results.

[Figure 1 about here]

The expected scores are quite close to the observed scores but show some shrinkage towards the overall mean (a slight gap at the lower and higher score levels). The shrinkage towards the overall mean is expected because the expected scores are empirical Bayes estimates. The EAP estimates show a similar pattern. The correlation between the expected scores and the observed scores is 0.994, and the correlation between the EAP estimates and the observed scores is 0.999.

5 Concluding remarks

A new IRT package **flirt** has been introduced in this paper. A major attraction of **flirt** is its convenient modular approach with which users can easily build, explore, and estimate a multitude of IRT models using various modeling options. Another and unique strength of **flirt** is its ability to incorporate explanatory features in traditional measurement models. Thus, with **flirt**, different hypotheses on item and person parameters can be modeled and tested.

The R package **flirt** includes several useful features that were little discussed in this paper. For instance, **flirt** allows the user to specify different person covariates in different dimensions in the multidimensional and bifactor family models. Suppose in a two-dimensional IRT model with mathematics and science test items, one is interested in finding out the effects of students' socio-economic-status (SES) on science abilities but also the effects of gender and private tutoring experience for science abilities. **flirt** can easily specify SES for the mathematics dimension and gender and private tutoring for the science dimension as person covariates.

There are some types of IRT models that are not currently available in **flirt**; for example, IRT models that include guessing parameters, such as the three parameter logistic model (Birnbaum, 1968), models with discrete mixtures (e.g., Rost, 1990), models with random coefficients of item predictors (e.g., Rijmen & De Boeck, 2002), models with hierarchical structures on the person side (e.g., multilevel IRT models) and on the item sides (e.g., higher-order IRT models), or models with random item parameters (e.g., De Boeck, 2008). A future **flirt** release may include some of the modeling options that are listed above.

Acknowledgements

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R305D110027 to Educational Testing Service. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variable regression. *Journal of Educational and Behavioral Statistics*, *22*, 47–76.

Adams, R. J., Wu, M. L., & Wilson, M. (2012). *ACER ConQuest 3.01: Generalized Item Response Modelling Software [Computer software and manual]*. Melbourne, Victoria, Australia: Australian Council for Educational Research.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-6. URL http://CRAN.R-project.org/package=lme4.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinees's ability. In F. M. Lord, & M. R. Novick (Eds.) *Statistical theories of mental test scores*, (pp. 397–479). Reading, MA: Addison-Wesley.

Bock, R., & Zimowski, M. (1997). Multiple group IRT. In W. van der Linden, & R. Hambleton (Eds.) *Handbook of modern item response theory*, (pp. 433–448). New York: Springer-Verlag.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51.

Cai, L. (2012). *flexMIRT TM version 1.86: A numerical engine for multilevel item factor analysis and test scoring. [Computer software]*. Seattle, WA: Vector Psychometric Group.

Cai, L., du Toit, S. H. C., & Thissen, D. (2011). *IRTPRO: Flexible, Multidimensional, Multiple Categorical IRT Modeling*. Lincolnwood, IL: Scientific Software International.

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*, 1–29.

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

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*, 1–28.

De Boeck, P., & Wilson, M. (2004). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York: Springer-Verlag.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179–197.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*, 495–515.

Fischer, G. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374.

Fox, J. P. (2007). Multilevel IRT modeling in practice with the package mlirt. *Journal of Statistical Software*, *20*, 1–16.

Haberman, S., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika*, *75*, 209–227.

Jackson, C. H. (2011). Multi-state models for panel data: The msm package for R. *Journal of Statistical Software*, *38*, 1–29.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2013). Modeling differential item functioning using a generalization of the multiple-group bifactor model. *Journal of Educational and Behavioral Statistics*, *38*, 32–60.

Kelderman, H., & Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika*, *56*, 149–176.

Kiefer, T., Robitzsch, A., & Wu, M. (2014). *TAM: Test Analysis Modules*. R package version 1.0-3.18-1. URL http://CRAN.R-project.org/package=TAM.

Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, *19*, 191–201.

Mair, P., Hatzinger, R., & M.J., M. (2014). *eRm: Extended Rasch Modeling*. R package version 0.15-4. URL http://erm.r-forge.r-project.org/.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

McCullagh, P., & Nelder, J. (1989). *Generalized Linear Models*. New York: Chapman and Hall.

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society, Series B*, *51*, 127–138.

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, *19*, 91–100.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Muthén, L., & Muthén, B. (2012). *Mplus Version 7 User's Guide*. Angeles, CA: Muthen & Muthen.

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

R Development Core Team (2013). *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/.

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

Rijmen, F. (2006). BNL: A Matlab toolbox for Bayesian networks with logistic regression. Tech. rep., Vrije Universiteit Medical Center, Amsterdam.

Rijmen, F. (2009). An efficient EM algorithm for multidimensional IRT models: Full information maximum likelihood estimation in limited time. ETS Research Report (RR0903).

Rijmen, F., & De Boeck, P. (2002). The random weights linear logistic test model. *Applied Psychological Measurement*, *26*, 269–283.

Rijmen, F., & Jeon, M. (2013). *BNLflirt*. Matlab code exchange. http://faculty.psy.ohio-state.edu/jeon/lab/flirt.php.

Rijmen, F., Jeon, M., Rabe-Hesketh, S., & Matthias, V. (2014). A third order item response theory model for modeling the effects of domains and subdomains in large-scale educational assessment surveys. *Journal of Educational and Behavioral Statistics*, *39*, 235–256.

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

Rijmen, F., Vansteelandt, K., & De Boeck, P. (2008). Latent class models for diary method data: Parameter estimation by local computations. *Psychometrika*, *73*, 167–182.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and Item Response Theory analyses. *Journal of Statistical Software*, *17*, 1–25.

Robitzsch, A. (2013). *sirt: Supplementary Item Response Theory Models R package version 0.36-30*. Http://CRAN.R-project.org/package=sirt.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271–282.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, *34*, 100–114.

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

Thissen, D. (1991). *MULTILOG [Software manual]*. Lincolnwood, IL: Scientific Software.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*, 39–55.

Vansteelandt, K. (2000). Formal models for contextualized personality psychology. Unpublished doctoral dissertation, K.U. Leuven, Belgium.

Wolfinger, R. D. (2008). Fitting non-linear mixed models with the new NLMIXED procedure. Tech. rep., SAS Institute, Cary, NC.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2006). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items [Computer software]*. Chicago: Scientific Software International.

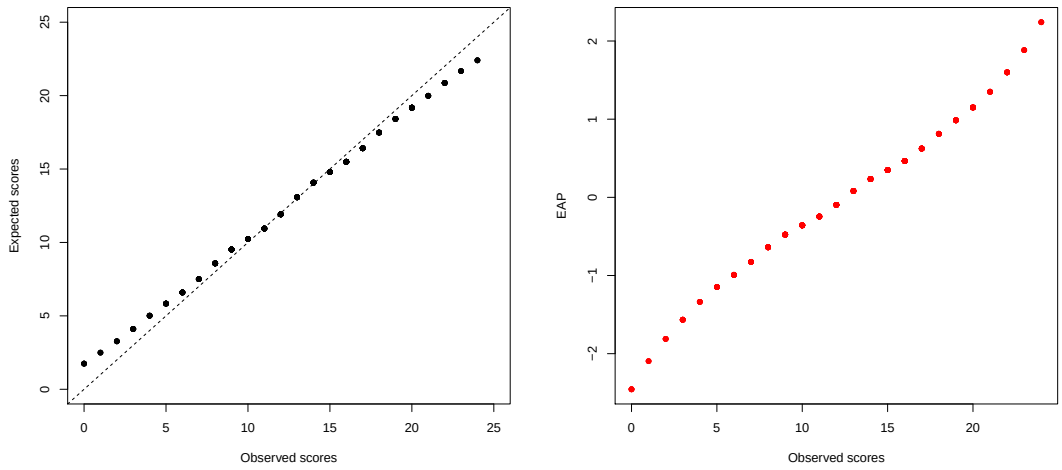

Figure 1: Expected scores compared with observed scores (left) and EAP estimates compared with observed scores (right).