A regularization approach for the detection of differential item functioning in generalized partial credit models



Gunther Schauberger^{1,2} • Patrick Mair³

Published online: 18 March 2019 © The Psychonomic Society, Inc. 2019

Abstract

Most common analysis tools for the detection of differential item functioning (DIF) in item response theory are restricted to the use of single covariates. If several variables have to be considered, the respective method is repeated independently for each variable. We propose a regularization approach based on the lasso principle for the detection of uniform DIF. It is applicable to a broad range of polytomous item response models with the generalized partial credit model as the most general case. A joint model is specified where the possible DIF effects for all items and all covariates are explicitly parameterized. The model is estimated using a penalized likelihood approach that automatically detects DIF effects and provides trait estimates that correct for the detected DIF effects from different covariates simultaneously. The approach is evaluated by means of several simulation studies. An application is presented using data from the children's depression inventory.

Keywords Differential item functioning · DIF · Generalized partial credit model · Regularization · Lasso · GPCMlasso

Introduction

Whenever items are found that show different endorsement probabilities with respect to certain characteristics of participants with the same ability, this phenomenon is termed *differential item functioning* (DIF). Several procedures exist to detect DIF in dichotomous and polytomous items with respect to a single, categorical

During the work on this research project, Gunther Schauberger was supported by the grants SCHA2053/1-1 and TU62/8-1 from the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG).

Gunther Schauberger gunther.schauberger@tum.de

¹ Department of Sport and Health Sciences, Chair of Epidemiology, Technical University of Munich, Munich, Germany

- ² Department of Statistics, LMU Munich, Munich, Germany
- ³ Department of Psychology, Harvard University, Cambridge, MA, USA

covariate.¹ Among the most frequently used procedures for dichotomous items are the Mantel–Haenszel test (Mantel & Haenszel, 1959; Mantel, 1963; Holland & Thayer, 1988), the likelihood-ratio test (Thissen et al., 1993), and logistic regression (Swaminathan & Rogers, 1990; Zumbo, 1999). For polytomous items, Choi et al. (2011) proposed a hybrid approach that combines models from item response theory (IRT) and the method of logistic regression to identify DIF. Welch and Hoover (1993) and Chang et al. (1996) published earlier approaches on DIF in polytomous items.

The approach presented in this article overcomes several limitations of the DIF methods proposed in the literature. First, it works for dichotomous as well as polytomous items. Second, it allows researchers to include several covariates of potentially mixed scale levels simultaneously. Technically, it extends the idea proposed by Tutz and Schauberger (2015) where DIF detection in Rasch models is translated into a parameter selection problem. In their model, DIF effects are represented by parameters and L_1 penalties are used for variable selection. Such a model-based approach has the advantages that the number of DIF-inducing covariates is not limited and that both metric and categorical covariates

¹Many approaches restrict the covariate to be binary.

can be used simultaneously. A simultaneous treatment of several covariates becomes important in multicollinear situations, where applying a method repeatedly to different variables is not appropriate. Also, if one is interested in latent trait estimates that are corrected for possible DIF effects, simultaneous treatment of several variables is necessary. Current methods are only able to correct for DIF induced by one single variable. At the end of the manuscript, we present an application to data from the Children's Depression Inventory (Kovacs, 2004) as an example for possible cases where corrected trait estimates might replace sum scores when applying established scales in clinical psychology.

While the approach of Tutz and Schauberger (2015) is restricted to DIF detection based on the Rasch model for dichotomous items, the current work aims at the more general *generalized partial credit model* (GPCM) proposed by Muraki (1992). Therefore, their approach is generalized both to polytomous items and to the incorporation of different item discrimination parameters.

In addition, as Penfield (2007) and Penfield et al. (2009) pointed out, DIF analysis in polytomous settings can also suffer from the phenomenon of *differential step functioning* (DSF). DSF implies differences at the level of each step of a polytomous item with respect to different covariate values.

Komboz et al. (2018) proposed an alternative to conventional test procedures using recursive partitioning to detect DIF and DSF for polytomous items. The big advantage of recursive partitioning is that it allows to include metric covariates without the need to discretize them beforehand. Instead, the splits in the variables are found to be data-driven. Also, the method allows to include several covariates simultaneously. After all, the focus of the approach is somewhat different because it attempts to find groups of persons where DIF is present while our approach will focus on the detection of DIF items. Bollmann et al. (2018) also proposed a tree-based approach, but here for each item where DIF is detected a separate tree is built. In both cases, no differing discrimination parameters across items are allowed.

The paper is structured as follows. In Section "The generalized partial credit model and its special cases" we introduce IRT models which fit into the GPCM class. In Section "A DIF model for generalized partial credit models" we develop our regularization-based DIF approach for this general model class, which we call *GPCMlasso*. Subsequently, we give details on the penalized likelihood approach for parameter estimation, including the determination of the optimal tuning parameter and posterior trait estimation. Section "Simulation study" compares the GPCMlasso to alternative DIF methods by means of several simulation studies. Finally, in Section "Application to youth depression data", the method is applied to data from the

Children's Depression Inventory (CDI). The GPCMlasso approach is implemented in an R package (R Core Team, 2018) of the same name, available from the CRAN repository.

The generalized partial credit model and its special cases

Let us formally introduce the GPCM as the most general model to be considered in our DIF approach. Let $Y_{pi} \in$ $\{0, 1, ..., k_i\}$ denote the response of person $p, p \in$ $\{1, ..., P\}$, to item $i, i \in \{1, ..., I\}$. Each item i can differ with respect to the number of response categories k_i . Then, the GPCM can be denoted by

$$\log\left(\frac{P(Y_{pi}=r)}{P(Y_{pi}=r-1)}\right) = \beta_i(\theta_p - \delta_{ir}), \quad r = 1, \dots, k_i.$$
(1)

The model consists of three different types of parameters: the trait parameters θ_p , the item step parameters δ_{ir} , and the item discrimination parameters β_i . The trait parameters θ_p measure the latent trait one wants to capture with the items and scales the persons with respect to that trait. In contrast, δ_{ir} and β_i characterize the different items: δ_{ir} $(r = 1, ..., k_i)$ is the location of step *r* in item *i*, whereas β_i denotes the discrimination of item *i*.

The model needs two restrictions to be identifiable, one for location and one for scale. Commonly, for the person/trait parameters θ_p the assumption of a normal distribution is made. This assumption will later be used to estimate the model via marginal maximum likelihood (MML) estimation. To ensure identifiability, for the sequel of the paper we will assume $\theta_p \sim N(0, 1)$.

All further models considered in this paper can be seen as special cases of the GPCM. For example, the GPCM simplifies to the *partial credit model* (PCM) as proposed by Masters (1982) if all discrimination parameters are restricted to be equal, $\beta_1 = \ldots = \beta_I = \beta$. Here, one assumes that all items share an exactly equal discriminatory power.

For polytomous items, alternatively the *rating scale* model (RSM; Andrich 1978) and, in case of item-specific discrimination parameters, the generalized rating scale model (GRSM; Muraki 1990) can be chosen. Both RSM and GRSM are restricted to the case of equal numbers of response categories across all items where $k_1, = \ldots = k_I = k$. In addition, as opposed to the (G)PCM, the differences between coefficients for adjacent steps in the items are restricted to be equal. Therefore, the restriction $\delta_{ir} - \delta_{is} = \delta_{i^*r} - \delta_{i^*s}$ for $i, i^* \in \{1, \ldots, I\}$ and $r, s \in \{1, \ldots, k\}$ holds. In particular, one typically uses the parameterization $\delta_{ir} = \delta_i + \alpha_r$ to formulate the RSM. Analogously to the GPCM, the GRSM evolves from the RSM if one allows for item-specific discrimination parameters. It can be denoted by

$$\log\left(\frac{P(Y_{pi}=r)}{P(Y_{pi}=r-1)}\right) = \beta_i(\theta_p - \delta_i - \alpha_r), \quad r = 1, \dots, k.$$

For the GRSM, another identifiability issue arises because the item location parameters δ_i and the step parameters α_r are only identified up to a constant. Therefore, the additional restriction $\alpha_1 = 0$ is imposed.

In the most simple case of dichotomous items (k = 1), both the RSM and the PCM automatically simplify to the Rasch model (RM). Again, the generalized RSM/PCM versions with item-specific discrimination parameters lead to the *two-parameter logistic model* (2-PL; Birnbaum 1968).

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

What follows is the introduction of a general method for DIF detection in GPCMs. For notation, we will always use this most general model class. Nevertheless, by applying any of the restrictions described above, the method is applicable to all models we presented in this section.

A DIF model for generalized partial credit models

The starting point of our elaborations is the model proposed in Tutz and Schauberger (2015) that parameterizes DIF effects in Rasch models. Applying the notation of this work and considering our restriction $\theta_p \sim N(0, 1)$, the model can be written as

$$\log\left(\frac{P(Y_{pi}=1)}{P(Y_{pi}=0)}\right) = \beta(\theta_p - \delta_i - \boldsymbol{x}_p^{\mathrm{T}}\boldsymbol{\gamma}_i),$$
(2)

where x_p is an *m*-dimensional covariate vector of person *p*, and γ_i an *m*-dimensional (item-specific) parameter vector. The idea of the model is to have parameters representing DIF instead of examining DIF with various testing approaches. The corresponding parameters γ_i represent DIF if they are unequal to zero after lasso regularization. For example, if for gender the respective γ parameter is unequal to zero for a specific item, this represents different item difficulties for males and females (assuming equal abilities). In total, $m \times I$ additional parameters are introduced, each of them representing DIF for a specific item-covariate pair.

The idea of parameterizing DIF can be generalized to the GPCM. We will call this model *GPCMlasso*, it can be denoted by

$$\log\left(\frac{P(Y_{pi}=r)}{P(Y_{pi}=r-1)}\right) = \beta_i \left(\theta_p - \delta_{ir} - \boldsymbol{x}_p^{\mathrm{T}} \boldsymbol{\gamma}_{i(r)}\right), \quad (3)$$

where $r = 1, ..., k_i$. The base model (1) is extended by (possibly) $m \times \sum_{i=1}^{I} k_i$ additional parameters. The additional parameters $\gamma_{i(r)}$ represent DIF or, potentially, DSF if there are different steps in relation to a covariate for an item. If the analysis of DSF is not of interest, (3) can be reduced to

$$\log\left(\frac{P(Y_{pi}=r)}{P(Y_{pi}=r-1)}\right) = \beta_i \left(\theta_p - \delta_{ir} - \boldsymbol{x}_p^{\mathrm{T}} \boldsymbol{\gamma}_i\right).$$
(4)

By restricting $\boldsymbol{\gamma}_{i(r)} = \boldsymbol{\gamma}_i$, only regular DIF is parameterized (with $m \times I$ additional parameters).

In both cases, the basic interpretation of the additional parameters is similar to the simpler Rasch case by Tutz and Schauberger (2015). The parameters in $\boldsymbol{\gamma}_{i(r)}$ represent changing item step parameters for different values of the respective covariates \boldsymbol{x}_p and, therefore, indicate uniform DIF. The new person-specific item step parameters can be denoted by $\delta_{irp} = \delta_{ir} + \boldsymbol{x}_p^{\mathsf{T}} \boldsymbol{\gamma}_{i(r)}$.

Furthermore, we propose to use additional main effect parameters α for all covariates in x_p . The main effects have the purpose to capture the possible effect that all γ parameters referring to the same variable could be unequal to zero. If we include main effects for all variables, this automatically captures these effects (which are no DIF effects because they are not item-specific). For example, such a main effect for the variable gender can capture possible gender differences between the overall trait levels of males and females and clearly separates these effects both from the trait parameters θ_p and the item-specific effects $\gamma_{i(r)}$. By including main effects into models (3) and (4) we assume the models

$$\log\left(\frac{P(Y_{pi}=r)}{P(Y_{pi}=r-1)}\right) = \beta_i \left(\theta_p + \boldsymbol{x}_p^{\mathrm{T}} \boldsymbol{\alpha} - \delta_{ir} - \boldsymbol{x}_p^{\mathrm{T}} \boldsymbol{\gamma}_{i(r)}\right), \quad (5)$$

or

$$\log\left(\frac{P(Y_{pi}=r)}{P(Y_{pi}=r-1)}\right) = \beta_i \left(\theta_p + \boldsymbol{x}_p^{\mathrm{T}} \boldsymbol{\alpha} - \delta_{ir} - \boldsymbol{x}_p^{\mathrm{T}} \boldsymbol{\gamma}_i\right),\tag{6}$$

respectively.

It should be noted that model (5) has additional identifiability issues compared to the base model (1). It would not make sense if, for one covariate, all corresponding γ parameters would be unequal to zero. Then, at least the average of these effects could also be included in the main effect parameters α . In general, this identifiability problem could be resolved by restricting all γ -parameters from one item to be zero, e.g., $\gamma_{I(r)} = 0$. After all, as for the estimation a penalized likelihood approach is chosen, this is not necessary. The penalized likelihood guarantees for unique estimates because the penalty can be seen as a restriction on the DIF parameters. Therefore, as long as the respective penalty term corresponds to a restriction that is strong enough, this identifiability issue can be ignored.

An attractive feature of model-based DIF detection is the fact that (assuming, among other assumptions, that the model is correctly specified) it that can provide trait estimates that are DIF-free with respect to the covariates used in the model. Certainly, assuming that a model is correctly specified will always be wrong, or as Box put it: "All models are wrong, but some are useful". In the same sense, the DIF effects detected by the proposed model can also be very useful if one wants to come closer to good estimates of an underlying latent trait despite the presence of DIF. Following Reise et al. (1993), even if DIF between groups is discovered in specific items the items after all measure the same latent trait for the respective groups. The additional DIF parameters account for group-specific (or rather covariate-adjusted) item difficulties and, therefore, the resulting trait estimates are corrected for possible DIF bias. Hence, purification approaches are not necessary. In a similar way, Crane et al. (2006) developed group-specific parameter estimates that account for DIF effects within the context of DIF detection using (ordinal) logistic regression. While this is performed in an iterative manner due to the fact that DIF detection and estimation of ability estimates are separated, GPCMlasso automatically provides corrected estimates when calculating posterior estimates of the trait parameters. Additionally, in contrast to Crane et al. (2006), GPCMlasso is not restricted to single covariates but can account for DIF effects across several variables simultaneously. Although the method to detect DIF proposed by Crane et al. (2006) could be applied repeatedly to different variables, for each variable different corrected estimates of the trait estimates would be calculated. In contrast, GPCMlasso allows to integrate DIF effects from different variables and, therefore, allows for correction of all DIF effects simultaneously.

Model estimation

The GPCMlasso is estimated using a marginal penalized likelihood approach, which will be elaborated on in this section. The starting point is the marginal maximum likelihood (MML) expression where the person parameters are treated as random effects and are integrated out of the likelihood. By collecting all parameters from model formula (5) as vectors $\boldsymbol{\theta}^{T} = (\theta_{1}, \dots, \theta_{P}), \boldsymbol{\alpha}^{T} = (\alpha_{1}, \dots, \alpha_{m}), \boldsymbol{\beta}^{T} = (\beta_{1}, \dots, \beta_{I}), \boldsymbol{\delta}^{T} = (\delta_{11}, \dots, \delta_{Ik_{I}}), \text{ and } \boldsymbol{\gamma}^{T} = (\boldsymbol{\gamma}_{1(1)}^{T}, \dots, \boldsymbol{\gamma}_{I(k_{I})}^{T})$, the marginal likelihood can be denoted by

$$L(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{p=1}^{P} \int P(Y_{p1}, \dots, Y_{pI}) f(\theta_p) d\theta_p,$$

with the typical assumption of $\theta_p \sim N(0, 1)$. After integrating out the person parameters, only the item step parameters, the main effect parameters, the discrimination parameters, and the DIF parameters have to be estimated.

Penalized likelihood estimation

An unrestricted MML solution is of no practical interest. This is due to the fact that it has to be assumed that the majority of all pairs between items and covariates will not be affected by DIF. Therefore, the majority of the additional DIF parameters collected in γ should be equal to zero. Also, as mentioned in the previous section, the model would not be identified.

The goal of the estimation procedure elaborated on in this section is to obtain a model that is as sparse as possible with respect to the DIF parameters γ . This can be achieved by applying a penalized likelihood principle. Let $\boldsymbol{\xi} = (\theta, \alpha, \delta, \beta, \gamma)$, that is, a vector containing all GPCMlasso model parameters. Instead of maximizing the regular version of the log-likelihood $l(\boldsymbol{\xi})$, we use the following penalized version:

$$l_p(\boldsymbol{\xi}) = l(\boldsymbol{\xi}) - \lambda J(\boldsymbol{\xi})$$

Here, $J(\boldsymbol{\xi})$ represents a specific penalty term on the parameters of the respective model, and λ is the *tuning* parameter. A popular penalty is the so-called lasso penalty $J(\boldsymbol{\xi}) = \sum_{i} |\xi_{i}|$ proposed by Tibshirani (1996). Lasso penalizes the sum of the absolute values of all parameters of a linear regressions model. By using the absolute values instead of the squared values as in the ridge penalty (Hoerl & Kennard, 1970), the penalty term not only shrinks the penalized parameters but is also able to set certain parameters equal to zero. Therefore, lasso allows for automatic variable selection. The amount of penalty is driven by the tuning parameter λ . For $\lambda = 0$, the penalty term vanishes and regular (unpenalized) maximum likelihood estimation is performed. For $\lambda \rightarrow \infty$, the influence of the penalty parameter increases and (in case of lasso) more and more parameters are set equal to zero.

Let us adapt the general principle of the lasso penalty to the estimation of our model to detect either DIF or DSF. We start with the simplified GPCMlasso from (6) (i.e., DIF only). We propose to use the penalty term

$$J(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{I} \sum_{j=1}^{m} w_{ij} |\gamma_{ij}|, \qquad (7)$$

where w_{ij} is a weight parameter. In contrast to the lasso penalty introduced above, here only a subset of the parameters is penalized. The penalty is solely applied to the additional DIF parameters γ_{ij} because only some of them are expected to be unequal to zero and, therefore, to represent DIF. None of the parameters collected in θ , α , δ and β are penalized or eliminated from the model. Therefore, if the tuning parameter λ is chosen as very large, one ends up with the chosen base model, but without any DIF effects (and possibly extended by the main effects contained in α).

We recommend to specify the weight parameter w_{ij} according to the principle of adaptive lasso as proposed by Zou and Hastie (2005) and Zou (2006). They use weights corresponding to the inverse absolute values of the parameter estimates from maximum likelihood (ML) estimation (or ridge estimation, as in our case). Therefore, parameters with small ML (or ridge) estimates will be penalized more severely than parameters with large estimates.

Figure 1 shows the parameter paths of the DIF parameters for an exemplary item along the tuning parameter when the penalty term (7) is used. Three DIF parameters are associated to the item, one per covariate. With decreasing tuning parameter λ (from left to right), sequentially all three variables show DIF and the shrinkage of the respective parameter diminishes. If, for example, the final model would be chosen at $\log(\lambda + 1) = 5.5$ (symbolized by grey dashed line), this item would be found to suffer from DIF with respect to variable Var1.

If instead of model (6), the more general GPCMlasso model as given in (5) (i.e., DIF and DSF detection) is assumed, the penalty can be extended to detect DSF. In this case, the penalty term can be denoted by

$$I(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{I} \sum_{j=1}^{m} \sum_{r=1}^{k_i} w_{ij(r)} |\gamma_{ij(r)}| + \sum_{i=1}^{I} \sum_{j=1}^{m} \sum_{r < s} w_{ij(r,s)} |\gamma_{ij(r)} - \gamma_{ij(s)}|.$$
(8)



Fig. 1 Parameter paths of DIF parameters along $log(\lambda + 1)$ for an exemplary item to illustrate penalty term (7). The *dashed grey line* represents an exemplary model where the item suffers from DIF with respect to variable Var1

It consists of two parts, which serve different purposes. The first part corresponds to the penalty term (7) and simply penalizes the absolute values of all (possible) DIF parameters and possibly sets them to zero. The second part is able to fuse the groups of all k - 1 parameters corresponding to the same pair between item *i* and covariate *j*. In total, $I \times m$ of these groups exist. Fusing is done by penalizing all absolute differences between the respective parameters. For λ large enough, this penalty can (within such a group of parameters) produce parameters that are exactly equal. If this is the case, the effect of the respective covariate on the item is not varying across the different categories and DIF instead of DSF is present.

Figure 2 illustrates how the penalty works. For an exemplary item, the plot depicts the parameter paths of all DIF parameters corresponding to this item. In the example, three potential DIF-inducing variables are used and the response is measured on a five-point scale. Accordingly, for the specific item $m \times (k-1) = 3 \times (5-1) = 12$ parameters are used to parameterize all possible DIF (or DSF) effects. With λ decreasing (from left to right), the model becomes more complex as the restrictions induced by the penalty term are relaxed. Clearly, the fusion property of the penalty can be seen, for large values of λ all parameters referring to the same variable are equal. Therefore, penalty (8) can be referred to as fusion penalty. The grey dashed line in Fig. 2 indicates the model corresponding to $log(\lambda + 1) =$ 3. If this model would be chosen, the model would have detected DIF induced by variable 1. Because all parameters corresponding to variable 1 are unequal to zero but equal among themselves, there is no DSF but only DIF. All other variables cause neither DSF nor DIF. As the penalty is further decreased, the parameter clusters of the single variables dissolve and one would detect DSF. Similar to penalty (7), the following weight parameters are used, here denoted by $w_{ij(r)}$ and $w_{ij(r,s)}$. Again, $w_{ij(r)}$ represents the



Fig. 2 Parameter paths of DIF/DSF parameters along $log(\lambda + 1)$ for an exemplary item to illustrate penalty term (8). This penalty is applied if also DSF is investigated

inverse of the absolute values of ML or ridge estimates of the respective parameters $\gamma_{ij(r)}$. Accordingly, $w_{ij(r,s)}$ consists of the inverse absolute difference between $\gamma_{ij(r)}$ and $\gamma_{ij(s)}$. Therefore, differences are penalized stronger if the ML or ridge estimates are very different.

Note that in order to apply the proposed penalties in a sensible way, all variables need to be transformed to equal scales. Only then the respective DIF parameters are of comparable size and the penalty terms can work properly.

Finding the optimal tuning parameter

An important issue in penalized likelihood approaches is to determine the optimal value of the tuning parameter λ . Typically, either model selection criteria or cross-validation (CV) can be used to determine the best value for λ out of a pre-specified grid of possible values. Classical model selection criteria include the AIC (Akaike, 1974) and the BIC (Schwarz, 1978). To enforce a more conservative parameter selection and because the BIC is consistent with respect to parameter selection, we will in the following use the BIC. For our model, it is defined as

$$BIC(\lambda) = -2l_{\lambda}(\cdot) + df(\lambda)\log(n)$$

where $l_{\lambda}(\cdot)$ denotes the likelihood for the parameters estimated with tuning parameter λ and $df(\lambda)$ denotes the total number of parameters estimated (uniquely) unequal to zero.

Alternatively, also *k*-fold CV could be used to select the final model. In CV, the data set is divided into *k* subsets containing one *k*-th of all persons from the whole data set. Iteratively, each subset is eliminated once from the data and used as test data set. The remaining subsets are used as training data. The model is fitted to the training data using a pre-defined grid of λ values. Afterwards, the fitted model is evaluated on the test data using a specific loss function, separately for each value of the tuning parameter. This procedure allows one to evaluate the out-of sample performance of the model along the tuning parameter λ and can, therefore, be used to detect the optimal value of the tuning parameter. As loss function, we propose to use the deviance of the test data based on the marginal log-likelihood.

Both alternatives are implemented in the package GPCMlasso. In the remainder of the paper, we will use the BIC, however. This has both technical and theoretical reasons. First, CV is more time-consuming because the procedure has to be executed repeatedly for several different training data. Second, CV is designed to choose the optimal model with respect to prediction. However, in the case of DIF detection optimality in parameter selection is much more important than prediction. Third, CV is known to be

a rather unstable method that can produce highly different results in rather similar situations.

Posterior trait estimation

The trait parameters, corrected for the DIF effects, can be estimated by taking the mean of the posterior distribution of the trait parameters. Here, the estimates obtained in the previous section are treated as known quantities.

$$\hat{\theta}_p = E(\theta_p | \boldsymbol{Y}_p) = \int \theta_p f(\theta_p | \boldsymbol{Y}_p, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\gamma}}) d\theta_p$$

where $\boldsymbol{Y}_p = \{Y_{p1}, \dots, Y_{pI}\}$ and $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\gamma}}$ contain all estimated $\beta_i, \alpha_j, \delta_{ir}$ and $\gamma_{ij(r)}$ parameters, respectively.

Mainly, GPCMlasso is designed as a method for DIF detection in GPCMs. After all, in some applications it may be the case that estimates for the trait parameters are of central interest. Then, the posterior estimates described here can provide trait estimates for the participants that already account for possible DIF effects. In contrast to other DIF-detection methods, GPCMlasso can take into account several variables simultaneously. Therefore, the traits estimated are corrected for possible DIF effects from several different variables.

Simulation study

To compare the method to existing approaches, several different simulation studies were performed. For simplicity, we focus on the detection of DIF ignoring the more general case of DSF. We will only consider the case where no main effects of the covariates are present in the data-generating process. However, the models we estimate refer to the more general case described in model formula (6) where main effects are estimated for all variables. We also conducted equivalent simulations as described below where main effects are present in the data-generating process. As the results are extremely similar to the results presented here, we abstain from including them in this manuscript.

Mainly, we will compare GPCMlasso using the GPCMlasso package (Schauberger, 2018) to the method proposed in Choi et al. (2011), which combines the logistic regression method for DIF detection (Swaminathan & Rogers, 1990) with IRT models for trait score estimation. A preliminary version of lordif can be found in Crane et al. (2006). It is implemented in the R package lordif (Choi et al., 2016). It has the feature to be applicable both to dichotomous and polytomous items. In the case of several covariates as considered in settings 2–5, the procedure has to be executed separately for each variable. In lordif, the user can choose between the GRM and the GPCM as the underlying IRT model. We will only focus on GPCM for

the simulation study. In general, with lordif no continuous covariates can be used. Therefore, continuous covariates will be split into two groups using the median as split point.

The first setting treats the easiest possible case of DIF in Rasch models (RM) with a single binary covariate. In this case, we also compare our method to the classical Mantel-Haenszel method (Holland & Thayer, 1988) as implemented in the R package difR (Magis et al., 2010) (using the builtin iterative item purification algorithm). In all remaining settings, polytomous items are used with either the PCM or the GPCM as the base model. Here, binary and continuous variables are alternated (i.e., in the case of five variables three of them are binary and two are continuous). All simulations use I = 20 items and P = 500 persons and are based on 100 replications. Both for Mantel-Haenszel and for lordif the significance level $\alpha = 0.01$ is used, which is the default in the R-package lordif. Besides the true underlying model, also the maximal response category k and the number of (possibly) DIF-inducing variables m are varied. For simplicity, we restrict the simulations to cases with equal numbers of response categories across all items. In the case of several variables, we also consider cases with high correlations between these variables. Table 1 summarizes the settings of the five different simulations that were performed.

Parameter estimation

As a first step, we empirically evaluate the ability of the method to reproduce the true underlying parameters of the simulation. For that purpose, we focus on setting 4 (strong) because it is the most general one. Exemplarily for the first ten items, Fig. 3 shows boxplots for the estimates of all (fixed) model parameters, separately for item step parameters (3a), discrimination parameters (3b), main effect parameters (3c), and DIF parameters (3d). True values are indicated by red crosses.

While the item step parameters and the main effect parameters are estimated very accurately, the discrimination parameters are minimally overestimated. In contrast to item step parameters, main effect parameters, and discrimination parameters, the DIF parameters are penalized. As we use 20 items and five variables, in total there are 100 possible DIF effects. Four of these effects actually cause DIF in our simulation (compare Table 1). With only two exceptions, the parameters corresponding to the true DIF effects are estimated to be unequal to zero and therefore correctly identify these DIF effects. Nevertheless, as expected, the penalty shrinks them toward zero, and on average the size of the DIF effects is underestimated. The remaining parameters are correctly set to zero most of the time with only a few false positives.

Additionally, we also inspect the estimation accuracy of the posterior mean estimates of the latent trait as sketched in Section "Posterior trait estimation". Figure 4 depicts the scatterplot of the true and the estimated trait parameters for an exemplary data set from simulation setting 4 (strong). It can be seen that most of the points are very close to the angle bisector and therefore the accuracy of the posterior mean estimates is very high. Overall, both Figs. 3 and 4 show very satisfying results regarding the estimation accuracy of the method.

DIF detection

The performance of the methods in terms of DIF detection is compared with respect to true-positive rates (TPR) and false-positive rates (FPR). For TPR, we take the ratio of correctly detected DIF effects among all true DIF effects. For FPR, we take the ratio of all incorrectly detected DIF effects among all DIF parameters that are simulated to be exactly zero.

Although the main focus of the method is to handle problems where multiple variables are to be considered

 Table 1
 Summary of parameter variations for all simulation settings

Setting		Model	k	m	Corr.	DIF parameters
1	strong	RM	1	1		$\gamma_1 = -\gamma_5 = 0.9$
	weak	RM	1	1		$\gamma_1 = -\gamma_5 = 0.3$
2	strong	PCM	5	3	0	$\gamma_{11} = \gamma_{22} = -\gamma_{33} = 0.3$
	weak	PCM	5	3	0	$\gamma_{11} = \gamma_{22} = -\gamma_{33} = 0.15$
3	strong	PCM	5	3	0.7	$\gamma_{11} = \gamma_{22} = -\gamma_{33} = 0.3$
	weak	PCM	5	3	0.7	$\gamma_{11} = \gamma_{22} = -\gamma_{33} = 0.15$
4	strong	GPCM	5	5	0	$\gamma_{11} = \gamma_{22} = -\gamma_{33} = -\gamma_{44} = 0.3$
	weak	GPCM	5	5	0	$\gamma_{11} = \gamma_{22} = -\gamma_{33} = -\gamma_{44} = 0.15$
5	strong	GPCM	5	5	0.7	$\gamma_{11} = \gamma_{22} = -\gamma_{33} = -\gamma_{44} = 0.3$
	weak	GPCM	5	5	0.7	$\gamma_{11} = \gamma_{22} = -\gamma_{33} = -\gamma_{44} = 0.15$













(d) Parameter estimates and true values for DIF parameters γ_{ij} (for items 1-10)

Fig. 3 Boxplots displaying parameter estimates from method GPCMlasso for 100 simulated data sets according to simulation setting 4 (strong). Boxplots are separated for **a** item step parameter, **b** discrimination parameters, **c** main effect parameters, and **d** DIF parameters. Only items 1-10 are shown. True values are indicated by *crosses*



Fig. 4 Scatterplot for true vs. estimated trait parameters θ_p for an exemplary data set from simulation setting 4 (strong). The *solid line* represents exact accordance between true and estimated values

for DIF, we first want to demonstrate its applicability and performance in the most simple case of DIF detection. Table 2 collects the TPR and FPR for setting 1 for all three compared methods.

It can be seen that GPCMlasso performs similar or slightly worse than lordif and MH. After all, it seems not to be reasonable to use GPCMlasso in simple situations like this because the classical methods are much less computationally demanding. The strength of GPCMlasso lies in much more complex data situations, especially in the case of several (possibly correlated) covariates. Therefore, we focus on the simulation results of setting 2–5.

Table 3 shows the results of DIF detection in setting 2–5 in terms of TPR and FPR.

There are several general trends that can be seen from Table 3. First, the TPR for strong DIF is very similar for both methods. For weak DIF, lordif has a slightly increased TPR compared to GPCMlasso except for setting 2. Second, GPCMlasso has lower (better) values for the FPR than lordif. In particular, in settings 3 and 5 with the presence of correlation between the variables we observe

Table 2 Selection results in terms of TPR and FPR for simulation setting $\boldsymbol{1}$

	GPCMlasso	lordif	MH
(a) Strong DIF			
TPR	0.975	0.975	0.965
FPR	0.019	0.014	0.008
(b) Weak DIF			
TPR	0.125	0.165	0.125
FPR	0.006	0.014	0.008

highly increased FPR values for lordif compared to the intended size 0.01 according to the α -level. For example, in setting 5 (strong), the FPR is 0.105, which is about ten times as high as intended. In contrast, GPCMlasso deals much better with correlations with constantly low values for FPR across all settings.

Certainly, the comparison of TPR and FPR in Table 3 is flawed in some sense because lordif (as it is a test-based method) explicitly restricts the FPR by the choice

Table 3 Results of DIF detection in terms of TPR and FPR for simulation settings 2--5

Setting		GPCMlasso	lordif
(a) Strong DIF			
2	TPR	1.000	1.000
	FPR	0.006	0.013
3	TPR	1.000	0.997
	FPR	0.012	0.060
4	TPR	0.995	0.975
	FPR	0.007	0.013
5	TPR	0.965	0.968
	FPR	0.013	0.105
(b) Weak DIF			
2	TPR	0.650	0.580
	FPR	0.006	0.010
3	TPR	0.593	0.600
	FPR	0.012	0.034
4	TPR	0.360	0.378
	FPR	0.006	0.011
5	TPR	0.335	0.482
	FPR	0.009	0.032



Fig. 5 ROC curves (*lines*) and realized results (*crosses*) for DIF detection for setting 2 (**a**), setting 3 (**b**), setting 4 (**c**), and setting 5 (**d**)

Table 4Results for DIF detection in terms of FDR for simulationsettings 2-5

Setting	GPCMlasso	lordif
(a) Strong DIF		
2	0.102	0.202
3	0.187	0.534
4	0.150	0.241
5	0.250	0.722
(b) Weak DIF		
2	0.148	0.243
3	0.273	0.521
4	0.298	0.401
5	0.385	0.616

of the α -level while this is not the case for GPCMlasso, which is controlled by the tuning parameter λ . Therefore, we additionally try to compare the TPR and FPR of both methods with the help of ROC curves as depicted in Fig. 5.

For both methods, the TPR and the FPR are plotted along a dense grid of the respective parameters that control the selection rates, i.e., the α -level for lordif and the tuning parameter λ for GPCMlasso. Crosses indicate the actual values of TPR and FPR as summarized in Table 3. Each ROC curve starts from the point (0,0) because both for $\lambda = \infty$ and $\alpha = 0$ no DIF can be detected. Subsequently, λ is decreased and α is increased, respectively. In the optimal case, the ROC curve has a very steep ascent and is as close as possible to the point (0,1), which represents a perfect selection of DIF effects. The ROC curves for GPCMlasso dominate the curves of lordif in all settings. In particular, this is the case for correlated variables (settings 3 and 5).

As a final measure of the performance in DIF detection, we want to consider the false discovery rates (FDR) of 289

both methods. The FDR represents the probability (or rather relative frequency) that effects that were detected as DIF effects are true DIF effects. When looking at TPR and FPR values, it is obvious that higher values of TPR will generally lead to higher values of FPR. In contrast, FDR takes into account both true and false positives and therefore gives interesting further insights into the performance of both methods. Table 4 collects the FDR values for settings 2–5.

The FDR values for GPCMlasso outperform the FDR of lordif for all settings. In contrast to lordif, GPCMlasso does not have highly increased values of FDR in the case of correlated variables. For lordif, in all cases with correlations, the FDR is greater than 50%, which (for a single found DIF effect) means that most likely it is a false positive. Overall, we see from the simulation studies that GPCMlasso performs slightly better than lordif in case the variables are not correlated but clearly outperforms lordif otherwise.

Application to youth depression data

For illustration, we will apply the proposed model to data from a study by Vaughn-Coaxum et al. (2016) on the Children's Depression Inventory (CDI; see Kovacs 2004). While the original study solely focused on racial/ethnic differences, we will include several different potential DIFinducing covariates simultaneously. The following variables will be treated as possibly DIF-inducing covariates.

Age	Age of the child
Gender	0: female; 1: male
Race	1: white; 2: black; 3: asian; 4: hispanic
Educ_Parents	1: both parents graduated from college; 0:
	otherwise



Fig. 6 Mosaic plot for relative frequencies of variable Educ_Parents conditional on variable race



Fig. 7 Coefficient paths of all DIF parameters (from model A) for GPCMlasso applied to CDI data, separately for each item. *Dashed vertical lines* represent the optimal model according to BIC



Fig. 8 Coefficient paths of all DIF parameters (*left plots*) and DSF parameters (*right plots*) for GPCMlasso applied to CDI data, exemplarily for the items 1, 4, and 5. Separately for the left and right plots, the *dashed vertical lines* represent the optimal model according to BIC

Item	Race	Race		Gender		Educ_Parents		Age	
	DIF	DSF	DIF	DSF	DIF	DSF	DIF	DSF	
1	√ ×	✓	√ √	✓	X X	x	x ×	x	
2	\checkmark \checkmark	\checkmark	××	×	× ×	×	××	×	
3	× 🗸	×	× ×	×	× ×	×	××	×	
4	x 🗸	×	× ×	×	× ×	×	××	×	
5	\checkmark \checkmark	×	\checkmark ×	×	× ×	×	\checkmark \checkmark	×	
6	x 🗸	×	\checkmark \checkmark	\checkmark	× ×	×	××	×	
7	\checkmark ×	\checkmark	× ×	×	× ×	×	××	×	
8	\checkmark \checkmark	\checkmark	× ×	×	××	×	××	×	
10	\checkmark ×	\checkmark	\checkmark \checkmark	\checkmark	××	×	××	×	
11	\checkmark \checkmark	\checkmark	××	×	\checkmark \checkmark	\checkmark	××	×	
12	\checkmark \checkmark	\checkmark	××	×	××	×	××	×	
13	× ×	×	\checkmark \checkmark	\checkmark	\checkmark \checkmark	\checkmark	××	×	
14	\checkmark ×	\checkmark	\checkmark \checkmark	\checkmark	××	×	××	×	
15	\checkmark \checkmark	\checkmark	\checkmark \checkmark	×	× ✓	×	××	×	
16	\checkmark ×	\checkmark	××	×	\checkmark \checkmark	×	\checkmark \checkmark	×	
17	\checkmark \checkmark	\checkmark	××	×	× ✓	×	××	×	
18	\checkmark \checkmark	\checkmark	××	×	× ✓	×	\checkmark ×	×	
19	\checkmark \checkmark	\checkmark	××	×	××	×	××	×	
20	××	×	××	×	××	×	××	×	
21	\checkmark \checkmark	×	\checkmark \checkmark	\checkmark	××	×	\checkmark ×	×	
22	××	\checkmark	××	×	××	×	××	×	
23	\checkmark \checkmark	×	××	×	\checkmark \checkmark	\checkmark	××	×	
24	\checkmark \checkmark	\checkmark	××	×	××	×	××	×	
25	\checkmark ×	\checkmark	\checkmark ×	×	××	×	××	×	
26	\checkmark \checkmark	\checkmark	\checkmark ×	×	\checkmark \checkmark	\checkmark	\checkmark ×	×	
27	\checkmark \checkmark	×	X X	×	X X	×	x ×	x	
Σ	20 17	17	10 7	6	5 8	4	5 2	0	

 Table 5
 Comparison of detected DIF items (model A) and DSF items (model B) of GPCMlasso (*black*) to DIF items detected by lordif (*gray*) for CDI data

The variable race is encoded by dummy coding with *white* as the reference category because the CDI was developed in a sample of youths that was predominantly white (similar to our sample at hand).

An interesting feature of the data is the rather high correlation between the variables race and education of parents, which is illustrated in Fig. 6. It shows a mosaic plot of the relative frequencies of education of parents while conditioning on the different values of race. For the majority of the white and Asian children, both parents have graduated from college while this is not the case for the majority of the black and Hispanic children. The data set contains data from 1471 participants on 26 of the 27 items of the CDI. Analogous to the original study by Vaughn-Coaxum et al. (2016), item nine about suicidal ideation was removed.

A main focus of this section is to illustrate the differences we find between DIF detection and DSF detection of the CDI data using GPCMlasso. Therefore, we will apply two different models to the data. In the remainder of this section, we will refer to the model for DIF detection as model A and to the model for DSF detection as model B. We start by fitting model A. Figure 7 shows the coefficient paths for all variables and all items along (a transformation of) the tuning parameter λ . The paths are plotted separately for each item. The red dashed lines represent the optimal model according to BIC.

Overall, in model A, only items 3, 4, 8, 20, and 22 are diagnosed to be completely DIF-free. Besides the DIF parameters, item step parameters, main effect parameters, and item discrimination parameters are also estimated, which are not presented here.

Due to the relatively high number of items, the single paths are rather hard to detect in Fig. 7. For model B, the number of parameters is even higher and paths would be even harder to distinguish. Therefore, we abstain from showing the paths of all items for model B. Instead, in Fig. 8 we exemplarily show the parameter paths (for DSF detection) for items 1, 4, and 5 and compare them to the parameters paths for model A. The paths on the left side refer to model A while the paths of the right side refer to model B. Therefore, all paths from one column in Fig. 8 refer to the same model (but different items) and paths from the same row refer to the same item (but different models).

While in model A both in items 1 and 5, DIF is detected, model B only detects DSF for item 1, and both models find no DIF or DSF effects for item 4. Within item 1 we see several differences between models A and B. In model A, DIF is found for gender and the race groups Hispanic and black. In model B, DSF is found for gender and race group Hispanic only for step 1, respectively. The effect of race group black is very similar for both models, the fusion penalty in model B fuses the parameters for both steps to be equal. Overall, the paths of the DSF parameters again illustrate the fusion property of the penalty we apply in model B. In many cases (for example for every variable in item 5), the two parameters corresponding to the same variable are equal at least for certain ranges of the tuning parameter λ .

Besides the comparison of the GPCMlasso models A and B, we also want to compare these models to the DIFdetection results we would get by applying the method of Choi et al. (2011) called lordif. Analogous to the simulation study, lordif is separately applied to each of the four variables as otherwise it cannot handle several variables simultaneously. Table 5 summarizes all variables whether DIF was detected in the single items by model A and lordif. Additionally, it shows whether model B detected DSF for at least one of the two corresponding steps.

The three models agree for the majority of the cases. However, there are also several differences between the results. Overall, model B detected the lowest number of items. However, that does not necessarily imply that model B is sparser than model A because by definition model B has much more (possible) parameters than model A. For age, model B found no DSF effect at all while model A found DIF in five items.

A remarkable result is that lordif identifies much more DIF items with respect to the variable education of parents than GPCMlasso. One reason for this result may be the fact that GPCMlasso is able to account for the high correlation between race and education of parents. By considering race and education simultaneously, GPCMlasso is supposed to be able to avoid duplicated DIF effects from race and education.

Conclusions

This work introduced the GPCMlasso approach for polytomous DIF detection in the presence of multiple, potentially correlated covariates of mixed scale levels. DIF is parameterized in a joint model that can be seen as an extension of a base IRT model, with the GPCM as the most general base model. The model is estimated using a penalized likelihood approach. Based on the detected DIF effects, one can also estimate trait estimates that are automatically corrected for DIF effects in different items and caused by different variables.

In general, the method could also be extended to cover both uniform and non-uniform DIF. This would require a second set of possible DIF effects. While the DIF effects for uniform DIF change the location of certain items, DIF effects for non-uniform DIF would change the discriminatory power of items. Therefore, the second set of DIF effects would have to be included as an extension of the item discrimination parameters β_i . Then, a new personspecific discrimination parameter could be parameterized as $\beta_{ip} = \beta_i + \boldsymbol{x}_p^{\mathrm{T}} \boldsymbol{\psi}_i$. Again, one would have one additional parameter for each pair between items and covariates. However, penalized likelihood estimation would become more complicated because the selection of the additional parameters ψ_i requires the introduction of a second tuning parameter. It cannot be assumed that, even if variables are scaled, the parameters for uniform and non-uniform DIF are on comparable scales. Accordingly, a two-dimensional grid search for the optimal tuning parameters becomes necessary.

In addition, it could be interesting for future research to intensively study the best criterion to select the optimal value of the tuning parameter and to develop alternatives to the BIC that is currently used. One alternative that is already discussed in the paper is CV. It has the disadvantage that it is not focused on optimal variable selection and tends to have higher numbers of false-positive effects than BIC. Another possible alternative could be stability selection as proposed by Meinshausen and Bühlmann (2010) and Shah and Samworth (2013). Here, the method would be applied repeatedly to new data sets that are generated by sub-sampling and only variables that are selected for almost all data sets would be used for the final model. However, this method would increase the running time disproportionately.

Acknowledgements We thank Rachel Vaughn-Coaxum and John Weisz for providing the data of the Children's Depression Inventory and for comments that helped to improve the manuscript.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561–573.
- Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F M Lord, F., & Novick, M. (Eds.) Statistical theories of mental rest score, reading, (pp. 397–479). Boston: Addison-Wesley.
- Bollmann, S., Berger, M., & Tutz, G. (2018). Item-focused trees for the detection of differential item functioning in partial credit models. *Educational and Psychological Measurement*, 78(5), 781–804. https://doi.org/10.1177/0013164417722179.
- Chang, H. H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items an adaptation of the SIBTEST procedure. *Journal of Educational Measurement*, 33(3), 333–353.
- Choi, S., Gibbons, L., & Crane, P. (2011). Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software*, 39(1), 1–30. https://www.jstatsoft.org/index.php/jss/article/view/ v039i08. https://doi.org/10.18637/jss.v039.i08
- Choi, S. W., with contributions from Laura E Gibbons, & Crane, P. K. (2016). Lordif: Logistic ordinal regression differential item functioning using IRT. https://CRAN.R-project.org/package=lordif, r package version 0.3-3.
- Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. *Medical Care*, 44(11), 115–123.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Bias estimation for nonorthogonal problems. *Technometrics*, 12, 55–67.
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel–Haenszel procedure. Test Validity, 129–145.
- Komboz, B., Strobl, C., & Zeileis, A. (2018). Tree-based global model tests for polytomous Rasch models. *Educational and Psychological Measurement*, 78, 128–166. https://doi.org/10.1177/00131644 16664394
- Kovacs, M. (2004). Children's Depression Inventory (CDI). Toronto: Multi-Health Systems Inc.
- Magis, D., Bèland, 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(3), 847–862. https://doi.org/10.3758/BRM.42.3.847
- Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel–Haenszel procedure. *Journal of the American Statistical Association*, 58(303), 690–700.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4), 719–748. http://jnci.oxfordjournals. org/content/22/4/719.short, http://jnci.oxfordjournals.org/content/ 22/4/719.full.pdf+html. https://doi.org/10.1093/jnci/22.4.719
- Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174.
- Meinshausen, N., & Bühlmann, P. (2010). Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(4), 417–473.

- Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, *14*(1), 59–71. https://doi.org/10.1177/014662169001400106
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159–176.
- Penfield, R. D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of Educational Measurement*, 44(3), 187–210. http://www.jstor.org/ stable/20461856.
- Penfield, R. D., Gattamorta, K., & Childs, R. A. (2009). An NCME instructional module on using differential step functioning to refine the analysis of DIF in polytomous items. *Educational Measurement: Issues and Practice*, 28(1), 38–49.
- R Core Team (2018). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/.
- Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*(3), 552–566.
- Schauberger, G. (2018). GPCMLasso: Differential item functioning in generalized partial credit models. R package version 0.1-1.
- Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics, 6, 461–464.
- Shah, R. D., & Samworth, R. J. (2013). Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1), 55– 80. https://doi.org/10.1111/j.1467-9868.2011.01034.x
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27(4), 361–370. https://doi.org/10.1111/j. 1745-3984.1990.tb00754.x
- Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In Holland, P. W., & Wainer, H. (Eds.) *Differential item functioning*, (pp. 67-113).
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 58, 267–288.
- Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika*, 80(1), 21–43. https://doi.org/10.1007/s11336-013-9377-6
- Vaughn-Coaxum, R. A., Mair, P., & Weisz, J. R. (2016). Racial/ethnic differences in youth depression indicators. *Clinical Psychological Science*, 4(2), 239–253. https://doi.org/10.1177/2167702615591768
- Welch, C., & Hoover, H. (1993). Procedures for extending item bias detection techniques to polytomously scored items. *Applied Measurement in Education*, 6(1), 1–19. https://doi.org/10.1207/ s15324818ame0601_1
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal* of the American Statistical Association, 101(476), 1418–1429.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*), 67, 301–320.
- Zumbo, B. (1999). A handbook on the theory and methods of differential item functioning (DIF). Ottawa: National Defense Headquarters.