Behav Res (2016) 48:742–755
DOI 10.3758/s13428-015-0606-z

A modular approach for item response theory modeling with the R package flirt

Minjeong Jeon[1] · **Frank Rijmen**[2]

Published online: 15 July 2015

Abstract The new R package **flirt** is introduced for flexible 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, such as parametric forms, number of dimensions, item and person covariates, person groups, link functions, etc. In this paper, we describe major features of **flirt** and provide examples to illustrate how **flirt** works in practice.

Keywords Modular approach · R software · Item response theory · Explanatory models · Multidimensional models · Bifactor models · DIF

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, structural equation modeling, or generalized linear mixed modeling software packages 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 recent years, a multitude of free IRT packages have been developed in the R environment (R Development Core Team, 2013), such as **eRm** (Mair et al., 2014) and **plRasch** (Li & Hong, 2007) for Rasch family models, **ltm** (Rizopoulos, 2006) for general parametric models, **mlirt** (Fox, 2007) for multilevel item response models, **mirt** (Chalmers, 2012) for uni- and multidimensional parametric models. **TAM** (Kiefer et al., 2014) and **sirt** (Robitzsch, 2013) offer estimation of special-purpose models as well as standard models using various existing functions in R. Reviews of some of the present R packages are provided in Section "Current R packages for item response analysis".

A major limitation of the current commercial and non-commercial IRT software packages is that they typically provide a set of pre-defined models that researchers need to choose from; that is, researchers have less freedom to build their own customized models. In addition, most software packages focus on descriptive IRT models, which are unsuitable for explaining sources for item properties as well as individual differences. Such questions can be answered by utilizing covariates on the item and person sides, which leads to explanatory IRT models. Explanatory models can also be useful for investigating construct validity (Embretson, 1983) and modifying distributional assumptions on the latent variables (Bock & Zimowski, 1997). General-purpose software packages, such as Mplus, SAS

✉ Minjeong Jeon
jeon.117@osu.edu

[1] Department of Psychology, Faculty of Quantitative Psychology, The Ohio State University, 228 Lazenby Hall 1827 Neil Avenue, Columbus, OH 43210 USA

[2] CTB/McGraw-Hill, New York, USA

nlmixed, gllamm, and **lme4** R package (Bates et al., 2014) can be used to fit (explanatory) IRT models, but with these general packages, IRT specification and output interpretation are not always straightforward.

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 modeling modules, such as the 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, which is more accurate than the ordinary Gauss–Hermite quadrature method.

The structure of this article is organized as follows: In the "Preparation for fitting IRT models with flirt" section, we introduce the package installation, data set-up, and its major fitting function. In the "Fitting IRT models with flirt" section, we illustrate how a variety of IRT models can be specified with **flirt** using an empirical data example. In the "Statistical framework" section, we describe the general statistical modeling and estimation frameworks that **flirt** is based on. In the "Current R packages for item response analysis" section, we review some of the current R packages for item response modeling. In the "Final remarks" section, we end with some discussions on **flirt** contributions, limitations, and future development.

Preparation for fitting IRT models with flirt

Installation

The package **flirt** requires a MATLAB Compiler Runtime (MCR).[1] After downloading the package source file from http://faculty.psy.ohio-state.edu/jeon/lab/flirt.php or by contacting the first author, **flirt** for 64-bit Windows computers can be installed and loaded in an R console as follows:[2]

```
R> install.packages("flirt_1.15.tar.gz", type="source", repos=NULL)
R> library(flirt)
```

Data set-up

Mandatory, input-item response data should be organized in wide form (subjects placed in rows and items placed in columns). A subset of persons and/or a subset of items can be selected using `subset` and `select` options.

To utilize person covariates in a model, users have two options: (1) include the person covariates in the input item response data, or (2) prepare external variables for the person covariates. With option 1, users can utilize the column numbers that correspond to the person covariates in the input data. With option 2, users need to prepare an external variable matrix that is organized in wide form (persons in rows and covariates in columns). To utilize item covariates, option 1 is unavailable because item covariates cannot be placed at the same time with regular item responses in the wide-form input data. Therefore, users need to prepare an external item design matrix (items in rows and covariates in columns). Use of item and person covariates will be further explained in Sub-sections "Person covariates" and "Item covariates".

Main function `flirt`

The package **flirt** includes one major fitting function `flirt`. The next R code illustrates how the arguments of the function `flirt` are organized for fitting a two-parameter logistic (2PL) multiple-group model for non-uniform DIF analysis.

```
R> model0 <- flirt(data =Data, loading=list(on=TRUE),
 +            mg=list(on=TRUE, group_maxtrix=race),
 +            dif=list(on=TRUE, dif_beta=c(1,2), dif_alpha=c(1,2)))
```

A `flirt` specification is based on three principles: (1) a model is built up by adding the modeling modules that are desired to be used (e.g., `mg`, `dif`), (2) each modeling

[1] The MCR can be freely downloaded from http://www.mathworks.com/products/compiler/mcr/

[2] For 32-bit Windows machines, the file *flirt.x32_1.15.tar.gz* should be used.

module is specified by 'turning on' its switch (`on=TRUE`),and (3) with no modules specified, a standard one-parameter logistic model is fitted.

In Section "Fitting IRT models with flirt", we illustrate in detail how a variety of IRT models can be constructed through combinations of **flirt**'s modeling modules.

Fitting IRT models with flirt

To illustrate fitting IRT models with **flirt**, we utilize 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) to 24 items. Each item describes a scenario characterized by a combination of one of each from: two situation types (self-to-blame vs. others-to-blame), three behavior types (cursing, scolding, and shouting), and two behavior modes (doing vs. wanting). For each scenario, 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 created by combining Perhaps with Yes categories: No (0), Perhaps (0), and Yes (1). Four-item covariates were used: (1) Do (vs. Want) (2) Others-to-blame (vs. Self-to-blame) (3) Blame (Curse, Scold vs. Shout), and (4) Express (Scold vs. Curse, Shout) as well as one person covariate: Male (vs. Female)

We begin by describing a one-parameter logistic (1PL) model and a two-parameter logistic (2PL) model for binary items. Various extensions of the 2PL model are illustrated with person covariates, item covariates, multiple groups, multidimensionality, and bifactor structures. Extensions to polytomous items are described at the end of this section.

1PL model

Let us denote binary response y_{pi} for person p $(= 1, ..., N)$ to item i $(= 1, ..., I)$. The 1PL model is then formulated as follows:

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

where β_i is the item intercept (or location) parameter that represents the mean of item i in the logit scale given the latent trait θ_p (or ability, proficiency) for person p. The latent trait θ_p is assumed to follow a normal distribution, with $\theta_p \sim N\left(0, \sigma^2\right)$. Typically, β_i is referred to as the item 'easiness' parameter (therefore, $-\beta_i$ denotes item difficulty) because when $P\left(y_{pi}=1|\theta_p\right) = 0.5$, we have $\theta_p = -\beta_i$, meaning that the probability of success can be predicted by comparing θ_p and $-\beta_i$ on the same scale.

With **flirt**, the 1PL model can be specified as follows:

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

In the first line, `verb2` is loaded, which is the 316×24 size verbal aggression response data matrix; in the second line, the data are specified with `data=verb2` in the `flirt` code.

The `flirt` object `model1` returns a short summary of the estimation:

```
R> model1
Estimation of Unidimensional 1PL Model Family
using 20 quadrature points, convergence = 1e-04 and total iterations= 35
with logit link function

Log-likelihood = -4036.263 with npar= 25
AIC = 8122.527
BIC = 8216.42
```

The output includes the estimation information (the number of quadrature points, convergence criterion, the total number of iterations, and the link function) as well as the model fit information (the log-likelihood, the number of estimated parameters, AIC, and BIC). Utilizing the `summary` function gives a detailed summary of the

estimated result, including data information, model fit, parameter estimates, and standard errors.[3]

```
> summary(model1)
Estimation of Unidimensional 1PL Model Family

Data:
  nobs  nitem maxcat ngroup
   316     24      2      1

Model fit:
  npar    AIC    BIC loglik
    25   8123   8216  -4036

Parameter estimates:
            Est     SE
th_sd  1.405673 0.0749
bet1  -1.227176 0.1604
bet2  -0.570722 0.1516
...
bet23  0.378194 0.1506
bet24  1.995842 0.1827
```

In the parameter estimates, `th_sd` indicates the standard deviation of the latent trait (σ) and `bet1` to `bet24` correspond to the item easiness parameters for items 1 to 24 (β_i) in Eq. 1. For more information on the structure and content of the output, we refer readers to the package manual (Jeon et al., 2015) and the vignette that are found on http://faculty.psy.ohio-state.edu/jeon/lab/flirt.php.[4]

2PL model

The 2PL model that includes the item discrimination parameters as well as the item intercept parameters can be formulated as follows:

$$\text{logit}\left(P(y_{pi}=1|\theta_p)\right)=\alpha_i\theta_p+\beta_i, \tag{2}$$

where α_i is the item discrimination (or slope) parameter and β_i is the item intercept (or location) parameter for item i. Equation 2 *item-intercept* parameterization can be formulated with an *item-easiness* parameterization as follows:

$$\text{logit}(P(y_{pi}=1|\theta_p))=\alpha_i\left(\theta_p+\beta_i^*\right), \tag{3}$$

Here β_i^* is referred to as the item easiness parameter (therefore, $-\beta_i^*$ denotes item difficulty). It can be shown that $\beta_i^*=\beta_i/\alpha_i$, where β_i is the item intercept parameter from Eq. 2.

With **flirt**, the 2PL model can be specified using the `loading` option, which, by default, is in the *item-intercept* parameterization

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

The *item-easiness* parameterization can be specified by utilizing the `inside` sub-option as follows:

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

Person covariates

Suppose R person covariates W_{pr} ($r=1,...,R$) are used to explain sources of individual differences in the latent trait θ_p. Then we specify a regression model for θ_p as follows:

$$\theta_p=\sum_{r=1}^{R}\gamma_r W_{pr}+\zeta_p, \tag{4}$$

where γ_r is the regression coefficient for the rth person covariates W_{pr} ($r=1,...,R$), and ζ_p is the residual for θ_p after being explained by the person covariates, with $\zeta_p\sim N\left(0,\sigma^{*2}\right)$. By plugging the regression model (4) to the 2PL measurement model (2), we obtain

$$\begin{aligned}\text{logit}(P(y_{pi}=1|\zeta_p)) &= \beta_i+\alpha_i\left(\sum_{r=1}^{R}\gamma_r W_{pr}+\zeta_p\right),\\ &= \beta_i+\alpha_i\sum_{r=1}^{R}\gamma_r W_{pr}+\alpha_i\zeta_p. \end{aligned}\tag{5}$$

An IRT model with person covariates is referred to as a latent regression model (Adams et al., 1997). With **flirt**, the 2PL latent regression model (5) can be constructed by using the `person_cov` module as follows:

```
R> data(person_design)
R> male <- person_design
R> model4 <- flirt(data=verb2, loading=list(on=TRUE),
 +            person_cov=list(on=TRUE, person_matrix=male) )
```

[3] Some results have been truncated with ellipsis (...) due to space limitations.

[4] The **flirt** output presents a similar structure for any specified model. Hence we will not include any additional **flirt** outputs.

In the first line, `person_design` is loaded, which includes the person covariate that indicates whether person p is male (or female). In the second line, the`person_design` variable is renamed as `male` and the `male` variable is specified as the person covariate in the latent regression model (`person_matrix=male`). **flirt** allows for more than one continuous (e.g., age, grades) or categorical (e.g., grade, race) covariate.

Note that in Eq. 5, the regression coefficient γ_r is multiplied by the item discrimination parameter α_i. That is, the effects of the covariates W_{pr} are moderated by the value of α_i. We can re-formulate the model such that the effects of the person covariates W_{pr} are estimated as the 'main' effects. That is,

$$\text{logit}(P(y_{pi}=1|\zeta_p)) = \beta_i + \sum_{r=1}^{R} \gamma_r W_{pr} + \alpha_i \zeta_p. \quad (6)$$

Observe that γ_r is no longer moderated (i.e., multiplied) by the item slope parameter α_i. That is, γ_r represents the average effect of the person covariate W_{pr} on the probability of success conditional on β_i and $\alpha_i \zeta_p$.

With **flirt**, the main effects of person covariates can be specified by simply choosing the `main` sub-option as follows:

```
R> model5 <- flirt(data=verb2, loading=list(on=TRUE),
 +            person_cov=list(on=TRUE, person_matrix=male, main=TRUE) )
```

Item covariates

Suppose Q item covariates (or stimulus features) are considered to explain item easiness (or difficulty). We can then specify a regression model for the item easiness parameter β_i as follows:

$$\beta_i = \sum_{q=1}^{Q} \beta_q^* X_{iq}, \quad (7)$$

where β_q^* is the regression coefficient for the qth item covariate X_{iq} ($q = 1, ..., Q$, $Q < I$), representing the effect of X_{iq} on the item-easiness parameters.

By plugging the item regression model (7) to the 2PL measurement model (2), we obtain

$$\text{logit}(P(y_{pi}=1|\theta_p)) = \sum_{q=1}^{Q} \beta_q^* X_{iq} + \alpha_i \theta_p. \quad (8)$$

If X_{i1} is unity ($X_{i1} = 1$ for all i), β_1^* represents the intercept of the item regression model.

The 1PL IRT model that includes item covariates for the item easiness parameters is referred to as the linear logistic test model (LLTM; Fischer, 1973). With **flirt**, the 2PL LLTM model with unknown item discrimination parameters can be specified by adding the `item_cov` option as follows:

```
R> data(item_design_bin)
R> model6 <- flirt(data=verb2, loading=list(on=TRUE),
 +            item_cov=list(on=TRUE, item_matrix_beta=item_design_bin) )
```

Here, `item_design_bin` is the item design matrix that includes five item covariates (intercept, do-want, others/self-to-blame, blame, express). In the `flirt` code, `item_matrix_beta=item_design_bin` means that the item intercept parameters (β_i) are explained by the five variables in `item_design_bin`.

The LLTM model is useful for understanding sources of item difficulty (or easiness). Suppose we also want to explain sources of item discrimination with S item covariates. Then we specify the item regression model for α_i as follows:

$$\alpha_i = \sum_{s=1}^{S} \alpha_s^* Z_{is}, \quad (9)$$

where α_s^* is the regression coefficient for the sth item covariate Z_{is} ($s = 1, ..., S$, $S < I$), representing the effect of Z_{is} on the item slope parameter α_i. If Z_{is} is unity, α_1^* becomes the intercept of the item regression model.

By combining model (9) with model (8), we obtain

$$\text{logit}(P(y_{pi}=1|\theta_p)) = \sum_{q=1}^{Q} \beta_q^* X_{iq} + \sum_{s=1}^{S} \alpha_s^* Z_{is} \theta_p, \quad (10)$$

which can be seen as an extension of the LLTM, as it includes item regression models for both item difficulty and item discrimination. Embretson (1999) referred to this model as the 2PL constrained model in the context of modeling the cognitive complexity of test items. With **flirt**, the 2PL constrained model can be specified as follows:

```
R> model7 <- flirt(data=verb2, loading=list(on=TRUE),
 +            item_cov=list(on=TRUE, item_matrix_beta=item_design_bin,
 +              item_matrix_alpha=item_design_bin) )
```

Notice that `item_design_bin` is specified in `item_matrix_beta` and `item_matrix_alpha`. This means that the same set of five item covariates are used to explain the item intercept parameters (β_i) and the item slope parameters (α_i). **flirt** allows users to utilize a different set of item covariates to explain β_i and α_i.

Multiple-group model

Standard IRT models assume that subjects are random samples of a common population distribution. This assumption can be relaxed by allowing for 'multiple groups' in the population that are characterized by group-specific means and variances. The basic 2PL model can be extended by allowing for multiple population groups as follows:

$$\text{logit}(P(y_{pi} = 1|\theta_p)) = \beta_i + \alpha_i \theta_{pg}, \quad (11)$$

where θ_{pg} represents the latent trait for person p that belongs to group g. The group-specific latent trait θ_{pg} is assumed to follow a normal distribution with $\theta_{pg} \sim N\left(\mu_g, \sigma_g^2\right)$. For a reference group (usually $g = 1$), the mean and variance are fixed to 0 and 1, respectively, while the other focal groups of interest ($g = 2, ..., G$) have group-specific means (μ_g) and variances $\left(\sigma_g^2\right)$ that are freely estimated.

With **flirt**, the multiple-group 2PL model can be specified by using the `mg` module

```
R> model8 <- flirt(data=verb2, loading=list(on=TRUE),
 +             mg=list(on=TRUE, group_matrix=male )
```

Here `group_matrix=male` means that subjects' gender (the variable `male`) is used to identify subjects' group membership (male is the focal group and female is the reference group).

Differential item functioning analysis

Suppose we are interested in examining group differences in the item parameters. Differential item functioning (DIF) is defined when an item is more difficult and/or discriminating to a particular group of people than to other groups of people that have the same ability levels. DIF can be quantified as a model parameter that represents the interaction between a person group membership and an item indicator variable; the estimated DIF parameter means then an additional level of difficulty (or discrimination) that the item shows to the particular group of people. Note that to examine DIF, it is critical to take into account possible distributional differences (in the means and variances) between the person groups (same as a multiple-group model); otherwise, the estimated DIF effects may mainly reflect the distributional differences between the groups, which is referred to as 'impacts' rather than DIF.

Suppose we want to investigate item f for DIF on the item intercept parameter (which is referred to as 'uniform' DIF) between two groups (group 1 is the reference and group 2 is the focal group). The 2PL model for analyzing the uniform DIF for item f can be specified as follows:

$$\text{logit}(P(y_{pi} = 1|\theta_p)) = \beta_i + \gamma_\beta G_{pg} X_{if} + \alpha_i \theta_{pg}, \quad (12)$$

where γ_β represents DIF, which is the regression coefficient for the interaction variable $G_{pg}X_{if}$, where G_{pg} is the person membership variable (1 if person p belong to the focal group ($g = 2$), and 0 otherwise ($g = 1$)) and X_{if} is the indicator variable for item f (1 if $i = f$ and 0 otherwise). Thus, the product of the two variables, $G_{pg}X_{if}$ takes value 1 if person p that belongs to the focal group ($g = 2$) answers item f, else $G_{pg}X_{if} = 0$. Notice that the group-specific latent trait θ_{pg} is specified as $\theta_{pg} \sim N\left(\mu_g, \sigma_g^2\right)$; that is, the differences both in the means and variances between groups are adjusted in modeling DIF. If more than one item is investigated for DIF, additional DIF parameters are needed as the effects of the interaction variables $G_{pg}X_{ih}$, where X_{ih} represents the indicator variable for item h ($h = 1, ..., H$).

With **flirt**, the 2PL uniform DIF model can be constructed using the `dif` and `mg` modules as follows:

```
R> model9 <- flirt(data=verb2, loading=list(on=TRUE),
 +             dif=list(on=TRUE, dif_beta=c(1,2) ),
 +             mg=list(on=TRUE, group_matrix=male) )
```

Here, `dif_beta=c(1,2)` indicates that DIF parameters for items 1 and 2 (in the first and second columns of the item response data) are estimated and `group_matrix=male` indicates that DIF is investigated between male and female subjects. Observe that the module for a multiple-group analysis `mg` is also specified in order to take care of possible distributional differences between the two groups.

DIF may exist not only for the item intercept parameters but also for the item slope parameters, which is referred to as 'non-uniform' DIF. The 2PL model for analyzing the non-uniform DIF for item f can be specified as follows:

$$\begin{aligned}\text{logit}(P(y_{pi}=1|\theta_p)) &= \beta_i + \gamma_\beta G_{pg} X_{if} \\ &\quad +(\alpha_i + \gamma_\alpha G_{pg} X_{if})\theta_{pg}, \end{aligned} \tag{13}$$

where γ_β represents the DIF for the intercept parameter β_i and γ_α represents the DIF for the slope parameter α_i for item f.

With **flirt**, the 2PL non-uniform DIF model can be constructed as follows:

```
R> model10 <- flirt(data=verb2, loading=list(on=TRUE),
 +             dif=list(on=TRUE, dif_beta=c(1,2), dif_alpha=c(1,2)),
 +             mg=list(on=TRUE, group_matrix=male) )
```

Here `dif_beta=c(1,2)` and `dif_alpha=c(1,2)` indicate that items 1 and 2 are investigated for the non-uniform DIF between male and female students (`group_matrix=male`).

Multiple dimensions

Suppose a test consists of K sub-scales that are intended to measure related latent traits. Then the standard 2PL model can be extended with K latent traits (or dimensions) as follows:

$$\text{logit}(P(y_{pi}=1|\boldsymbol{\theta}_p)) = \beta_i + \sum_{k=1}^{K} \alpha_{ik}\theta_{pk}, \tag{14}$$

where α_{ik} is the item slope parameter for item i in dimension k that item i belongs to and θ_{pk} is the kth latent variable ($k = 1, ..., K$, where K is the total number of dimensions). If item i belongs to only one dimension, the model is referred to as a between-dimension multidimensional model, whereas if item i belongs to more than one dimension, the model is referred to as a within-item multidimensional model. The following two loading matrices illustrate between- and within-item multidimensional models for the case of six items with three dimensions:

$$\Lambda_1 = \begin{bmatrix} \alpha_{11} & 0 & 0 \\ \alpha_{21} & 0 & 0 \\ 0 & \alpha_{32} & 0 \\ 0 & \alpha_{42} & 0 \\ 0 & 0 & \alpha_{53} \\ 0 & 0 & \alpha_{63} \end{bmatrix}, \qquad \Lambda_2 = \begin{bmatrix} \alpha_{11} & 0 & \alpha_{13} \\ \alpha_{21} & 0 & 0 \\ 0 & \alpha_{32} & \alpha_{33} \\ 0 & \alpha_{42} & 0 \\ \alpha_{51} & 0 & \alpha_{53} \\ 0 & 0 & \alpha_{63} \end{bmatrix}.$$

The rows represent items and the columns represent dimensions. In the loading matrix Λ_1, the first and second items belong to dimension 1, the third and fourth items belong to dimension 2, and the fifth and six items belong to dimension 3. In the loading matrix Λ_2, the first item belongs to dimensions 1 and 3, the third item belongs to dimensions 2 and 3, the fifth item belongs to dimensions 1 and 3 simultaneously, while the other items belong to only one dimension.

The K latent variables are allowed to be correlated with each other and assumed to follow a multivariate normal distribution, $\boldsymbol{\theta}_p = (\theta_{p1}, \ldots, \theta_{pK})' \sim N(\mathbf{0}, \Sigma)$. The package **flirt** estimates the elements of a lower triangular Cholesky matrix L for the covariance matrix Σ, where $\Sigma = L \cdot L^\top$, and $L^\top$ the transpose of L. For example, with $K = 3$

$$\begin{aligned} &\underbrace{\begin{bmatrix} c1 & 0 & 0 \\ c4 & c2 & 0 \\ c5 & c6 & c3 \end{bmatrix}}_{L} \cdot \underbrace{\begin{bmatrix} c1 & c4 & c5 \\ 0 & c2 & c6 \\ 0 & 0 & c3 \end{bmatrix}}_{L^\top} \\ &= \underbrace{\begin{bmatrix} c1^2 & c1 \cdot c4 & c1 \cdot c5 \\ c1 \cdot c4 & c4^2 + c2^2 & c4 \cdot c5 + c2 \cdot c6 \\ c1 \cdot c5 & c4 \cdot c5 + c2 \cdot c6 & c5^2 + c6^2 + c3^2 \end{bmatrix}}_{\Sigma}. \end{aligned} \tag{15}$$

Once the Cholesky elements $c1$ to $c6$ are estimated, the estimated covariance matrix Σ is then constructed. For identification of 2PL multidimensional models, the diagonal elements of L ($c1$, $c2$, and $c3$) are fixed to 1 as all loading parameters are freely estimated.

With **flirt**, the 2PL multidimensional model can be specified using the `mul` module as follows:

```
R> model11 <- flirt(data=verb2, loading=list(on=TRUE),
 +             mul=list(on=TRUE, dim_info=list(dim1=1:12, dim2=13:24 ) ) )
```

Here, `dim_info=list(dim1=1:12, dim2=13:24)` indicates that two dimensions are specified and the first dimension (dim1) includes the first 12 items (`1:12`) and the second dimension (dim2) includes the next 12 items (`13:24`). That is, the model specified here is a between-item, two-dimensional model. A within-item model can

be specified if there are overlapping items between dimensions. For instance, suppose items 1 to 4 belong to dimensions 1 and 2 at the same time. Then a within-item multidimensional can be specified as follows:

```
R> model12 <- flirt(data=verb2, loading=list(on=TRUE),
 +            mul=list(on=TRUE, dim_info=list(dim1=c(1:12), dim2=c(13:24,1:4) ) ) )
```

Bifactor models

A bifactor model is a useful psychometric model for a test with several sub-domains or testlets (or item clusters). A bifactor model consists of the general factor that represents the primary latent trait and a set of specific factors that capture dependence within the sub-domains (representing the sub-domain latent traits). Each item is therefore loaded on the general factor and one of the specific factors in a bifactor model; the general and specific factors are assumed to be independent of each other. Figure 1 illustrates a path diagram for an item bifactor model with three specific factors.

For a test with K testlets, a 2PL bifactor model can be formulated as follows:

$$\text{logit}\left(P(y_{pi}=1|\theta_p^G,\theta_{pk}^S)\right)=\beta_i+\alpha_{ik}^S\theta_{pk}^S+\alpha_i^G\theta_p^G, \quad (16)$$

where α_{ik}^S and α_i^G are the item slope parameters for the kth specific factor θ_{pk}^S ($k=1,...,K$, where K is the number of specific dimensions) and the general factor θ_p^G, respectively. That is, item i has two loadings, one for the general factor and another for the specific factor that item i belongs to.

The package **flirt** fits a bifactor model using the `bifac` module as follows:

```
R> model13 <- flirt(data=verb2, loading=list(on=TRUE),
 +             bifac=list(on=TRUE, dim_info=list(dim1=1:8, dim2=9:16, dim3=17:24)) )
```

Similar to the multidimensional model, `dim_info` is used to provide dimensional information of the three specific dimensions. That is, `dim_info=list(dim1=1:8, dim2=9:16, dim3=17:24)` means that there are three specific dimensions (`dim1`, `dim2`, `dim3`) and each dimension includes eight items.

Models for polytomous responses

flirt handles polytomous item responses using two different logit link functions, such as cumulative and adjacent-category logits. Here we describe how different models for polytomous items can be constructed based on the two logit link functions and how those models can be fitted with **flirt**.

1) Cumulative logits

For each response category j ($j=0,...,J$), 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), which can be written as

$$P(y_{pi}\geq j|\theta_p)=\frac{\exp\left(\alpha_i\theta_p+\beta_{ij}\right)}{1+\exp\left(\alpha_i\theta_p+\beta_{ij}\right)}, \quad (17)$$

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 a category greater than j, i.e., $P(y_{pi}=j|\theta_p)=P(y_{pi}\geq j|\theta_p)-P(y_{pi}>j|\theta_p)$.

With **flirt**, a graded response model can be specified by choosing the cumulative link in the `control` option as follows:

```
R> data(verb3)
R> model15 <- flirt(data=verb3, loading=list(on=TRUE),
 +            control=list(link="cumulative") )
```

In the first line, `verb3` is loaded which includes the item response data with the original three response categories: No(0), Perhaps(1), Yes(2). Note that for mixed-type item responses that include polytomous and binary items, specifying the cummulative link function implies that a graded response model is applied to polytomous items, while an ordinary 2PL model is applied to binary items.

2) Adjacent-category logits

The adjacent-category logit function (Mellenbergh, 1995) contrasts each category j with the 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

$$P(y_{pi}=j|\theta_p)=\frac{\exp\sum_{m=0}^{j}\left(\alpha_i\theta_p+\beta_{ij}\right)}{\sum_{r=0}^{J}\exp\sum_{m=0}^{r}\left(\alpha_i\theta_p+\beta_{ij}\right)}, \qquad (18)$$

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.

With **flirt**, a partial credit model can be specified by choosing the adjacent link as follows:

```
R> model16 <- flirt(data=verb3, loading=list(on=TRUE),
 +            control=list(link="adjacent") )
```

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

$$P(y_{pi}=j|\theta_p)=\frac{\exp\sum_{m=0}^{j}\left(\alpha_i\theta_p+\beta_i+\delta_j\right)}{\sum_{r=0}^{J}\exp\sum_{m=0}^{r}\left(\alpha_i\theta_p+\beta_i+\delta_j\right)}, \qquad (19)$$

where β_i is the intercept for item i and δ_j is the step parameter for category j, and $\sum_{r=0}^{0}(\alpha_i\theta_p+\beta_i+\delta_j)\equiv 0$. Observe that the $(I\times J)$ step parameters β_{ij} in the generalized partial credit model (18) is simplified into I item intercept parameters β_i and J step parameters δ_j (that are equal across items) in the 2PL rating scale model (19).

With **flirt**, a rating scale model is treated as a constrained version of the partial credit model and the constraints can be imposed by manipulating the item design matrix. For instance, a 2PL rating scale model is specified as follows:

```
R> data(item_design_rating)
R> model17 <- flirt(data=verb3, loading=list(on=TRUE),
 +            item_cov=list(on=TRUE, item_matrix_beta=item_design_rating) ,
 +            control=list(link="adjacent") )
```

In the first line, the 48 × 25 item design matrix `item_design_rating` is loaded, which defines the constraints on the step parameters. Specifically, the design matrix includes 24 item indicator variables (to estimate I item intercept parameters β_i) and one category indicator variable (to estimate one step parameter δ_j). Similar to the partial credit model, the adjacent link function is specified in the `control` option. Observe that the item design matrix to impose the constraints on the step parameters is incorporated using the `item_cov` module.

Constructing customized models

An important and unique strength of **flirt** is that users can build customized IRT models by freely utilizing `flirt` modeling modules. An example of a customized model is a multidimensional latent regression 2PL graded response model with item covariates for the item slope parameters. This seemingly complex model can be constructed by simply adding the modules for (1) multidimensionality, (2) person covariates, (3) item covariates for the item slope parameters, and (4) cummulative link function. That is,

```
R> model18 <- flirt(data=verb3,  loading=list(on=TRUE),
 +            mul=list(on=TRUE, dim_info=list(dim1=1:12, dim2=13:24),
 +                     cov_info=list(dim1=1, dim2=0)),
 +            person_cov=list(on=TRUE, person_matrix=male),
 +            item_cov=list(on=TRUE, item_matrix_alpha=item_design_bin),
 +            control=list(link="cumulative") )
```

Note that `cov_info=list(dim1=1, dim2=0)` means that the person covariate `male` is used to explain the first dimension only. When more than one person covariate is available, users can choose which covariates will be used to explain each dimension. Also observe that `item_cov` is specified so that item covariates (`item_design_bin`) explain the item slope parameters α_i.

Another useful example of a customized model is a bifactor model with differential feature functioning (DFF) parameters; here DFF represents the degree to which the effects of item features on the item intercept and/or slope

parameters differ between different groups of people with the same ability level. This bifactor DFF model can be specified by using `bifac`, `item_cov`, `dif`, and `mg` modules as follows:

```
R> model19 <- flirt(data=verb2,
 +            bifac=list(on=TRUE, dim_info=list(dim1=1:12, dim2=13:24)),
 +            item_cov=list(on=TRUE, item_matrix_beta=item_design_bin,
 +                      item_matrix_alpha=item_design_bin),
 +            dif=list(on=TRUE, dif_beta=c(2,3), dif_alpha=c(2,3) ),
 +            mg=list(on=TRUE, group_matrix=male) )
```

Here, `dif=list(on=TRUE, dif_beta=c(2,3), dif_alpha=c(2,3) )` means that the DFF parameters for the second and third item features (in the second and third columns of the item design matrix) are estimated for the item intercept parameters and the item slope parameters (separately for the general and specific factors).

These are only two examples of the many customized models that can be constructed with **flirt**. With its convenient model building system, **flirt** can be utilized to construct e.g., various longitudinal IRT models (e.g., Andersen, 1985; Embretson, 1991) and item-response tree models (e.g., Jeon and de Boeck, 2015a in press; Jeon et al., 2015b). Examples of `flirt` codes for longitudinal item analysis and item-response tree modeling can be found on http://faculty.psy.ohio-state.edu/jeon/lab/flirt.php.

Practical options

flirt provides several practical options. For instance, empirical Bayes prediction (or expected a posteriori; EAP) and its standard errors can be obtained by using `post=TRUE`. The `post` option also provides expected scores and the IRT reliability coefficient (Haberman and Sinharay, 2010). Item characteristic curves as well as item/test information curves for unidimensional models can be drawn using the `IRF` and `Item_info` functions. For multidimensional models, standardization of the item loading parameters and the covariance matrix (between dimensions) can be obtained using `std_coef` and `std_cov` functions. With the `control` option, a variety of estimation options can be modified, such as the number of quadrature points (`nq`), maximum iteration numbers (`max_it`), convergence criterion (`conv`), adaptive quadrature (`adapt`) and so on. The minimum percentage of category frequencies for polytomous items can be modified using `minpercent`; an item response category that has frequency rates lower than the specified `minpercent` is collapsed with an adjacent lower category. User-specified starting values can be provided using `start` and user-specified linear constraints can be imposed using `constraint`. Also, `evaluate` can be utilized to evaluate the log-likelihood given fixed values of model parameters. In addition, frequency (or sampling) weights can be incorporated with the `weight` option.

More information on these and other practical options can be found in the package manual (Jeon et al., 2015) and the example R script that are available on http://faculty.psy.ohio-state.edu/jeon/lab/flirt.php.

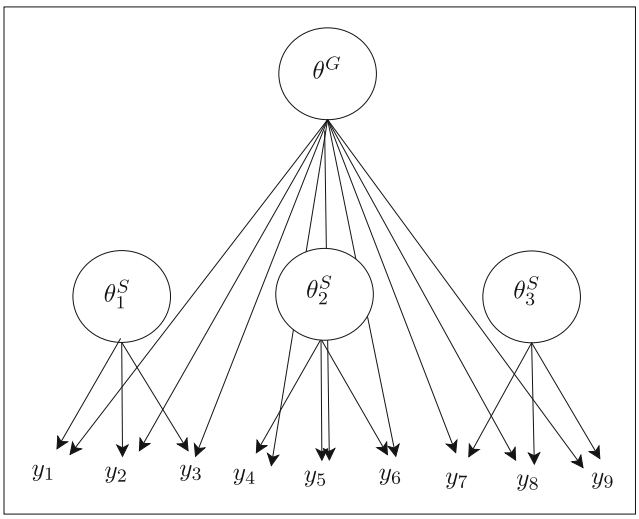

Fig. 1 A path diagram for an item bifactor model with three specific factors (the subscript p for person is suppressed in the figure)

Statistical framework

Generalized linear and nonlinear mixed models

The flexibility of **flirt** comes from its underlying general statistical modeling framework-generalized linear and nonlinear mixed models (GLNMMs). Generalized linear mixed models (GLMMs) are a class of statistical models used to analyze 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) and De Boeck and Wilson (2004) among others. Here we briefly describe a GLNMM framework and show how IRT models can be specified in this framework. Binary responses y_{pi} for person p $(=1, ..., N)$ to item i $(=1, ..., I)$ are assumed to have a Bernoulli distribution with conditional probability π_{pi} given the 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

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

In GLNMMs, the linear predictor can be written as follows:

$$\text{logit}(P(y_{pi} = 1|\theta_p) = \nu_{pi} = \sum_{q=1}^{Q} \beta_q X_{iq} + \sum_{r=1}^{T} \alpha_r Z_{ir}\theta_p, \tag{20}$$

where β_q is the regression coefficient for the qth observed covariate X_{iq} $(q=1, ..., Q)$, α_r is the regression coefficient for the rth covariate Z_{ir} $(r=1, ..., R)$, which is multiplied by the random effect θ_p. The random effect for person p, θ_p is typically assumed to follow a normal distribution, $\theta_p \sim N\left(0, \sigma^2\right)$. Model (20) is a nonlinear model because the fixed-effect parameter α_r is multiplied by the random-effect parameter θ_p. Some of the regression coefficients β_q can vary across people (i.e., random coefficients), e.g., $\beta_1 = \beta_1^I + \theta_p^I$; then the model involves two correlated latent traits that follow a multivariate normal distribution, $\left(\theta_p, \theta_p^I\right)' \sim N(\mathbf{0}, \Sigma)$.

Note that when the covariates $\mathbf{X}$ and $\mathbf{Z}$ consist of all item indicator variables, GLNMM (20) becomes a standard 2PL IRT model, in which β_q and α_r correspond to the item intercept and slope parameters, respectively.

Therefore, various IRT models described in the "Fitting IRT models with `flirt`" section can be formulated as extensions or modifications of the GLNMM (20). For instance, the 1PL model is obtained by choosing no covariates for Z_{ir} or by constraining α_r to 1. Including item covariates and person covariates are straightforward in Eq. 20. Multidimensional models can be formulated by utilizing random coefficients for item covariates (that indicate item groups or dimensions).

Estimation

The package **flirt** adopts an efficient modified expectation-maximization (EM) algorithm (Lauritzen, 1995; Rijmen et al., 2008), where the E-step is modified based on a graphical model framework. Specifically, 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), and Rijmen et al. (2014).

The gains from the efficient E-step can be considerable for multidimensional models where computationally demanding high-dimensional numerical integration is required over the joint space of all latent variables. 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. Furthermore, the package **flirt** provides the adaptive quadrature option, which is more accurate than the ordinary Gauss–Hermite quadrature (Pinheiro and Bates, 1995; Rabe-Hesketh et al., 2005).

It is worth noting that the package **flirt** is faster in general than the general software packages such as SAS nlmixed, gllamm, and **lme4** that require long-form input data. For example, to fit a three-dimensional 1PL model with 108 items (36 items per dimension) and 1,069 subjects, the R package **lme4** (Bates et al., 2014) took nearly 4 h (14,264 s) whereas **flirt** took 809 s on an Intel Pentium Dual-Core 2.5-GHz processor computer with 3.2 GB of memory.

Current R packages for item response analysis

Here we review some current R packages whose main purpose is to provide maximum likelihood estimation of IRT models with continuous latent traits. Those packages can roughly be categorized into three classes.

The first class of packages focus on fitting Rasch family models. For instance, **eRm** (Mair et al., 2014) offers conditional maximum likelihood estimation of the Rasch model and the linear logistic test model (LLTM) for binary items (i.e., allowing for item covariates for item difficulty) and the rating scale model and the partial credit model for polytomous items. The package **plRasch** provides conditional maximum likelihood estimation of the Rasch family model with a log-linear-by-linear association (LLLA; Anderson et al., 2007).

The second class of packages handle general parametric models. For instance, **ltm** (Rizopoulos, 2006) fits 1PL, 2PL, and 3PL models for binary items using marginal maximum likelihood estimation. For polytomous items, the graded response model and the generalized partial credit model can be fitted with **ltm**.

The third class of packages fit multidimensional parametric models. For example, **ltm** fits 2PL multidimensional models, but up to two dimensions only. Using a Metropolis-Hastings Robbins-Monro algorithm as well as a regular EM method, **mirt** (Chalmers, 2012) provides a broader range of multidimensional models, including exploratory models and confirmatory models as well as bifactor models, two-tier models, and partially compensatory models. The latest version of **mirt** (version 1.8) allows for some parametric models to incorporate fixed and random effects of item covariates for item and slope parameters and person covariates. In addition, **TAM** (Kiefer et al., 2014) provides joint and marginal maximum likelihood estimation of various parametric uni- and multidimensional models, including 2PL, 3PL models, and bifactor models. Person covariates as well as item covariates (both for item and slope parameters) can be included for some models.

The fourth class of packages provide estimation of a wide range of special-purpose IRT models utilizing existing R packages and functions. For instance, **TAM** uses other packages (e.g., **lavaan** (Rosseel, 2012)) to fit various special-purpose models such as multi-facets models, structured latent class analysis, finite-mixture IRT models, located latent class models, and cognitive diagnostic models. The package **sirt** (Robitzsch, 2013) is designed to provide a common interface to fit a variety of IRT models by utilizing various existing packages and functions in R.

Our package **flirt** takes a position between the third and fourth classes in the sense that (1) **flirt** fits various multidimensional parametric models, and (2) it allows users to develop customized, special-purpose models but without relying on other packages.

Compared to other existing packages, **flirt** shows a particular strength in developing customized bifactor models. For example, **flirt** allows users to specify linear regression models for the item slope parameters that are associated with the general factor and/or (some or all) specific factors. Similarly, a different set of person covariates can be specified for the general factor and/or each of the specific factors, while the person covariates can also be entered as 'main effects' (as explained in Section "Person covariates").

In addition, **flirt** provides great advantages in analyzing differential item functioning (DIF). Specifically, **flirt** allows not only for estimating DIF for the item intercept parameters and for the item slope parameters (in each dimension for multidimensional models), but also for estimating differential feature functioning (DFF) (both for item intercept and slope parameters). Second, **flirt** can produce more accurate DIF/DFF estimates by adjusting for group differences more precisely. Specifically, **flirt** takes into account group differences in both means and variances (in all dimensions for multidimensional models), whereas most packages consider only the mean differences between groups (for multiple-dimensional models, only the overall mean or the mean of the general factor is typically adjusted). Third, for bifactor models, **flirt** offers a generalized multiple-group bifactor model which relaxes the typical independence assumption to conditional independence of specific factors given the general factor (Jeon et al., 2013). That is, **flirt** adjusts for possible group differences in the means, variances, and correlations (between the general factor and the specific factors), which is critical for accurately estimating DIF parameters in bifactor models. For more information on the generalized multiple-group bifactor DIF model, see e.g., Jeon et al. (2013).

Final remarks

Contribution

In this article, we introduced the new IRT software package **flirt**. A major attraction of **flirt** is its convenient modular approach that allows researchers to easily build, explore, and estimate various customized IRT models. Specifically, **flirt** allows users to (1) utilize item and person explanatory variables in order to investigate sources of item properties as well as individual differences in latent traits, (2) develop customized bifactor models, and (3) examine differential item and feature functioning for various types of models. Hence, **flirt** contributes to the field with these strengths that are complementary to other existing R packages.

Limitations and future development

There are some operational shortcomings of using **flirt** in practice. First, **flirt** (version 1.15) can only be installed manually on Windows 32/64-bit computers. Mac and Linux

versions as well as more convenient ways of installing and updating the package will be devised in the near future. Second, **flirt** calls the MATLAB program, BNLflirt (Rijmen and Jeon, 2013) from R, which somewhat decreases **flirt**'s speed (despite the package implements computationally efficient estimation methods). In general, faster computation can be achieved using the BNLflirt program in MATLAB. Interested readers are encouraged to explore BNLflirt that can be obtained by contacting the first author.

The R package **flirt** is currently under active development. A variety of new modeling modules will continue to be incorporated in future versions, such as guessing item parameters (e.g., Birnbaum, 1968), second-order and third-order item response models (e.g., Rijmen et al., 2014), discrete latent variables (e.g., Rost, 1990), hierarchical structures on the person side (e.g., multilevel IRT models), random coefficients of item predictors (e.g., Rijmen and De Boeck, 2002), and random item parameters (e.g., De Boeck, 2008). Combining multiple link functions will also be possible for a more flexible modeling of polytomous item responses.

Acknowledgments 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: Australian Council for Educational Research.

Andersen, E.B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*, 3–16.

Anderson, C., Li, Z., & Vermunt, J. (2007). Estimation of models in a Rasch family for polytomous items and multiple latent variables. *Journal of Statistical Software*, *20*, 1–36.

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. 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.

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

Cai, L., du Toit, S.H.C., & Thissen, D. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling*. Lincolnwood: 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., & Wilson, M. (2004). *Explanatory item response models: a generalized linear and nonlinear approach*. New York: Springer.

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.

Embretson, S.E. (1999). Generating items during testing: psychometric issues and models. *Psychometrika*, *64*, 407–433.

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.

Jeon, M., & de Boeck, P. (2015a). (in press). A generalized item response tree model for psychological assessments. *Behavior Research Methods*.

Jeon, M., de Boeck, P., & van der Linden, W. (2015b). (submitted). A multidimensional tree approach for modeling item response and change behavior.

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.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2015). *flirt: R package for flexible item response theory modeling*. Manual, http://faculty.psy.ohio-state.edu/jeon/lab/flirt.php

Kiefer, T., Robitzsch, A., & Wu, M. (2014). *TAM: Test Analysis Modules*, R package version 1.0-3.18-1. 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.

Li, Z., & Hong, F. (2007). *lRasch: Log linear by linear association models*, R package version 0.1, http://CRAN.R-project.org/package=plRasch

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

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

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: 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, 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. (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

Rosseel, Y. (2012). lavaan: an R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36.

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.

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

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.