



Using Lasso and Adaptive Lasso to Identify DIF in Multidimensional 2PL Models

Chun Wang, Ruoyi Zhu & Gongjun Xu

To cite this article: Chun Wang, Ruoyi Zhu & Gongjun Xu (2022): Using Lasso and Adaptive Lasso to Identify DIF in Multidimensional 2PL Models, *Multivariate Behavioral Research*, DOI: [10.1080/00273171.2021.1985950](https://doi.org/10.1080/00273171.2021.1985950)

To link to this article: <https://doi.org/10.1080/00273171.2021.1985950>



Published online: 28 Jan 2022.



Submit your article to this journal [↗](#)



Article views: 46




View related articles [↗](#)



View Crossmark data [↗](#)



Using Lasso and Adaptive Lasso to Identify DIF in Multidimensional 2PL Models

Chun Wang^a , Ruoyi Zhu^a, and Gongjun Xu^b

^aUniversity of Washington; ^bUniversity of Michigan

ABSTRACT

Differential item functioning (DIF) analysis refers to procedures that evaluate whether an item's characteristic differs for different groups of persons after controlling for overall differences in performance. DIF is routinely evaluated as a screening step to ensure items behave the same across groups. Currently, the majority DIF studies focus predominately on unidimensional IRT models, although multidimensional IRT (MIRT) models provide a powerful tool for enriching the information gained in modern assessment. In this study, we explore regularization methods for DIF detection in MIRT models and compare their performance to the classic likelihood ratio test. Regularization methods have recently emerged as a new family of methods for DIF detection due to their advantages: (1) they bypass the tedious iterative purification procedure that is often needed in other methods for identifying anchor items, and (2) they can handle multiple covariates simultaneously. The specific regularization methods considered in the study are: lasso with expectation-maximization (EM), lasso with expectation-maximization-maximization (EMM) algorithm, and adaptive lasso with EM. Simulation results show that lasso EMM and adaptive lasso EM hold great promise when the sample size is large, and they both outperform lasso EM. A real data example from PROMIS depression and anxiety scales is presented in the end.

KEYWORDS

IRT; differential item functioning

Introduction

The increasing availability of rich survey data and the emerging needs of assessing complex latent traits pose great challenges to existing techniques used to handle and analyze the data, in particular when the data are collected from heterogeneous populations. Different forms of multilevel, multidimensional item response theory (MIRT) models have been proposed to extract meaningful information from complex survey data. In addition to item calibration, items for large-scale standardized testing are routinely scrutinized for differential item functioning (DIF) to ensure equitable comparison of assessment outcomes among different examinee groups.

Differential item functioning (DIF) analysis refers to procedures that evaluate whether an item's characteristic differs for different groups of examinees after controlling for overall differences in performance. Two types of DIF are often differentiated: uniform DIF and non-uniform DIF. The former refers to an item having a constant advantage for a particular group, whereas the latter refers to the advantage

varying in magnitude and/or direction across the latent trait continuum (Penfield & Camilli, 2006; Woods & Grimm, 2011). DIF is routinely evaluated for any new items added in large-scale assessments as a quality control step. For example, in the National Assessment of Educational Progress (NAEP), screening for item DIF often involves three comparisons: male vs female, White vs. Black, and White vs. Hispanic. In the English Language Proficiency Assessment for the 21st Century (ELPA21), DIF is evaluated across gender, ethnicity, economic status, English learner status, and disability status. In addition to these demographic or time variables, DIF may also be caused by cognitive variables that are relevant for item solution processes (Walker & Beretvas, 2003) or other item and test mode effects.

Early research on detecting multidimensional DIF is mostly extensions of unidimensional DIF indices that are based on the difference of examinees' performance in the focal group versus the reference group. Oshima et al. (1997) first proposed an index for DIF detection in two-dimensional models, namely,

the multidimensional differential item functioning of items and tests (DFIT). They assumed that the examinees' true scores (defined as the sum of model-based probabilities) would be independent of group membership, and DIF is quantified by the expected difference in true scores between groups. Because their test statistics require the calculation of model-based probabilities, linking is needed to find two sets of item parameters for the focal and reference groups, respectively. This has to be done with iterative linking based on matching test response functions, which is inefficient and likely to be confounded with DIF itself. Stout et al. (1997) extended simultaneous item bias test (SIBTEST) (Chang et al., 1996; Shealy & Stout, 1993) to the two-dimensional case and proposed MULTISIB. Because their methods use only observed total score, they are generalizable to other IRT models and are not susceptible to model misfit. In the 2-dimensional case, a matching subtest is needed for each dimension, and because matching is based on the combination of total scores on both subtests, it would be very complicated to expand the procedure to tests beyond two dimensions.

Three other more flexible approaches are the multiple indicators multiple causes (MIMIC) models, multiple-group IRT modeling, and logistic regression (Choi et al., 2011; Swaminathan & Rogers, 1990; Zumbo, 1999). These three approaches share great similarity. In a MIMIC model, at least one observed variable (i.e., group variable), often called a casual indicator, predicts a latent variable (Jöreskog & Goldberger, 1975). In the multiple-group IRT approach, rather than regressing latent θ on the grouping variable, one fits two multiple-group IRT models with different equality constraints on target items to the data and conducts a likelihood ratio test to test for DIF (Suh & Cho, 2014). Logistic regression, as the name entails, recasts the IRT model as a logistic regression model such that uniform DIF is modeled by including group variable as a predictor in addition to θ , whereas non-uniform DIF has θ -by-group interaction as an additional predictor. Of note, while MIMIC and multiple-group IRT model can naturally handle θ distribution differences by groups (i.e., impact), in logistic regression however, one does not distinguish the distribution of θ for different groups. Hence, logistic regression does not usually account for impact. All three approaches perform better using a free-baseline designated-anchor approach. That is, while holding the anchor item parameters the same across groups, the full model that allows all studied items to have DIF is fitted first. Then a constrained

model is fitted, one for each item. Because the analysis proceeds one item at a time, testing for multiple studied items usually require additional control for Type I error rate, such as the Benjamini-Hochberg procedure (BH) (Benjamini & Hochberg, 1995; Lee et al., 2017; Raykov et al., 2013). Moreover, the selection of the designated anchor items is critical to the success of the methods. Various methods for identifying anchors have been suggested, and most of them are iterative purification procedures (Bolt et al., 2004; Edelen et al., 2006; Kopf et al., 2015).

Procedurally for all these three approaches, it is cumbersome to perform a likelihood ratio test (or Wald test) separately for one item at a time. For instance, in Woods (2009) study, if there are 10 studied items, then at least 12 different models need to be fitted separately. In educational assessment with a large item pool, this procedure can be prohibitively time consuming, especially if the model is high dimensional. Instead, in this paper, we propose to use statistical regularization methods that can handle multiple group comparisons simultaneously, making DIF detection extremely efficient.

Regularization is a process of adding information with the purpose of solving an ill-posed problem or to prevent overfitting. The regularization term, known as penalty, imposes a cost on the optimization function to remove parameters that have little influence on the fit of the model (Bauer et al., 2020; Belzak & Bauer, 2020; Magis et al., 2015; Tutz & Schauburger, 2015). The penalty can take on different forms depending on the research purpose, but the two most widely used types of penalty are the l_2 penalty (Hoerl & Kennard, 1970) and the l_1 penalty (Tibshirani, 1996). The former one considers the sums of the squared parameters whereas the latter one considers the sum of the absolute values of the parameters. In the context of DIF detection, an item level DIF parameter is introduced for each covariate and item parameter type. For instance, a uniform DIF due to a binary group variable would result in one DIF parameter per item that indicates how an item's difficulty differs across two groups. Then a penalty is imposed on the DIF parameters, and with appropriate regularization algorithms, they will either shrink to 0 implying no DIF or remain non-zero implying DIF.

Specifically, Magis et al. (2015) proposed a logistic regression least absolute shrinkage and selection operator DIF method (LR-lasso), which aims to identify uniform DIF in Rasch model using total score as the matching criterion. In their method, a lasso penalty is put on all DIF parameters, which are the regression

Using Lasso and Adaptive Lasso to Identify DIF in Multidimensional 2PL Models

Chun Wang
University of Washington
Ruoyi Zhu
University of Washington
Gongjun Xu
University of Michigan

Correspondence concerning this manuscript should be addressed to Chun Wang at:

312E Miller Hall
Measurement and Statistics
College of Education, University of Washington
2012 Skagit Ln, Seattle, WA 98105
e-mail: wang4066@uw.edu
phone: 217-722-7037

Acknowledgement: The project was supported by IES R305D200015

Using Lasso and Adaptive Lasso to Identify DIF in Multidimensional 2PL Models

Abstract

Differential item functioning (DIF) analysis refers to procedures that evaluate whether an item's characteristic differs for different groups of persons after controlling for overall differences in performance. DIF is routinely evaluated as a screening step to ensure items behavior the same across groups. Currently, the majority DIF studies focus predominately on unidimensional IRT models, although multidimensional IRT (MIRT) models provide a powerful tool for enriching the information gained in modern assessment. In this study, we explore regularization methods for DIF detection in MIRT models and compare their performance to the classic likelihood ratio test. Regularization methods have recently emerged as a new family of methods for DIF detection due to their advantages: (1) they bypass the tedious iterative purification procedure that is often needed in other methods for identifying anchor items, and (2) they can handle multiple covariates simultaneously. The specific regularization methods considered in the study are: lasso with expectation-maximization (EM), lasso with expectation-maximization-maximization (EMM) algorithm, and adaptive lasso with EM. Simulation results show that lasso EMM and adaptive lasso EM hold great promise when the sample size is large, and they both outperform lasso EM. A real data example from PROMIS depression and anxiety scales is presented in the end.

1 Introduction

The increasing availability of rich survey data and the emerging needs of assessing complex latent traits pose great challenges to existing techniques used to handle and analyze the data, in particular when the data are collected from heterogeneous populations. Different forms of multilevel, multidimensional item response theory (MIRT) models have been proposed to extract meaningful information from complex survey data. In addition to item calibration, items for large-scale standardized testing are routinely scrutinized for differential item functioning (DIF) to ensure equitable comparison of assessment outcomes among different examinee groups.

Differential item functioning (DIF) analysis refers to procedures that evaluate whether an item's characteristic differs for different groups of examinees after controlling for overall differences in performance. The term DIF was first defined by Holland and Thayer (1988), and two types of DIF are often differentiated: uniform DIF and non-uniform DIF. The former refers to an item having a constant advantage for a particular group, whereas the latter refers to the advantage varying in magnitude and/or direction across the latent trait continuum (Penfield & Camilli, 2006; Woods & Grimm, 2011). DIF is routinely evaluated for any new items added in large-scale assessments as a quality control step. For example, in the National Assessment of Educational Progress (NAEP), screening for item DIF often involves three comparisons: male vs female, White vs. Black, and White vs. Hispanic. In the English Language Proficiency Assessment for the 21st Century (ELPA21), DIF is evaluated across gender, ethnicity, economic status, English learner status, and disability status. In addition to these demographic or time variables, DIF may also be caused by cognitive variables that are relevant for item solution processes (Walker & Beretvas, 2003) or other item and test mode effects.

Early research on detecting multidimensional DIF is mostly extensions of unidimensional DIF indices that are based on the difference of examinees' performance in the focal group versus the reference group. Oshima, Raju, and Flowers (1997) first proposed an index

for DIF detection in two-dimensional models, namely, the multidimensional differential item functioning of items and tests (DFIT). They assumed that the examinees' true scores (defined as the sum of model-based probabilities) would be independent of group membership, and DIF is quantified by the expected difference in true scores between groups. Because their test statistics require the calculation of model-based probabilities, linking is needed to find two sets of item parameters for the focal and reference groups, respectively. This has to be done with iterative linking based on matching test response functions, which is inefficient and likely to be confounded with DIF itself. Stout, Li, Nandakumar, and Bolt (1997) extended simultaneous item bias test (SIBTEST) (Chang, Mazzeo, & Roussos, 1996; Shealy & Stout, 1993a, 1993b) to the two-dimensional case and proposed MULTISIB. Because their methods use only observed total score, they are generalizable to other IRT models and are not susceptible to model misfit. In the 2-dimensional case, a matching subtest is needed for each dimension, and because matching is based on the combination of total scores on both subtests, it would be very complicated to expand the procedure to tests beyond two dimensions.

Three other more flexible approaches are the multiple indicators multiple causes (MIMIC) models, multiple-group IRT modeling, and logistic regression (Choi, Gibbons, & Crane, 2011; Swaminathan & Rogers, 1990; Zumbo, 1999). These three approaches share great similarity. In a MIMIC model, at least one observed variable (i.e., group variable), often called a casual indicator, predicts a latent variable (Jöreskog & Goldberger, 1975). In the multiple-group IRT approach, rather than regressing latent θ on the grouping variable, one fits two multiple-group IRT models with different equality constraints on target items to the data and conducts a likelihood ratio test to test for DIF (Suh & Cho, 2014). Logistic regression, as the name entails, recasts the IRT model as a logistic regression model such that uniform DIF is modeled by including group variable as a predictor in addition to θ , whereas non-uniform DIF has θ -by-group interaction as an additional predictor. Of note, while MIMIC and multiple-group IRT model can naturally handle θ distribution differences by groups

(i.e., impact), in logistic regression however, one does not distinguish the distribution of θ for different groups. Hence, logistic regression does not usually account for impact. All three approaches perform better using a free-baseline designated-anchor approach. That is, while holding the anchor item parameters the same across groups, the full model that allows all studied items to have DIF is fitted first. Then a constrained model is fitted, one for each item. Because the analysis proceeds one item at a time, testing for multiple studied items usually require additional control for Type I error rate, such as the Benjamini-Hochberg procedure (BH) (Benjamini & Hochberg, 1995; Lee, Bulut, & Suh, 2017; Raykov, Marcoulides, Lee, & Chang, 2013). Moreover, the selection of the designated anchor items is critical to the success of the methods. Various methods for identifying anchors have been suggested, and most of them are iterative purification procedures (Bolt, Hare, Vitale, & Newman, 2004; Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006; Kopf, Zeileis, & Strobl, 2015).

Procedurally for all these three approaches, it is cumbersome to perform a likelihood ratio test (or Wald test) separately for one item at a time. For instance, in Woods (2009) study, if there are 10 studied items, then at least 12 different models need to be fitted separately. In educational assessment with a large item pool, this procedure can be prohibitively time consuming, especially if the model is high dimensional. Instead, in this paper, we propose to use statistical regularization methods that can handle multiple group comparisons simultaneously, making DIF detection extremely efficient.

Regularization is a process of adding information with the purpose of solving an ill-posed problem or to prevent overfitting. The regularization term, known as penalty, imposes a cost on the optimization function to remove parameters that have little influence on the fit of the model (Bauer, Belzak, & Cole, 2020; Belzak & Bauer, 2020; Magis, Tuerlinckx, & De Boeck, 2015; Tutz & Schauberger, 2015). The penalty can take on different forms depending on the research purpose, but the two most widely used types of penalty are the l_2 penalty (Hoerl & Kennard, 1970) and the l_1 penalty (Tibshirani, 1996). The former one considers the sums of the squared parameters whereas the latter one considers the sum of the absolute values of

the parameters. In the context of DIF detection, an item level DIF parameter is introduced for each covariate and item parameter type. For instance, a uniform DIF due to a binary group variable would result in one DIF parameter per item that indicates how an item’s difficulty differs across two groups. Then a penalty is imposed on the DIF parameters, and with appropriate regularization algorithms, they will either shrink to 0 implying no DIF or remain non-zero implying DIF.

Specifically, Magis et al. (2015) proposed a logistic regression least absolute shrinkage and selection operator DIF method (LR-lasso), which aims to identify uniform DIF in Rasch model using total score as the matching criterion. In their method, a lasso penalty is put on all DIF parameters, which are the regression coefficient in front of the item-group interaction in a logistic regression model. They found that for small samples, the LR-lasso method outperforms the classic LR method or Mantel-Haenszel method in terms of false positive and true positive rates. The advantage of LR-lasso seems to diminish with larger sample sizes. Basing on Rasch model, Tutz and Schauberger (2015) further studied the penalty approach in detecting uniform DIF when there are multiple, potentially correlated covariates, that collectively cause DIF. Compared to Magis et al. (2015)’s study, two major difference are (1) they used latent trait θ as the matching criterion instead of total score; and (2) they used group lasso penalty (Yuan & Lin, 2006) by grouping DIF parameters of an item on all covariates as a unit. This way, when the “unit” shrinks to 0, it implies the item is DIF-free, otherwise, the item is considered having DIF regardless of the number of actual non-zero DIF parameters within a unit. Without recourse to cumbersome pair-wise hypothesis tests, their group lasso approach arrives at a model in which DIF is included only for those covariates and only on those items where DIF inclusion meaningfully increases model fit. Their simulation results demonstrate both the feasibility and promise of the penalty approach in detecting DIF caused by multiple covariates. Along a similar line of inquiry, Bauer et al. (2020) also studied the regularized DIF detection approach (reg-DIF) in the presence of multiple covariates, albeit within the moderated nonlinear factor analysis (MNLFA) framework. They studied

both uniform and non-uniform DIF and rather than using group lasso, they simply put l_1 penalty on all DIF parameters (i.e., sum of absolute values). Then an item is considered having DIF when at least one of the DIF parameters is non-zero. They found that reg-DIF performs the best when both DIF magnitude and sample size are large, whereas the performance is comprised when DIF is particularly pervasive. Within the same MNLFA framework, Belzak and Bauer (2020) delved deeper into the classic two-group setting and provide a thorough empirical investigation of Reg-DIF against “business-as-usual” IRT-LR-DIF. They found that Reg-DIF shows far better Type I error control than IRT-LR-DIF, particularly when there are considerable amounts of DIF and the sample size is sufficiently large.

Building on the preliminary evidence demonstrating the promise of the regularized DIF approach, the contribution of this study is two-fold. First, we will explore DIF detection in the context of the simple-structure¹ multidimensional two-parameter logistic model that is pervasively used in applied settings. We will consider both uniform and non-uniform DIF in a three-group comparison scenario (i.e., one reference and two focal groups), such that the conclusions are more generalizable to future complex applications. Second, we provide technical details regarding the coordinate-wise soft-thresholding within expectation-maximization (EM) algorithm for DIF detection at individual parameter level, along with self-written code for implementing the algorithms. As both Bauer et al. (2020) and Belzak and Bauer (2020) mentioned, their way of implementing reg-DIF is non-standard and their reliance on SAS NLMIXED is relatively inefficient, which calls for the needs of developing computationally sound algorithms for reg-DIF. Another note to make is, in Belzak and Bauer (2020), their benchmark IRT-LR-DIF method proceeds by treating all items except the studied item as anchors. Instead, we use a version of the IRT-LR-DIF method in which all items except the studied item and anchor items have DIF. Using a least constrained model as a baseline for LR comparison should help control Type I error rate. Third, we propose to use

¹Simple-structure refers to the factor structure in which each item only loads on one factor; in contrast, complex structure indicates that items in a test can load on more than one factor.

two variations of the lasso method, namely the lasso expectation-maximization-maximization (EMM) algorithm and the adaptive lasso method. As the paper unfolds, it is clear that these two alternatives perform better than the vanilla lasso EM algorithm.

The rest of the paper is organized as follows. First, the M2PL model with group covariates is introduced, illustrating the uniform and non-uniform DIF parameterizations. Second, model estimation via l_1 regularization (i.e., lasso) is described, along with a discussion about adaptive lasso algorithm and a modified EMM algorithm. Then, a comprehensive simulation study is presented and a real data analysis example is provided at the end.

2 Methods

2.1 Multidimensional 2PL Model with Categorical Group Covariates

Let J denote test length and K denote the total number of dimensions. For a dichotomously scored item j , the probability that person i with a latent trait vector $\boldsymbol{\theta}_i$ giving a correct response to item j is

$$P_j(u_{ij} = 1|\boldsymbol{\theta}_i) = \frac{1}{1 + e^{-[\mathbf{a}_j^T \boldsymbol{\theta}_i + d_j + (\mathbf{X}_i \boldsymbol{\gamma}_j) \boldsymbol{\theta}_i + \mathbf{X}_i \boldsymbol{\beta}_j]}} \quad (i = 1, \dots, N; j = 1, 2, \dots, J). \quad (1)$$

In Equation 1, \mathbf{a}_j is a K -by-1 vector of discriminations for item j , d_j is an intercept of item j which can be interpreted as item easiness, and $\boldsymbol{\theta}_i$ is a K -by-1 vector of latent trait for person i . In addition, \mathbf{X}_i is a 1-by- P vector including all the grouping information related to DIF, and $\boldsymbol{\beta}_j$ is also a P -by-1 vector of regression coefficients implying the effect of grouping variables on correct item response probability. Similarly, $\boldsymbol{\gamma}_j$ is a P -by- K matrix of regression coefficients that denote the interaction effects of $\boldsymbol{\theta}$ and grouping variable on item responses. Please note that in a confirmatory MIRT model, if $a_{jk} = 0$, then the k th column of $\boldsymbol{\gamma}_j$ will be zero by default. Take NAEP analysis for ethnicity DIF as an example. Since it includes

two comparisons (i.e., White vs. Black, and White vs. Hispanic), $P=2$ in this case such that a White student will have $\mathbf{X}_i = (0, 0)$, a Black student will have $\mathbf{X}_i = (1, 0)$ and a Hispanic student will have $\mathbf{X}_i = (0, 1)$. Then we can spell out $\boldsymbol{\gamma}_j$ as follows:

$$\boldsymbol{\gamma}_j = (\boldsymbol{\gamma}_{j1}, \boldsymbol{\gamma}_{j2})^T = \begin{pmatrix} \gamma_{j11} & \gamma_{j12} & \cdots & \gamma_{j1k} & \cdots & \gamma_{j1K} \\ \gamma_{j21} & \gamma_{j22} & \cdots & \gamma_{j2k} & \cdots & \gamma_{j2K} \end{pmatrix} \quad (k = 1, \dots, K),$$

where $\boldsymbol{\gamma}_{j1}$ is the non-uniform DIF effect for the first focal group and $\boldsymbol{\gamma}_{j2}$ is the non-uniform DIF effect for the second focal group. $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2})^T$ denotes the uniform DIF effect, where β_{j1} is the uniform DIF effect for the first focal group and β_{j2} is the uniform DIF effect for the second focal group. Including person covariates in the model also follows the same spirit of the “person explanatory model” (Wilson, De Boeck, & Carstensen, 2008). By way of this parameterizations, if item j does not have DIF, then $\boldsymbol{\gamma}_j = \mathbf{0}$ and $\boldsymbol{\beta}_j = \mathbf{0}$. If item j has uniform DIF, then $\boldsymbol{\gamma}_j = \mathbf{0}$. Similar to the multiple-group IRT approach, $\boldsymbol{\theta}_i$ in Equation 1 can be written as $\boldsymbol{\theta}_{i(p)}$ to reflect that the distribution of $\boldsymbol{\theta}$ is allowed to differ across different groups.

Model Identifiability. Before we proceed to model estimation, an important premise to check is model identifiability. Here, we extend Tutz and Schauberger (2015)’s conclusion to multidimensional models and non-uniform DIF conditions. Specifically, for model defined in Equation 1 to be identifiable, three assumptions need to be satisfied:

1. The P -by-2 matrix, $(\mathbf{1}, \mathbf{X}_i^T)$, has full rank.
2. $\boldsymbol{\theta}$ in the reference group has mean of 0 and variance of 1 for all dimensions.
3. When the test displays a simple multidimensional structure, we need q DIF-free anchor items, one for each dimension separately.

While the first two assumptions are exactly the same as in Tutz and Schauberger (2015), the third assumption is unique to multidimensional models. It is needed because the reason

of model in Equation 1 being non-identifiable, after we fix the scale of $\boldsymbol{\theta}$, is due to the indeterminacy of the two parts:

$$(d_j + \mathbf{X}_i\boldsymbol{\beta}_j) + [\mathbf{a}_j^T + (\mathbf{X}_i\boldsymbol{\gamma}_j)]\boldsymbol{\theta}_i$$

For the first part, we can write

$$d_j + \mathbf{X}_i\boldsymbol{\beta}_j = (d_j + \mathbf{X}_i\mathbf{c}) + \mathbf{X}_i(\boldsymbol{\beta}_j - \mathbf{c}) = \tilde{d}_j + \mathbf{X}_i\tilde{\boldsymbol{\beta}}_j, \quad (2)$$

where \mathbf{c} is an arbitrary constant vector. Setting $\boldsymbol{\beta}_1 = 0$ (i.e., arbitrarily assuming item 1 is the anchor item), that means for Equation 2 to hold for any value of \mathbf{X}_i , $\tilde{\boldsymbol{\beta}}_1$ has to be 0 as well. This implies that $\boldsymbol{\beta}_1 - \mathbf{c} = 0$, hence $\mathbf{c} = \boldsymbol{\beta}_1 = 0$. So setting one anchor item's intercept parameter to 0 will remove indeterminacy in part I.

For the second part, we have

$$\mathbf{a}_j^T + (\mathbf{X}_i\boldsymbol{\gamma}_j) = (\mathbf{a}_j^T + \mathbf{X}_i\mathbf{c}) + \mathbf{X}_i(\boldsymbol{\gamma}_j - \mathbf{c}) = \tilde{\mathbf{a}}_j^T + \mathbf{X}_i\tilde{\boldsymbol{\gamma}}_j, \quad (3)$$

where both \mathbf{c} and $\boldsymbol{\gamma}_j$ are P -by- K matrix and they contain columns of systematic 0's depending on the loading structure in \mathbf{a}_j . Say, if item j loads on the first dimension, then both \mathbf{c} and $\boldsymbol{\gamma}_j$ have a potentially non-zero first column, whereas the remaining columns are all 0 by default. In this regard, even setting $\boldsymbol{\gamma}_j = 0$ will only result in the first column of \mathbf{c} being 0. Hence, if the test displays a simple structure, we need K anchor items, one for each dimension. If the test displays a complex structure, we may only need 1 anchor item which loads on all three dimensions.²

²As we will make clearer in the rest of the paper, pre-selecting anchor items is not necessary for regularization methods as non-DIF items will have their DIF parameters shrunk to 0 so that they will serve as anchor items.

2.2 Model Estimation and DIF Detection via l_1 Regularization

Estimation of the model in Equation 1 can proceed using the state-of-art EM algorithm. Specifically, let u_{ij} denote the response of the i th person to the j th item, and assume person i belongs to group g . Note that throughout the paper, we assume only categorical covariates \mathbf{X}_i are considered such that each person is assigned to one and only one group defined by the collection of the covariates. For instance, if two levels of gender and three levels of ethnicity covariates are considered, then each person can be uniquely assigned to one of the six groups. We consider categorical covariates with the intention to show that the $\boldsymbol{\theta}$ distribution can be estimated separately per group. If one intends to include continuous covariates and handle impact at the same time, both the model and the estimation methods need to be updated, as we will mention in the discussion session.

Let Δ denote the set of model parameters, i.e., item parameters (\mathbf{a} , \mathbf{d} and $\boldsymbol{\beta}$) and latent trait distribution parameters ($\boldsymbol{\mu}_p$ and $\boldsymbol{\Sigma}_p$). The marginal likelihood given covariates \mathbf{X} and response \mathbf{u} is

$$L(\Delta) \equiv \int L(\Delta | \mathbf{X}, \mathbf{u}, \boldsymbol{\theta}) \partial \boldsymbol{\theta} = \prod_{g=1}^G \prod_{i=1}^{N_g} \int \prod_{j=1}^J L(\mathbf{a}_j, d_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j | \mathbf{X}_i, u_{ij}, \boldsymbol{\theta}) f(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g | \boldsymbol{\theta}) \partial \boldsymbol{\theta}, \quad (4)$$

where

$$L(\mathbf{a}_j, d_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j | \mathbf{X}_i, u_{ij}, \boldsymbol{\theta}) = P_j(\boldsymbol{\theta})^{u_{ij}} (1 - P_j(\boldsymbol{\theta}))^{1-u_{ij}}$$

is the likelihood of item parameters and

$$f(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p | \boldsymbol{\theta}) = (2\pi)^{-\frac{K}{2}} |\boldsymbol{\Sigma}_g|^{-\frac{1}{2}} e^{-0.5(\boldsymbol{\theta} - \boldsymbol{\mu}_g)^T |\boldsymbol{\Sigma}_g|^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_g)}$$

is the density function of $\boldsymbol{\theta}$ in the population group g . $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ are the population mean and covariance matrix respectively. G is the total number of groups and N_g is the sample size for group g . One maximizes the marginal likelihood to obtain the maximum likelihood

estimator (MLE)

$$(\hat{\mathbf{a}}, \hat{\mathbf{d}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = \underset{\mathbf{a}, \mathbf{d}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma}}{\operatorname{argmax}} \log(L(\boldsymbol{\Delta})). \quad (5)$$

However, the MLE does not serve the purpose of DIF detection. Instead, we consider a l_1 regularized estimator by maximizing the following objective function

$$\log L(\boldsymbol{\Delta}) - \eta_1 \|\boldsymbol{\beta}\|_1 - \eta_1 \|\boldsymbol{\gamma}\|_1, \quad (6)$$

where

$$\|\boldsymbol{\beta}\|_1 = \sum_j^J \sum_p^P |\beta_{jp}|, \quad \|\boldsymbol{\gamma}\|_1 = \sum_j^J \sum_p^P \sum_k^K |\gamma_{jpk}| \mathbf{1}_{a_{jk} \neq 0},$$

and $\eta_1 > 0$ is regularization parameters that controls sparsity.

Directly maximizing the marginal likelihood in Equation 4 is challenging, instead, the EM algorithm provides a viable computational tool (Bock & Aitkin, 1981; Wang, Chen, & Jiang, 2020). The EM algorithm alternates between the E-step and M-step. In the E-step, we construct the conditional expectation of the complete data log-likelihood with respect to missing data (i.e., $\boldsymbol{\theta}$). Suppose at the $(t + 1)$ th EM cycle, then we have

$$\begin{aligned} Q(\boldsymbol{\Delta} | \boldsymbol{\Delta}^{(t)}) &= E_{h(\boldsymbol{\theta} | \mathbf{X}, \mathbf{u}, \boldsymbol{\Delta}^{(t)})}(\log(L(\boldsymbol{\Delta} | \mathbf{X}, \mathbf{u}, \boldsymbol{\theta}))) \\ &= \sum_g^G \sum_i^{N_g} \left[\int l(\mathbf{a}, \mathbf{d}, \boldsymbol{\beta}, \boldsymbol{\gamma} | \mathbf{X}_i, \mathbf{u}_i, \boldsymbol{\theta}) h(\boldsymbol{\theta} | \mathbf{X}_i, \mathbf{u}_i, \boldsymbol{\Delta}^{(t)}) \partial \boldsymbol{\theta} + \int \log f(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p | \boldsymbol{\theta}) h(\boldsymbol{\theta} | \mathbf{X}_i, \mathbf{u}_i, \boldsymbol{\Delta}^{(t)}) \partial \boldsymbol{\theta} \right] \\ &= \sum_g^G \sum_i^{N_g} \sum_j^J \left[\int l(\mathbf{a}_j, d_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j | \mathbf{X}_i, \mathbf{u}_i, \boldsymbol{\theta}) h(\boldsymbol{\theta} | \mathbf{X}_i, \mathbf{u}_i, \boldsymbol{\Delta}^{(t)}) \partial \boldsymbol{\theta} \right] \\ &+ \sum_g^G \sum_i^{N_g} \left[\int \log f(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g | \boldsymbol{\theta}) h(\boldsymbol{\theta} | \mathbf{X}_i, \mathbf{u}_i, \boldsymbol{\Delta}^{(t)}) \partial \boldsymbol{\theta} \right] \\ &\equiv \sum_j^J Q_j(\mathbf{a}_j, d_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j | \boldsymbol{\Delta}^{(t)}) + Q(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \boldsymbol{\Delta}^{(t)}). \end{aligned} \quad (7)$$

When the number of dimensions is not high (i.e., 2 or 3), Gauss-Hermite quadrature can be used to approximate the integrals in Equation 7, otherwise, either Monte Carlo integration

(Chen, Wang, Xin, & Chang, 2017; Newman & Barkema, 1999) or other variational methods (Cho, Wang, Zhang, & Xu, 2021) could be used. In the M-step, Equation 7 is maximized with respect to each parameter.

First, taking derivatives of Equation 7 with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ result in closed-form solutions for both parameters. That is, denote M_0 as the number of points we evenly take from each coordinate dimension, resulting in $M = (M_0)^K$ total quadrature samples, and each sample \mathbf{q}_m is a K -dimensional vector. Then it can be shown that, the population mean of a focal group g is

$$\hat{\boldsymbol{\mu}}_g = \frac{\sum_{m=1}^M n_{gm} \mathbf{q}_m}{N_g}, \quad (8)$$

where $n_{gm} = \sum_{i=1}^{N_g} h(\mathbf{q}_m | \mathbf{X}_i, \mathbf{u}_i, \boldsymbol{\Delta}^{(t)})$ is the expected number of persons in group g and m th quadrature bin. The population covariance matrix of the reference group is (without loss of generality, let us assume the first group is the reference group, i.e., $g = 1$)

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{\sum_{m=1}^M n_{1m} (\mathbf{q}_m)^t (\mathbf{q}_m)}{N_1}. \quad (9)$$

For the purpose of model identifiability, the variances of $\boldsymbol{\theta}$ in the reference group are fixed at 1, whereas the covariances (hence correlation) can be freely estimated. To achieve this constraint, we standardize the covariance matrix $\hat{\boldsymbol{\Sigma}}_1$ by rescaling the quadrature vectors as follows

$$\mathbf{q}_m^* = \left(\frac{q_{m1}}{(\hat{\boldsymbol{\Sigma}}_1)_{11}}, \dots, \frac{q_{mK}}{(\hat{\boldsymbol{\Sigma}}_1)_{KK}} \right).$$

These rescaled quadrature points are used for computing the estimated covariance matrix for focal groups, i.e.,

$$\hat{\boldsymbol{\Sigma}}_g = \frac{\sum_{m=1}^M n_{gm} (\mathbf{q}_m - \hat{\boldsymbol{\mu}}_g)^t (\mathbf{q}_m - \hat{\boldsymbol{\mu}}_g)}{N_g}. \quad (10)$$

Second, for the item parameters, we can maximize $Q_j(\mathbf{a}_j, d_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j | \boldsymbol{\Delta}^{(t)})$ separately for item j . Specifically, for \mathbf{a}_j and d_j , as there are not closed-form solutions for the gradient of these parameters, we use Newton-Raphson method to find the maximum numerically. Take

d_j as an example. At the $(r + 1)$ th iteration (of M-step) within the $(t + 1)$ th EM cycle, we update d_j using

$$d_j^{(r+1)} = d_j^{(r)} - \frac{\partial_{d_j} Q_j(\mathbf{a}_j, \mathbf{d}_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j | \boldsymbol{\Delta}^{(t)})}{\partial_{d_j}^2 Q_j(\mathbf{a}_j, \mathbf{d}_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j | \boldsymbol{\Delta}^{(t)})}. \quad (11)$$

The parameters \mathbf{a}_j can be updated similarly. Below is a description of how the DIF parameters are updated in the M-step.

2.2.1 Uniform DIF

To detect uniform DIF, $\boldsymbol{\gamma}$ parameters vanish from the previous exposition, and the focus is to update $\boldsymbol{\beta}_j$ for each item separately within the M-step by maximizing the following objective function

$$\hat{\boldsymbol{\beta}}_j^{(t+1)} = \operatorname{argmax}_{\boldsymbol{\beta}_j} \left[Q_j(\mathbf{a}_j, d_j, \boldsymbol{\beta}_j | \boldsymbol{\Delta}^{(t)}) - \eta_1 \|\boldsymbol{\beta}_j\|_1 \right]. \quad (12)$$

We use soft-thresholding technique within coordinate descent algorithm for estimating $\boldsymbol{\beta}_j$. Coordinate descent updates one parameter at a time while treating all other parameters as constant, and cycles through all parameters in each iteration of the optimization routine, i.e., we update β_{jp} one at a time.

Following Sun, Chen, Liu, Ying, and Xin (2016), we first employ a quadratic approximation of $Q_j(\mathbf{a}_j, d_j, \boldsymbol{\beta}_j | \boldsymbol{\Delta}^{(t)}) \equiv Q_j(\boldsymbol{\beta}_j | \boldsymbol{\Delta}^{(t)})$ as a function of β_{jp} . That is,

$$Q_j(\boldsymbol{\beta}_j | \boldsymbol{\Delta}^{(t)}) \approx Q_j(\beta_{jp}^{(r)} | \boldsymbol{\Delta}^{(t)}) + \partial_{\beta_{jp}} Q_j(\beta_{jp} | \boldsymbol{\Delta}^{(t)}) \times (\beta_{jp} - \beta_{jp}^{(r)}) + \frac{\partial_{\beta_{jp}}^2 Q_j(\beta_{jp} | \boldsymbol{\Delta}^{(t)})}{2} (\beta_{jp}^{(r)} - \beta_{jp})^2, \quad (13)$$

where $\beta_{jp}^{(r)}$ is the r th iteration within the M-step of the $(t + 1)$ th EM cycle. The l_1 -penalized maximization with the approximated Q -function aims to maximize the following objective function

$$\left[Q_j(\beta_{jp}^{(r)} | \boldsymbol{\Delta}^{(t)}) + \partial_{\beta_{jp}} Q_j(\beta_{jp} | \boldsymbol{\Delta}^{(t)}) \times (\beta_{jp} - \beta_{jp}^{(r)}) + \frac{\partial_{\beta_{jp}}^2 Q_j(\beta_{jp} | \boldsymbol{\Delta}^{(t)})}{2} (\beta_{jp}^{(r)} - \beta_{jp})^2 - \eta_1 |\beta_{jp}| \right]. \quad (14)$$

Setting the first derivative of Equation 14 with respect to β_{jp} equal to 0 yield the following

updated rule

$$\beta_{jp}^{(k+1)} = -\frac{\text{soft}\left(-\partial_{\beta_{jp}}^2 Q_j(\beta_{jp}|\Delta^{(t)}) \times \beta_{jp}^{(r)} + \partial_{\beta_{jp}} Q_j(\beta_{jp}|\Delta^{(t)}), \eta_1\right)}{\partial_{\beta_{jp}}^2 Q_j(\beta_{jp}|\Delta^{(t)})}, \quad (15)$$

where the soft threshold operator is defined as $\text{soft}(S, \eta) = \text{sign}(S)(|S| - \eta)_+$ (Donoho & Johnstone, 1995). After the updated values of all parameters of item j are smaller than a pre-specified convergence tolerance, cyclical coordinate descent is applied again to estimate the next item $j + 1$. After all J items' parameters are updated, the M step ends and the estimating process moves to the next EM cycle. To ensure proper convergence, we use the estimator for $\eta = 0$ (the MLE) as starting value for the estimation of later η values.

2.2.2 Non-uniform DIF

For the non-uniform DIF, we consider the condition in which DIF occurs on both slopes and intercepts. Other parameters \mathbf{a} , \mathbf{d} , $\boldsymbol{\mu}_p$ and $\boldsymbol{\Sigma}_p$ are estimated in the same way as in the previous case. Again, the EM algorithm with Gauss-Hermite quadrature approximation is used, and in the M-step, the objective function to be maximized is

$$\hat{\boldsymbol{\gamma}}_j^{(t+1)} = \underset{\boldsymbol{\gamma}_j}{\text{argmax}} \left[Q_j(\mathbf{a}_j, d_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j | \Delta^{(t)}) - \eta_1 \|\boldsymbol{\beta}_j\|_1 - \eta_2 \|\boldsymbol{\gamma}_j\|_1 \right]. \quad (16)$$

The cyclical coordinate descent with soft-thresholding is again used to obtain the estimates.

2.3 Tuning Parameter Selection, EMM, and Adaptive Lasso

The selection of the tuning parameter is a key component of the reg-DIF method. Taking uniform-DIF as an example. As η_1 keeps increasing, more DIF parameters β_{jp} shrink to 0. Denote $\hat{\boldsymbol{\beta}}_{\eta_1}$ as the regularized estimator at tuning parameter value η_1 . Since the regularized estimators are biased, we perform a re-estimation step without penalty for each tuning parameter value η_1 based on the DIF item pattern identified by non-zero $\hat{\boldsymbol{\beta}}_{\eta_1}$. Then, we apply

two information criterion, Akaike information criterion (AIC) and Bayesian information criterion (BIC), to select the best-fitting tuning parameter. Specifically, with the re-estimated model parameters using EM algorithm without penalty, resulting in $\hat{\Delta}_{\eta_1}^* \equiv (\hat{\mathbf{a}}_{\eta_1}^*, \hat{\mathbf{d}}_{\eta_1}^*, \hat{\boldsymbol{\beta}}_{\eta_1}^*)$, AIC and BIC can be calculated by

$$AIC \equiv -2 \times \log L(\hat{\Delta}_{\eta_1}^*) + 2 \|\hat{\boldsymbol{\beta}}_{\eta_1}\|_0, \quad (17)$$

and

$$BIC \equiv -2 \times \log L(\hat{\Delta}_{\eta_1}^*) + \log N \times \|\hat{\boldsymbol{\beta}}_{\eta_1}\|_0 \quad (18)$$

where the l_0 norm is calculated by $\|\hat{\boldsymbol{\beta}}\|_0 = \sum_{j,p} I(\beta_{jp} \neq 0)$ ³.

In the lasso estimation, while $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ of non-DIF items shrink to 0 due to the soft-thresholding operator, those non-zero parameters of DIF items also shrink and hence their estimates are biased. Such bias may propagate via iterative EM cycles and eventually the DIF items may not be properly identified. To overcome this bias, two solutions are considered.

The first solution is to use an adaptive lasso algorithm (Zou, 2006). The primary idea is to use different weights for penalizing different coefficient in the l_1 penalty. Take uniform DIF detection as an example, the objective function that needs to be maximized now becomes

$$Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma} | \hat{\mathbf{a}}^{(t)}, \hat{\mathbf{d}}^{(t)}, \hat{\boldsymbol{\gamma}}^{(t)}, \hat{\boldsymbol{\mu}}_p^{(t)}, \hat{\boldsymbol{\Sigma}}_p^{(t)}) - \eta_a \sum_j \hat{\mathbf{w}}_j \|\boldsymbol{\gamma}_j\|_1, \quad (19)$$

where η_a is the adaptive-lasso tuning parameter and $\hat{\mathbf{w}}_j$ is the weight vector. When $\eta_a = 0$, we have the MLE of $\boldsymbol{\gamma}$. Denote the MLE of $\boldsymbol{\gamma}$ as $\hat{\boldsymbol{\gamma}}^{(\text{ML})}$. Then the weight vector $\hat{\mathbf{w}}_j = 1/|\hat{\boldsymbol{\gamma}}_j^{(\text{ML})}|^\lambda$, where λ is a tuning constant which is set to 1 in the current study. In adaptive lasso, coefficients with larger absolute values (i.e., higher absolute MLE) are assigned lower weight of penalty, resulting in lower bias than the lasso estimators.

³Note that $L(\hat{\Delta}_{\eta_1}^*)$ is the marginal likelihood evaluated at $\hat{\Delta}_{\eta_1}^*$ by definition, but operationally, we replaced this quantity by the Q-function defined in Equation 7 upon convergence to save computation time. They are not exactly the same though because the Q-function approximates the marginal log-likelihood.

The second solution is to re-estimate non-zero estimators during each EM cycle and use the re-estimated coefficients to update the next E-step. That is, we can perform one more M-step without penalty but using the DIF detection results from the previous M-step within each EM cycle. Hence, each E-step is followed by two M-steps, first with penalty and second without penalty. We call this method “EMM” algorithm. In the Appendix, we present algorithmic details for three algorithms.⁴

3 Simulation Studies

Two simulation studies were conducted to evaluate the performance of the regularization methods in terms of detecting uniform DIF (study I) and non-uniform DIF (study II). The likelihood ratio test (LRT) was used as a reference. For both studies, two-dimensional two-parameter logistic (2PL) IRT model was used. The total number of items was fixed at 20. Two discrimination parameters were generated from Uniform(1.5, 2.5) and the boundary parameters were generated from N(0,1). Each item measured only one of the two latent traits. The true item parameters are given in Table 1.

Table 1: Simulated True Item Parameters

Item	1	2	3	4	5	6	7	8	9	10
\mathbf{a}_1	2.17	0	2.41	2.45	2.34	1.84	1.85	1.92	1.94	1.90
\mathbf{a}_2	0	2.46	0	0	0	0	0	0	0	0
\mathbf{d}	0.03	-1.28	0.58	-2.06	0.12	3.25	-0.41	-0.51	0.89	1.33
Item	11	12	13	14	15	16	17	18	19	20
\mathbf{a}_1	1.92	0	0	0	0	0	0	0	0	0
\mathbf{a}_2	0	2.43	1.82	2.22	1.93	1.88	1.84	2.12	2.42	2.15
\mathbf{d}	0.85	0.82	-0.37	-0.99	-0.27	0.19	1.73	0.05	-1.86	-0.63

Three factors were manipulated. The total sample size had two levels, 1500 and 3000; the DIF proportion had two levels, 20% and 60%. This choice was consistent with prior studies (Suh & Cho, 2014). In addition, two levels of correlations between two factors were

⁴The R code for the three algorithms are available at <https://github.com/wang4066/MIRT-RegDIF>

considered: low level at 0.25 and high level at 0.85 (Jiang, Wang, & Weiss, 2016). The total sample size was evenly divided into three groups, one reference group and two focal groups.

3.1 Simulation Study I

The first focal group has small magnitude DIF ($\beta_{j1} = 0.5$) and the second focal group has large magnitude DIF ($\beta_{j2} = 1$), where $j = 4, 5, 12, 13$ for the 20% DIF condition and $j = 4, 5, 6, 7, 8, 9, 12, 13, 14, 15, 16, 17$ for the 60% DIF condition. Low DIF proportion condition was the same as Tutz and Schauberger (2015), whereas the high DIF proportion was the same as Bauer et al. (2020). To further exemplify the DIF magnitude, the average area between the expected item score curves from focal and reference groups, weighted by the normal distribution was computed, and such area is called wABC (Edelen, Stucky, & Chandra, 2015). We used standard normal distribution in computing wABC, and the wABC values of the true DIF items are given in table ???. As shown, the DIF magnitudes were selected to follow the convention in literature (Suh & Cho, 2014). θ follows a multivariate normal distribution with means zero and variances one. 50 independent datasets were generated from the model in Equation 1 for each condition. Similar to the simulation design used in Belzak and Bauer (2020), we kept DIF effects constant across replications within a given condition to avoid mixing within-and between-condition variability in the magnitude of DIF.

The reg-DIF methods do not need pre-specified anchor items, but LRT requires a pre-specified set of designated anchors to link the metric of θ for different groups (Woods, 2009). Unfortunately, in most real-data scenarios, researchers, or even content experts, have difficulty identifying anchor items. Hence, we adopt an iterative procedure to find a set of anchor items. The procedure began with assuming all items are DIF-free. Conduct a LRT assuming all items except the studied item having no DIF. Cycle through all items in one round, and items displaying DIF were excluded from the living set of anchors. LRT was conducted again using the new reduced anchor set. These steps were repeated until two successive steps suggest the same sets of anchors (Kopf, Zeileis, & Strobl, 2014), and this final set was

considered a purified set. Then, in LRT, each non-DIF item was tested individually for DIF. For a particular studied item, an analysis begins with a test of the null hypothesis that target parameters (i.e., intercept for uniform DIF or slope and intercept for non-uniform DIF) for the studied item were group invariant. Two nested models were compared: a model with all parameters for the studied item constrained equal across groups versus a model with target parameters for the studied item allowing to vary across groups. In both models, the anchor item parameters were constrained equal across groups. Different versions of LRT can be used to test DIF between reference group and one focal group yielding a typical two-group comparison scenario, or DIF between reference group and two focal groups omnibusly yielding a three-group comparison scenario. Benjamini-Hochberg (BH) false discovery rate control (Benjamini & Hochberg, 1995; Lee et al., 2017; Raykov et al., 2013) was used to control for family-wise error rate. Different from Bauer et al. (2020) and Belzak and Bauer (2020), all parameters of the remaining items (i.e., non-studied items, non-anchor items) in both models were also permitted to vary among groups. This yields a free-baseline designated anchor approach, which should outperform a constrained baseline approach that assumes all non-studied non-anchor items are DIF-free (Woods, 2009). Regularization methods was carried out using self-written R code, whereas LRT was conducted using the ‘mirt’ package (Chalmers et al., 2020).

Table 2: Uniform DIF magnitude measured by wABC

Item	4	5	6	7	8	9
Focal 1	0.06	0.07	0.03	0.08	0.08	0.07
Focal 2	0.12	0.13	0.05	0.16	0.15	0.13
Item	12	13	14	15	16	17
Focal 1	0.06	0.08	0.07	0.08	0.08	0.06
Focal 2	0.12	0.16	0.14	0.15	0.15	0.11

Evaluation criterion includes Type I error and power at both omnibus level and group level, as well as the DIF parameter recovery. At omnibus level, the null hypothesis for LRT is that the item has no DIF for neither of the focal groups, whereas the alternative hypothesis is the item has DIF for *both* focal groups. This is due to the intrinsic multiple-group IRT

model estimation as part of the LRT approach. That is, for the studied item, the full model assumes the parameters of this item vary across all three groups whereas the constrained model assumes the parameters of this item are the same among all three groups. In contrast, at group level, say for the first focal group, the null hypothesis for LRT becomes that the item has no DIF between the reference and focal group and the alternative hypothesis is that the item has DIF between these two groups. As the second focal group is irrelevant in this context, only a subset of item responses from reference and first focal group were fed into the ‘multipleGroup’ function in the ‘mirt’ package. In contrast, in the reg-DIF method, a distinct DIF parameter is designated for each studied item parameter and each focal group. Therefore, an item is said to have omnibus DIF if at least one focal group shows DIF on that item, i.e., when there is at least one non-zero element in β_j that item j is considered to have uniform-DIF at omnibus level. The group level DIF is then flagged based on the specific non-zero β_{jp} . By this definition, it is not surprising to note that, for the reg-DIF method, the power at omnibus level is nothing but the maximum of the two group level powers.

Table 3 summarizes the Type I error (i.e., false positive) of all four different methods under each simulated condition. The values in the parenthesis are standard deviation across 50 replications. First of all, consistent with the findings that LRT tends to be too liberal when DIF is pervasive and sample size is large (Belzak & Bauer, 2020; Finch, 2005; Stark, Chernyshenko, & Drasgow, 2006), LRT produces quite inflated Type I error when DIF proportion is 60%. When DIF proportion is low, the LRT performs well and sometimes better than reg-DIF. This is because the correct anchor items can be determined in this scenario. All three reg-DIF methods produce well-behaved Type I error rate, except lasso EM under the 60% DIF condition. When sample size is large, the omnibus error rate could go up to 0.4, which is unacceptable. There is no appreciable difference between the lasso EMM and adaptive lasso methods.

Table 4 presents the power of all four methods. When DIF is not pervasive and sample size is moderate, all methods produce high power at omnibus level, although detecting DIF at

Table 3: Study I Type I error (standard deviation) of detecting uniform DIF

Corr	N	DIF%	Group	LRT	Lasso EM	Lasso EMM	Adaptive lasso
0.85	1500	20%	Omnibus DIF	0.028 (0.006)	0.043 (0.007)	0.021 (0.005)	0 (0)
			Low DIF	0.013 (0.004)	0.018 (0.004)	0.013 (0.004)	0 (0)
			High DIF	0.031 (0.006)	0.028 (0.005)	0.011 (0.003)	0 (0)
		60%	Omnibus DIF	0.893 (0.022)	0.200 (0.023)	0.035 (0.011)	0.008 (0.008)
			Low DIF	0.308 (0.043)	0.098 (0.016)	0.025 (0.009)	0.005 (0.005)
			High DIF	0.88 (0.022)	0.153 (0.02)	0.013 (0.005)	0.005 (0.005)
	3000	20%	Omnibus DIF	0.033 (0.007)	0.031 (0.007)	0.026 (0.006)	0.006 (0.002)
			Low DIF	0.024 (0.006)	0.018 (0.005)	0.021 (0.005)	0.004 (0.002)
			High DIF	0.034 (0.007)	0.021 (0.005)	0.006 (0.003)	0.003 (0.002)
		60%	Omnibus DIF	0.91 (0.036)	0.345 (0.027)	0.060 (0.015)	0.035 (0.035)
			Low DIF	0.498 (0.035)	0.218 (0.02)	0.058 (0.015)	0.035 (0.035)
			High DIF	0.90 (0.035)	0.268 (0.027)	0.008 (0.004)	0.008 (0.008)
0.25	1500	20%	Omnibus DIF	0.028 (0.005)	0.029 (0.007)	0.016 (0.005)	0.013 (0.004)
			Low DIF	0.013 (0.004)	0.011 (0.003)	0.005 (0.002)	0.005 (0.002)
			High DIF	0.027 (0.004)	0.019 (0.007)	0.011 (0.004)	0.008 (0.003)
		60%	Omnibus DIF	0.86 (0.03)	0.084 (0.014)	0.038 (0.013)	0.020 (0.008)
			Low DIF	0.31 (0.028)	0.076 (0.013)	0.035 (0.013)	0.020 (0.008)
			High DIF	0.86 (0.03)	0.033 (0.010)	0.005 (0.004)	0 (0)
	3000	20%	Omnibus DIF	0.029 (0.007)	0.023 (0.006)	0.005 (0.002)	0.003 (0.002)
			Low DIF	0.02 (0.005)	0.010 (0.004)	0.005 (0.002)	0.003 (0.002)
			High DIF	0.030 (0.007)	0.015 (0.004)	0.003 (0.002)	0 (0)
		60%	Omnibus DIF	0.91 (0.027)	0.128 (0.015)	0.103 (0.024)	0.071 (0.011)
			Low DIF	0.468 (0.048)	0.125 (0.015)	0.095 (0.022)	0.071 (0.011)
			High DIF	0.90 (0.026)	0.024 (0.008)	0.010 (0.005)	0.006 (0.003)

low-DIF group level is understandably hard. When DIF is pervasive, LRT cannot correctly detect DIF with extremely low power in almost all conditions. Overall, increasing sample size leads to higher power, whereas changing the correlation between two factors does not alter the results. The general take-away message is LRT can only be used if DIF is not pervasive unless a purified set of anchor items are known in advance, whereas the lasso EMM and adaptive lasso can be used in all conditions. Our findings about lasso EM are in concert with Bauer et al. (2020)’s conclusion that it “performs best at identifying true DIF when DIF is large in magnitude and the sample size is large and is more likely to identify DIF erroneously when DIF is particularly pervasive.” Our proposed two variations of the lasso

Table 4: Study I Power (standard deviation) of detecting uniform DIF

Corr	N	DIF%	Group	LRT	Lasso EM	Lasso EMM	Adaptive lasso
0.85	1500	20%	Omnibus DIF	0.985 (0.008)	0.965 (0.013)	0.96 (0.017)	0.985 (0.009)
			Low DIF	0.615 (0.045)	0.455 (0.036)	0.55 (0.043)	0.470 (0.05)
			High DIF	0.988 (0.008)	0.965 (0.012)	0.96 (0.017)	0.985 (0.009)
		60%	Omnibus DIF	0.147 (0.022)	0.678 (0.027)	0.885 (0.019)	0.885 (0.024)
			Low DIF	0.027 (0.007)	0.232 (0.019)	0.208 (0.024)	0.193 (0.021)
			High DIF	0.163 (0.022)	0.677 (0.027)	0.885 (0.019)	0.885 (0.024)
	3000	20%	Omnibus DIF	1 (0)	1.000 (0)	1.000 (0)	1.000 (0)
			Low DIF	0.915 (0.022)	0.786 (0.039)	0.84 (0.029)	0.845 (0.032)
			High DIF	1 (0)	1.000 (0)	1.000 (0)	1.000 (0)
		60%	Omnibus DIF	0.115 (0.016)	0.943 (0.010)	0.998 (0.002)	1 (0)
			Low DIF	0.017 (0.006)	0.467 (0.023)	0.632 (0.032)	0.44 (0.038)
			High DIF	0.120 (0.016)	0.942 (0.010)	0.998 (0.002)	1 (0)
0.25	1500	20%	Omnibus DIF	0.97 (0.012)	0.955 (0.013)	0.965 (0.013)	0.985 (0.009)
			Low DIF	0.62 (0.041)	0.430 (0.037)	0.490 (0.040)	0.510 (0.041)
			High DIF	0.975 (0.012)	0.955 (0.013)	0.965 (0.013)	0.985 (0.009)
		60%	Omnibus DIF	0.167 (0.020)	0.728 (0.035)	0.885 (0.022)	0.859 (0.024)
			Low DIF	0.035 (0.009)	0.228 (0.018)	0.197 (0.023)	0.238 (0.017)
			High DIF	0.175 (0.020)	0.728 (0.035)	0.885 (0.022)	0.859 (0.024)
	3000	20%	Omnibus DIF	1 (0)	1 (0)	1 (0)	1 (0)
			Low DIF	0.895 (0.022)	0.806 (0.036)	0.907 (0.022)	0.878 (0.028)
			High DIF	1 (0)	1 (0)	1 (0)	1 (0)
		60%	Omnibus DIF	0.123 (0.018)	0.944 (0.009)	0.998 (0.002)	1 (0)
			Low DIF	0.023 (0.009)	0.356 (0.015)	0.513 (0.033)	0.317 (0.024)
			High DIF	0.132 (0.018)	0.944 (0.009)	0.998 (0.002)	1 (0)

methods, lasso EMM and adaptive lasso, alleviate this issue hence they outperform lasso EM by a large margin when DIF is pervasive.

Table 5 summarizes the mean absolute bias of the DIF parameters. For each method under each condition, the bias was calculated as the estimated DIF parameter minus the true DIF parameter for only the true DIF items that were correctly flagged. The true DIF items that were missed by each method at both group levels were not considered in computing this statistic because we want to evaluate if the DIF magnitude can be precisely recovered without the contamination of Type II error. DIF items that were correctly flagged from at least one group were included. Moreover, the mean absolute bias was reported at only group

Table 5: Simulation I Mean absolute bias (standard deviation) of DIF parameter estimates

Corr	N	DIF%	Group	LRT	Lasso EM	Lasso EMM	Adaptive lasso
0.85	1500	20%	Low DIF	0.114 (0.009)	0.119 (0.012)	0.116 (0.010)	0.121 (0.012)
			High DIF	0.141 (0.010)	0.202 (0.008)	0.180 (0.008)	0.199 (0.009)
		60%	Low DIF	<i>0.212 (0.068)</i>	0.108 (0.006)	0.111 (0.009)	0.113 (0.008)
			High DIF	<i>0.311 (0.024)</i>	0.288 (0.011)	0.225 (0.007)	0.235 (0.010)
	3000	20%	Low DIF	0.102 (0.005)	0.099 (0.006)	0.101 (0.005)	0.098 (0.005)
			High DIF	0.111 (0.007)	0.145 (0.008)	0.135 (0.008)	0.131 (0.009)
		60%	Low DIF	<i>0.126 (0.097)</i>	0.143 (0.008)	0.100 (0.005)	0.108 (0.006)
			High DIF	<i>0.541 (0.116)</i>	0.233 (0.012)	0.157 (0.005)	0.192 (0.006)
0.25	1500	20%	Low DIF	0.108 (0.009)	0.115 (0.011)	0.116 (0.012)	0.116 (0.012)
			High DIF	0.140 (0.009)	0.194 (0.010)	0.192 (0.010)	0.200 (0.011)
		60%	Low DIF	<i>0.272 (0.082)</i>	0.107 (0.006)	0.113 (0.005)	0.117 (0.010)
			High DIF	<i>0.565 (0.107)</i>	0.248 (0.010)	0.220 (0.007)	0.238 (0.010)
	3000	20%	Low DIF	0.098 (0.005)	0.095 (0.005)	0.096 (0.008)	0.096 (0.005)
			High DIF	0.115 (0.007)	0.141 (0.008)	0.129 (0.010)	0.131 (0.007)
		60%	Low DIF	<i>0.077 (0.044)</i>	0.120 (0.005)	0.113 (0.007)	0.108 (0.008)
			High DIF	<i>0.484 (0.087)</i>	0.202 (0.008)	0.157 (0.005)	0.189 (0.008)

The italicized values need to be interpreted with caution because the power in the corresponding cells is too low.

level (no omnibus level bias) because we want to separately show the recovery of DIF size when true DIF size is either small or large. One caveat to note when reading the results is that for LRT, we estimated DIF size at different focal group levels by pooling together data from reference group and respective focal group and fitted a two-group MIRT model. In contrast, we estimated DIF size from the reg-DIF approach by essentially fitting a three-group MIRT model. Then if an item only shows DIF at high-DIF group level but shows DIF-free (type II error) at low-DIF group level, this item will have only one non-zero β . We could have fitted the three-group MIRT model in LRT as well, but that would have given us almost the same results as those from the reg-DIF methods because both approaches use EM algorithm without penalty for final parameter estimation. The results reported in Table 5 are more interesting to illuminate that when DIF occurs at one focal group level, albeit small, if ignored, it would bias the DIF size at the other focal group level. This conclusion is supported by the observation that the mean absolute biases from three regularization

methods are higher than those from LRT for the high DIF group in most conditions.⁵ The results from LRT under 60% condition also needs to be interpreted with caution because the power is extremely low in this condition such that the items entered in this calculation are too few.

3.2 Simulation Study II

The second simulation study focuses on detecting non-uniform DIF. The item parameters were kept to be the same as in simulation study I. For the DIF parameters, we still had the first focal group with smaller magnitude of DIF and the second focal group with larger magnitude of DIF. For the first focal group, we had $\gamma_{j1} = (-0.4, 0)$ and $\beta_j = 0.25$ for $j = 4, 5$ and $\gamma_{j1} = (0, -0.4)$ and $\beta_j = 0.25$ for $j = 12, 13$ for the 20% DIF condition, and $\gamma_{j1} = (-0.4, 0)$ and $\beta_j = 0.25$ for $j = 4, 5, 6, 7, 8, 9$ and $\gamma_{j1} = (0, -0.4)$ and $\beta_j = 0.25$ for $j = 12, 13, 14, 15, 16, 17$ for the 60% DIF condition. Similarly, for the second focal group, we had $\gamma_{j2} = (-0.6, 0)$, $\gamma_{j2} = (0, -0.6)$ and $\beta_j = 0.6$ on the respective DIF items. The wABC of DIF size per item is given in table 6. Compared to table ??, the DIF size in study II is somewhat smaller, hence it will not be surprising if power drops.

Table 6: Non-uniform DIF magnitude measured by wABC

Item	4	5	6	7	8	9
Focal 1	0.02	0.05	0.04	0.05	0.04	0.06
Focal 2	0.04	0.10	0.06	0.11	0.10	0.12
Item	12	13	14	15	16	17
Focal 1	0.05	0.05	0.03	0.05	0.06	0.06
Focal 2	0.11	0.11	0.08	0.11	0.12	0.11

Type I error and power from study II are summarized in Tables 7 and 8. In general, the same pattern preserves as compared to study I. That is, all three reg-DIF methods appear to have good control of Type I error whereas LRT has hugely inflated Type I error when

⁵In this case, for some items included in the calculation, the smaller DIF at ‘low DIF group’ level may be missed, and they further contribute to the bias of DIF estimate at high DIF group level. Whereas at low DIF group level, items that show large DIF at ‘high DIF group’ will always be detected, so they will not contribute to additional bias.

DIF proportion is high. Power is uniformly lower in study II, which may be an artifact of smaller wABC in this case. In addition, Lasso EMM seems to slightly outperform all other methods in all conditions. Table 9 and 10 present the mean absolute bias of DIF parameter estimates, γ and β . As shown, the bias is relatively higher for γ , which is consistent with prior finding that detecting DIF on loading parameter is a lot harder (Bauer et al., 2020).

Table 7: Study II Type I error (standard deviation) of detecting non-uniform DIF

Corr	N	DIF%	Group	LRT	Lasso EM	Lasso EMM	Adaptive lasso
0.85	1500	20%	Omnibus DIF	0.029 (0.007)	0.045 (0.009)	0.036 (0.006)	0.042 (0.008)
			Low DIF	0.013 (0.012)	0.020 (0.006)	0.020 (0.004)	0.023 (0.005)
			High DIF	0.018 (0.010)	0.030 (0.007)	0.020 (0.005)	0.026 (0.007)
		60%	Omnibus DIF	0.407 (0.015)	0.037 (0.009)	0.035 (0.014)	0.02 (0.006)
			Low DIF	0.105 (0.009)	0.023 (0.007)	0.017 (0.007)	0.015 (0.005)
			High DIF	0.4 (0.014)	0.032 (0.008)	0.022 (0.013)	0.007 (0.004)
	3000	20%	Omnibus DIF	0.02 (0.005)	0.032 (0.007)	0.035 (0.007)	0.035 (0.007)
			Low DIF	0.014 (0.007)	0.018 (0.005)	0.017 (0.005)	0.015 (0.004)
			High DIF	0.015 (0.007)	0.032 (0.004)	0.026 (0.005)	0.028 (0.005)
		60%	Omnibus DIF	0.752 (0.035)	0.042 (0.009)	0.015 (0.006)	0.05 (0.013)
			Low DIF	0.207 (0.025)	0.020 (0.006)	0.005 (0.003)	0.017 (0.006)
			High DIF	0.777 (0.031)	0.037 (0.009)	0.012 (0.006)	0.045 (0.011)
0.25	1500	20%	Omnibus DIF	0.018 (0.004)	0.036 (0.007)	0.038 (0.006)	0.032 (0.006)
			Low DIF	0.006 (0.003)	0.018 (0.005)	0.026 (0.006)	0.017 (0.005)
			High DIF	0.015 (0.004)	0.025 (0.005)	0.021 (0.004)	0.017 (0.004)
		60%	Omnibus DIF	0.415 (0.036)	0.077 (0.016)	0.030 (0.009)	0.062 (0.014)
			Low DIF	0.11 (0.019)	0.030 (0.008)	0.020 (0.005)	0.03 (0.009)
			High DIF	0.415 (0.037)	0.070 (0.014)	0.013 (0.008)	0.047 (0.011)
	3000	20%	Omnibus DIF	0.026 (0.005)	0.026 (0.006)	0.046 (0.010)	0.048 (0.007)
			Low DIF	0.016 (0.004)	0.015 (0.004)	0.026 (0.008)	0.028 (0.006)
			High DIF	0.022 (0.005)	0.012 (0.004)	0.032 (0.006)	0.03 (0.005)
		60%	Omnibus DIF	0.792 (0.031)	0.105 (0.015)	0.026 (0.008)	0.08 (0.015)
			Low DIF	0.262 (0.030)	0.06 (0.012)	0.018 (0.006)	0.035 (0.011)
			High DIF	0.807 (0.031)	0.08 (0.015)	0.021 (0.006)	0.06 (0.012)

4 Real Data Analysis

A real data set from patient reported outcome measures (PROMIS) was used to illustrate the performance of the three reg-DIF methods in comparison to LRT. The sample contains

Table 8: Study II Power (standard deviation) of detecting non-uniform DIF

Corr	N	DIF%	Group	LRT	Lasso EM	Lasso EMM	Adaptive lasso
0.85	1500	20%	Omnibus DIF	0.665 (0.043)	0.645 (0.035)	0.730 (0.029)	0.69 (0.037)
			Low DIF	0.325 (0.037)	0.120 (0.024)	0.175 (0.026)	0.135 (0.028)
			High DIF	0.66 (0.043)	0.645 (0.035)	0.730 (0.029)	0.69 (0.037)
		60%	Omnibus DIF	0.218 (0.072)	0.325 (0.029)	0.396 (0.039)	0.371 (0.039)
			Low DIF	0.007 (0.043)	0.060 (0.012)	0.065 (0.013)	0.007 (0.003)
			High DIF	0.216 (0.063)	0.315 (0.028)	0.395 (0.039)	0.368 (0.039)
	3000	20%	Omnibus DIF	0.905 (0.015)	0.890 (0.019)	0.915 (0.017)	0.91 (0.020)
			Low DIF	0.535 (0.033)	0.235 (0.032)	0.335 (0.039)	0.24 (0.036)
			High DIF	0.905 (0.015)	0.890 (0.019)	0.915 (0.017)	0.91 (0.020)
		60%	Omnibus DIF	0.165 (0.019)	0.688 (0.027)	0.845 (0.019)	0.83 (0.025)
			Low DIF	0.004 (0.009)	0.158 (0.015)	0.176 (0.016)	0.133 (0.016)
			High DIF	0.181 (0.021)	0.685 (0.027)	0.845 (0.019)	0.825 (0.025)
0.25	1500	20%	Omnibus DIF	0.645 (0.041)	0.695 (0.036)	0.74 (0.031)	0.69 (0.039)
			Low DIF	0.285 (0.036)	0.130 (0.027)	0.235 (0.030)	0.13 (0.025)
			High DIF	0.645 (0.041)	0.695 (0.036)	0.74 (0.031)	0.69 (0.039)
		60%	Omnibus DIF	0.181 (0.017)	0.365 (0.025)	0.423 (0.036)	0.466 (0.035)
			Low DIF	0.036 (0.008)	0.071 (0.011)	0.073 (0.013)	0.086 (0.014)
			High DIF	0.18 (0.017)	0.356 (0.025)	0.42 (0.028)	0.451 (0.035)
	3000	20%	Omnibus DIF	0.89 (0.017)	0.875 (0.018)	0.895 (0.019)	0.925 (0.016)
			Low DIF	0.52 (0.037)	0.265 (0.031)	0.355 (0.042)	0.26 (0.038)
			High DIF	0.88 (0.018)	0.875 (0.018)	0.895 (0.019)	0.92 (0.016)
		60%	Omnibus DIF	0.16 (0.019)	0.701 (0.025)	0.785 (0.024)	0.771 (0.027)
			Low DIF	0.028 (0.007)	0.161 (0.017)	0.131 (0.015)	0.165 (0.021)
			High DIF	0.188 (0.022)	0.693 (0.026)	0.785 (0.024)	0.77 (0.027)

Table 9: Simulation II Mean absolute bias (standard deviation) of DIF parameter γ estimates

Corr	N	DIF%	Group	LRT	LASSO EM	LASSO EMM	Adaptive LASSO
0.85	1500	20%	Low DIF	0.369 (0.035)	0.371 (0.077)	0.559 (0.058)	0.501 (0.078)
			High DIF	0.324 (0.025)	0.212 (0.056)	0.343 (0.047)	0.255 (0.050)
		60%	Low DIF	<i>0.443 (0.068)</i>	0.610 (0.234)	0.424 (0.114)	0.287 (-)
			High DIF	<i>0.371 (0.035)</i>	0.284 (0.115)	0.225 (0.024)	0.165 (0.024)
	3000	20%	Low DIF	0.220 (0.018)	0.317 (0.059)	0.323 (0.037)	0.281 (0.035)
			High DIF	0.185 (0.011)	0.133 (0.016)	0.142 (0.020)	0.154 (0.018)
		60%	Low DIF	<i>0.263 (0.048)</i>	0.370 (0.084)	0.232 (0.026)	0.283 (0.072)
			High DIF	<i>0.392 (0.041)</i>	0.166 (0.032)	0.121 (0.018)	0.151 (0.017)
0.25	1500	20%	Low DIF	0.186 (0.028)	0.787 (0.120)	0.727 (0.081)	0.837 (0.114)
			High DIF	0.148 (0.021)	0.294 (0.105)	0.398 (0.058)	0.381 (0.075)
		60%	Low DIF	<i>0.121 (0.069)</i>	0.680 (0.153)	0.631 (0.101)	0.558 (0.074)
			High DIF	<i>0.166 (0.035)</i>	0.332 (0.105)	0.267 (0.042)	0.306 (0.068)
	3000	20%	Low DIF	0.107 (0.017)	0.241 (0.062)	0.349 (0.039)	0.403 (0.067)
			High DIF	0.117 (0.011)	0.110 (0.018)	0.191 (0.031)	0.190 (0.021)
		60%	Low DIF	<i>0.201 (0.043)</i>	0.432 (0.094)	-	0.306 (0.058)
			High DIF	<i>0.124 (0.026)</i>	0.142 (0.019)	-	0.154 (0.012)

The italicized values need to be interpreted with caution because the power in the corresponding cells is too low. The two empty cells imply no item was flagged to have DIF on slope parameter.

5,219 cancer patients' responses to the two PROMIS scales, depression and anxiety scales. We focused on detecting age DIF because in the sample, age is a categorical variable with three levels and prior researches have studied age DIF on these items using the same sample (Teresi, Ocepek-Welikson, Kleinman, Ramirez, & Kim, 2016a, 2016b).

Among the three age groups, the reference group is 'Age 21-49' (sample size $n = 1,143$), and two focal groups are 'Age 50-64' ($n = 1,935$) and 'Age 65-84' ($n = 2,141$) respectively. The original data set contains 21 polytomous items with five response categories (1 = never, 2 = Rarely, 3 = Sometimes, 4 = Often, 5 = Always). As we focused on M2PL model throughout the paper, we artificially combined response categories to create a dichotomous data set. Given the proportion of the 'never' response falls between 50%-65% in most items, we combined the other four response categories and made all 21 items dichotomous. That is, the patient response to each item is either yes or no. This treatment is similar to Bauer et al. (2020). The first 10 items measure depression and the other 11 items measure anxiety.

Table 10: Simulation II Mean absolute bias (standard deviation) of DIF parameter β estimates

Corr	N	DIF%	Group	LRT	LASSO EM	LASSO EMM	Adaptive LASSO
0.85	1500	20%	Low DIF	0.208 (0.021)	0.287 (0.016)	0.288 (0.020)	0.346 (0.024)
			High DIF	0.172 (0.015)	0.129 (0.008)	0.137 (0.010)	0.131 (0.009)
		60%	Low DIF	<i>0.269 (0.073)</i>	0.309 (0.047)	0.338 (0.048)	0.699 (0.142)
			High DIF	<i>0.188 (0.019)</i>	0.141 (0.010)	0.148 (0.009)	0.136 (0.007)
	3000	20%	Low DIF	0.113 (0.009)	0.146 (0.013)	0.159 (0.026)	0.176 (0.021)
			High DIF	0.113 (0.006)	0.122 (0.007)	0.132 (0.018)	0.125 (0.006)
		60%	Low DIF	<i>0.136 (0.033)</i>	0.142 (0.016)	0.136 (0.014)	0.160 (0.019)
			High DIF	<i>0.228 (0.016)</i>	0.150 (0.005)	0.144 (0.004)	0.148 (0.005)
0.25	1500	20%	Low DIF	0.213 (0.025)	0.269 (0.062)	0.285 (0.045)	0.330 (0.063)
			High DIF	0.143 (0.010)	0.126 (0.026)	0.144 (0.016)	0.126 (0.026)
		60%	Low DIF	<i>0.273 (0.050)</i>	0.255 (0.076)	0.355 (0.092)	0.501 (0.110)
			High DIF	<i>0.189 (0.022)</i>	0.161 (0.015)	0.151 (0.009)	0.159 (0.014)
	3000	20%	Low DIF	0.108 (0.009)	0.145 (0.011)	0.150 (0.013)	0.161 (0.016)
			High DIF	0.109 (0.005)	0.122 (0.005)	0.126 (0.005)	0.128 (0.005)
		60%	Low DIF	<i>0.119 (0.026)</i>	0.166 (0.033)	0.114 (0.009)	0.206 (0.038)
			High DIF	<i>0.144 (0.017)</i>	0.159 (0.005)	0.145 (0.005)	0.157 (0.005)

The italicized values need to be interpreted with caution because the power in the corresponding cells is too low.

Table 11 presents the item content.

To start the analysis, we first used three reg-DIF methods without any anchor items to perform DIF detection (1) on discrimination parameter only (2) on intercept only, and (3) on discrimination and intercept parameters simultaneously. From all three analyses, items 3, 14, and 15 did not show any DIF across all three regularization methods. The results, on one hand, are consistent with the conclusions in Teresi et al. (2016b) and Teresi et al. (2016a); and on other hand, imply that these three items can very well serve as anchor items for LRT. To ensure fair comparison across methods, we also re-conducted the three reg-DIF methods again using the same anchor items as in LRT. The results stay the same with and without anchor items.

Figure 1 presents the flagged uniform DIF items and their corresponding DIF magnitude on intercepts from all four methods. Items 3 and 15 served as anchor items in this case. As shown, the LRT flagged eight items, whereas lasso EMM flagged 6 items, adaptive lasso

Table 11: PROMIS depression and anxiety imputed data set: Item description

1	I felt worthless
2	I felt that I had nothing to look forward to
3	I felt helpless
4	I felt sad
5	I felt like a failure
6	I felt depressed
7	I felt unhappy
8	I felt hopeless
9	I felt discouraged about the future
10	I felt disappointed in myself
11	I felt fearful
12	I felt anxious
13	I felt worried
14	I found it hard to focus on anything other than my anxiety
15	I felt nervous
16	I felt uneasy
17	I felt tense
18	My worries overwhelmed me
19	I felt like I needed help for my anxiety
20	Many situations made me worry
21	I had difficulty calming down

flagged 5 items, and lasso EM flagged 4 items. Because the items detected by the reg-DIF methods are a subset of the items detected by LRT, the result is somewhat consistent with the findings from simulation study I. That is, LRT may yield inflated Type I error when DIF prevalence is high. Indeed, the two unique items flagged by LRT have relatively smaller DIF size, which may be false detection by LRT. Note there are two small DIF sizes around .07 from LRT, which may be false detection.

Figure 2 shows the flagged non-uniform DIF items and their corresponding DIF magnitude on discrimination parameters from all four methods. Items 3 and 14 served as anchor items. Quite surprisingly, LRT did not flag any items, whereas the items detected by the three regularization methods were consistent, and their estimated DIF sizes were close. One reason could be that in this analysis, we assumed that DIF only occurred on discrimination parameters. Therefore, the regularization methods may attribute the differences in item characteristic curves across groups to discrimination differences when the intercepts were

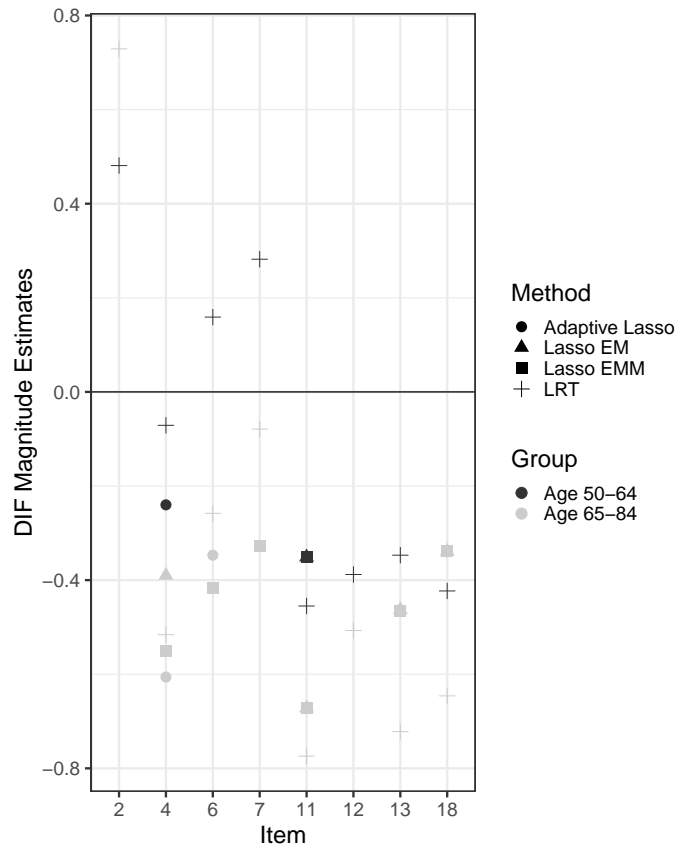


Figure 1: Items detected to exhibit uniform DIF

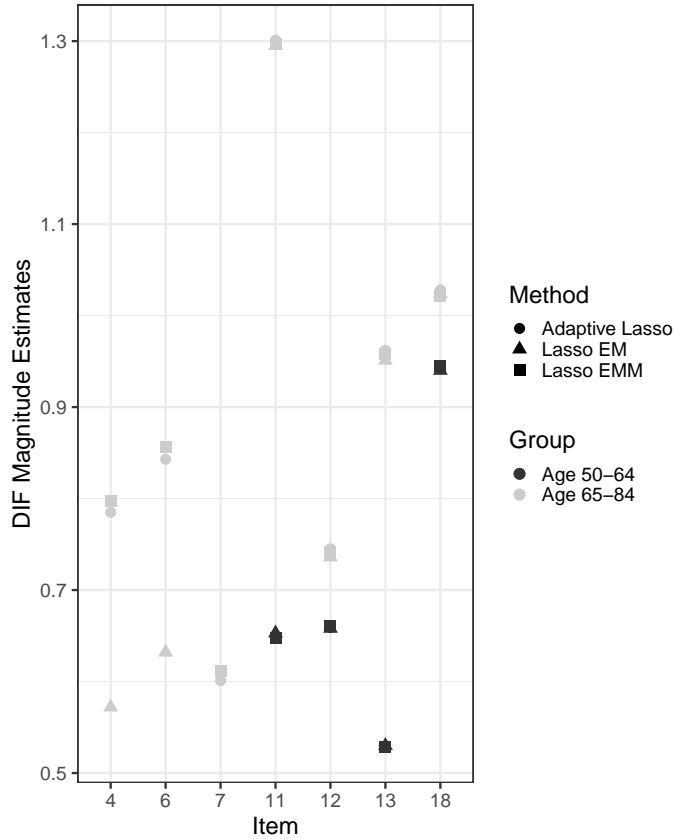


Figure 2: Items detected to exhibit non-uniform DIF

constrained equal across groups. In fact, Figure 3 shows the flagged DIF items when we assumed DIF could happen on both intercept and discrimination parameters. Now almost all detected DIF from regularization methods appear on the intercept parameters. By the nature of LRT, it tests both parameters of a studied item together for DIF, but as reflected, some of the DIF sizes on discrimination parameters are close to 0. Note that we cannot compare our DIF detection results directly to those in Teresi et al. (2016a, 2016b) because they did not take impact into consideration. However, we found that impact on means is relatively high that it cannot be ignored. Table 12 presents the estimated mean and covariance matrix of the two focal groups from the analysis where both uniform and non-uniform DIF were considered together, as this is the most flexible analysis. Results from the other two analyses are rather similar. For the reference group, the means and variances of θ were

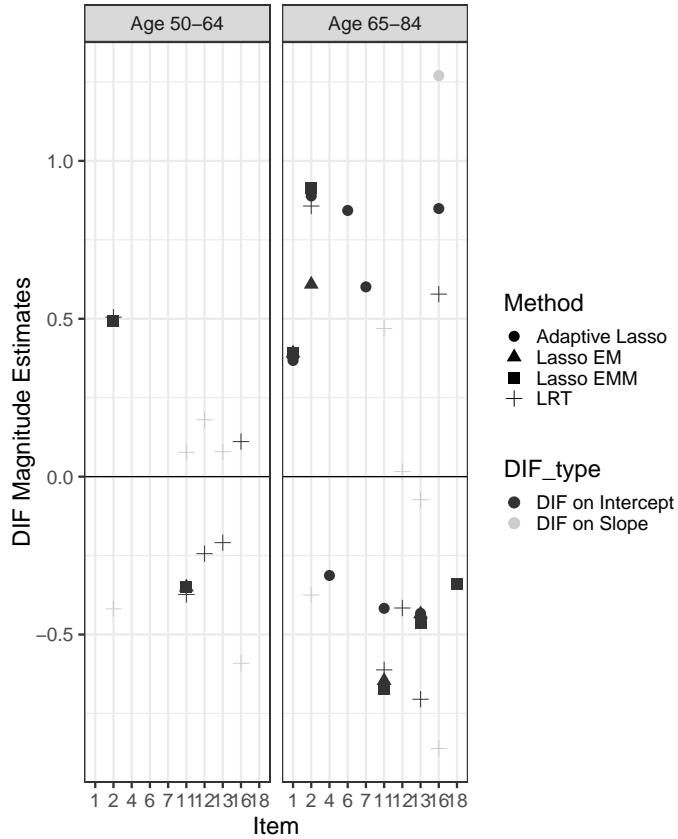


Figure 3: Items detected to exhibit omnibus DIF

fixed at 0 and 1 respectively. The correlation between the two factors was freely estimated.

5 Discussion

The idea of using regularization methods for DIF detection started to emerge in psychometrics literature a decade ago, although the statistical methods per se are around for much longer. The regularization approach is fundamentally different from traditional DIF detection approaches which often involves hypothesis testing of some kind. The main purpose of this study is to introduce the lasso regularization method within MIRT context and evaluate its performance in detecting both uniform DIF and non-uniform DIF. Up to date, there are only a handful of studies that explored DIF with MIRT models (Bolt & Johnson, 2009; Bulut

Table 12: Estimated mean and covariance matrix (Impact) from PROMIS analysis

	LRT			Lasso EM		
	Age 21-49	Age 50-64	Age 65-84	Age 21-49	Age 50-64	Age 65-84
μ_1	0	-.157	-.42	0	-.319	-.560
μ_2	0	-.116	-.40	0	-.317	-.571
σ_1^2	1	1.367	1.298	1	1.084	.979
σ_{12}	.907	1.330	1.198	.915	1.036	.925
σ_2^2	1	1.52	1.304	1	1.183	1.044
	Lasso EMM			Adaptive Lasso		
	Age 21-49	Age 50-64	Age 65-84	Age 21-49	Age 50-64	Age 65-84
μ_1	0	-.327	-.564	0	-.324	-.550
μ_2	0	-.316	-.570	0	-.321	-.576
σ_1^2	1	1.060	.966	1	1.062	.951
σ_{12}	.915	1.026	.919	.915	1.028	.908
σ_2^2	1	1.185	1.046	1	1.186	1.036

& Suh, 2009; Fukuhara & Kamata, 2011; Lee et al., 2017; Mazor, Hambleton, & Clauser, 1998; Suh & Cho, 2014), and this study will add to the growing literature on this topic and meanwhile, expand the applications of the reg-DIF methods.

Aside from MIRT applications, several unique features of our study are worth highlighting. First, Bauer et al.(2020) and Belzak & Bauer (2020) relied on general-purpose optimization routines in SAS NLMIXED. Because l_1 regularization criterion is non-differentiable, their approach may not efficiently find a (local) maximizer. Instead, we directly programmed a soft-thresholding operator within the coordinate descent algorithm that is more authentic for l_1 optimization. Specifically, the quadratic approximation to the marginal likelihood enables the direct uses of the soft-thresholding operator (Sun et al., 2016). In the event when the number of dimensions is high, the marginal likelihood could be replaced by its variational lower bound (Cho et al., 2021) to further speed up the computation. Second, when DIF proportion is high, the inflated Type I error of reg-DIF is likely due to the bias from using the l_1 penalty. Therefore, we propose to use two variants of the lasso method, namely, the EMM algorithm and adaptive lasso. Both tend to perform much better than the original

lasso method. In the future, alternative penalties could be considered, such as the smoothly clipped absolute deviation (SCAD) penalty (Fan & Li, 2001), or Minimax concave penalty (MCP) (Zhang, 2010). These penalties generally serve to mitigate the l_1 bias by lessening the strength of the penalty on estimates that are large in absolute value (Hastie, Tibshirani, & Tibshirani, 2017). Hence, they should better distinguish large DIF effects from small DIF effects. Third, Tutz & Schauburger (2015) and Magis (2015) only considered one focal group as most of the DIF studies did, we intentionally constructed two focal groups just to exemplify the advantage of reg-DIF in handling multiple sources of DIF. Indeed, reg-DIF will be extremely efficient to hone in on DIF items and specific covariates that cause DIF simultaneously as such information is encapsulated in either β or γ . In contrast, it is procedurally cumbersome to perform a likelihood ratio test separately for one item and one covariate at a time. For instance, if there are 10 studied items and one covariate with two levels, then at least 10 different models need to be fitted separately. In educational assessment with a large item pool, this procedure can be prohibitively time consuming and error prone, especially if the model is high dimensional.

Our simulation results reveal that reg-DIF performs better when DIF proportion is low. This conclusion, on one hand, coincides with Bauer et al (2019) that reg-DIF performance starts to deteriorate when DIF proportion is 40%. On the other hand, it is not too surprising because regularization methods will in general perform better when data is truly sparse. When the DIF is too pervasive, there will be too many non-zero elements in β and γ , yielding a non-sparse scenario. Even so, reg-DIF still greatly outperforms LRT because when DIF proportion is high, it is almost impossible to identify a purified set of DIF-free anchors, and LRT performance is greatly compromised as a result. A follow-up study is underway where DIF is caused by multiple categorical variables simultaneously and DIF proportion is low. This is exactly the scenario where reg-DIF approach will shine, especially in multicollinear situations. Needless to say, simultaneous treatment of several covariates is necessary to obtain accurate θ estimates that correct for possible DIF effects.

We present the detailed identification conditions where MIRT DIF analysis can proceed. In essence, for simple structure MIRT, we need at minimum one anchor item per dimension to ensure model identifiability. However, “identifiable” should be separated from “estimable.” That is, for reg-DIF, in the presence of sufficient penalty, no anchor items are required as some DIF parameters will be shrunk to 0, “making the model estimable even without prior selection of anchor items” (Bauer et al., 2020). Similarly, Schauberg and Mair (2019) also argued that “as long as the respective penalty term corresponds to a restriction that is strong enough, this identifiability issue can be ignored.” On the other hand, the selection of the designated anchor items is critical to the success of LRT, which can be conceived as a limitation of this method as finding the right anchors may not be easy when DIF proportion is high (Bolt, Hare, Vitale, & Newman, 2004; Edelen, Thissen, Teresi, et al., 2006). For practitioners, we recommend using reg-DIF when no prior knowledge of DIF-free items is available. The advantage of reg-DIF methods is especially salient when DIF proportion is high. In the reg-DIF method, not only are DIF items detected, but also item parameters and non-zero DIF parameters are estimated simultaneously. Then they can be used in Equation 1 to estimate person parameters that correct for DIF effects.

Given the limitation of the simulation study design, the current study can be extended in a few directions. Aside from the aforementioned ongoing study that considers multiple categorical covariates, the method can also be generalized to include continuous covariates. In that case, instead of estimating $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ per subgroup as shown in Equation 4, 8-10, one can update Equation 1 as follows

$$\log\left(\frac{P_j(\boldsymbol{\theta}_i)}{1 - P_j(\boldsymbol{\theta}_i)}\right) = \boldsymbol{\alpha}_j^T \boldsymbol{\theta}_i + d_j + (\mathbf{X}_i \boldsymbol{\gamma}_j) \boldsymbol{\theta}_i + \mathbf{X}_i \boldsymbol{\beta}_j + \mathbf{X}_i \boldsymbol{\alpha}, \quad (i = 1, \dots, N; j = 1, 2, \dots, J).$$

where $\boldsymbol{\alpha}$ is a P -by-1 regression coefficients implying the impact of covariates on $\boldsymbol{\theta}$ distribution. In addition, we only considered between-item two-dimensional 2PL model. Future study could extend to within-item multidimensional design and polytomous items. As reg-

DIF requires multiple rounds of estimation to find the best tuning parameter based on BIC, generalizing reg-DIF to more than two dimensions of course calls for a more efficient estimation algorithm than the currently used quadrature-based EM algorithm. Alternatives, such as Metropolis-Hastings Robbins-Monro algorithm (Cai, 2010) or variational methods (Cho et al., 2021) can be plausible candidates for future studies.

References

- Bauer, D. J., Belzak, W. C. M., & Cole, V. T. (2020). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 43-55.
- Belzak, W. C. M., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods*, *25*(6), 673-690.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289-300.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, *46*(4), 443-459.
- Bolt, D. M., Hare, R. D., Vitale, J. E., & Newman, J. P. (2004). A multigroup item response theory analysis of the psychopathy checklist-revised. *Psychological assessment*, *16*(2), 155-168.
- Bolt, D. M., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied psychological measurement*, *33*(5), 335-352.
- Bulut, O., & Suh, Y. (2009). Detecting multidimensional differential item functioning with

- the multiple indicators multiple causes model, the item response theory likelihood ratio test, and logistic regression. *Front. Educ*, 33(5), 335-352.
- Cai, L. (2010). High-dimensional exploratory item factor analysis by a metropolis-hastings robbins-monro algorithm. *Psychometrika*, 75, 33-57.
- Chalmers, P., Pritikin, J., Robitzsch, A., Zoltak, M., Kim, K., Falk, C., ... Oguzhan, O. (2020). Multidimensional item response theory [Computer software manual]. Retrieved from <https://github.com/philchalmers/mirt>
- 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.
- Chen, P., Wang, C., Xin, T., & Chang, H. (2017). Developing new online calibration methods for multidimensional computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 70, 81-117.
- Cho, A. E., Wang, C., Zhang, X., & Xu, G. (2021). Gaussian variational estimation for multidimensional item response theory. *British Journal of Mathematical and Statistical Psychology*, 74, 52-85.
- Choi, S. W., 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(8), 1-30.
- Donoho, D. L., & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432), 1200-1224.
- Edelen, M. O., Stucky, B., & Chandra, A. (2015). Quantifying 'problematic' dif within an irt framework: application to a cancer stigma index. *Qual Life Res*, 24(1), 95-103.
- Edelen, M. O., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: application to the mini-mental state examination. *Medical Care*, 44(11), 134-142.

- Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, *96*(456), 1348-1360.
- Finch, H. (2005). The mimic model as a method for detecting dif: Comparison with mantel-haenszel, sibtest, and the irt likelihood ratio. *Applied psychological measurement*, *29*(4), 278-295.
- Fukuhara, H., & Kamata, A. (2011). A bifactor multidimensional item response theory model for differential item functioning analysis on testlet-based items. *Applied psychological measurement*, *35*(8), 604-622.
- Hastie, T., Tibshirani, R., & Tibshirani, R. J. (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. <https://arxiv.org/abs/1707.08692>.
- Hoerl, A., & Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67.
- Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in Psychology*, *7*, 109.
- Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American statistical Association*, *70*(351), 631-639.
- Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for dif analysis: review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*, 22-56.
- Lee, S., Bulut, O., & Suh, Y. (2017). Multidimensional extension of multiple indicators multiple causes models to detect dif. *Educational and Psychological Measurement*, *77*(4), 545-569.
- Magis, D., Tuerlinckx, F., & De Boeck, P. (2015). Detection of differential item functioning using the lasso approach. *Journal of Educational and Behavioral Statistics*, *40*(2), 111-135.

- Mazor, K. M., Hambleton, R. K., & Clauser, B. E. (1998). Multidimensional dif analyses: The effects of matching on unidimensional subtest scores. *Applied psychological measurement, 22*(4), 357-367.
- Newman, M. E., & Barkema, G. T. (1999). *Monte carlo methods in statistical physics*. Oxford, UK: Clarendon Press.
- Oshima, T., Raju, N. S., & Flowers, C. P. (1997). Development and demonstration of multidimensional irt-based internal measures of differential functioning of Items and tests. *Journal of Educational Measurement, 34*(3), 253-272.
- Penfield, R. D., & Camilli, G. (2006). 5 differential item functioning and item bias. *Handbook of Statistics, 26*, 125-167.
- Raykov, T., Marcoulides, G. A., Lee, C.-L., & Chang, C. (2013). Studying differential item functioning via latent variable modeling: A note on a multiple-testing procedure. *Educational and Psychological Measurement, 73*(5), 898-908.
- Schauberger, G., & Mair, P. (2019). A regularization approach for the detection of differential item functioning in generalized partial credit models. *Behavior research methods*, 1-16.
- Shealy, R., & Stout, W. F. (1993a). Differential item performance and the mantel-haenszel procedure. *Test validity*, 197-239.
- Shealy, R., & Stout, W. F. (1993b). A model-based standardization approach that separates true bias/dif from group ability differences and detects test bias/dtt as well as item bias/dif. *Psychometrika, 58*, 159-194.
- Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292-1306.
- Stout, W., Li, H.-H., Nandakumar, R., & Bolt, D. (1997). Multisib: A procedure to investigate dif when a test is intentionally two-dimensional. *Applied psychological measurement, 21*(3), 195-213.
- Suh, Y., & Cho, S.-J. (2014). Chi-square difference tests for detecting differential func-

- tioning in a multidimensional irt model: A monte carlo study. *Applied psychological measurement*, *38*(5), 359-375.
- Sun, J., Chen, Y., Liu, J., Ying, Z., & Xin, T. (2016). Latent variable selection for multidimensional item response theory models via l1 regularization. *Psychometrika*, *81*(4), 921-939.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361-370.
- Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Ramirez, M., & Kim, G. (2016a). Measurement equivalence of the patient reported outcomes measurement information system (promis) anxiety short forms in ethnically diverse groups. *Psychological test and assessment modeling*, *58*(1), 183-219.
- Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Ramirez, M., & Kim, G. (2016b). Psychometric properties and performance of the patient reported outcomes measurement information system (promis) depression short forms in ethnically diverse groups. *Psychological test and assessment modeling*, *58*(1), 141-181.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society (Series B)*, *58*(1), 267-288.
- Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in rasch models. *Psychometrika*, *80*(1), 21-43.
- Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional irt as a diagnostic aid. *Journal of Educational Measurement*, *40*(3), 255-275.
- Wang, C., Chen, P., & Jiang, S. (2020). Item calibration methods with multiple subscale multistage testing. *Journal of Educational Measurement*, *57*, 3-28.
- Wilson, M., De Boeck, P., & Carstensen, C. (2008). Explanatory item response models: a brief introduction. In P. De Boeck and M. Wilson *Assessment of Competencies in Educational Contexts*, Kirkland, WA: Hogrefe and Huber Publishers.

- Woods, C. M. (2009). Evaluation of mimic-model methods for dif testing with comparison to two-group analysis. *Multivariate behavioral research*, *44*(1), 1-27.
- Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied psychological measurement*, *35*(5), 339-361.
- Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49-67.
- Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, *38*, 894-942.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical Association*, *101*(476), 1418-1429.
- Zumbo, B. (1999). *A handbook on the theory and methods of differential item functioning (dif): Logistic regression modelling as a unitary framework for binary and likert-type (ordinal) item scores*. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

6 Appendix

Algorithm 1: Uniform DIF Detection via Adaptive Lasso EM

Input : $\mathbf{a}_0, \mathbf{d}_0, \boldsymbol{\beta}_0, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \mathbf{u}, \eta_l, \mathbf{w}, \varepsilon_1, \varepsilon_2, \mathbf{X}$

Output: $\hat{\mathbf{a}}^*, \hat{\mathbf{d}}^*, \hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\mu}}^*, \hat{\boldsymbol{\Sigma}}^*$

set $t = 1$, $\mathbf{a}^{(0)} = \mathbf{a}_0$, $\mathbf{d}^{(0)} = \mathbf{d}_0$, $\boldsymbol{\beta}^{(0)} = \boldsymbol{\beta}_0$, $\boldsymbol{\mu}^{(0)} = \boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}^{(0)} = \boldsymbol{\Sigma}_0$, $\delta_1^{(t-1)} = 1$;

while $\delta_1^{(t-1)} > \varepsilon_1$ **do**

 Calculate $n_{gm} = \sum_{i=1}^{N_g} h(\mathbf{q}_m | \mathbf{X}_i, \mathbf{u}_i, \boldsymbol{\Delta}^{(t-1)})$ and

$r_{gjm} = \sum_{i=1}^{N_g} u_{ij} h(\mathbf{q}_m | \mathbf{X}_i, \mathbf{u}_i, \boldsymbol{\Delta}^{(t-1)})$;

 Update $\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\Sigma}^{(t)}$;

for $j=1, \dots, m$ **do**

 set $k = 1$, $\delta_2^{(k-1)} = 1$;

while $\delta_2^{(k-1)} > \varepsilon_2$ **do**

$a_{jr}^{(k)} = a_{jr}^{(k-1)} - \frac{\partial_{a_{jr}} Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\beta})}{\partial_{a_{jr}}^2 Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\beta})}$;

$d_j^{(k)} = d_j^{(k-1)} - \frac{\partial_{d_j} Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\beta})}{\partial_{d_j}^2 Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\beta})}$;

$\beta_{jp}^{(k)} = \text{soft}(\beta_{jp}^{(k-1)} - \frac{\partial_{\beta_{jp}} Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\beta})}{\partial_{\beta_{jp}}^2 Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\beta})}, -\frac{\eta_l w_j}{\partial_{\beta_{jp}}^2 Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\beta})})$;

$\delta_2^{(k)} = \|\mathbf{a}_j^{(k)} - \mathbf{a}_j^{(k-1)}\| + \|\mathbf{d}_j^{(k)} - \mathbf{d}_j^{(k-1)}\| + \|\boldsymbol{\beta}_j^{(k)} - \boldsymbol{\beta}_j^{(k-1)}\|$;

$k = k + 1$;

$\delta_1^{(t)} = \|\mathbf{a}^{(t)} - \mathbf{a}^{(t-1)}\| + \|\mathbf{d}^{(t)} - \mathbf{d}^{(t-1)}\| + \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}\|$;

$t = t + 1$;

Set $\eta_a = 0$ and re-estimate all none-zero estimates.

Algorithm 2: Uniform DIF Detection via Lasso EMM

Input : $\mathbf{a}_0, \mathbf{d}_0, \boldsymbol{\beta}_0, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \mathbf{u}, \eta_l, \varepsilon_1, \varepsilon_2, \mathbf{X}$ **Output:** $\hat{\mathbf{a}}^*, \hat{\mathbf{d}}^*, \hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\mu}}^*, \hat{\boldsymbol{\Sigma}}^*$ set $t = 1$, $\mathbf{a}^{(0)} = \mathbf{a}_0$, $\mathbf{d}^{(0)} = \mathbf{d}_0$, $\boldsymbol{\beta}^{(0)} = \boldsymbol{\beta}_0$, $\boldsymbol{\mu}^{(0)} = \boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}^{(0)} = \boldsymbol{\Sigma}_0$, $\delta_1^{(t-1)} = 1$;**while** $\delta_1^{(t-1)} > \varepsilon_1$ **do** Calculate n_{gm} and r_{gjm} ; Update $\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\Sigma}^{(t)}$; **for** $j=1, \dots, m$ **do** set $k = 1$, $\delta_2^{(k-1)} = 1$; **while** $\delta_2^{(k-1)} > \varepsilon_2$ **do**

$$a_{jr}^{(k)} = a_{jr}^{(k-1)} - \frac{\partial a_{jr} Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\beta})}{\partial a_{jr}^2 Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\beta})}, d_j^{(k)} = d_j^{(k-1)} - \frac{\partial d_j Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\beta})}{\partial d_j^2 Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\beta})};$$

$$\beta_{jp}^{(k)} = \text{soft}\left(\beta_{jp}^{(k-1)} - \frac{\partial \beta_{jp} Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\beta})}{\partial \beta_{jp}^2 Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\beta})}, -\frac{\eta_l}{\partial \beta_{jp}^2 Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\beta})}\right);$$

$$\delta_2^{(k)} = \|\mathbf{a}_j^{(k)} - \mathbf{a}_j^{(k-1)}\| + \|\mathbf{d}_j^{(k)} - \mathbf{d}_j^{(k-1)}\| + \|\boldsymbol{\beta}_j^{(k)} - \boldsymbol{\beta}_j^{(k-1)}\|;$$

 $k = k + 1$; **for** $j=1, \dots, m$ **do** set $k = 1$, $\delta^{(k-1)} = \text{any value greater than } \varepsilon_2$; **while** $\delta_2^{(k-1)} > \varepsilon_2$ **do**

$$a_{jr}^{(k)} = a_{jr}^{(k-1)} - \frac{\partial a_{jr} Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\beta})}{\partial a_{jr}^2 Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\beta})}, d_j^{(k)} = d_j^{(k-1)} - \frac{\partial d_j Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\beta})}{\partial d_j^2 Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\beta})},$$

$$\beta_{jp}^{(k)} = \beta_{jp}^{(k-1)} - \frac{\partial \beta_{jp} Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\beta})}{\partial \beta_{jp}^2 Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\beta})};$$

$$\delta_2^{(k)} = \|\mathbf{a}_j^{(k)} - \mathbf{a}_j^{(k-1)}\| + \|\mathbf{d}_j^{(k)} - \mathbf{d}_j^{(k-1)}\| + \|\boldsymbol{\beta}_j^{(k)} - \boldsymbol{\beta}_j^{(k-1)}\|;$$

 $k = k + 1$;

$$\delta_1^{(t)} = \|\mathbf{a}^{(t)} - \mathbf{a}^{(t-1)}\| + \|\mathbf{d}^{(t)} - \mathbf{d}^{(t-1)}\| + \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}\|;$$

 $t = t + 1$;

Algorithm 3: Non-Uniform DIF Detection via Adaptive Lasso EM

Input : $\mathbf{a}_0, \mathbf{d}_0, \boldsymbol{\gamma}_0, \boldsymbol{\beta}_0, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \mathbf{u}, \eta_l, \mathbf{w}, \varepsilon_1, \varepsilon_2, \mathbf{X}$ **Output:** $\hat{\mathbf{a}}^*, \hat{\mathbf{d}}^*, \hat{\boldsymbol{\gamma}}^*, \hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\mu}}^*, \hat{\boldsymbol{\Sigma}}^*$ set $t = 1$, $\mathbf{a}^{(0)} = \mathbf{a}_0$, $\mathbf{d}^{(0)} = \mathbf{d}_0$, $\boldsymbol{\gamma}^{(0)} = \boldsymbol{\gamma}_0$, $\boldsymbol{\beta}^{(0)} = \boldsymbol{\beta}_0$, $\boldsymbol{\mu}^{(0)} = \boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}^{(0)} = \boldsymbol{\Sigma}_0$, $\delta_1^{(t-1)} = 1$;**while** $\delta_1^{(t-1)} > \varepsilon_1$ **do** Calculate n_{gm} and r_{gjm} ; Update $\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\Sigma}^{(t)}$; **for** $j=1, \dots, m$ **do** set $k = 1$, $\delta_2^{(k-1)} = 1$; **while** $\delta_2^{(k-1)} > \varepsilon_2$ **do**

$$a_{jr}^{(k)} = a_{jr}^{(k-1)} - \frac{\partial a_{jr} Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\gamma}, \boldsymbol{\beta})}{\partial a_{jr}^2 Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\gamma}, \boldsymbol{\beta})}; d_j^{(k)} = d_j^{(k-1)} - \frac{\partial d_j Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\gamma}, \boldsymbol{\beta})}{\partial d_j^2 Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\gamma}, \boldsymbol{\beta})};$$

$$\gamma_{jpr}^{(k)} = \text{soft}\left(\gamma_{jpr}^{(k-1)} - \frac{\partial \gamma_{jpr} Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\gamma}, \boldsymbol{\beta})}{\partial \gamma_{jpr}^2 Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\gamma}, \boldsymbol{\beta})}, -\frac{\eta_l w_j}{\partial \gamma_{jpr}^2 Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\gamma}, \boldsymbol{\beta})}\right);$$

$$\beta_{jp}^{(k)} = \text{soft}\left(\beta_{jp}^{(k-1)} - \frac{\partial \beta_{jp} Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\gamma}, \boldsymbol{\beta})}{\partial \beta_{jp}^2 Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\gamma}, \boldsymbol{\beta})}, -\frac{\eta_l w_j}{\partial \beta_{jp}^2 Q(\mathbf{a}, \mathbf{d}, \boldsymbol{\gamma}, \boldsymbol{\beta})}\right);$$

$$\delta_2^{(k)} = \|\mathbf{a}_j^{(k)} - \mathbf{a}_j^{(k-1)}\| + \|\mathbf{d}_j^{(k)} - \mathbf{d}_j^{(k-1)}\| + \|\boldsymbol{\gamma}_j^{(k)} - \boldsymbol{\gamma}_j^{(k-1)}\| + \|\boldsymbol{\beta}_j^{(k)} - \boldsymbol{\beta}_j^{(k-1)}\|;$$

 $k = k + 1$;

$$\delta_1^{(t)} = \|\mathbf{a}^{(t)} - \mathbf{a}^{(t-1)}\| + \|\mathbf{d}^{(t)} - \mathbf{d}^{(t-1)}\| + \|\boldsymbol{\gamma}^{(t)} - \boldsymbol{\gamma}^{(t-1)}\| + \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}\|;$$

 $t = t + 1$;Set $\eta_a = 0$ and re-estimate all none-zero estimates.