British Journal of Mathematical and Statistical Psychology (2016), 69, 80–103
© 2015 The British Psychological Society

The British Psychological Society

www.wileyonlinelibrary.com

Detection of differential item functioning in Rasch models by boosting techniques

Gunther Schauberger* and Gerhard Tutz

Department of Statistics, Ludwig-Maximilians-University, Germany

Methods for the identification of differential item functioning (DIF) in Rasch models are typically restricted to the case of two subgroups. A boosting algorithm is proposed that is able to handle the more general setting where DIF can be induced by several covariates at the same time. The covariates can be both continuous and (multi-)categorical, and interactions between covariates can also be considered. The method works for a general parametric model for DIF in Rasch models. Since the boosting algorithm selects variables automatically, it is able to detect the items which induce DIF. It is demonstrated that boosting competes well with traditional methods in the case of subgroups. The method is illustrated by an extensive simulation study and an application to real data.

1. Introduction

In the early days of item response theory (IRT) the focus was on the Rasch model (Rasch, 1960) and its extensions to the two-parameter logistic (2PL) and three-parameter logistic (3PL) models by Birnbaum (1968). The Rasch model assumes that every person has a fixed latent ability and every item has a fixed difficulty. The difference between ability and difficulty determines the probability that a person solves an item. The extensions by Birnbaum (1968) attenuated this assumption by introducing two additional item parameters, for discrimination and guessing. Since then, IRT has been a topic of intensive research and has been extended in various ways.

A well-known problem in item response models is that the probability of scoring on an item might vary over persons with the same latent ability. This may be caused by certain characteristics of the persons such as gender, age, or race, or by other unknown (latent) classes within the population tested. This phenomenon is known under the name differential item functioning (DIF). If an item is detected to have DIF, one option is to remove the item because it does not provide a fair measurement of the respective trait.

There is a wide range of literature on DIF in general; see, for example, Holland and Wainer (1993) and Millsap and Everson (1993). A very popular choice for detecting DIF is the Mantel–Haenszel (MH) method. This is based on a test statistic proposed by Mantel and Haenszel (1959) and was used to detect DIF in IRT by Holland and Thayer (1988). Various other methods to identify items which induce DIF have been proposed; see, for example, Swaminathan and Rogers (1990) and Lord (1980). Magis, Béland, Tuerlinckx, and Boeck (2010) set up a framework for the existing DIF methods and gave an excellent overview on

*Correspondence should be addressed to Gunther Schauberger, Department of Statistics, Ludwig-Maximilians-Universität Munich, Akademiestraße 1, 80799 Munich, Germany (email: gunther.schauberger@stat.uni-muenchen.de).

DOI:10.1111/bmsp.12060

currently available methods along with a software implementation (Magis, Beland, & Raiche, 2013).

The essential drawback of the MH method and most of the other existing methods is that they are limited to identifying DIF between two subgroups; for example, for male and female participants. Some methods for multiple subgroups have been developed; see Somes (1986), Penfield (2001), Magis, Raîche, Béland, and Gérard (2011), and Kim, Cohen, and Park (1995). Gonçalves, Gamerman, and Soares (2013) set up a quite general Bayesian multifactor model for the detection of DIF in the 3PL model (Birnbaum, 1968). But methods that are able to handle DIF induced by continuous covariates or by a whole vector of covariates at the same time are scarce. Recently, Strobl, Kopf, and Zeileis (2015) proposed the use of tree methodology, whereas Magis, Tuerlinckx, and De Boeck (2015) and Tutz and Schauberger (2015) used penalization techniques.

The aim of the paper is to propose a new and efficient method for the detection of DIF in Rasch models that can deal with several (continuous and categorical) covariates as well as interactions between the covariates simultaneously. The method is based on boosting techniques which have been developed more recently in the machine learning community (Freund & Schapire, 1996) and in statistics (Bühlmann and Hothorn, 2007a), but their potential has not yet been exploited to uncover structures in item response models.

In Section 2 a DIF model is given in which DIF is explicitly represented by parameters. Section 3 introduces the idea of boosting in general, and Section 4 describes in detail the proposed estimation algorithm. Sections 5 and 6 illustrate the method with applications to both simulated and real data sets and compare it to existing approaches.

2. Differential item functioning model

In the binary Rasch model the probability of a person scoring on an item is determined by a parameter for the latent ability of the person and a parameter for the item difficulty. In the case of P persons and I items, the Rasch model is given by

$$P(Y_{pi}=1)=\frac{\exp(\theta_p-\beta_i)}{1+\exp(\theta_p-\beta_i)},\quad p=1,\ldots,P, i=1,\ldots,I, \tag{1}$$

where Y_{pi} represents the response of person p on item i. It is coded by $Y_{pi}=1$ if person p solves item i and $Y_{pi}=0$ otherwise. Both the person parameters, θ_p, $p=1,\ldots,P$, and the item parameters, β_i, $i=1,\ldots,I$, are unknown and have to be estimated. Alternatively, model (1) can be given in the form

$$\log\left(\frac{P(Y_{pi}=1)}{P(Y_{pi}=0)}\right)=\theta_p-\beta_i, \tag{2}$$

where the left-hand side specifies the so-called log-odds or logits. As model (2) is not identifiable in this general form, a restriction on the parameters is needed. A common choice, which is also used in what follows, is $\theta_P=0$. Alternatively, one item parameter or the sum of all item parameters could be restricted to zero.

In item response models, DIF appears if an item has different difficulties depending on characteristics of the person which tries to solve the item. Therefore, DIF changes the item difficulty depending on covariates of the participants. This concept can be formalized by

$$\log\left(\frac{P(Y_{pi}=1)}{P(Y_{pi}=0)}\right) = \theta_p - (\beta_i + \mathbf{x}_p^T\gamma_i), \tag{3}$$

where $\mathbf{x}_p^T = (x_{p1}, \ldots, x_{pm})$ denotes a person-specific covariate vector of length m and, again, the restriction $\theta_P = 0$ is used. This general DIF model is an extension of the Rasch model (2) allowing for person-specific item difficulties $\beta_i + \mathbf{x}_p^T\gamma_i$. The item-specific parameters $\gamma_i^T = (\gamma_{i1}, \ldots, \gamma_{im})$ determine how the covariates $x_{p1}, \ldots, x_{pm}$ influence the difficulty of item i for person p. The original Rasch model corresponds to the special case where $\gamma_i = \mathbf{0}$ for all items. The general model (3) was proposed by Tutz and Schauberger (2015), and a special case of the model was considered by Paek and Wilson (2011). However, estimation methods were quite different from the approach suggested here.

The main problem with the general DIF model is that $m{\cdot}I$ additional parameters (compared to the Rasch model) have to be estimated. Since each item has its own parameter for each covariate, the number of parameters in the model can be huge. As the full DIF model is not identifiable, maximum likelihood (ML) estimation is not an option in this case. One possibility to overcome this problem are penalization methods where a penalized likelihood is maximized. For example, the ridge estimator (Hoerl & Kennard, 1970) or the lasso estimator (Tibshirani, 1996) can still be calculated when regular ML estimation fails. Penalization methods of this type were used by Tutz and Schauberger (2015). Here we propose a quite different method, namely, boosting. Boosting is an algorithmic procedure with origins in machine learning; see, for example, Freund and Schapire (1997).

Boosting as a method of statistical learning was developed by Friedman, Hastie, and Tibshirani (2000) and extended, for example, by Bühlmann and Yu (2003), Tutz and Binder (2006), Bühlmann (2006) and Bühlmann and Hothorn (2007a). Boosting in basic regression methods is available for the R statistical software (R Core Team, 2014), which will be used for all following calculations. It is, for example, implemented in the add-on package `mboost` (see Hothorn, Buehlmann, Kneib, Schmid, & Hofner, 2013), which is also used for our computations.

One strength of boosting is that it is able to select relevant terms in the predictor even in very high-dimensional settings. This establishes the link to DIF in item response models. The general assumption for our model is that only some of the items show DIF and that only for these items item-specific parameters γ_i have to be estimated. Therefore, detection of DIF means selection of variables, or, in parametric models, selection of parameters that should be included in the model and, therefore, have non-zero estimates. If a whole vector γ_i is set to zero, the difficulty of item i does not depend on the covariates and no DIF is present.

Generally, in the following all covariates are assumed to be standardized. This has the advantage that the covariates have the same scale and, therefore, can be compared directly. In particular, estimates for the item-specific covariates γ_{ip} can be compared directly and represent the size of the respective DIF effect.

3. Basic boosting procedures

Before developing boosting procedures for DIF models, in this section we briefly consider the basic concept of boosting and the choice of tuning parameters. The adaptation to DIF models will be considered subsequently. We start with the linear model, where boosting is much easier to conceptualize, and then proceed to boosting for the generalized linear model (GLM).

Let us first consider a linear regression model

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \quad i = 1, \ldots, n,$$

for p covariates. If p is very large and it is suspected that not all covariates are influential, ML estimation is a bad choice because of its instability in high-dimensional settings. In contrast, boosting is able to fit additive structures even in high-dimensional settings by successively fitting only parts of the model.

A basic ingredient of boosting is the specification of the so-called base learners $\hat{f}(\cdot)$. The base learners specify the structure that is fitted within one step of the procedure. Since we want to fit a linear model, the base learners are the ordinary least squares estimators

$$\hat{f}(x) = \hat{\beta}_s x_s$$

for *single* covariates $s = 1, \ldots, p$. This means that within one step only one covariate is used. Although all the updates of all covariates are evaluated, in each boosting step a specific covariate s^* is selected that yields the greatest reduction of the residual sum of squares given the previous estimate. Let $\hat{\eta}^{(l-1)}$ represent the current linear predictor (the current model fit) from the previous boosting step l–1; then the residual of the ith observation is $u_i = y_i - \hat{\eta}^{(l-1)}$. In the lth step one fits a linear model for only one covariate on the data (u_i, x_{is}), $i = 1, \ldots, n$, and then selects the best predictor, which is used to update the linear predictor. Thus boosting is a stepwise procedure, which iteratively improves the fit by fitting a model to the current residuals. Tukey (1977) proposed the so-called 'twicing', which means that after fitting a model one fits it again on the residuals. Boosting means not just two but many iterative fits. The following algorithm can be seen as the basic boosting procedure for linear models.

L$_2$ boost in linear models

Step 1 (Initialization)

Given data $\{y_i, \mathbf{x}_i\}$, fit the base procedure to yield the function estimate $\eta^{(0)}(\mathbf{x}_i)$. Typically one fits an intercept model obtaining $\eta^{(0)}(\mathbf{x}_i) = \hat{\beta}_0$.

Step 2 (Iteration: Fitting of base learners and selection)

For $l = 1, 2, 3, \ldots$, compute the residuals $u_i = y_i - \hat{\eta}^{(l-1)}(\mathbf{x}_i)$ and fit the base learners to the current data $\{u_i, \mathbf{x}_i\}$.
(a) One fits by minimizing least squares, that is, for fixed j one minimizes $\sum_{i=1}^{n}(u_i - \beta_j x_{ij})^2$, obtaining $\hat{\beta}_j$,
(b) Selection means that one determines s^* such that $s^* = \arg\min_{j} \sum_{i=1}^{n}(u_i - \hat{\beta}_j x_{ij})^2$.
(c) The improved fit is obtained by the update $\hat{\eta}^{(l)}(\mathbf{x}_i) = \hat{\eta}^{(l-1)}(\mathbf{x}_i) + \nu\hat{\beta}_{s^*}x_{is^*}$.

Step 3 (Stop)

Iterate step 2 until $l = l_{\text{stop}}$ is reached.

In step 2, the linear predictor is updated by $\hat{\eta}^{(l)}(\mathbf{x}_i) = \hat{\eta}^{(l-1)}(\mathbf{x}_i) + \nu\hat{\beta}_{s^*}x_{is^*}$, which serves to compute the residuals in the following boosting step. It should be noted that the linear structure is maintained and only one component of the linear predictor is updated.

Let $\hat{\eta}^{(l-1)}(\mathbf{x}_i)$ have the linear form $\sum_{j=1}^{p} \hat{\beta}_j^{(l-1)} x_{ij}$; then the addition of $\nu\hat{\beta}_{s^*} x_{is^*}$ changes only the weight on the variable s^*. The parameter ν, $0 < \nu \leq 1$, is used as a shrinkage parameter. This shrinkage parameter makes the base learners 'weak' and, therefore, prevents overfitting because only small steps towards the optimal solution are made. For this purpose, ν has to be chosen sufficiently small; $\nu = .1$ is a common choice. The procedure corresponds to a stepwise fitting of the linear model. In every step, one of the coefficients is updated by a rather small amount. The weakness of the learner is important because only then is the fit efficient (Bühlmann and Yu, 2003; Bühlmann, 2006). The smaller ν is chosen, the weaker the learner gets but the more boosting steps are required. If one does not stop, boosting is a complicated way of obtaining the ML estimate. The selection effect is obtained by stopping the procedure before it converges. Then, only the variables that obtained non-zero weights are included in the model and one obtains a regularized estimate. Bühlmann (2006) showed that the procedure is consistent for underlying regression functions that are sparse in terms of the L_1-norm.

Boosting can also be seen as a stepwise optimization of a specific loss function. For the linear regression model, the optimized loss function is the L_2 loss between the response and the linear predictor. In this context, boosting can be seen as a gradient descent method and sometimes is called *gradient boosting*. For the (slightly modified) L_2 loss function

$$L(y, \eta) = \frac{1}{2}(y - \eta)^2,$$

the gradient is given by the residuals

$$\frac{\partial L(y, \eta)}{\partial \eta} = y - \eta.$$

Therefore, instead of stepwise fitting of the residuals, boosting can be seen as repeated fitting of the response with a so called offset, which is a known constant. In our case it is given by the estimate of the previous step $\hat{\eta}^{(l-1)}(\mathbf{x}_i)$. The least squares estimate uses the criterion $\sum_{i=1}^{n}(u_i - \beta_j x_{ij})^2 = \sum_{i=1}^{n}(y_i - (\hat{\eta}^{(l-1)}(\mathbf{x}_i) + \beta_j x_{ij}))^2$. In the latter form it is seen that one minimizes the least squares criterion for the original data y_i, but including the known constant $\hat{\eta}^{(l-1)}(\mathbf{x}_i)$ in the fit.

The iterative fitting with an offset offers a way to obtain boosting estimates also for GLMs. A GLM is in particular determined by the structure $\mu_i = E(y_i|\mathbf{x}_i) = h(\eta_i)$, where $h(\cdot)$ is a known response function and the linear predictor has the form $\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$. One difference between the L_2 boost and a GLM boosting is that in the boosting step one cannot fit a GLM to the residuals because, for example, with binary data, residuals are not from $\{0, 1\}$. The role of the residuals is taken by the offset.

Typically, the boosting algorithm is repeated for a large predefined number of steps l_{stop}. After the end of the algorithm, an appropriate criterion is used to determine the optimal number of steps l_{opt}. This can either be done by information criteria such as the Akaike (AIC) or Bayesian (BIC) or by the method of cross-validation. For the example of the linear model, this corresponds to a model selection between l_{stop} possible models. The first model simply represents a null model where no covariates are included. With every boosting step, a new covariate is added or (if the respective covariate has been selected before) the parameter of a covariate is updated. As the base learners are assumed to be 'weak', successive models only differ slightly from each other. This makes it more likely for the optimal model to be found. Implicitly, this model selection corresponds to variable selection. Typically, in the finally chosen model l_{opt}, not all of the possible predictors have been chosen and, therefore, are excluded from the final model. Thus, l_{opt} is the most important regularization parameter for the boosting algorithm. A quite different approach to bypassing the problem of overfitting is stability selection (see Section 4.4), which will be applied to our DIFboost algorithm.

4. Boosting in DIF

4.1. The DIF model as a GLM

The Rasch model and also the more general DIF model (3) can be embedded into the GLM framework. Let the data be given by $(Y_{pi}, \mathbf{x}_p)$, $p = 1, \ldots, P$, $i = 1, \ldots, I$. For simplicity, we use the notation $\mathbf{1}^T_{P(p)} = (0, \ldots, 0, 1, 0, \ldots, 0)$ and $\mathbf{1}^T_{I(i)} = (0, \ldots, 0, 1, 0, \ldots, 0)$, where $\mathbf{1}_{P(p)}$ and $\mathbf{1}_{I(i)}$ have lengths $P-1$ and I and have the value 1 at positions p and i, respectively. Therefore, the vectors are constructed in such a way that they can be seen as dummy variables for the corresponding persons and items, respectively. Then, model (3) can be represented as

$$\begin{aligned}\log\left(\frac{P(Y_{pi}=1)}{P(Y_{pi}=0)}\right) &= \theta_p - \beta_i - \mathbf{x}_p^T \boldsymbol{\gamma}_i \\ &= \mathbf{1}^T_{P(p)}\boldsymbol{\theta} - \mathbf{1}^T_{I(i)}\boldsymbol{\beta} - \mathbf{x}_p^T\boldsymbol{\gamma}_i = \mathbf{z}_{pi}^T\boldsymbol{\alpha}.\end{aligned} \tag{4}$$

Here, $\boldsymbol{\alpha}^T = (\boldsymbol{\theta}^T, \boldsymbol{\beta}^T, \boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_I^T)$ denotes the complete parameter vector containing $\boldsymbol{\theta}^T = (\theta_1, \ldots, \theta_{P-1})$ and $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_I)$. The design vector for person p and item i is denoted by $\mathbf{z}_{pi}^T = (\mathbf{1}^T_{P(p)}, -\mathbf{1}^T_{I(i)}, 0, \ldots, 0, -\mathbf{x}_p^T, 0, \ldots, 0)$. In $\mathbf{z}_{pi}$, the position of the component $-\mathbf{x}_p$ corresponds to the parameter $\boldsymbol{\gamma}_i$ in $\boldsymbol{\alpha}$.

In general, model (4) represents the structural component of a GLM for binary response with logit link. GLMs are extensively investigated in McCullagh and Nelder (1989), and introductions with the focus on categorical data can be found in Agresti (2002) and Tutz (2012).

Of course, the regular Rasch model can also be represented in the GLM framework by

$$\log\left(\frac{P(Y_{pi}=1)}{P(Y_{pi}=0)}\right) = \theta_p - \beta_i = \mathbf{1}^T_{P(p)}\boldsymbol{\theta} - \mathbf{1}^T_{I(i)}\boldsymbol{\beta}, \tag{5}$$

where the design vector and the parameter vector reduce to $(\mathbf{1}_{P(p)}, -\mathbf{1}_{I(i)})$ and $(\boldsymbol{\theta}^T, \boldsymbol{\beta}^T)$, respectively.

4.2. The DIFboost algorithm

The objective of our approach is to detect DIF by boosting the logistic DIF model (3). Because selection refers to DIF effects only, it is sensible to start the boosting selection procedure after the basic Rasch model has been fitted. The initial step is to fit the regular Rasch model (5). This step results in parameter estimates for the person and item parameters. The model fit from this first step is used as starting point for further steps where boosting techniques are used to select potential DIF effects. A similar approach was used by Boulesteix and Hothorn (2010) in a quite different context. In what follows, our algorithm is described in detail.

The starting point for the algorithm is to fit a regular Rasch model to our data. This is done by embedding the Rasch model into the logistic regression model (5). It can be estimated by standard software, such as the R function `glm` (R Core Team, 2014). Then one obtains estimates $\hat{\boldsymbol{\theta}}^T = (\hat{\theta}_1, \ldots, \hat{\theta}_{P-1})$ for the person parameters and $\hat{\boldsymbol{\beta}}^T = (\hat{\beta}_1, \ldots, \hat{\beta}_I)$ for the item difficulties. For a single observation, a linear predictor $\hat{\eta}_{pi} = \hat{\theta}_p - \hat{\beta}_i$ can be calculated which can be used to predict the probability of person p scoring on item i as

$$P(Y_{pi} = 1) = \frac{\exp(\hat{\eta}_{pi})}{1 + \exp(\hat{\eta}_{pi})}.$$

The linear predictors from the Rasch model for all person-item combinations are collected in $\hat{\boldsymbol{\eta}}_{\mathrm{RM}} = (\hat{\eta}_{11}, \hat{\eta}_{12}, \ldots, \hat{\eta}_{\mathrm{IP}})$ and are passed on to the subsequent steps of the algorithm.

For the boosting steps, the Rasch model (2) is extended to the more general DIF model (3). The parameters of the DIF model determine the base learners that are used. In our case, the model consists of three components, namely the person parameters, the item parameters, and the item-specific covariate parameters. Therefore, each of these components serves as a possible base learner:

$$\tilde{\eta}(\mathbf{x}_p, p, i) = \begin{cases} \tilde{\theta}_p, & p = 1, \ldots, P-1 \\ \tilde{\beta}_i, & i = 1, \ldots, I \\ \mathbf{x}_p^T \tilde{\gamma}_i, & i = 1, \ldots, I. \end{cases} \tag{6}$$

It is noteworthy that all base learners are linear. Nevertheless, they refer to different types of components that contain differing numbers of parameters (e.g., $\tilde{\gamma}_i$ vs. $\tilde{\beta}_i$). In cases like this, it is essential to ensure that all base learners share the same complexity so that the chances of being chosen are balanced. The complexity of base learners is determined by their degrees of freedom, which can be adapted by using internal penalty terms. In the case of linear base learners typically ridge penalties are used. Therefore, all the base learners presented above are restricted to have one degree of freedom by applying a ridge penalty when fitting the model. For more details on the complexity of base learners, see Hofner, Hothorn, Kneib, and Schmid (2011).

In every boosting step, only one of the base learners is updated, namely the one which yields the strongest reduction of an adequate loss function. The loss function that is used,

$$L(Y_{pi}, \tilde{\pi}_{pi}) = -(Y_{pi} \log(\tilde{\pi}_{pi}) + (1 - Y_{pi}) \log(1 - \tilde{\pi}_{pi})), \tag{7}$$

is the negative log-likelihood of a logit model with binary response. For boosting step l, this can be denoted by

$$\tilde{\eta}^*(\mathbf{x}_p, p, i) = \underset{\tilde{\theta}_p, \tilde{\beta}_i, \mathbf{x}^T\tilde{\gamma}_i}{\arg\min} \sum_{p,i} L(Y_{pi}, \tilde{\pi}_{pi}),$$

where the fitted probability $\tilde{\pi}_{pi}$ is calculated by fitting the model

$$\tilde{\pi}_{pi} = \frac{\exp(\tilde{\eta}^{(l)})}{\exp(1 + \exp(\tilde{\eta}^{(l)}))}, \quad \text{with predictor } \tilde{\eta}^{(l)} = \tilde{\eta}^{(l-1)} + \tilde{\eta}(\mathbf{x}_p, p, i),$$

separately for every base learner from (6).

The estimates for the single candidates of the base learner are obtained by fitting logit models where the linear predictor from the current model fit is used as known offset and the respective base learner is the only predictor. Therefore, based on the current model fit, in every step only the base learner with the highest information gain is updated. An additional parameter ν, $0 < \nu < 1$, regulates the step size of the parameter updates. It is chosen sufficiently small (typically $\nu = .1$) and only allows for small changes in every step. The parameter ν makes the base learners ‘weak’ and is used to prevent quick overfitting. This procedure is repeated for a predefined number of steps l_{stop}.

For the first boosting step, the offset is chosen to be the linear predictor $\hat{\boldsymbol{\eta}}_{\text{RM}}$ from the Rasch model, $\tilde{\eta}^{(0)} = \hat{\boldsymbol{\eta}}_{\text{RM}}$. This provides two advantages: First, the person parameters θ and item parameters β are, in contrast to the item-specific covariate parameters γ_i, essential for the interpretability of model (3). Therefore, it is sensible to prevent those parameters from being excluded from the model. From this point of view, the offset provides starting values for the person and item parameters. Second, the object of our approach is to detect the improvement of the model fit by extending the Rasch model to the DIF model. Therefore, we start from the model fit of the regular Rasch model. The boosting steps (possibly) add the information from the covariates. At some point during the boosting procedure, it can become necessary to adapt the person or the item parameters. Consequently, they can also be chosen as base learners within the boosting algorithm.

Typically, the model fitted after l_{stop} steps is overfitted and, therefore, not desirable. Two different strategies exist to finally identify the optimal model. One possibility is early stopping. Here, an optimal boosting step l_{opt} has to be found, either by an information criterion or by cross-validation. By early stopping, the boosting algorithm has the desirable effect of variable selection. The final model will only contain some of the possible parameters from model (3), namely, those that have at least once been found to be the best base learner before the optimal step l_{opt}. The second option is stability selection (discussed in detail in Section 4.4). For our analysis, we tried both early stopping using the BIC and stability selection with similar results. As stability selection provided slightly more stable results, the option of early stopping is omitted for the rest of this paper.

To conclude this subsection, the DIFboost algorithm outlined is briefly sketched below:

DIFboost

```
Step 1 (Initialization)
    Fit (5) for given scores $Y_{pi}$ and initialize the offset $\tilde{\eta}^{(0)} = \hat{\eta}_{\text{RM}}$.
    Initialize $\tilde{\theta}_p = 0, p = 1, \ldots, P-1, \tilde{\beta}_i = 0$ and $\tilde{\boldsymbol{\gamma}}_i = \mathbf{0}, i = 1, \ldots, I$.
    Set $l = 0$.
Step 2 (Iteration)
    $l \to l + 1$.
    Fit a logit model for every possible base learner where $\tilde{\eta}^{(l-1)}$ is used as offset.
    Select the best base learner $\eta^*(\mathbf{x}_p, p, \boldsymbol{i})$.
    Update the linear predictor by $\tilde{\eta}^{(l)} = \tilde{\eta}^{(l-1)} + \nu\tilde{\eta}^*(\mathbf{x}_p, p, \boldsymbol{i})$.
Step 3 (Stop)
    Iterate step 2 until $l = l_{\text{stop}}$ is reached.
```

4.3. Illustrative example

By way of illustration, first a single simulated data set will be considered. The data set is randomly drawn from setting 2 (medium) of the simulation study in Section 5.2. We have $P = 500$ persons, $I = 20$ items (four items with DIF, 16 without DIF) and $m = 5$ covariates; $l_{\text{stop}} = 500$ boosting steps are performed.

Figure 1 shows the coefficient paths along the boosting steps from $l = 0$ to $l = l_{\text{stop}} = 500$. The solid black lines represent the paths of the four DIF items, while the dashed grey lines represent the DIF-free items. Every item is represented by five paths because $m = 5$ covariates are used to find DIF. This makes the plot hard to digest as it is hard to distinguish between the different items. Figure 2 reduces the plot to one path per item. Here, a path represents the Euclidean norm of the item-specific parameter vectors γ_i of the corresponding item i. This plot is much clearer and easier to interpret than Figure 1, although some information is suppressed. The DIF items (black solid lines) can clearly be separated from the other items (dashed grey lines) because they are updated much earlier in the boosting algorithm and, therefore, seem to be much more informative for the response. The dashed vertical line represents the theoretically optimal model where all DIF items are in the model and all DIF-free items are excluded.

4.4. Stability selection

Choosing the optimal number of boosting steps via the BIC (or any other information criterion) has some drawbacks and may, therefore, not always be the best choice. One drawback is that the variable selection implied by the BIC can be unstable. Variables (or, more precisely, base learners) only have to be chosen in one single boosting step to be part of the final model. Therefore, it may happen that some items are diagnosed to have DIF although they have minimal coefficient estimates. This could lead to an increased false positive rate (FPR). Another drawback is that information criteria such as the BIC or AIC use the degrees of freedom. For example, the degrees of freedom can be determined using the hat matrix of the boosting algorithm, as proposed by Bühlmann and Hothorn (2007a) and Hofner *et al.* (2011). Yet this is very

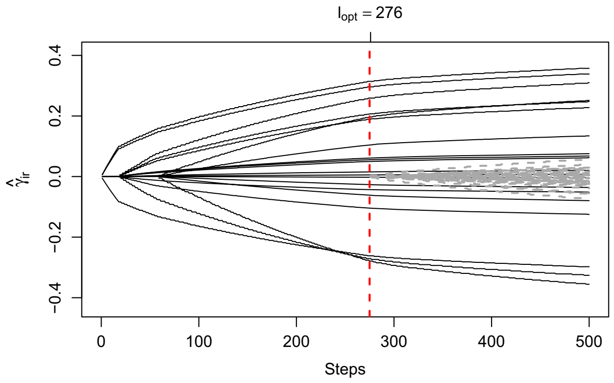


Figure 1. Boosting paths of item-specific parameters $\hat{\gamma}_{ir}$ for the example data set; solid paths represent differential item functioning (DIF) items, dashed paths represent non-DIF items; the dashed vertical line represents the theoretically optimal boosting step l_{opt}.

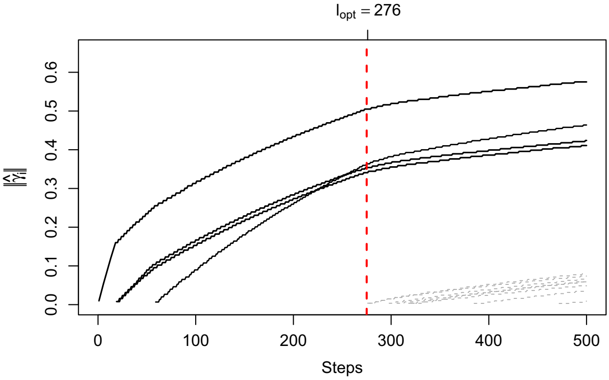


Figure 2. Boosting paths of Euclidian norms of item-specific parameter vectors $\hat{\gamma}_i$ for the example data set; solid paths represent differential item functioning (DIF) items, dashed paths represent non-DIF items; the dashed vertical line represents the theoretically optimal boosting step l_{opt}.

time-consuming and has also led to controversial methodological discussions; see Hastie (2007) and Bühlmann and Hothorn (2007b).

These drawbacks can be avoided by the concept of stability selection which was developed by Meinshausen and Bühlmann (2010). This is a very general approach which can be applied to a broad range of methods that include variable selection. It is based on the common idea of model/variable selection by subsampling. This can be computationally beneficial because it allows for parallelized computations. Furthermore, it addresses the problem of unstable variable selection by pooling over many subsamples.

For the DIF model (3), stability selection can be obtained in the following way. For a predefined number of replications B, $\lfloor P/2 \rfloor$ persons are drawn randomly from the original data set. The data set for one replication consists only of the observations in this subsample of persons. For each of the subsamples, the boosting algorithm is executed until l_{stop}. Then one counts how often a specific base learner was selected at each specific step $l = 0, \ldots, l_{\text{stop}}$. This gives the probabilities $\hat{\Pi}_i^l$ (or rather the relative frequencies over the B replications) of the base learner i being in the model at a specific boosting step l. The probabilities are illustrated by so-called stability paths along the boosting steps, as displayed in Figure 3. Finally, all base learners are selected with stability paths beyond a certain threshold value. These base learners represent the most frequent elements within the selected active set and, therefore, have to be considered as influential. In our application, we want to know which items have DIF and, therefore, we are only interested in the stability paths for γ_i for all items.

Stability selection is mainly determined by two parameters. The first parameter is q, which denotes how many distinct base learners are taken into the model when boosting the subsamples. As soon as q base learners have been selected, the procedure is stopped for the respective subsample. If fewer than q base learners are selected at $l = l_{\text{stop}}$, l_{stop} has to be increased. In the following, we choose 60% as a reasonable upper bound of the percentage of DIF items within a test and, therefore, $q = .6 \cdot I$. The second parameter is the threshold value for the selection probabilities of the single base learners, which is denoted by π_0. It is used to finally determine the set $\hat{S}^{\text{stable}}$ of stable base learners. This set is defined by

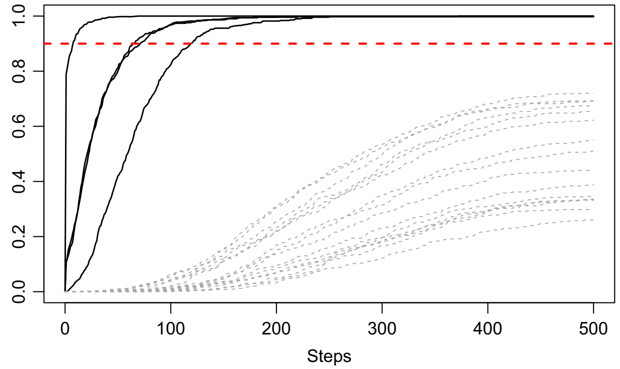


Figure 3. Stability paths for the example data set; solid paths represent differential item functioning (DIF) items, dashed paths represent non-DIF items; the dashed horizontal line represents the threshold values $\pi_0 = .9$.

$$\hat{S}^{\text{stable}} = \left\{ i : \max_{l=1,\ldots,l_{\text{stop}}} \left(\hat{\Pi}_i^l\right) \geq \pi_0 \right\}.$$

According to Meinshausen and Bühlmann (2010), the threshold value should be chosen within the range $\pi_0 \in (.6, .9)$, also depending on the choice of q and the desired sparseness of the final model.

Although two parameters have to be determined in advance, stability selection proved to be very stable. The choice of q in particular turned out not to be crucial as long as it is chosen in a reasonable range. The main tuning parameter of the procedure is the threshold parameter π_0. In our analysis, $\pi_0 = .9$ turned out to be a good choice. The threshold parameter π_0 is comparable to the level of significance in test-based procedures. In the simulation studies presented in the following section, $\pi_0 = .9$ led to FPRs of about 5% if no DIF was present, which is a popular choice for the level of significance in test-based procedures.

We use stability selection as a method of variable selection, but it does not provide parameter estimates. Estimates for the identified DIF effects are obtained by fitting a final DIF model for the selected items by ML estimation. By way of illustration, Figure 3 shows the stability paths for the simulated data set from Section 3 where four out of 20 items have DIF. We used $q = .6{\cdot}I = 12$ and $B = 500$ subsamples. The stability paths for the four DIF items are drawn with solid lines. They can clearly be separated from the stability paths of the DIF-free items which are drawn with dashed lines. The threshold value $\pi_0 = .9$ is depicted by a dashed horizontal line. With the given threshold value, all DIF items are identified; all DIF-free items are not selected.

4.5. Identifiability

Without any further constraints, the DIF model (3) is not identifiable. If person p tries to solve item i, let the linear predictor be denoted by $\eta_{pi} = \theta_p - \beta_i - \mathbf{x}_p^T\gamma_i$. We set $\theta_P = 0$, which is a common constraint to obtain identifiability in simple Rasch models. However, in the DIF model a fixed vector $\mathbf{c}$ allows us to reparameterize the linear predictor to obtain

$$\eta_{pi} = \theta_p - \beta_i - \mathbf{x}_p^T \gamma_i = \underbrace{\theta_p - \mathbf{x}_p^T \mathbf{c}}_{\tilde{\theta}_p} - \beta_i - \mathbf{x}_p^T \underbrace{(\gamma_i - \mathbf{c})}_{\tilde{\gamma}_i}.$$

Thus, the parameter sets $\{\theta_p, \beta_i, \gamma_i\}$ and $\{\tilde{\theta}_p, \beta_i, \tilde{\gamma}_i\}$ describe the identical model. This identification problem could be solved by restricting at least one item (the so-called reference item R) to have parameters $\gamma_R = \mathbf{0}$. But, by definition this item cannot have DIF and, therefore, would have to be chosen carefully. In particular, the choice of the reference item (or the corresponding $\mathbf{c}$) determines how many items show DIF (see also Tutz & Schauberger, 2015). A sensible strategy is to select the constraints in such a way that only few items show DIF. In this respect the boosting approach offers a natural solution. The starting point of the algorithm is the Rasch model and, therefore, the best model fit if no DIF is permitted. Step by step, the DIF parameters are updated. During the boosting algorithm, every item which has not yet been chosen as a DIF item can be used as reference item. Therefore, the models are identifiable as long as at least one item is left out. In practice, one of the left-out items is chosen to be the reference item R, and, for reasons of simplicity, we then use the additional restriction $\beta_R = 0$ instead of $\theta_P = 0$.

5. Simulation study

A simulation study is carried out to illustrate the performance of the method in terms of identification of DIF items. First, the method is compared to established methods of DIF detection. This is done by simulation settings with only one binary or multi-categorical covariate which can also be handled by existing methods. The second part of the simulation deals with settings with several (both continuous and categorical) covariates. These settings cannot be compared directly to established methods and are compared to the recently published DIFlasso approach of Tutz and Schauberger (2015).

5.1. Comparison to established methods

5.1.1. Methods

Typically, in the literature DIF is considered only for two groups, namely a reference group and a focal group. The standard method for this purpose is the MH method proposed by Holland and Thayer (1988). The methods involve computing a χ^2 test that compares the performances of the groups separately for all items, conditional on the total test score.

Alternative methods include Lord's χ^2 test (Lord, 1980) and the logistic regression method (Swaminathan & Rogers, 1990). In Lord's χ^2 test, for each group the parameters are estimated separately. Afterwards, a χ^2 test is used that tests the null hypothesis of equal item parameters for both groups. The logistic regression method for the detection of uniform DIF uses the model

$$\log\left(\frac{P(Y_{pi} = 1)}{P(Y_{pi} = 0)}\right) = \beta_0 + \beta_1 s_p + \beta_2 x_p \tag{8}$$

for every item i, where s_p is the total test score of person p and x_p encodes the group membership. Uniform DIF is tested by a likelihood ratio test (α = .05) on the null hypothesis $H_0 : \beta_2 = 0$. Model (8) can be extended by including a parameter for the interaction between the total test score and the group membership. This parameter could be used to test for non-uniform DIF. As the focus of this work is on uniform DIF, this extension will not be considered here.

For the more general case of multi-group comparisons, the methods presented have been extended by Somes (1986) and Penfield (2001) for MH, Kim *et al.* (1995) for Lord's χ^2 test, and Magis *et al.* (2011) for logistic regression.

All results from the present paper, including this simulation study, have been obtained by the R statistical software (R Core Team, 2014). The three reference methods for the simulation study are implemented in the add-on package `difR`; see Magis *et al.* (2010, 2013). A significance level of $\alpha = .05$ was chosen for all tests.

5.1.2. Settings

The simulation study encompasses five different settings. Each setting is performed for different strengths of DIF, where the strength is measured by

$$\frac{1}{I_{\text{DIF}}}\sum_{i=1}^{I_{\text{DIF}}}\left(\frac{1}{m}\sqrt{\sum_{j=1}^{m}\gamma_{ij}^2}\right),$$

and I_{DIF} encodes the number of DIF items. The term $\sum_{j=1}^{m}\gamma_{ij}^2$ represents the variance of the item difficulties $\beta_i + \mathbf{x}_p^T\gamma_i$ for standardized covariates, where m again encodes the number of covariates. Therefore, the DIF strength in the simulations is measured as the mean of the variance of the item difficulties while accounting for the number of covariates. For details on measuring the DIF strength, see Tutz and Schauberger (2015). The DIF strength in the simulation varies between .3 (very strong), .15 (strong), .1125 (medium), and .075 (weak).

For each setting, $P = 500$ persons and $I = 20$ items were generated; abilities θ and difficulties β were drawn from standard normal distributions. The number of groups and the number of DIF items are varied. Below we present the five settings used and the parameters used for 'strong' DIF (for a different DIF strength the parameters are simply multiplied by an appropriate factor):

1. $I_{\text{DIF}} = 4$ DIF items, $k = 2$ groups, $\gamma_1 = .15$, $\gamma_2 = -.15$, $\gamma_3 = .1$, $\gamma_4 = -.2$, $\gamma_5, \ldots, \gamma_{20} = 0$;
2. $I_{\text{DIF}} = 8$ DIF items, $k = 2$ groups, $\gamma_1 = \gamma_5 = .15$, $\gamma_2 = \gamma_6 = -.15$, $\gamma_3 = \gamma_7 = .1$, $\gamma_4 = \gamma_8 = -.2$, $\gamma_9, \ldots, \gamma_{20} = 0$;
3. as setting 1, but the abilities are highly correlated with the group membership: $\theta_i|x_i = 0 \sim N(0, 1)$, $\theta_i|x_i = 1 \sim N(1, 1)$;
4. $I_{\text{DIF}} = 4$ DIF items, $k = 5$ groups, $\gamma_1 = (.4, 0, .3, -.3)$, $\gamma_2 = (.5, .4, -.2, 0)$, $\gamma_3 = (0, -.2, .4, .3)$, $\gamma_4 = (-.2, .4, 0, .4)$, $\gamma_5 = \ldots = \gamma_{20} = (0, 0, 0, 0)$;
5. $I_{\text{DIF}} = 8$ DIF items, $k = 5$ groups, $\gamma_1 = \gamma_5 = (.4, 0, .3, -.3)$, $\gamma_2 = \gamma_6 = (.5, .4, -.2, 0)$, $\gamma_3 = \gamma_7 = (0, -.2, .4, .3)$, $\gamma_4 = \gamma_8 = (-.2, .4, 0, .4)$, $\gamma_9 = \ldots = \gamma_{20} = (0, 0, 0, 0)$.

In addition, the general settings 1, 3, and 4 were run under the assumption that no DIF is present ($I_{\text{DIF}} = 0$). The only difference between the corresponding settings 1 and 3 is that in setting 3 the abilities correlate with the group membership.

5.1.3. Results

For every setting, 100 replications were simulated. Table 1 shows the results for DIFboost ($q = 12$ and $\pi_0 = .9$) and the three reference methods in terms of true positive rate (TPR) and FPR. The TPR is determined by the rate of correctly identified DIF items. Therefore, higher values represent better performance. The FPR represents the rate of DIF-free items which have been assigned to be DIF items by mistake. Higher values represent worse performance.

For weak or medium DIF in settings 1–3, DIFboost outperforms MH and Lord in terms of TPR, with similar FPR. Logistic regression shows both higher TPR and FPR. For strong and very strong DIF, DIFboost shows lower FPR than the competitors. In the multi-group settings 4 and 5, DIFboost again shows very low FPR but also partly lower TPR. All in all, all methods show rather similar results; DIFboost compares well to its competitors. This also holds for the settings where no DIF is present. Again, Lord shows the lowest FPR and does not come close to attaining the intended α-level of 5%.

As an additional investigation, we compare the methods by using receiver operating characteristic (ROC) curves where the TPR is plotted against the FPR; see also Magis *et al.* (2015) who also use ROC curves as diagnostic tools. For that purpose, the settings presented above were used, but with varying parameters. For the reference methods we varied the level of significance, while for DIFboost we varied the threshold parameter π_0. The goal was to provide a comparison of the methods that is not confounded by the choice of these parameters. Figure 4 shows the ROC curves for all weak settings; the ROC curves for the other strengths show similar tendencies and are omitted for the sake of brevity. Again, it can be seen that in general the performance of all methods is very similar. Two tendencies can be seen from the curves. First, DIFboost tends to handle situations with many DIF items better than its competitors. Consequently, it outperforms its competitors in setting 2 and especially in setting 5. Relative to its competitors, it improves from setting 1 to setting 2 and also from setting 4 to setting 5 when eight items rather than four have DIF. Second, DIFboost seems to perform better in more complex situations with more than two groups. Relative to its competitors, it improves from setting 1 to setting 4 and from setting 2 to setting 5 (when $k = 2$ is changed to $k = 5$). Finally, in setting 5 (with both $k = 5$ and $I_{\mathrm{DIF}} = 8$) DIFboost clearly outperforms its competitors.

5.2. Simulations with many covariates

5.2.1. Methods

As DIFboost can include many covariates at the same time and is able to handle continuous covariates, the method can be used in much more general settings than explored in the previous subsection. Here, we present a simulation study for settings where several possibly DIF-inducing covariates are available. The reference methods from the previous subsection cannot be used in these situations. Consequently, we compare the methods to the DIFlasso method of Tutz and Schauberger (2015).

The method of Rasch trees (Strobl, Kopf, & Zeileis, 2015) can also handle several (possibly continuous) variables simultaneously. It only provides groups within the respondents with equal item parameters. It does not provide an actual identification of DIF items as, between different groups, all item parameters are different. Therefore, this method cannot be used for comparison when it comes to identification of DIF items and will not be used in the simulation study.

Table 1. True positive rates (TPR) and false positive rates (FPR) from five different simulation settings comparing DIFboost to the reference methods Mantel–Haenszel (MH), Lord and logistic regression

Setting				DIFboost	MH	Lord	Logistic
		Very strong	TPR	.725	.765	.733	.810
			FPR	.030	.037	.025	.049
	P = 500	Strong	TPR	.305	.292	.260	.343
	I = 20		FPR	.041	.034	.026	.048
1	I_{DIF} = 4	Medium	TPR	.190	.168	.147	.203
	k = 2		FPR	.041	.034	.026	.046
		Weak	TPR	.117	.087	.085	.140
			FPR	.041	.037	.026	.048
	I_{DIF} = 0	No DIF	FPR	.041	.037	.024	.445
		Very strong	TPR	.705	.782	.757	.823
			FPR	.019	.044	.033	.051
	P = 500	Strong	TPR	.281	.300	.258	.347
2	I = 20		FPR	.029	.034	.026	.047
	I_{DIF} = 8	Medium	TPR	.198	.179	.161	.217
	k = 2		FPR	.036	.035	.027	.045
		Weak	TPR	.114	.095	.080	.133
			FPR	.040	.037	.024	.042
		Very strong	TPR	.677	.685	.692	.735
			FPR	.034	.044	.031	.062
	P = 500	Strong	TPR	.212	.195	.185	.258
	I = 20		FPR	.045	.041	.031	.059
3[a]	I_{DIF} = 4	Medium	TPR	.150	.128	.120	.170
	k = 2		FPR	.048	.040	.031	.059
		Weak	TPR	.075	.082	.065	.100
			FPR	.051	.043	.029	.059
	I_{DIF} = 0	No DIF	FPR	.048	.041	.029	.056
		Strong	TPR	.990	1.000	.993	1.000
	P = 500		FPR	.027	.049	.017	.058
	I = 20	Medium	TPR	.875	.910	.845	.927
4	I_{DIF} = 4		FPR	.026	.051	.015	.056
	k = 5	Weak	TPR	.570	.593	.470	.608
			FPR	.031	.049	.016	.053
	I_{DIF} = 0	No DIF	FPR	.047	.052	.017	.051
		Strong	TPR	.976	.999	.995	1.000
	P = 500		FPR	.008	.072	.027	.077
	I = 20	Medium	TPR	.866	.944	.884	.942
5	I_{DIF} = 8		FPR	.008	.062	.020	.063
	k = 5	Weak	TPR	.552	.624	.471	.645
			FPR	.012	.052	.017	.055

Note. DIF = differential item functioning.

[a]The person abilities for setting 3 are highly correlated with the group membership.

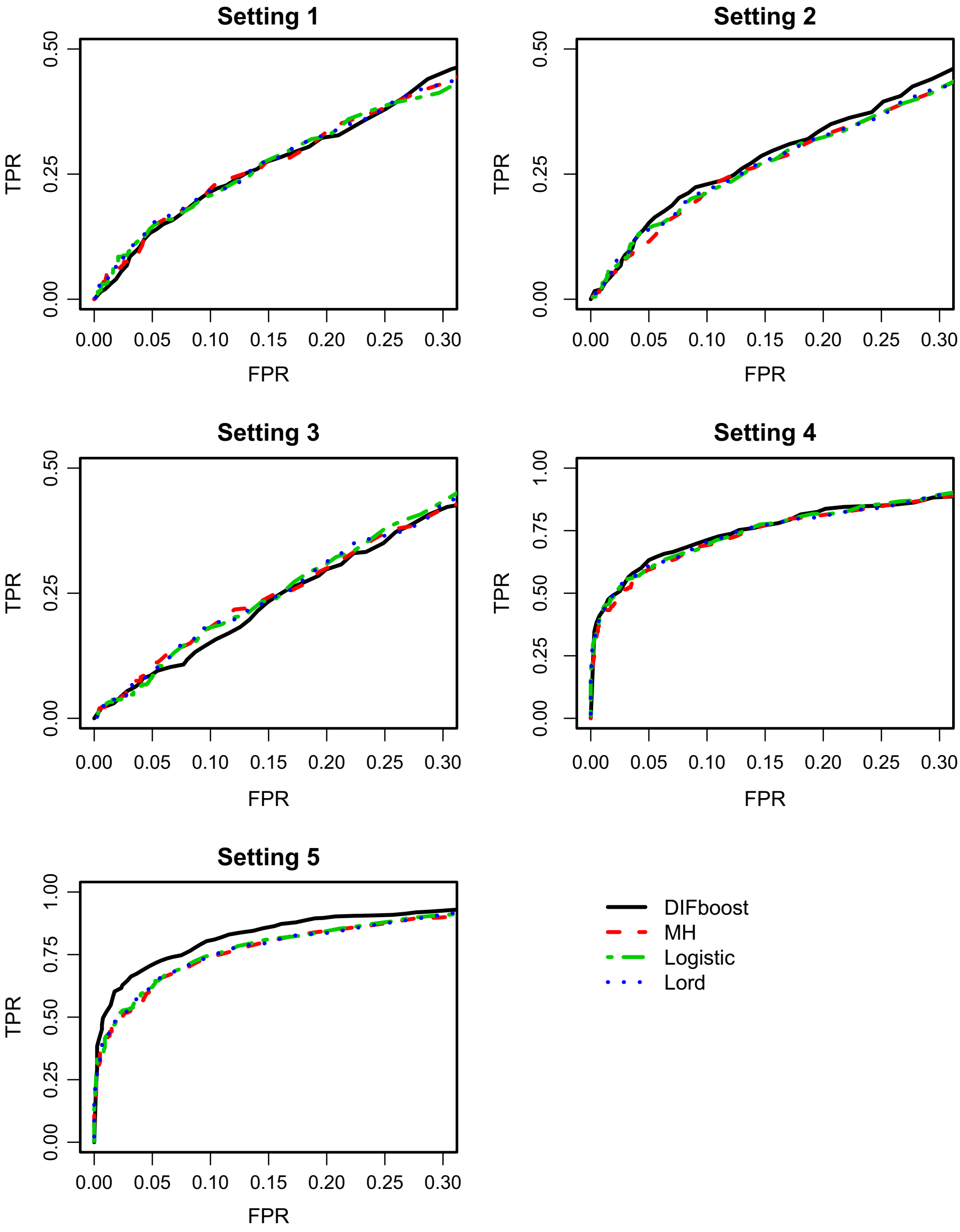


Figure 4. Receiver operating characteristic (ROC) curves for all weak settings in the simulation study comparing DIFboost to the reference methods Mantel–Haenszel (MH), Lord and logistic regression.

5.2.2. Settings

Four different settings are considered, each with $I = 20$ items and $m = 5$ covariates (two binary, three continuous). Again, abilities θ and difficulties β are drawn from standard normal distributions. The number of persons and the number of DIF items are varied. For each setting, ‘strong’, ‘medium’, and ‘weak’ DIF are used with DIF strengths .3, .15 and .1125. Below we present the four settings and the parameters used for ‘medium’ DIF (for a different DIF strength the parameters are simply multiplied by an appropriate factor):

1. $P = 250$ persons, $I_{\text{DIF}} = 4$ DIF items, $\gamma_1 = (-.5, .4, 0, 0, .5)$, $\gamma_2 = (0, .5, -.4, 0, .3)$, $\gamma_3 = (.4, 0, .5, -.5, 0)$, $\gamma_4 = (0, 0, .5, .4, -.2)$, $\gamma_5, \ldots, \gamma_{20} = (0, 0, 0, 0, 0)$;
2. as setting 1, but with $P = 500$ persons;
3. as setting 2, but with $I_{\text{DIF}} = 8$ DIF items, items 5–8 same as items 1–4;
4. as setting 2, but the abilities are highly correlated with the group membership, $\theta_i|x_i = 0 \sim N(0, 1), \theta_i|x_i = 1 \sim N(1, 1)$.

Again, for settings 1, 2, and 4 also settings with no DIF are simulated, where setting 2 differs from setting 4 as in the latter the abilities are correlated with the group membership.

5.2.3. Results

Table 2 shows the results for 100 replications of the different simulation settings in terms of TPR and FPR. For medium and especially for weak DIF, DIFboost clearly outperforms DIFlasso in terms of TPR. Also, DIFlasso shows increased FPR in some settings, whereas DIFboost is very stable regarding FPR. Therefore, DIFboost proved to be a very interesting alternative with regard to DIF detection for several covariates. For the settings with no DIF, it is no surprise that DIFlasso has a lower FPR than DIFboost. Still, the chosen parameters for DIFboost provide FPRs around 5% and, therefore, if no DIF is present the procedure can be compared to a test procedure with a level of significance $\alpha = .05$.

6. DIF in the Intelligence-Structure-Test 2000 R

In the following, the method is applied to data from the Intelligence-Structure-Test 2000 R (I-S-T 2000 R),[1] developed by Amthauer, Brocke, Liepmann, and Beauducel (2001). The test is a fundamentally revised version of its predecessors I-S-T 70 (Amthauer, Brocke, Liepmann, & Beauducel, 1973) and I-S-T 2000 (Amthauer, Brocke, Liepmann, & Beauducel, 1999). Generally, its aim is to measure deductive reasoning ability. It consists of three basic modules on verbal intelligence, numerical intelligence, and figural intelligence. Each of these modules is divided into three subtests, each consisting of 20 items. For example, the module for numerical intelligence consists of the subtests numerical calculations, number series, and numerical signs. Further details on the I-S-T 2000 R and its predecessors can be found, for example, in Schmidt-Atzert, Hommers, and Heß (1995), Brocke, Beauducel, and Tasche (1998), and Schmidt-Atzert (2002).

The data originate from a test on 273 students from different faculties from the University of Marburg, Germany, aged between 18 and 39 years. The data have already been analysed in Bühner, Ziegler, Krumm, and Schmidt-Atzert (2006), where they were used to test whether the I-S-T 2000 R is Rasch scalable using mixed Rasch models (Rost, 1990).

[1] Data from Testzentrale Göttingen; http://www.testzentrale.de.

Table 2. True positive rates (TPR) and false positive rates (FPR) for four different simulation settings comparing DIFboost to DIFlasso

Setting				DIFboost	DIFlasso
1	$P = 250$ $I = 20$ $I_{\mathrm{DIF}} = 4$ $m = 5$	Strong	TPR	1.000	1.000
			FPR	.024	.024
		Medium	TPR	.873	.228
			FPR	.028	.000
		Weak	TPR	.642	.030
			FPR	.029	.000
	$I_{\mathrm{DIF}} = 0$	No DIF	FPR	.053	.000
2	$P = 500$ $I = 20$ $I_{\mathrm{DIF}} = 4$ $m = 5$	Strong	TPR	1.000	1.000
			FPR	.011	.036
		Medium	TPR	1.000	.983
			FPR	.029	.004
		Weak	TPR	.948	.383
			FPR	.026	.000
	$I_{\mathrm{DIF}} = 0$	No DIF	FPR	.051	.000
3	$P = 500$ $I = 20$ $I_{\mathrm{DIF}} = 8$ $m = 5$	Strong	TPR	1.000	1.000
			FPR	.002	.118
		Medium	TPR	.990	.993
			FPR	.007	.021
		Weak	TPR	.900	.294
			FPR	.008	.001
4[a]	$P = 500$ $I = 20$ $I_{\mathrm{DIF}} = 4$ $m = 5$	Strong	TPR	1.000	1.000
			FPR	.016	.028
		Medium	TPR	.968	.890
			FPR	.031	.006
		Weak	TPR	.873	.228
			FPR	.033	.000
	$I_{\mathrm{DIF}} = 0$	No DIF	FPR	.065	.000

Note. DIF = differential item functioning.

[a] The person abilities for setting 4 are highly correlated with one of the binary covariates.

We will analyse the items of the sentence completion subtest from the verbal intelligence module. Three covariates were used as possibly DIF-inducing covariates, gender (0, male; 1, female), age (in years) and the interaction between gender and age.

Figure 5 shows the stability paths for DIFboost, where, in accordance with the simulation study, the parameters $q = .6{\cdot}I = 12$ and $\pi_0 = .9$ are chosen. It can be seen that four items (8, 9, 11, and 15) are identified as having DIF.

We illustrate the coefficients of the DIF items by effect stars (Tutz & Schauberger, 2013). Since the logit link is used, the exponentials of the coefficients represent the effects of the covariates on the odds $P(Y_{pi} = 1)/P(Y_{pi} = 0)$. The length of the rays corresponds to the exponentials of the respective coefficients. The circle around each star has a radius of $\exp(0) = 1$ and, therefore, represents the no-effect case. Both gender and age were standardized prior to the analysis so that the size of the coefficient estimates is

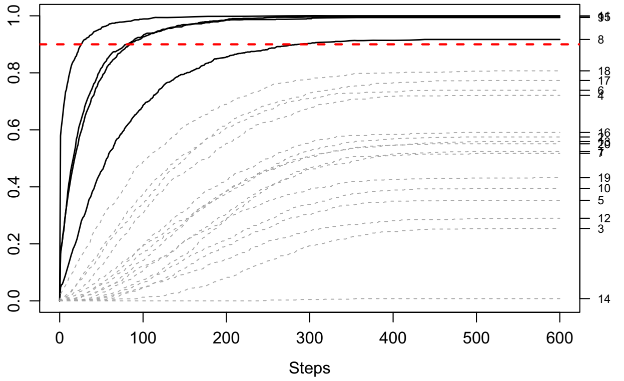
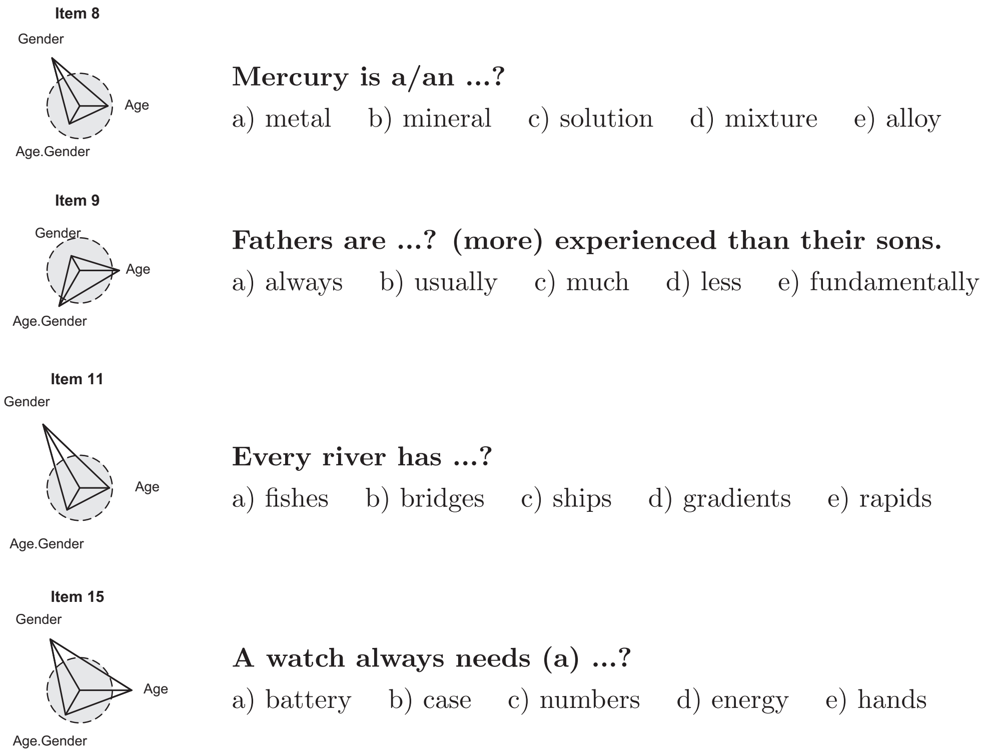


Figure 5. Stability paths for DIFboost for the items of the sentence completion subtest; the dashed line represents the threshold π_0 = .9; items 8, 9, 11 and 15 are diagnosed as differential item functioning (DIF) items.



Mercury is a/an ...?

a) metal b) mineral c) solution d) mixture e) alloy

Fathers are ...? (more) experienced than their sons.

a) always b) usually c) much d) less e) fundamentally

Every river has ...?

a) fishes b) bridges c) ships d) gradients e) rapids

A watch always needs (a) ...?

a) battery b) case c) numbers d) energy e) hands

Figure 6. Effect stars and item descriptions for items with differential item functioning (DIF) in the sentence completion subtest (IST 2000 R; Amthauer *et al.*, 2001) detected by DIFboost.

comparable. Figure 6 shows the effect stars for the estimated coefficients and the item descriptions of the DIF items.

Generally, a ray beyond the circle represents positive coefficients. With positive coefficients, the difficulty of the respective item is increased if the corresponding

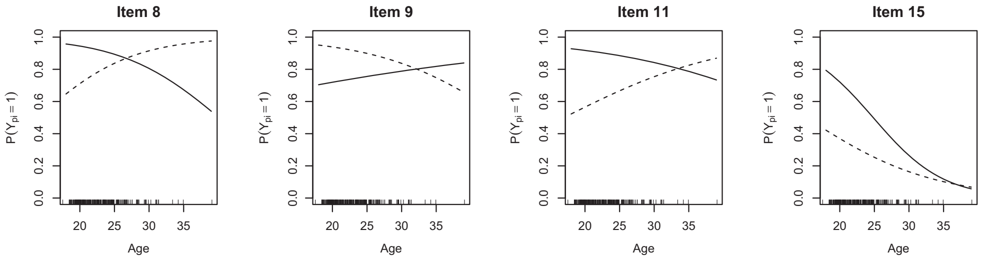


Figure 7. Probabilities of scoring on items depending on gender and age for all differential item functioning (DIF) items. Solid lines represent male, dashed lines represent female participants.

covariate is increased while the probability of solving the item is decreased. Item 9, for example, has a negative coefficient for gender. Therefore, this item is easier for female participants as female is encoded by 1. Since also the interaction between gender and age is considered, one has to look at all coefficients at once. With growing age, the difficulty increases for female participants.

Figure 7 shows for each DIF item the effects of both gender and age on the probability of scoring on the respective item. Separately for male (solid lines) and female (dashed lines) participants, the probability of scoring on the respective item is depicted along the covariate age. For simplicity, the plots refer to a person with a 'mean' ability according to the estimates of the θ parameters. Figure 7 demonstrates the effect of the interaction term. As the probabilities of scoring on an item can intersect, the main effects of age and gender should not be interpreted separately but always with respect to the interaction term. The ability to include interaction terms in this manner can be seen as a big improvement compared to existing methods of DIF detection, allowing for new insights into the occurrence of DIF. In extreme cases, both the main effects for gender and age could even be negligible but the interactions term could still be influential.

Therefore, item 9 cannot generally be assumed to be easier for female participants. This holds only for participants younger than 30 years, but the order changes for older participants. Items 11 and 15 are, in general, easier for male participants, in particular if they are rather young. For growing age this difference slowly vanishes; in item 11 the effect is even reversed for higher age.

For comparison, the data were also analysed with the method of Rasch trees (Strobl *et al.*, 2015). The corresponding Rasch tree is plotted in Figure 8.

By recursive partitioning of the covariate space, one tries to find groups within the observations which have the same item parameters. In our case, only one partition was found in the data, namely male and female participants. For age, no significant difference was found. When using recursive partitioning, the item parameters are estimated separately within the groups. The estimates are also shown in Figure 8. The estimates for the items diagnosed as DIF items from the other methods are highlighted. But, by far the highest difference between both groups seems to be for item 14 which is the hardest item for male participants and the easiest item for female participants. However, all other methods did not identify item 14 to be a DIF item. That means, Rasch trees may yield quite different results than other methods when trying to identify DIF items. To our knowledge, no systematic investigation that compares Rasch trees and alternative methods is available.

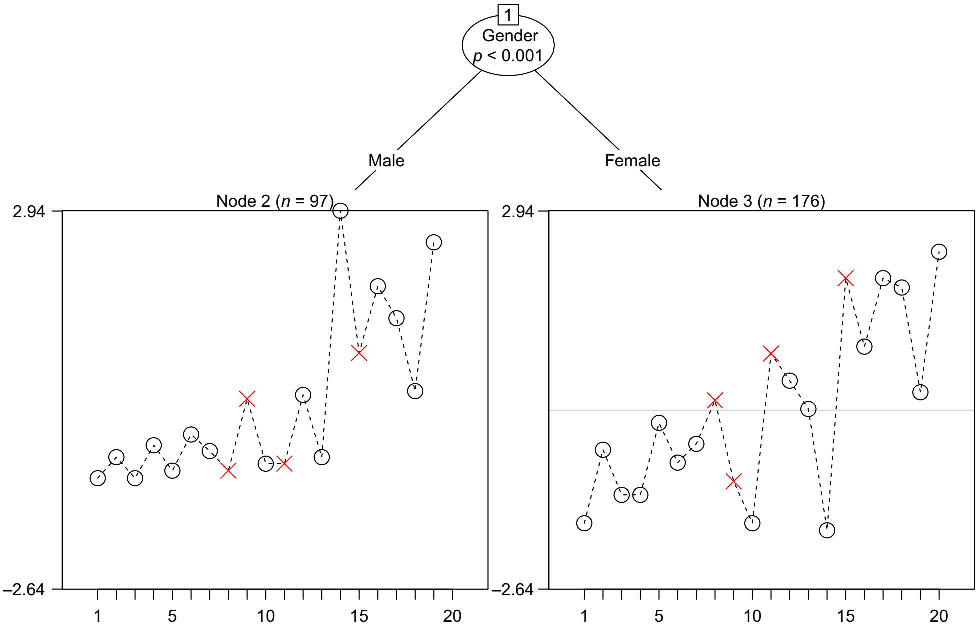


Figure 8. Rasch tree for sentence completion subtest. Highlighted items represent items diagnosed as differential item functioning (DIF) items by DIFboost.

7. Concluding remarks

A new method called DIFboost is proposed for detecting DIF induced by several covariates simultaneously. In the case of DIF in subgroups, the method competes well with established methods for DIF detection. For the more general case of several, possibly continuous covariates, it outperforms the competitive DIFlasso approach.

In contrast to the established test procedures, DIFboost is able to identify DIF items without the specification of anchor items. In most other methods, one assumes that the other items have no DIF and, therefore, all items besides the one investigated serve as anchor items. Besides this strategy, there exist other possibilities for finding anchor items; see, for example, Kopf, Zeileis, and Strobl (2014) or Woods (2009). The need for anchor items remains problematic, especially if many possible covariates have to be considered.

DIFboost is a model-based method. This provides two further advantages over test-based methods. First, the problem of multiple testing is avoided. Generally, DIF tests perform one test per item and covariate. A test is designed to restrict the probability of a Type I error to a certain level. If there are many covariates and many items, there are many tests and the problem of multiple testing arises. To control for this, correction strategies such as Bonferroni adjustment become necessary.

Second, unlike tests, the DIFboost method provides parameter estimates which allow for a deeper look into the data structure and gives interpretible results. The linear effects of the covariates can be complemented by the incorporation of interaction effects or by using smooth functions for the covariate effects. Therefore, model-based methods not only tell us which items have DIF but also provide valuable information about the underlying covariate effects.

For simplicity, the approach presented is limited to the Rasch model. However, extensions to other models are possible and should be investigated in future research. For example, the boosting algorithm seems well suited to an extension to the 2PL model. Parameter estimation in the 2PL model is rather complicated because of its multiplicative structure. By using boosting concepts, this problem can be tackled in a stepwise manner. In particular, the discrimination parameters can be used as further base learners that are updated only for those items that call for it.

References

Agresti, A. (2002). *Categorical data analysis*. New York, NY: Wiley.

Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (1973). *Intelligenz-Struktur-Test [Intelligence-Structure-Test] (IST 70)*. Göttingen, Germany: Hogrefe.

Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (1999). *Intelligenz-Struktur-Test 2000 [Intelligence-Structure-Test 2000] (IST 2000)*. Göttingen, Germany: Hogrefe.

Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (2001). *Intelligenz-Struktur-Test 2000 R [Intelligence-Structure-Test 2000 R] (IST 2000 R)*. Göttingen, Germany: Hogrefe.

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

Boulesteix, A.-L., & Hothorn, T. (2010). Testing the additional predictive value of high-dimensional molecular data. *BMC Bioinformatics*, *11*, 1–11. doi:10.1186/1471-2105-11-78

Brocke, B., Beauducel, A., & Tasche, K. (1998). Der Intelligenz-Struktur-Test: Analysen zur theoretischen Grundlage und technischen Güte [The Intelligence-Structure-Test: Analyses of the theoretical foundation and technical quality]. *Diagnostica*, *44*, 84–99.

Bühlmann, P. (2006). Boosting for high-dimensional linear models. *Annals of Statistics*, *34*, 559–583. doi:10.1214/009053606000000092

Bühlmann, P., & Hothorn, T. (2007a). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, *22*, 477–505. doi:10.1214/07-STS242

Bühlmann, P., & Hothorn, T. (2007b). Rejoinder: Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, *22*, 516–522. doi:10.1214/07-STS242RE

Bühlmann, P., & Yu, B. (2003). Boosting with the L_2 loss: Regression and classification. *Journal of the American Statistical Association*, *98*, 324–339. doi:10.1198/016214503000125

Bühner, M., Ziegler, M., Krumm, S., & Schmidt-Atzert, L. (2006). Ist der IST 2000 R Rasch-skalierbar? [Is the IST 2000 R Rasch-scaleable?] *Diagnostica*, *52*, 119–130. doi:10.1026/0012-1924.52.3.119

Freund, Y., & Schapire, R. E. (1996). *Experiments with a new boosting algorithm*. Proceedings of the Thirteenth International Conference on Medicine Learning (pp. 148–156). San Francisco, CA: Morgan Kaufmann.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*, 119–139. doi:10.1006/jcss.1997.1504

Friedman, J. H., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, *28*, 337–407.

Gonçalves, F., Gamerman, D., and Soares, T. (2013). Simultaneous multifactor DIF analysis and detection in item response theory. *Computational Statistics & Data Analysis*, *59*, 144–160. doi:10.1016/j.csda.2012.10.011

Hastie, T. (2007). Comment: Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, *22*, 513–515. doi:10.1214/07-STS242B

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Bias estimation for nonorthogonal problems. *Technometrics*, *12*, 55–67. doi:10.1080/00401706.1970.10488634

Hofner, B., Hothorn, T., Kneib, T., & Schmid, M. (2011). A framework for unbiased model selection based on boosting. *Journal of Computational and Graphical Statistics*, *20*, 956–971. doi:10.1198/jcgs.2011.09220

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel–Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum.

Holland, W., & Wainer, H. (Eds.) (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M., & Hofner, B. (2013). *mboost: Model-based boosting*. R package version 2.2-3.

Kim, S.-H., Cohen, A. S., & Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement*, *32*, 261–276.

Kopf, J., Zeileis, A., & Strobl, C. (2014). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*(1), 22–56. doi:10.1177/0013164414529792

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Magis, D., Beland, S., & Raiche, G. (2013). *difR: Collection of methods to detect dichotomous differential item functioning (DIF) in psychometrics*. R package version 4.4.

Magis, D., Béland, S., Tuerlinckx, F., & Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*, 847–862. doi:10.3758/BRM.42.3.847

Magis, D., Raîche, G., Béland, S., & Gérard, P. (2011). A generalized logistic regression procedure to detect differential item functioning among multiple groups. *International Journal of Testing*, *11*, 365–386. doi:10.1080/15305058.2011.602810

Magis, D., Tuerlinckx, F., & De Boeck, P. (2015). Detection of differential item functioning using the lasso approach. *Journal of Educational and Behavioral Statistics*, *40*, 111–135. doi:10.3102/1076998614559747

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719–748. doi:10.1093/jnci/22.4.719

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). New York, NY: Chapman & Hall.

Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society, Series B (Methodological)*, *72*, 417–473. doi:10.1111/j.1467-9868.2010.00740.x

Millsap, R., & Everson, H. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*, 297–334. doi:10.1177/014662169301700401

Paek, I., & Wilson, M. (2011). Formulating the Rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison with Mantel–Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement*, *71*, 1023–1046. doi:10.1177/0013164411400734

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel–Haenszel procedures. *Applied Measurement in Education*, *14*, 235–259. doi:10.1207/S15324818AME1403_3

R Core Team (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271–282. doi:10.1177/014662169001400305

Schmidt-Atzert, L. (2002). Intelligenz-Struktur-Test 2000 R (Testrezension) [Intelligence-Structure-Test 2000 R (test recension)]. *Zeitschrift für Personalpsychologie*, *1*, 50–56.

Schmidt-Atzert, L., Hommers, W., & Heß, M. (1995). Der I-S-T 70. Eine Analyse und Neubewertung [The I-S-T 70. An analysis and reassessment]. *Diagnostica*, *41*, 108–130.

Somes, G. W. (1986). The generalized Mantel–Haenszel statistic. *The American Statistician*, *40*, 106–108. doi:10.2307/2684866

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*, 289–316. doi:10.1007/s11336-013-9388-3

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361–370. doi:10.1111/j.1745-3984.1990.tb00754.x

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, *58*, 267–288. doi:10.1111/j.1467-9868.2011.00771.x

Tukey, J. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Tutz, G. (2012). *Regression for categorical data*. Cambridge, UK: Cambridge University Press.

Tutz, G., & Binder, H. (2006). Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics*, *62*, 961–971. doi:10.1111/j.1541-0420.2006.00578.x

Tutz, G., & Schauberger, G. (2013). Visualization of categorical response models: From data glyphs to parameter glyphs. *Journal of Computational and Graphical Statistics*, *22*(1), 156–177. doi:10.1080/10618600.2012.701379

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika*, *80*, 21–43. doi:10.1007/s11336-013-9377-6

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*(1), 42–57. doi:10.1177/0146621607314044

Received 9 December 2014; revised version received 9 January 2015