PSYCHOMETRIKA—VOL. 89, NO. 1, 267–295
MARCH 2024
https://doi.org/10.1007/s11336-024-09948-7

DIF ANALYSIS WITH UNKNOWN GROUPS AND ANCHOR ITEMS

GABRIEL WALLIN

DEPARTMENT OF MATHEMATICS AND STATISTICS, LANCASTER UNIVERSITY

YUNXIAO CHEN AND IRINI MOUSTAKI

LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Ensuring fairness in instruments like survey questionnaires or educational tests is crucial. One way to address this is by a Differential Item Functioning (DIF) analysis, which examines if different subgroups respond differently to a particular item, controlling for their overall latent construct level. DIF analysis is typically conducted to assess measurement invariance at the item level. Traditional DIF analysis methods require knowing the comparison groups (reference and focal groups) and anchor items (a subset of DIF-free items). Such prior knowledge may not always be available, and psychometric methods have been proposed for DIF analysis when one piece of information is unknown. More specifically, when the comparison groups are unknown while anchor items are known, latent DIF analysis methods have been proposed that estimate the unknown groups by latent classes. When anchor items are unknown while comparison groups are known, methods have also been proposed, typically under a sparsity assumption – the number of DIF items is not too large. However, DIF analysis when both pieces of information are unknown has not received much attention. This paper proposes a general statistical framework under this setting. In the proposed framework, we model the unknown groups by latent classes and introduce item-specific DIF parameters to capture the DIF effects. Assuming the number of DIF items is relatively small, an L_1-regularised estimator is proposed to simultaneously identify the latent classes and the DIF items. A computationally efficient Expectation-Maximisation (EM) algorithm is developed to solve the non-smooth optimisation problem for the regularised estimator. The performance of the proposed method is evaluated by simulation studies and an application to item response data from a real-world educational test.

Key words: differential item functioning, measurement invariance, latent DIF, latent class analysis, lasso.

Psychometric models to analyse data from instruments such as survey questionnaires and educational tests rely on an equivalence assumption on the item parameters across groups of respondents. That is, conditioning on the latent construct measured by the instrument, a respondent's response to each item is independent of their group membership. This assumption is known as measurement invariance. If violated, the psychometric property of the item(s) is not constant across groups, which can cause measurement bias (Millsap, 2012) . The measurement invariance assumption is typically investigated through differential item functioning (DIF) analysis, a class of statistical methods that compares respondent groups at the item level and detects non-invariant, i.e., DIF, items.

Traditional DIF detection methods assume that both the comparison groups and a set of non-DIF items, commonly referred to as the anchor set, are known a priori. The anchor items are used to identify the latent construct the instrument measures, and a DIF detection method compares the performances of the comparison groups, controlling for their performance on the anchor items as a proxy of the latent construct level. Depending on their specific assumptions, these DIF detection methods can be divided into Item-Response-Theory-based (IRT-based) methods (e.g., Kim et al., 1995; Lord, 1977; Lord, 1980; Steenkamp & Baumgartner, 1988; Tay et al.,

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11336-024-09948-7.

Correspondence should be made to Yunxiao Chen, Department of Statistics, London School of Economics and Political Science, Columbia House, Room 5.16 Houghton Street, London WC2A 2AE, UK. Email: y.chen186@lse.ac.uk

© 2024 The Author(s)

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

2012; Thissen et al., 1988; Thissen & Steinberg, 1998; Wainer, 2016; Woods et al., 2013) and non-IRT-based methods (e.g., Cao et al., 2017; Dorans & Kulick, 1986; Drabinová & Martinková, 2017; Holland & Thayer, 1986; Holland & Wainer, 1993; Shealy & Stout, 1993; Swaminathan & Rogers, 1990; Tay et al., 2015; Woods et al., 2013; Zwick et al., 2000); see Millsap (2012) for a review of traditional DIF analysis methods. Generally speaking, IRT-based methods tend to provide a clearer definition of DIF effects through a generative probabilistic model at the price of a risk of model misspecification.

Unfortunately, comparison groups and anchor items may not always be available in real-world applications, in which cases the aforementioned traditional methods are not applicable. Even if we have some information about anchor items, the result may be sensitive to the specific anchor items we use, when we only have a small number of such, and there will be a big issue if the anchor items are misspecified. Modern DIF analysis methods have been developed in situations where either the comparison groups or the anchor items are unknown. When anchor items are unknown, the latent construct is not identified, in which case, DIF detection is an ill-posed problem if no additional assumptions are made. A reasonable assumption in this situation is sparsity – the number of DIF items is relatively small, under which the detection of DIF items is turned into a model selection problem. To tackle the model selection problem, item purification methods have been proposed (e.g., Candell & Drasgow, 1988; Clauser et al., 1993; Fidalgo et al., 2000; Kopf et al., 2015a; 2015b; Wang et al., 2009; Wang & Su, 2004; Wang & Yeh, 2003), where stepwise model selection methods are used to detect DIF items. More recently, Lasso-type regularised estimation methods have been proposed to solve the model selection problem (Magis et al., 2015; Tutz & Schauberger, 2015; Belzak & Bauer, 2020; Bauer et al., 2020; Schauberger & Mair, 2020) . In these methods, the DIF effects are represented by item-specific parameters under an IRT model, where a zero coefficient encodes no DIF effect for an item, and Lasso-type penalties are imposed on the DIF parameters to obtain a sparse solution, i.e., many items are DIF-free. A drawback of regularised estimation methods is that, due to the bias brought by Lasso regularisation, they do not provide valid p-values for testing whether each item is DIF-free. Recently, Chen et al. (2023) considered a limiting case of a regularised estimator and showed that the estimator can simultaneously identify the latent construct and yield valid statistical inferences on the individual DIF effects. An alternative direction of DIF analysis without anchor items is based on the idea of differential item pair functioning. Under the Rasch model, Bechger and Maris (2015) showed that although a Rasch model with group-specific difficulty parameters is not identifiable, the relative difficulties of item pairs are identifiable and can be used for detecting DIF items. Based on this idea, Yuan et al. (2021) introduced visualisation methods for DIF detection. We lastly point out that there is related literature on L_1 regularisation for general mixture models such as Gaussian mixture models (e.g., Bhattacharya & McNicholas, 2014; Bouveyron & Brunet-Saumard, 2014; Luo et al., 2008), which also consider model-based clustering but are based on continuous instead of categorical data. However, it is worth noting that these works all consider a high-dimensional data setting, and the regularisation is used for dimension reduction. The current paper focuses on a relatively low-dimensional setting, and a regularised estimator is proposed for the purpose of model selection.

The comparison groups may sometimes be unavailable, and DIF analysis in this situation is typically referred to as latent DIF analysis (Cho et al., 2016; De Boeck et al., 2011) . As suggested in De Boeck et al. (2011), latent DIF analysis is needed when we do not know the crucial groups for comparison, we cannot observe the groups of interest, or there are validity concerns regarding the true group membership of the respondents. For example, for self-reported health and mental health instruments (Teresi & Reeve, 2016; Reeve & Teresi, 2016; Teresi et al., 2021), many covariates are collected, such as age, gender, ethnicity, and other background variables, but the crucial groups for DIF analysis are typically unclear. For another example, when analysing data from an educational test in which a subset of test takers have preknowledge on some leaked items

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

(Cizek & Wollack, 2017) , the two comparison groups of interest – the ones with and without item preknowledge – are not directly observable. Moreover, the observed group membership may sometimes poorly indicate the "true" group membership that causes the DIF pattern in the item response data (e.g., Bennink et al., 2014; Cho & Cohen, 2010; Finch & Hernández Finch, 2013; Von Davier et al., 2011). Most existing latent DIF analysis methods assume that an anchor set is known and use a mixture IRT model – a model that combines IRT and latent class analysis– to identify the unknown groups and detect the DIF items simultaneously (Cho & Cohen, 2010; Cohen & Bolt, 2005; De Boeck et al., 2011) ; see Cho et al. (2016) for a review.

In practice, both the comparison groups and the anchor set may be unknown. For example, besides the aforementioned challenges of identifying the crucial comparison groups, the DIF analysis of self-reported health and mental health instruments also faces the challenge of identifying anchor items (Teresi & Reeve, 2016; Reeve & Teresi, 2016) . In the item preknowledge example above, not only the comparison groups are unobserved, but also prior knowledge about non-leaked items is likely unavailable, and thus, correctly specifying an anchor set is a challenge (O'Leary et al., 2016) . Almost no general methods are available for latent DIF analysis when the anchor set is unavailable. Two notable exceptions are Chen et al. (2022) and Robitzsch (2022). In Chen et al. (2022), a Bayesian hierarchical model for latent DIF analysis is proposed and applied for the simultaneous detection of item leakage and preknowledge in educational tests. In this model, latent classes are imposed among the test takers to model the comparison groups, and also among the items to model the DIF and non-DIF item sets. In addition, both the person- and item-specific parameters are treated as random variables and inferred via a fully Bayesian approach. However, the inference of this model relies on a Markov chain Monte Carlo algorithm, which suffers from slow mixing. Moreover, as most traditional DIF analysis methods adopt a frequentist setting, it is of interest to develop a frequentist approach to latent DIF analysis when the anchor set is unknown. Robitzsch (2022) proposed a latent DIF procedure based on a regularised estimator under a mixture Rasch model. In this work, a nonconvex penalty called the Smoothly Clipped Absolute Deviation (SCAD) penalty (Fan & Li, 2001) other than the L_1 penalty is investigated. The methodology proposed in the current paper is similar in spirit to that of Robitzsch (2022) but developed independently. The proposed framework focuses on the two-parameter logistic (2-PL) model (Birnbaum, 1968) with an L_1 penalty and further provides a scope to generalise to other item response theory models.

This paper proposes a frequentist framework for DIF analysis when both the comparison groups and the anchor set are unknown. The proposed framework combines the ideas of mixture IRT modeling for latent DIF analysis and regularised estimation for manifest DIF analysis with unknown anchor items. More specifically, the unknown groups are modelled by latent classes, and the DIF effects are characterised by item-specific DIF parameters. An L_1-regularised marginal maximum likelihood estimator is proposed, assuming that the number of DIF items is relatively small. This estimator penalises the DIF parameters by a Lasso regularisation term so that the DIF items can be selected by the non-zero pattern of the estimated DIF parameters. Computing the L_1-regularised estimator involves solving a non-smooth optimisation problem. We propose a computationally efficient Expectation-Maximisation (EM) algorithm (Dempster et al., 1977; Bock & Aitkin, 1981) , where the non-smoothness of the objective function is handled by a proximal gradient method (Parikh & Boyd, 2014) . We evaluate the proposed method through simulation studies and an application to item response data from a real-world educational test. For the real-world application, we consider data from a midwestern university in the United States. This data set has been studied in Bolt et al. (2002), where end-of-test items are believed to cause DIF due to insufficient time. Both the comparison groups, i.e. the speeded and non-speeded respondents, and the anchor items are unknown. In Bolt et al. (2002), the DIF items and comparison groups are detected by borrowing information from an additional test form which is carefully designed so that the potential speededness-DIF items in the original form are administered at earlier locations,

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

and thus, are unlikely to suffer from speededness-DIF. Thanks to the proposed procedure, we are able to identify the unknown DIF items and comparison groups without utilising information from the additional test form, and our findings are consistent with those of Bolt et al. (2002).

The rest of the paper is organised as follows. In Sect. 1, we propose a modelling framework for latent DIF analysis with unknown groups and anchor items and a regularised estimator that simultaneously identifies the unknown groups and detects the DIF items. In Sect. 2, we propose a computationally efficient EM algorithm. The proposed method is evaluated by simulation studies in Sect. 3 and further applied to data from a real-world educational test in Sect. 4. We conclude with discussions in Sect. 5. Details about the computational algorithm are given in the Appendix.

1. Proposed Framework

1.1. Measurement Model

Consider N respondents answering J binary items. Let $Y_{ij} \in \{0, 1\}$ for $i = 1, \ldots, N$ and $j = 1, \ldots, J$ be a binary random variable recording individual i's response to item j. The response vector of individual i is denoted by $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})^\top$. We assume that the items measure a unidimensional construct, which is modelled by a latent variable θ_i. We further assume that the respondents are random samples from $K + 1$ unobserved groups, where the group membership is denoted by the latent variable $\xi_i \in \{0, 1, ..., K\}$. Given the latent trait θ_i and the latent class ξ_i, consider the two-parameter item response model with a logit link (2-PL) (measurement model) (Birnbaum, 1968)

$$\text{logit} P(Y_{ij} = 1|\theta_i, \xi_i) = a_j\theta_i + d_j + \delta_{j\xi_i}, \tag{1}$$

where a_j and d_j are known as the discrimination and easiness parameters respectively and $\delta_{j\xi_i}$ is referred to as the DIF-effect parameter, as it quantifies the DIF effect of latent class k on item j.

We treat $\xi_i = 0$ as the baseline group, also known as the reference group, and set $\delta_{j0} = 0$ for all $j = 1, \ldots, J$. In that case, $a_j\theta_i + d_j$ denotes the item response function for the reference group. When a_j is common across all items, the baseline model becomes the Rasch model (Rasch, 1960). We focus on the 2-PL model here, but the proposed method easily adapts to other baseline IRT models.

The parameter δ_{jk} characterises how respondents in group k differ from those in the reference group in terms of the item response behaviour on item j. For the reference group, the DIF parameter remains zero for all items, serving as a reference point. For the remaining latent classes, the DIF parameter can be non-zero for certain items. Crucially, the magnitude of this parameter is allowed to differ across these latent classes. This flexibility accounts for varying degrees of DIF effects across different latent groups, when comparing with the reference group. The DIF effect parameter can also be expressed in terms of log-odds. Specifically, under the 2-PL model,

$$\delta_{jk} = \log\left(\frac{P(Y_{ij} = 1|\theta_i = \theta, \xi_i = k)/(1 - P(Y_{ij} = 1|\theta_i = \theta, \xi_i = k))}{P(Y_{ij} = 1|\theta_i = \theta, \xi_i = 0)/(1 - P(Y_{ij} = 1|\theta_i = \theta, \xi_i = 0))}\right),$$

i.e., δ_{jk} is the log-odds-ratio when comparing two respondents from group k and the reference group given that they have the same latent construct level.

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

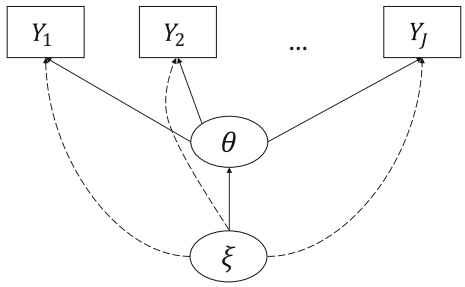


FIGURE 1.
Path diagram of the proposed model, where the dashed lines indicate the DIF effects.

1.2. Structural Model

The structural model specifies the joint distribution of the latent variables (θ_i, ξ_i). We assume that the latent classes follow a categorical distribution,

$$\xi_i \sim \text{Categorical}(\{0, 1, \ldots, K\}, (\nu_0, \nu_1, \ldots, \nu_K)),$$

where $P(\xi_i = k) = \nu_k$. There are consequently $K + 1$ latent classes with class probabilities $\boldsymbol{\nu} = (\nu_0, \nu_1, \ldots, \nu_K)^\top$ such that $\nu_k \geq 0$ and $\sum_{k=0}^{K} \nu_k = 1$. We further assume that conditional on ξ_i, the latent ability θ_i follows a normal distribution with class-specific mean and variance, i.e.,

$$\theta_i | \xi_i = k \sim \mathcal{N}(\mu_k, \sigma_k^2).$$

To ensure model identification, we fix the mean and variance of the reference group, i.e. $\mu_0 = 0$ and $\sigma_0^2 = 1$.

The path diagram of this model is given in Fig. 1. We note that the model coincides with a MIMIC model (Jöreskog & Goldberger, 1975) for manifest DIF analysis (Muthen & Lehman, 1985; Muthén, 1989; Woods, 2009; Woods & Grimm, 2011) when conditioning on the latent class ξ_i (i.e., viewing ξ_i as observed). However, since ξ_i is unobserved, the statistical inference of the proposed model differs substantially from that of the MIMIC model. More specifically, the inference of the proposed model will be based on the marginal likelihood function where both latent variables ξ_i and θ_i are marginalised out. When the baseline IRT model is the 2-PL model, the marginal likelihood function takes the form

$$L(\Delta) = \prod_{i=1}^{N} \sum_{k=0}^{K} \nu_k \int \left(\prod_{j=1}^{J} \left(\exp((a_j\theta + d_j + \delta_{jk})Y_{ij}) / (1 + \exp(a_j\theta + d_j + \delta_{jk})) \right) \right) \phi(\theta | \mu_k, \sigma_k^2) d\theta, \tag{2}$$

where $\phi(\theta|\mu_k, \sigma_k^2)$ denotes the density function of a normal distribution with mean μ_k and variance σ_k^2, and we use vector Δ to denote all the unknown parameters, including the item parameters a_j and d_j, δ_{jk}, ν_k, μ_k and σ_k^2, for $j = 1, ..., J$ and $k = 0, 1, ..., K$.

1.3. Model Identifiability

The current model suffers from two sources of unidentifiability. The first source of unidentifiability comes from not knowing the anchoring items, which occurs even if we condition on the latent class ξ_i, i.e., when the model becomes a MIMIC model for manifest DIF analysis. That

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

is, for any constants $c_1, ..., c_K$, if we simultaneously replace μ_k by $\mu_k + c_k$ and replace δ_{jk} by $\delta_{jk} - a_j \mu_k$ for all $j = 1, ..., J$ and $k = 1, ..., K$, the likelihood function value $L(\Delta)$ remains the same. This source of unidentifiability can be avoided when one or more anchor items are known a priori. Suppose that item j is known to be DIF-free. Under the proposed model framework, it implies the constraints $\delta_{jk} = 0$ for all k. Consequently, the aforementioned transformation can no longer apply, as otherwise, the zero constraints for the anchor items will be violated. As discussed in Sect. 1.4, this source of unidentifiability can be handled by a regularised estimation approach under a sparsity assumption that many DIF parameters δ_{jk} are zero.

The second source of unidentifiability is the label-switching phenomenon of latent class models (Redner & Walker, 1984), as a result of the exchangeability of the latent classes. Under the current model, the baseline class is uniquely identified through the constraints $\delta_{j0} = 0$, $\mu_0 = 0$ and $\sigma_0^2 = 1$. However, the remaining latent classes are exchangeable, and the likelihood function value remains the same when switching their labels. While label switching often causes trouble when inferring a latent class model with Bayesian Markov chain Monte Carlo (MCMC) algorithms (Stephens, 2000), it is not a problem for the estimator to be discussed in Sect. 1.4. Our estimator is proposed under the frequentist setting and computed by an EM algorithm. When the EM algorithm converges, it will reach one of the equivalent solutions in the sense of label switching.

1.4. Sparsity, Model Selection and Estimation

As explained above, the latent trait cannot be identified without anchor items. In that case, additional assumptions are needed to solve the DIF analysis problem. Specifically, we adopt the sparsity assumption, i.e., many DIF parameters δ_{jk} are zero. This is a common assumption in the manifest DIF literature, see for example (Magis et al., 2015; Tutz & Schauberger, 2015; Belzak & Bauer, 2020; Bauer et al., 2020; Schauberger & Mair, 2020). In many applications, for example, the detection of aberrant behaviour or parameter drift in educational testing, the number of DIF items is low, suggesting that this assumption is meaningful.

Under the above sparsity assumption, we propose an L_1 regularised estimator to simultaneously estimate the unknown model parameters and learn the sparsity pattern of the DIF-effect parameters. This estimator takes the form

$$\tilde{\Delta}^{(\lambda)} = \arg\min_{\Delta} -\log L(\Delta) + \lambda \sum_{j=1}^{J} \sum_{k=1}^{K} |\delta_{jk}|, \text{ s.t. } \nu_k \geq 0, k = 0, 1, ..., K, \text{ and } \sum_{k=0}^{K} \nu_k = 1, \tag{3}$$

where $L(\Delta)$ is the marginal likelihood function defined in (2), and $\lambda > 0$ is a tuning parameter. The computation of this estimator will be discussed in Sect. 2. Similar to Lasso regression (Tibshirani, 1996), the L_1 regularisation term $\lambda \sum_{j=1}^{J} \sum_{k=1}^{K} |\delta_{jk}|$ in (3) tends to shrink some of the DIF-effect parameters to be exactly zero. In the most extreme case where λ goes to infinity, all the DIF-effect parameters will shrink to zero. Under suitable regularity conditions and when λ is chosen properly (i.e., λ goes to infinity at a suitable speed), the L_1 regularised estimator yields both estimation and selection consistency (Zhao & Yu, 2006; van de Geer, 2008). In that case, the latent trait is consistently identified, and the consistently selected sparse patterns of the estimated DIF-effect parameters can be used to classify items as DIF and non-DIF items.

We select the tuning parameter λ based on the Bayesian Information Criterion (BIC; Schwarz, 1978) using a grid search approach. Specifically, we consider a pre-specified set of grid points for λ, denoted by

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

$\lambda_1,\ldots \lambda_M$. For each value of λ_m, we solve the optimisation problem (3) and obtain $\tilde{\Delta}^{(\lambda_m)}$. To compute the BIC value for the model encoded by $\tilde{\Delta}^{(\lambda_m)}$, we compute a constrained maximum likelihood estimator, fixing δ_{jk} to zero if $\tilde{\delta}_{jk}^{(\lambda_m)} = 0$. That is,

$$\begin{aligned} \hat{\Delta}^{(\lambda_m)} &= \arg\min_{\Delta} -\log L(\Delta), \\ &\text{s.t. } \nu_k \geq 0, k = 0, ..., K, \ \sum_{k=0}^{K} \nu_k = 1, \\ &\delta_{jk} = 0 \ \text{ if } \ \tilde{\delta}_{jk}^{(\lambda_m)} = 0, \ j = 1, ..., J, k = 1, ..., K. \end{aligned} \tag{4}$$

The BIC corresponding to λ_m is calculated as

$$\text{BIC}_{\lambda_m} = -2 \log L(\hat{\Delta}^{(\lambda_m)}) + \log(N)\text{Card}(\hat{\Delta}^{(\lambda_m)}), \tag{5}$$

where $\text{Card}(\hat{\Delta}^{(\lambda_m)})$ denotes the number of free parameters in $\hat{\Delta}^{(\lambda_m)}$ that equals to the total number of free parameters in Δ minus the corresponding number of zero constraints. The tuning parameter is then selected as

$$\hat{\lambda} = \arg\min_{\lambda_m, m=1,...,M} \text{BIC}_{\lambda_m}.$$

Thanks to the asymptotic properties of the BIC (Shao, 1997) , the true model will be consistently selected if it can be found by one of the tuning parameters.

We use the constrained maximum likelihood estimator $\hat{\Delta}^{(\hat{\lambda})}$ as the final estimator of the selected model and declare an item j to be a DIF item if $\|(\hat{\delta}_{j1}^{(\hat{\lambda})}, ..., \hat{\delta}_{jK}^{(\hat{\lambda})})^\top\| \neq 0$. We summarise this procedure in Algorithm 1 below.

Algorithm 1
Regularised estimation and model selection.

Input: Observed data Y_{ij}, $i = 1, ..., N$ and $j = 1, ..., J$, and a grid of tuning parameters $0 < \lambda_1 < \ldots < \lambda_M$.

For $m = 1, \ldots, M$,

1. Solve the optimisation problem (3) with tuning parameter λ_m and obtain solution $\tilde{\Delta}^{(\lambda_m)}$.
2. Given $\tilde{\Delta}^{(\lambda_m)}$, solve the constrained optimisation problem (4) and obtain $\hat{\Delta}^{(\lambda_m)}$.
3. Calculate the BIC value BIC_{λ_m} based on equation (5) for the model encoded by $\tilde{\Delta}^{(\lambda_m)}$.

Select $\hat{\lambda} = \arg\min_{\lambda_m, m=1,...,M} \text{BIC}_{\lambda_m}$.

Output: Estimate of the selected model, $\hat{\Delta}^{(\hat{\lambda})}$, and the selected DIF items $\{j : \|(\hat{\delta}_{j1}^{(\hat{\lambda})}, ..., \hat{\delta}_{jK}^{(\hat{\lambda})})^\top\| \neq 0\}$.

---

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

1.5. Other Inference Problems

The latent class membership can be inferred by an empirical Bayes procedure, i.e., by the maximum a posteriori (MAP) estimate under the estimated model. For the MAP estimator, the goal is to find the most probable latent class k for each respondent i, given their observed responses $\boldsymbol{y}_i$. This is done by maximizing the posterior probability of ξ_i being equal to k, conditioned on the observed responses $\boldsymbol{y}_i$:

$$\hat{\xi}_{\text{MAP},i} = \underset{k\in\{0,1,\ldots,K\}}{\arg\max}\ \hat{P}(\hat{\xi} = k|\mathbf{y}_i)$$

We get these posterior probabilities through

$$\hat{P}(\xi_i = k|\mathbf{y}_i) = \frac{\hat{P}(\mathbf{y}_i|\xi_i = k, \theta_i)\cdot\hat{P}(\xi_i = k)}{\hat{P}(\mathbf{y}_i)} \tag{6}$$

based on the estimated model parameters.

Lastly, the number of latent classes, i.e., the choice of K, can be determined using the BIC. That is, we solve the optimisation problem in (3) for different values of K, yielding $\hat{\Delta}^{(K)}$. We thereafter compute the BIC as

$$\text{BIC}(K) = -2\log L(\hat{\Delta}^{(K)}) + \log(N)\text{Card}(\hat{\Delta}^{(K)}), \tag{7}$$

and select the K that yields the smallest BIC:

$$K^* = \underset{K}{\arg\min}\,\text{BIC}(K) \tag{8}$$

1.6. Extensions

The proposed model is possible to extend in several ways to accommodate different data types and more than one factor. To make this clear, our model can be expressed as

$$P(Y_{ij} = 1|\theta_i, \xi_i) = f\Big(g(\theta_i, \beta_j) + \delta_{j\xi_i}\Big) \tag{9}$$

The function g determines the parametrisation of the person and item parameters, denoted by θ and β respectively. If for example the Rasch model (Rasch, 1960) is adopted, $\beta_j = (a, d_j)$ where $a_j = a$ for all $j = 1, \ldots, J$,

$$g(\theta_i, \beta_j) = a\theta_i + d_j,$$

and $f(x) = \exp(x)/(1 + \exp(x))$.

It is also possible to consider link functions other than the logistic function considered in this paper, such as the probit link:

$$f(x) = \int_{-\infty}^{x} \phi(z)dz.$$

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

DIF analysis using multidimensional IRT models with unknown anchor items has recently been considered in Wang et al. (2023). As in the unidimensional case, no method can handle situations where both the groups and the anchor items are missing. Our proposed framework can however be extended to handle such situations. Consider the extension of (9) where each respondent, in addition to the latent class membership ξ_i, is represented by an L-dimensional latent vector $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iL})^\top$. Each item is represented by an intercept parameter d_j and L loading parameters $\boldsymbol{a}_j = (a_{j1}, \ldots, a_{jL})^\top$. This extension of the model can be expressed as a multidimensional 2-PL model with an added DIF component, i.e.,

$$P(Y_{ij} = 1|\boldsymbol{\theta}_i, \xi_i) = f\left(d_j + \boldsymbol{a}_j^\top \boldsymbol{\theta}_i + \delta_{j\xi_i}\right).$$

The proposed modeling framework can also be extended to accommodate ordinal data. Denoting the response categories of Y_{ij} by $c = 1, \ldots, C$, such model can, using the logistic link, be expressed as

$$\text{logit}P(Y_{ij} \leq c) = d_{jc} - a_j\theta_i + \delta_{j\xi_i}$$

The model without the DIF parameter is known as the proportional odds model (Samejima, 1969) . Note the negative sign in front of the slope parameter so that if a_j is positive, increasing θ_i will increase the probability of higher-numbered levels of Y_{ij}.

Lastly, we mention the possibility of extending the model to accommodate for DIF effects in the discrimination parameter, known as non-uniform DIF. To consider such a case, we introduce a similar DIF-effect parameter for a_j, just as we have for d_j. Let's denote the DIF effect on the discrimination parameter $\alpha_{j\xi_i}$. Given that $\xi_i = 0$ is treated as the reference group, we set $\alpha_{j0} = 0$ for all $j = 1, \ldots, J$. In this case, the modified 2-PL model with a logit link which accounts for DIF in both a_j and d_j can be written as:

$$\text{logit}P(Y_{ij} = 1|\theta_i, \xi_i) = (a_j + \alpha_{j\xi_i})\theta_i + d_j + \delta_{j\xi_i}. \tag{10}$$

To include DIF in the L_1 regularised estimator, the penalty term needs to be modified to penalize both α_{jk} and δ_{jk} terms. The modified estimator is given by

$$\tilde{\Delta}^{(\lambda)} = \arg\min_{\Delta} -\log L(\Delta) + \lambda \left(\sum_{j=1}^{J}\sum_{k=1}^{K}|\delta_{jk}| + \sum_{j=1}^{J}\sum_{k=1}^{K}|\alpha_{jk}|\right),$$
$$\text{s.t. } \nu_k \geq 0, k = 0, 1, ..., K, \text{and} \sum_{k=0}^{K}\nu_k = 1, \tag{11}$$

2. Computation

The computation of the optimisation problems (3) and (4) is carried out using the EM algorithm (Dempster et al., 1977; Bock & Aitkin, 1981) . An EM algorithm is an iterative algorithm, alternating between an Expectation (E) step and a Maximisation (M) step. Optimisation problem (4) involves maximising the marginal likelihood function of a regular latent variable model, and thus, can be solved by a standard EM algorithm. Thus, its details are skipped here. However, the

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

optimisation problem (3) involves a non-smooth L_1 term. Consequently, the M step of the algorithm cannot be carried out using a gradient-based numerical solver, such as a Newton-Raphson algorithm. We develop an efficient proximal-gradient-based EM algorithm that uses a proximal gradient update (Parikh & Boyd, 2014) to carry out the non-smooth optimisation problem in the M-step. In what follows, we elaborate on this algorithm using the 2-PL model as the baseline IRT model, while pointing out that the algorithm easily extends to other baseline IRT models.

Suppose that t iterations of the algorithm have been run and let $\Delta^{(t)}$ be the current parameter value. In the E-step of the tth iteration, we construct a local approximation of the negative objective function at $\Delta^{(t)}$ in the form of

$$Q(\Delta|\Delta^{(t)}) = \sum_{i=1}^{N} \mathbb{E}\left[\log\left(\nu_{\xi_i}\prod_{j=1}^{J}\left(\frac{\exp((a_j\theta_i + d_j + \delta_{j\xi_i})Y_{ij})}{1+\exp(a_j\theta_i + d_j + \delta_{j\xi_i})}\right)\phi(\theta_i|\mu_{\xi_i}, \sigma^2_{\xi_i})\big|\mathbf{Y}_i, \Delta^{(t)}\right)\right] - \lambda\sum_{j=1}^{J}\sum_{k=1}^{K}|\delta_{jk}|. \tag{12}$$

We note that the expectation in (12) is with respect to the conditional distribution of the latent variables (θ_i, ξ_i) given $\mathbf{Y}_i$, evaluated at the current parameters $\Delta^{(t)}$.

In the M-step, we find $\Delta^{(t+1)}$ such that

$$Q(\Delta^{(t+1)}|\Delta^{(t)}) > Q(\Delta|\Delta^{(t)}),$$

or equivalently,

$$-Q(\Delta^{(t+1)}|\Delta^{(t)}) < -Q(\Delta|\Delta^{(t)}).$$

By Jensen's inequality, it consequently guarantees that the objective function of (3) decreases, i.e.,

$$-\log L(\Delta^{(t+1)}) + \lambda\sum_{j=1}^{J}\sum_{k=1}^{K}|\delta_{jk}^{(t+1)}| < -\log L(\Delta^{(t)}) + \lambda\sum_{j=1}^{J}\sum_{k=1}^{K}|\delta_{jk}^{(t)}|.$$

More specifically, we write $\Delta = (\Delta_1^\top, \Delta_2^\top)^\top$, where $\Delta_1 = (\nu_0, ..., \nu_K)^\top$ and Δ_2 contains the rest of the parameters. We notice that $-Q(\Delta|\Delta^{(t)})$ in (12) can be decomposed as the sum of a smooth function
$D_t(\Delta_1) = -\sum_{i=1}^{N}\mathbb{E}\left[\log\left(\nu_{\xi_i}\big|\mathbf{Y}_i, \Delta^{(t)}\right)\right]$, a smooth function $F_t(\Delta_2)$, defined as

$$F_t(\Delta_2) = -\sum_{i=1}^{N}\mathbb{E}\left[\log\left(\prod_{j=1}^{J}\left(\frac{\exp((a_j\theta_i + d_j + \delta_{j\xi_i})Y_{ij})}{1+\exp(a_j\theta_i + d_j + \delta_{j\xi_i})}\right)\phi(\theta_i|\mu_{\xi_i}, \sigma^2_{\xi_i})\big|\mathbf{Y}_i, \Delta_2^{(t)}\right)\right]$$

and a non-smooth function $G(\Delta_2) = \lambda\sum_{j=1}^{J}\sum_{k=1}^{K}|\delta_{jk}|$. We note that $\Delta_1^{(t+1)}$ can be obtained by solving the following constrained optimisation problem

$$\Delta_1^{(t+1)} = \arg\min_{\Delta_1} D_t(\Delta_1), \text{ s.t. } \nu_k \geq 0, k = 0, 1, ..., K, \text{ and } \sum_{k=0}^{K}\nu_k = 1.$$

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

Using the method of Lagrangian multiplier, this optimisation problem has a closed-form solution; see the Appendix for the details.

We then find $\Delta_2^{(t+1)}$ such that $F_t(\Delta_2^{(t+1)}) + G(\Delta_2^{(t+1)}) < F_t(\Delta_2^{(t)}) + G(\Delta_2^{(t)})$. Consider the optimisation problem min $F_t(\Delta_2) + G(\Delta_2)$. Due to the non-smoothness of G, this objective function is not differentiable everywhere. Consequently, gradient-based methods are not applicable. We find $\Delta_2^{(t+1)}$ by using a proximal gradient method (Parikh & Boyd, 2014) . Denote the dimension of Δ_2 by d, where d equals the number of free parameters which is determined by the choice of baseline model ($2J + 2K$ for the 2-PL model). We define the proximal operator of G at Δ_2 as

$$\text{Prox}_G(\Delta_2) = \underset{\tilde{\Delta}_2 \in \mathbb{R}^d}{\arg\min}\ G(\Delta_2) + \frac{1}{2}\|\tilde{\Delta}_2 - \Delta_2\|^2.$$

We update Δ_2 by

$$\Delta_2^{(t+1)} = \text{Prox}_{\alpha G}(\Delta_2^{(t)} - \alpha \nabla F_t(\Delta_2^{(t)})), \tag{13}$$

where $\nabla F_t(\Delta_2^{(t)})$ denotes the gradient of F_t at $\Delta_2^{(t)}$, and α is a step size. According to Parikh and Boyd (2014, sect.4.2), for a sufficiently small step size α, it is guaranteed that $F_t(\Delta_2^{(t+1)}) + G(\Delta_2^{(t+1)}) < F_t(\Delta_2^{(t)}) + G(\Delta_2^{(t)})$, unless $\Delta_2^{(t)}$ is already a stationary point. Thus, we select α by a line search procedure, whose details are provided in the Appendix. We note that this proximal operator has a closed-form solution. Specifically, each DIF-effect parameter δ_{jk} is updated by solving

$$\delta_{jk}^{(t+1)} = \underset{\tilde{\delta}_{jk}}{\arg\min} -\frac{1}{2}\left(\tilde{\delta}_{jk} - \left(\delta_{jk}^{(t)} - \alpha \frac{\partial F_t(\Delta_2)}{\partial \delta_{jk}}\Big|_{\Delta_2 = \Delta_2^{(t)}}\right)\right)^2 + \alpha\lambda|\tilde{\delta}_{jk}|,$$

which has a closed-form solution given by soft-thresholding (Chapter 3, Hastie et al., 2009). The rest of the parameters in Δ_2 do not appear in the non-smooth function $G(\Delta_2)$, and thus, the resulting update of (13) degenerates to a vanilla gradient descent update. For example,

$$d_j^{(t+1)} = d_j^{(t)} - \alpha \frac{\partial F_t(\Delta_2)}{\partial d_j}\Big|_{\Delta_2 = \Delta_2^{(t)}}.$$

Further details of the proximal gradient update can be found in the Appendix. We summarise the main steps of this EM algorithm in Algorithm 2 below.

Remark. While differentiable approximations, such as the smoothed Lasso, can allow the use of gradient-based methods, they often come with their own set of challenges. Introducing a smoothing parameter can make the method sensitive to its choice, and in some situations, the approximation might not be close enough to the original problem, especially when the smoothing parameter is not sufficiently small. We opt for an approach based on non-smooth optimisation. We believe that directly tackling the non-smoothness ensures that we do not compromise on the sparsity of the solution, which is critical for our analysis. We have designed our algorithm around the EM algorithm, which inherently possesses certain convergence properties. Specifically, the EM algorithm is guaranteed to increase the log-likelihood in each iteration, and under mild regularity conditions, it converges to at least a local maximum of the log-likelihood. While we acknowledge that there are potential pitfalls in the convergence of non-smooth optimisation algorithms, we have implemented strategies to ensure the stability of our algorithm, such as adaptive step sizes and convergence checks.

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

Algorithm 2
An EM algorithm for solving (3).

Input: The initial value $\Delta^{(0)} = (d_j^{(0)}, a_j^{(0)}, \delta_{jk}^{(0)}, \mu_k^{(0)}, \sigma_k^{(0)}, \nu_k^{(0)})$, $j = 1, \ldots J, k = 0, \ldots, K$.

For iterations $t = 0, 1, 2, ..., T$, we iterate between the following steps:

E-Step: Calculate the Q function in (12).
M-step: Calculate the parameter update for $\boldsymbol{\nu}$ according to

$$\nu_k^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N}\gamma_{ik}^{(t)}, \quad k = 1, \ldots K,$$

where

$$\gamma_{ik}^{(t)} = \frac{\nu_k^{(t)}\prod_{j=1}^{J}\pi_{ijk}^{y_{ij}}(1-\pi_{ijk}^{1-y_{ij}})\phi(\theta_i|\mu_k,\sigma_k^2)}{\sum_{k'=0}^{K}\nu_{k'}^{(t)}\prod_{j=1}^{J}\pi_{ijk}^{y_{ij}}(1-\pi_{ijk}^{1-y_{ij}})\phi(\theta_i|\mu_k,\sigma_k^2)}$$

and

$$\pi_{ijk} = P(Y_{ij} = 1|\theta_i = \theta, \xi_i = k).$$

Update the rest of the parameters according to

$$d_j^{(t+1)} = d_j^{(t)} - \alpha\frac{\partial F_t(\Delta_2)}{\partial d_j}\Big|_{\Delta_2=\Delta_2^{(t)}}, \quad j = 1, \ldots, J,$$

$$a_j^{(t+1)} = a_j^{(t)} - \alpha\frac{\partial F_t(\Delta_2)}{\partial a_j}\Big|_{\Delta_2=\Delta_2^{(t)}}, \quad j = 1, \ldots, J,$$

$$\mu_k^{(t+1)} = \mu_k^{(t)} - \alpha\frac{\partial F_t(\Delta_2)}{\partial \mu_k}\Big|_{\Delta_2=\Delta_2^{(t)}}, \quad j = 1, \ldots, J,$$

$$\sigma_k^{(t+1)} = \sigma_k^{(t)} - \alpha\frac{\partial F_t(\Delta_2)}{\partial \sigma_k}\Big|_{\Delta_2=\Delta_2^{(t)}}, \quad j = 1, \ldots, J,$$

and

$$\delta_{jk}^{(t+1)} = \begin{cases} \delta_{jk}^{(t)} - \alpha\lambda, & \text{if } \delta_{jk}^{(t)} > \alpha\lambda \\ \delta_{jk}^{(t)} + \alpha\lambda, & \text{if } \delta_{jk}^{(t)} < -\alpha\lambda, \\ 0, & \text{otherwise} \end{cases}$$

where α is a step size chosen by line search.

Iterate until convergence has been reached.

Output: $\Delta^{(T)} = (d_j^{(T)}, a_j^{(T)}, \delta_{jk}^{(T)}, \mu_k^{(T)}, \sigma_k^{(T)}, \nu_k^{(T)})$, $j = 1, \ldots J, k = 0, \ldots, K$.

3. Simulation Study

In this simulation, we evaluate the performance of the proposed method, treating the number of latent classes K as fixed. We consider cases with two and three latent classes, respectively. For each simulation scenario, we run $B = 100$ independent replications.

3.1. Settings

We examine 16 scenarios under the two-group setting and 8 scenarios under the three-group setting, considering $J \in \{25, 50\}$, $N \in \{500, 1000\}$, and $K \in \{1, 2\}$. Note that $K = 1$ represents a two-group setting and $K = 2$ indicates a three-group setting. In the two-group setting, the class

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

proportions are varied, with the reference group consisting of either 50%, 80%, or 90% of the respondents. In the three-group scenario, 50% of the respondents belong to the reference group, 30% belong to the second latent class, and 20% belong to the third latent class. The number of DIF items is set to 10 for all cases with $J = 25$ and 20 for all cases with $J = 50$, i.e., the proportion of DIF items remains the same.

The intercept parameters d_j are generated from the Uniform$(-2, 2)$ distribution and the slope parameters a_j from the Uniform$(0.5, 1.5)$ distribution. In the two-group setting, we consider three cases for the class proportions, where $\nu = (0.1, 0.9)$, $\nu = (0.2, 0.8)$, and $\nu = (0.5, 0.5)$, respectively. The latent construct θ within each latent class ξ follows

$$\theta_i|\xi = 0 \sim \mathcal{N}(0, 1) \quad \text{and} \quad \theta_i|\xi = 1 \sim \mathcal{N}(0.5, 1.5),$$

and for the three-group case, we additionally let $\theta_i|\xi = 2 \sim \mathcal{N}(1, 1.2)$. The DIF effect parameters are generated as $\delta_{jk} \sim$ Uniform$(0.5, 1.5)$ for the non-zero elements and set to 0 for the remaining items. For the three-group case, the DIF effects for the second and third latent class are generated from a Uniform$(0.5, 1)$ and Uniform$(1, 1.5)$ distribution, respectively. In the two-group setting, the DIF items are positioned at the beginning of the scale (items 1-10 when $J = 25$ and 1-20 when $J = 50$). In the three-group setting, the DIF items are positioned at the end of the scale (items 16-25 when $J = 25$ and 31-50 when $J = 50$). The true model parameters are given in the supplementary material.

3.2. Evaluation Criteria

Detection of DIF items. We check whether the DIF items can be accurately detected by the proposed method. In this analysis, we assume the number of latent classes K is known. We calculate the average True Positive Rate (TPR) as

$$\overline{\text{TPR}} = \frac{1}{B}\sum_{b=1}^{B}\frac{\sum_{j,k}\mathbb{1}_{\{\hat{\delta}_{jk}^{(b)}\neq 0, \delta_{jk}^{*}\neq 0\}}}{\sum_{j,k}\mathbb{1}_{\{\delta_{jk}^{*}\neq 0\}}}, \tag{14}$$

where $\{\hat{\delta}_{jk}^{(b)}\}_{J\times K}$ is the estimated DIF effect matrix in the b-th replication and $\{\delta_{jk}^{*}\}_{J\times K}$ denotes the true DIF effect matrix. Similarly, we calculate the average true negative rate ($\overline{\text{TNR}}$), which is the failure rate of identifying zero elements:

$$\overline{\text{FPR}} = \frac{1}{B}\sum_{b=1}^{B}\frac{\sum_{j,k}\mathbb{1}_{\{\hat{\delta}_{jk}^{(b)}\neq 0, \delta_{jk}^{*}= 0\}}}{\sum_{j,k}\mathbb{1}_{\{\delta_{jk}^{*}= 0\}}}. \tag{15}$$

To better evaluate the performance of the proposed method in detecting DIF items, we compare the FPR and TPR with the results under an oracle setting where the group membership ξ_i is known while anchor items are unknown. Under this oracle setting, we solve the following regularised estimation problem as in Bauer et al. (2020):

$$\min_{\Xi} -\log L^{ora}(\Xi) + \lambda\sum_{j=1}^{J}\sum_{k=1}^{K}|\delta_{jk}|, \tag{16}$$

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

where

$$L^{ora}(\Xi) = \prod_{i=1}^{N} \int \left(\prod_{j=1}^{J} \left(\exp((a_j\theta + d_j + \delta_{j\xi_i})Y_{ij})/(1 + \exp(a_j\theta + d_j + \delta_{j\xi_i})) \right) \right) \phi(\theta|\mu_{\xi_i}, \sigma^2_{\xi_i}) d\theta$$

and Ξ includes the parameters in Δ except for those in $\boldsymbol{\nu}$. The tuning parameter λ is chosen by BIC. The FPR and TPR for detecting DIF items are also calculated under this setting and compared with those from the proposed method where ξ_is are unknown.

Classification of respondents. We then consider the classification of respondents based on the MAP estimate. Again, we assume the number of latent focal groups K is known. We calculate the average classification error rate, given by the fraction of incorrect classifications to the overall number of classifications averaged over the number of replications:

$$\text{error}_{MAP} = \frac{1}{N} \sum_{i=1}^{N} [\hat{\xi}_{\text{MAP},i} \neq \xi_i],$$

where

$$\hat{\xi}_{\text{MAP},i} = \underset{k \in \{0,1,\ldots,K\}}{\arg\max} \ \hat{P}(\hat{\xi} = k|\mathbf{y}_i)$$

and $\hat{P}(\hat{\xi} = k|\mathbf{y}_i)$ is the posterior probability of category k given in (6). We notice that there is a label-switching problem under the setting with $K = 2$ when calculating the classification error. This problem is solved by a post-hoc label switching based on the estimated ν_1 and ν_2, using the ordering information that class 2 is larger than class 1, i.e., $\nu_2 > \nu_1$. We also calculate the MAP classification error under the true model and compare it with the classification error of the proposed method.

Parameter estimation accuracy. We further evaluate the accuracy of our final estimator $\hat{\Delta}^{(\hat{\lambda})}$, assuming that K is known. For each unknown parameter, the root mean square error (RMSE) and the absolute bias are calculated based on the 100 replications.

3.3. Results

The classification performance in the simulation study is presented in Tables 1 and 2, which display the respondent and item classification accuracy for the two-group and three-group settings, respectively. First, it is observed that the classification error is predominantly influenced by the number of items, with larger item sizes resulting in better respondent classification performance. This observation is consistent with previous literature on DIF detection using IRT models (Chen et al., 2022; Kuha & Moustaki, 2015) .

For respondent classification in the two-group setting, we observe in Table 1 that the classification error is small for all simulation scenarios, and always better than a naïve classifier that assigns all respondents to the reference group. This is true even when the focal group only consists of 10% of the respondents. As the proportion in the focal group (in Table 1 denoted by π) increases, the proposed method's enhancement over the naïve classifier, which assigns all respondents to the baseline group, becomes more pronounced. We also note that the AUC values increase when the proportion of respondents in the focal group increases, but the increments are small. Table 1 furthermore shows that the classification accuracy is only slightly worse when using the estimated parameters compared to when the true model parameters are used.

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

TABLE 1.
Respondent and item classification accuracy under different simulation scenarios for the two-group case.

	$N = 1,000$					
	$J = 25$			$J = 50$		
Evaluation measure	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.5.$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.5$
Classification error	0.078	0.141	0.256	0.071	0.113	0.179
Classification error true	0.077	0.140	0.252	0.069	0.111	0.175
AUC	0.796	0.817	0.824	0.896	0.903	0.904
AUC true	0.813	0.827	0.830	0.903	0.907	0.908
TPR	0.957	0.948	0.948	0.997	0.993	0.996
FPR	0	0	0	0	0	0
TPR oracle	1	1	1	1	1	1
FPR oracle	0	0	0	0	0	0
	$N = 5,000$					
	$J = 25$			$J = 50$		
Evaluation measure	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.5$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.5$
Classification error	0.087	0.161	0.259	0.074	0.124	0.177
Classification error true	0.086	0.160	0.253	0.073	0.122	0.174
AUC	0.809	0.812	0.821	0.897	0.901	0.904
AUC true	0.819	0.819	0.827	0.902	0.904	0.907
TPR	0.953	0.947	0.950	0.995	0.993	0.996
FPR	0	0	0	0	0	0
TPR oracle	1	1	1	1	1	1
FPR oracle	0	0	0	0	0	0

The classification error and AUCs present respondent classification performance, where 'AUC true' gives the results using the true parameter values. The TPRs and FPRs present the results for item classification, where 'TPR oracle' and 'FPR oracle' are the performance of the oracle estimator.

For item classification in the two-group setting, the true positive rates are very high, and with no item being falsely flagged as a DIF item, across all scenarios. This suggests that the proposed framework is effective in identifying DIF items and minimizing false positives for various numbers of items and proportions in the focal group. We also note that the oracle estimator performance is only slightly better, i.e., knowing the group membership of the respondents only leads to marginal improvement in item classification accuracy.

In the three-group setting, Table 2 shows that the classification error rates are generally higher than those observed in the two-group case, which is expected given the increased complexity of the DIF detection problem when more than two groups are involved. However, note that the classification error is always clearly smaller than the naïve classifier that assigns all of the respondents to the reference group. This increase in classification performance is particularly clear in the simulation scenarios with $J = 50$. We also observe that the AUC values for classes 2 and 3 are within reasonable ranges, suggesting that the proposed method is capable of adequately classifying respondents even in more challenging settings. The TPR and FPR values for item classification in the three-group setting are furthermore high for both class 2 and class 3 items and with no misclassified DIF-free item. This further supports the effectiveness of the proposed framework in identifying DIF items across different group configurations.

Figures 2 and 3 show the RMSEs (Root Mean Squared Errors) for all the item parameter estimates for the two-group setting ($K = 1$). The RMSEs are small for all estimates across the items, with the exemption of one or two items that show larger RMSEs for the estimated DIF

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

TABLE 2.
Respondent and item classification accuracy under different simulation scenarios for the three-group case.

Evaluation measure	$N = 1,000$		$N = 5,000$	
	$J = 25$	$J = 50$	$J = 25$	$J = 50$
Classification error	0.391	0.353	0.397	0.357
Classification error true	0.393	0.355	0.400	0.357
AUC class 2	0.725	0.777	0.723	0.764
AUC class 3	0.782	0.812	0.775	0.822
AUC class 2 true	0.743	0.794	0.742	0.779
AUC class 3 true	0.807	0.836	0.794	0.838
TPR class 2	0.925	0.971	0.931	0.960
TPR class 3	1	1	1	1
FPR class 2	0	0	0	0
FPR class 3	0	0	0	0
TPR class 2 oracle	1	0.95	1	1
TPR class 3 oracle	1	1	1	1
FPR class 2 oracle	0	0	0	0
FPR class 3 oracle	0	0	0	0

The classification error and AUCs present respondent classification performance, where 'AUC true' gives the results using the true parameter values. The TPRs and FPRs present the results for item classification, where 'TPR oracle' and 'FPR oracle' are the performance of the oracle estimator.

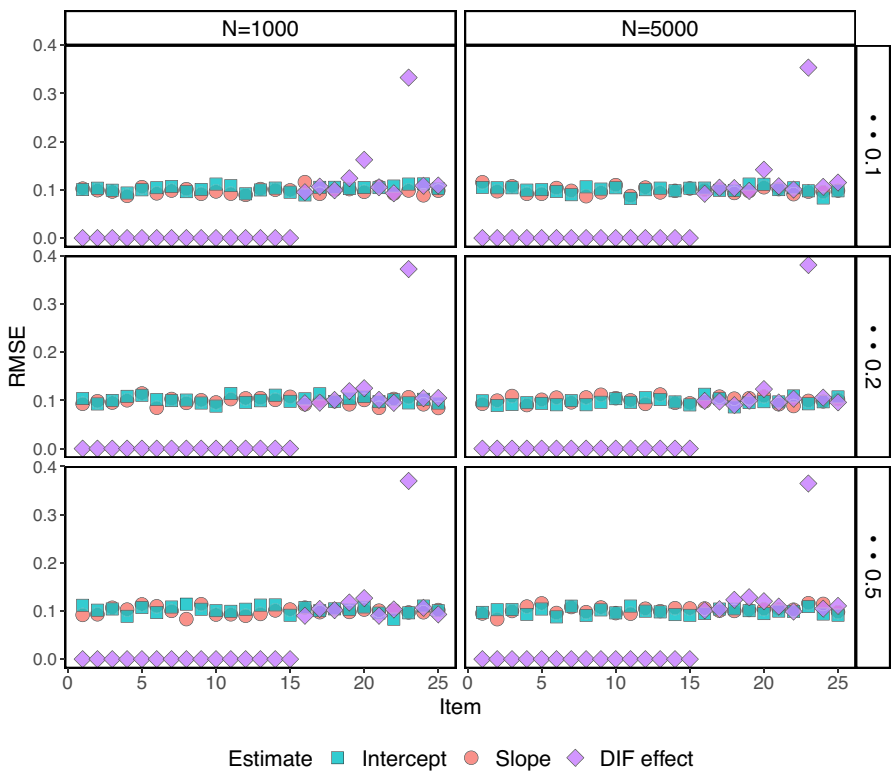


FIGURE 2.
RMSE for $J = 25$ under the 2-group setting.

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

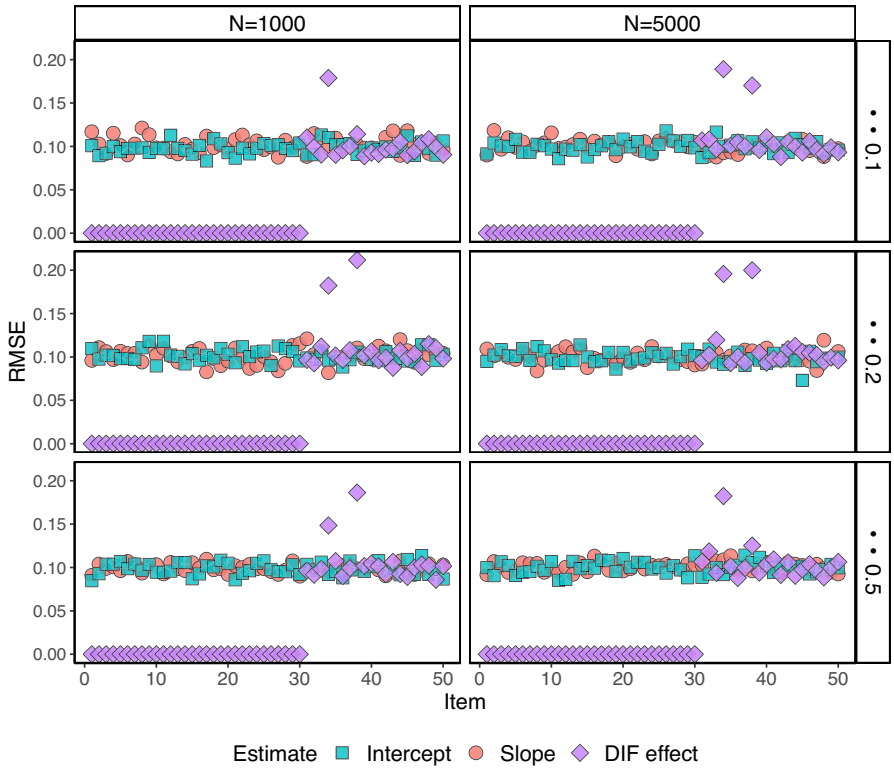


FIGURE 3.
RMSE for $J = 50$ under the 2-group setting.

parameter. The RMSEs for the estimated DIF parameters for the DIF-free items are zero as the proposed estimation procedure successfully classifies the DIF items. We also see that the number of items and proportion of respondents in the focal group have essentially no influence on the RMSE, for the configurations considered in this simulation study.

In Fig. 4, we can observe the RMSEs specifically for the three-group scenario. It is important to note that in this case, there exist two DIF effects, one for each focal group. For the first focal group, the true DIF effects are drawn from values in the range [0.5 – 1] and for the second focal group, they are drawn from the range [1 – 1.5]. The increased difficulty of estimating smaller DIF effects is reflected by larger RMSEs for the first focal group.

In Table 3 and Table 4, the absolute bias and RMSE under the 2-group setting, averaged over the number of items of the same type, are displayed for both sample sizes and every focal group proportion π considered. The absolute bias and RMSE are small for all parameters, and differences are very small between different values of π and different values of N. The most notable difference is seen for the estimate of the focal group proportion π, where the bias and RMSE clearly decrease when the sample size increases. We also see that the DIF effect parameter is estimated with only a small bias and RMSE. In Table 5, the absolute bias and RMSE are shown for the 3-group setting. As in the two-group setting, the bias and RMSE values are small under all settings.

In summary, the simulation results presented in Tables 1–5, and Figs. 2–4 demonstrate the potential of the proposed framework for DIF analysis with unknown anchor items and comparison groups. The framework performs well in terms of respondent and item classification accuracy

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

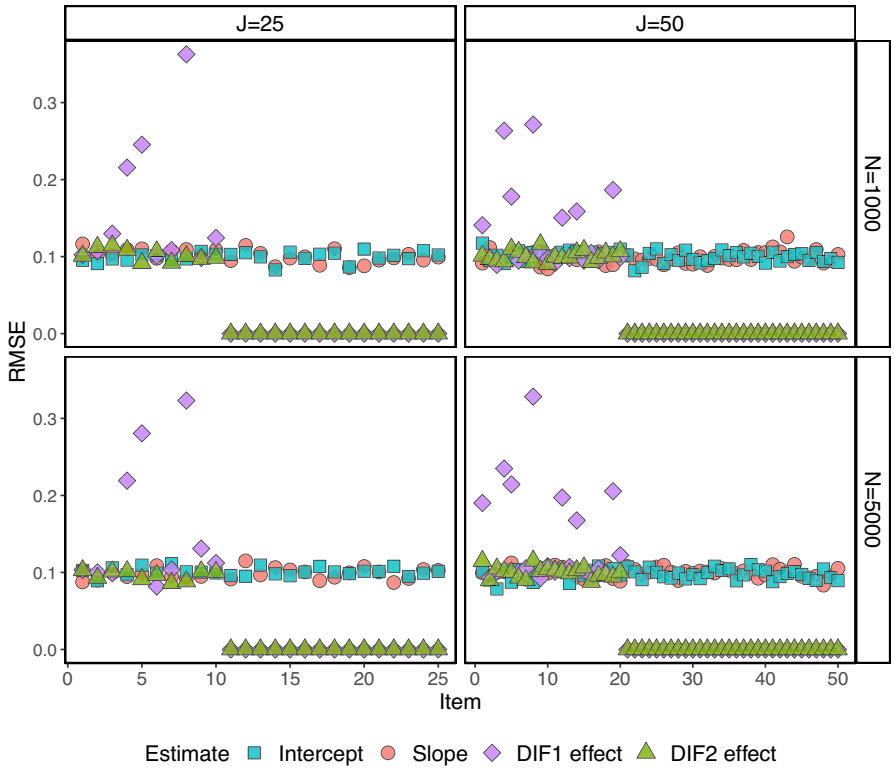


FIGURE 4.
The RMSEs for under the 3-group setting.

across a range of scenarios, and with good parameter recovery, suggesting its applicability in various real-world settings.

4. Real Data Analysis

To illustrate the proposed method, we analyse a mathematics test from a midwestern university in the United States. This data set has been analysed in both Bolt et al. (2002) and De Boeck et al. (2011). The data contains 3000 examinees answering 26 binary-scored items. The original dataset contains two test forms, with 8 items in common.[1] Six of the common items are of particular interest as they are positioned at the end of the test. Bolt et al. (2002) hypothesised the existence of two latent classes: one speeded class that answered end-of-test items with insufficient time, and another non-speeded class. The identification of speeded items was conducted using a two-form design. Specifically, they examined common items across two test administrations, where the common items were placed at the end of the test in one form and earlier in the other form. By estimating the item difficulty for the end-of-test common items and comparing it to the difficulty estimates from the other form, they were able to quantify the DIF effect. Our goal is to detect the DIF items, i.e., the speeded items, and classify respondents into latent classes. Thanks to our procedure, we can analyse only one form, without using information from the second form. A

[1]The original dataset contains 36 items, with 12 common items. We follow the procedure in Bolt et al. (2002) and only analyse items 1-18 and 29-36. In this analysis, items 19-26 therefore actually occupied item position 29-36.

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

TABLE 3.
Average absolute bias and RMSE over all items by estimated parameter type, when $J = 25$, $N = 1000$, and 5000, under the 2-group setting.

		$N = 1,000$		$N = 5,000$	
Scenario	Parameter	Abs. bias	RMSE	Abs. bias	RMSE
$\pi = 0.1$	d	0.078	0.098	0.079	0.099
	a	0.082	0.103	0.081	0.100
	δ	0.040	0.053	0.041	0.053
	μ	0.087	0.114	0.077	0.098
	σ	0.074	0.093	0.088	0.108
	π	0.035	0.040	0.017	0.022
$\pi = 0.2$	d	0.078	0.098	0.081	0.101
	a	0.081	0.101	0.078	0.098
	δ	0.041	0.053	0.041	0.052
	μ	0.081	0.105	0.082	0.107
	σ	0.082	0.101	0.086	0.106
	π	0.050	0.056	0.022	0.027
$\pi = 0.5$	d	0.079	0.100	0.081	0.102
	a	0.082	0.103	0.079	0.099
	δ	0.040	0.052	0.042	0.055
	μ	0.084	0.104	0.076	0.093
	σ	0.075	0.094	0.080	0.101
	π	0.040	0.049	0.030	0.037

similar analysis using simulated data was also conducted in Robitzsch (2022) but using a Rasch mixture model.

We start by fitting the proposed model to the data for different values of K, which determines the number of latent classes. The BIC for a model with $K = 0$, i.e. no latent classes other than the reference group, equals 117,552.2, the BIC for $K = 1$ equals 92,300.1, and for $K = 2$, the BIC equals 92,522.8. We therefore proceed with a model using $K = 1$, i.e., two latent classes. This aligns with the two-group model considered in Bolt et al. (2002). It took the proposed EM algorithm 37.04, 144.760, and 277.390 s to converge[2] for the 1-. 2-, and 3-class solutions, respectively.[3] The proposed model classifies 25.8% of the respondents into the second latent class. If we interpret the two classes as a speeded and non-speeded class, this means that about 26% of the respondents belong to the speeded class. The estimated mean ability in the speeded class equals -0.351 with the estimated standard deviation equal to 1.075. Since the reference group (the non-speeded class) has a prespecified ability mean and standard deviation equal to 0 and 1, respectively, our results therefore indicate that the speeded class has a lower ability on average compared to the non-speeded class. These findings align closely with the results presented in Bolt et al. (2002).

In Table 6 we give the estimated item parameters from the educational test data. The estimated item discrimination and easiness parameters, $\hat{a}$ and $\hat{d}$ respectively, are provided together with the estimated DIF effect $\hat{\delta}$. Common items are denoted by asterisks. For the majority of the items, the DIF effects are estimated to be zero, indicating that these items do not exhibit any significant measurement bias between different groups. However, items 20-26 exhibit non-zero DIF effects,

[2]In our implementation of the proposed EM algorithm, we stop the iterations when the increase in the log-likelihood is smaller than 10^{-4}.

[3]CPU configuration: 11th Gen Intel(R) Core(TM) i7-1165G7 2.80GHz 3200MHz.

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

TABLE 4.
Average absolute bias and RMSE over all items by estimated parameter type, when $J = 50$, $N = 1000$, and 5000 under the 2-group setting.

		$N = 1,000$		$N = 5,000$	
Scenario	Parameter	Abs. bias	RMSE	Abs. bias	RMSE
$\pi = 0.1$	d	0.082	0.102	0.080	0.100
	a	0.078	0.098	0.079	0.100
	δ	0.032	0.041	0.033	0.043
	μ	0.081	0.102	0.079	0.096
	σ	0.078	0.099	0.077	0.096
	π	0.026	0.030	0.013	0.016
$\pi = 0.2$	d	0.081	0.102	0.080	0.100
	a	0.081	0.102	0.079	0.099
	δ	0.034	0.044	0.034	0.044
	μ	0.081	0.100	0.077	0.101
	σ	0.082	0.097	0.084	0.106
	π	0.034	0.038	0.014	0.018
$\pi = 0.5$	d	0.079	0.099	0.081	0.101
	a	0.079	0.098	0.080	0.100
	δ	0.032	0.042	0.032	0.042
	μ	0.091	0.112	0.078	0.097
	σ	0.073	0.093	0.081	0.106
	π	0.030	0.038	0.021	0.025

TABLE 5.
Average absolute bias and RMSE over all items by estimated parameter type, $N = 1000$ and 5000, under the 3-group setting.

		$N = 1,000$		$N = 5,000$	
Scenario	Parameter	Abs. bias	RMSE	Abs. bias	RMSE
$J = 25$	d	0.081	0.101	0.078	0.098
	a	0.080	0.100	0.080	0.101
	δ_2	0.046	0.064	0.045	0.062
	δ_3	0.033	0.041	0.031	0.038
	μ_2	0.084	0.103	0.079	0.095
	μ_3	0.074	0.094	0.073	0.094
	σ_2	0.085	0.107	0.083	0.103
	σ_3	0.081	0.099	0.084	0.103
	π_2	0.060	0.075	0.045	0.057
	π_3	0.045	0.057	0.044	0.054
$J = 50$	d	0.080	0.099	0.079	0.100
	a	0.08	0.101	0.078	0.098
	δ_2	0.038	0.052	0.039	0.057
	δ_3	0.033	0.040	0.032	0.040
	μ_2	0.083	0.100	0.083	0.104
	μ_3	0.072	0.090	0.080	0.097
	σ_2	0.076	0.092	0.085	0.107
	σ_3	0.081	0.103	0.072	0.096
	π_2	0.064	0.079	0.055	0.067
	π_3	0.044	0.056	0.035	0.046

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

TABLE 6.
Estimated item easiness and DIF effects for the detected DIF items.

Item	$\hat{a}$	$\hat{d}$	$\hat{\delta}$	Item	$\hat{a}$	$\hat{d}$	$\hat{\delta}$
1	1.298	2.987	0	14	0.746	0.941	0
2	1.287	1.584	0	15	0.471	−1.646	0
3	0.552	2.968	0	16*	1.507	−0.526	0
4	0.878	0.707	0	17	1.271	1.843	0
5	1.336	2.247	0	18	1.249	−2.588	0
6	0.973	0.822	0	19	1.071	0.944	0
7	0.669	0.296	0	20*	1.017	0.455	−1.411
8	0.687	−0.882	0	21*	1.425	1.324	−1.808
9	1.268	0.908	0	22*	0.664	−0.608	−1.104
10	0.874	2.101	0	23*	0.767	−0.063	−0.624
11*	0.984	−0.575	0	24*	0.929	−0.639	−0.885
12	0.912	−0.804	0	25	1.247	−2.020	−1.019
13	1.040	0.917	0	26*	1.340	−0.079	−0.851

The asterisks denote the common items.

suggesting that these items might be functioning differently for the two latent classes. Among these, items 20, 21, 22, 23, and 24 are also common items, which may require further investigation to ensure fair assessment across test administrations. Since the DIF effect for these end-of-test items is all negative, it suggests that these items become more difficult for the second latent class. This class could therefore consist of respondents that ran out of time and had insufficient time to answer these items. This is known as a speededness effect. As a result, the item difficulty is inflated, which could lead to biased subsequent analyses. The presence of non-zero DIF effects for some end-of-test items highlights the need to scrutinize these items more closely and potentially revise the test administration to minimize the impact of speededness. For instance, increasing the allocated time for the test or redistributing the items more evenly throughout the test could help alleviate the speededness effect and create a more unbiased assessment.

5. Concluding Remarks

In this paper, we presented a comprehensive framework for DIF analysis that overcomes several limitations of existing methods. Our approach can deal with the situation in which both anchor items and comparison groups are unknown, a setting commonly encountered in real-world applications. The use of latent classes in our approach allows us to model heterogeneity among the observations. In this sense, our approach relates to an exploratory dimensionality analysis where there is, in addition to the primary latent dimension, a second dimension that is treated as unknown. In our empirical analysis, this additional dimension is labeled as a speededness effect. In addition to modeling the additional latent dimension(s), the proposed regularised estimator enables us to identify DIF items and quantify their effect on the intercept parameter of the model. We also propose an efficient EM algorithm for the estimation of the model parameters.[4] One merit of our framework is its flexibility. While focusing on the 2-PL model as the baseline model, our approach can be easily extended to accommodate other widely used IRT models, such as the Rasch model and the proportional odds model. We can also allow the baseline model to be a multidimensional IRT model, as shown in Sect. 1.6. Our framework is furthermore able to accommodate more than two comparison groups, allowing DIF effects to vary between the groups. Lastly, the proposed

[4]The R code for the proposed method is available from https://github.com/gabrieltwallin/LatentDIF/

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

method can be extended as shown in Section 2.6 to detect non-uniform DIF. This flexibility makes our framework applicable to a wide range of contexts.

Although our approach shows promising results, there are still several limitations to be addressed in future research. For example, we do not provide confidence intervals for the DIF effect parameters which would be useful for practitioners and researchers in interpreting the magnitude and significance of the DIF effects. In Chen et al. (2023) for example, where the comparison groups are known but the anchor items are unknown, the distribution of $\hat{\delta}_{jk} - \delta_{jk}$ is approximated by Monte Carlo simulation to yield valid statistical inference. This procedure does in essence apply to our case as well. In addition, we have not linked the latent classes to covariates, as in for example Vermunt (2010) and Vermunt and Magidson (2021). By doing so, researchers can gain insights into the underlying characteristics of the different classes and better understand the factors that may contribute to DIF. This would enhance the interpretability of the results and help identify potential sources of DIF that could be addressed in the development of assessment instruments. To address this limitation, future research could explore the integration of covariates within a structural equation modeling (SEM) framework. This would enable the simultaneous modelling of both the measurement model (i.e., the IRT model) and the structural model (i.e., relationships between latent variables and covariates). Incorporating covariates in this manner would not only improve the interpretability of the results but could also provide a more comprehensive understanding of the relationships between the items, latent traits, and potential sources of DIF. Our framework could also be extended to accommodate non-uniform DIF, i.e., DIF in the slope parameter, such as in Wang et al. (2023) that considers a multidimensional IRT model with known comparison groups and unknown anchor items.

In this study, we focus on the Lasso penalty for its simplicity, computational efficiency, and well-documented ability to perform both variable selection and regularisation. The Lasso's convex optimisation problem is easier to solve computationally than some non-convex penalties like the SCAD (Fan & Li, 2001) and the Minimax Concave Penalty (MCP; Zhang, 2010). We acknowledge that the Lasso penalty can introduce some bias into parameter estimates. However, in our proposed method we use the Lasso for model selection. As we thereafter refit the selected model there will be no bias, asymptotically, supposing that the model selection based on the Lasso is consistent (Zhao et al., 2021). Alternative penalties, including the adaptive Lasso (Zou, 2006), SCAD, and MCP, have their own merit. However, they also come with some challenges, especially in terms of computational complexity and algorithm stability. We, therefore, argue that the Lasso penalty is a suitable choice for the proposed model and its identifying assumptions. We believe it would be interesting in the future to compare the performance of estimators with different penalty functions under the current latent DIF setting.

Our proposed framework provides a powerful tool for DIF analysis with unknown anchor items and comparison groups. The framework has the potential to inform the development of fair and unbiased assessments. Future research can build upon our approach by addressing the limitations and exploring other applications. In terms of the potential impact of our work, the framework could be particularly beneficial in specific contexts, such as educational assessment, where identifying and addressing DIF is critical to ensure that tests fairly measure students' abilities across heterogeneous populations, thereby promoting equal access to educational opportunities. It could also be considered in employment selection, where unbiased assessment instruments are crucial to creating a diverse and inclusive workforce that complies with legal requirements related to fairness in employment practices (Ployhart & Holtz, 2008; Hough et al., 2001). Another application is psychological evaluations, where accurate identification of DIF can help improve diagnostic tools and treatment recommendations, leading to better outcomes for individuals from diverse backgrounds (Teresi et al., 2021). By addressing the limitations and further refining our approach, this framework has the potential to contribute to the development of more fair assessment practices in these and other domains, ultimately benefiting a wide range of stakeholders.

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

Funding Gabriel Wallin is supported by the Vetenskapsrådet (Grant No. 2020-06484).

Declarations

Conflict of interest statement The authors have no competing interests to declare that are relevant to the content of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

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

A. Gradients for the Proximal Gradient Descent

In the M-step of the proposed EM-algorithm, we implement a proximal gradient descent. This algorithm requires the gradients of the objective function, which for the proposed model can be expressed as

$$\frac{\partial l(\Delta)}{\partial \eta_j} = \sum_{i=1}^{N} \int \frac{\partial \varphi_{ij}}{\partial \eta_j} \Big[y_{ij} - P(Y_{ij} = 1|\theta_i, \xi_i) \Big] \phi(\theta_i; \mu_k, \sigma_k^2) d\theta_i, \tag{17}$$

where η_j is a generic notation for the item parameters (a_j, d_j, δ_{jk}), and $\varphi_{ij} = d_j + a_j\theta_i + \delta_{j\xi_i}$, To give an example, we give the explicit expressions for the two-group case. We start by parametrising the latent construct of the focal group as $\theta_2 = \mu_2 + \sigma_2\theta_1$, where μ and σ is the mean and standard deviation of the latent construct in the focal group, and $\theta_1 \sim \mathcal{N}(0, 1)$ is the latent construct in the reference group. We have that $\varphi_{ij}^{(1)} = d_j + a_j\theta_{i,1}$ for the reference group and $\varphi_{ij}^{(2)} = d_j + \delta_j + a_j\theta_{i,2}$ for the focal group. The partial derivatives in (17) are given by

$$\begin{aligned}
&\frac{\partial \varphi_{ij}^{(1)}}{\partial d_j} = 1 \quad \frac{\partial \varphi_{ij}^{(2)}}{\partial d_j} = 1 \frac{\partial \varphi_{ij}^{(1)}}{\partial a_j} = \theta_1 \quad \frac{\partial \varphi_{ij}^{(2)}}{\partial a_j} = \mu_2 + \sigma_2\theta_{i,1} \\
&\frac{\partial \varphi_{ij}^{(1)}}{\partial \delta_{j2}} = 0 \quad \frac{\partial \varphi_{ij}^{(2)}}{\partial \delta_{j2}} = 1 \frac{\partial \varphi_{ij}^{(1)}}{\partial \mu_2} = 0 \quad \frac{\partial \varphi_{ij}^{(2)}}{\partial \mu_2} = 1 \frac{\partial \varphi_{ij}^{(1)}}{\partial \sigma_2} = 0 \quad \frac{\partial \varphi_{ij}^{(2)}}{\partial \sigma_2} = \theta_1
\end{aligned} \tag{18}$$

B. The Line Search Procedure

The implemented line search algorithm attempts to find an appropriate step size for the proximal gradient descent implemented in the M-step of the proposed EM algorithm. It does so by iteratively adjusting the step size until the change in the objective function (i.e., the penalised log-likelihood) is within a specified tolerance. The algorithm starts with an initial step size and iteratively reduces it by a factor (in this case, dividing by 2) until the new objective function value satisfies the tolerance condition. If the maximum number of iterations is reached without finding a satisfactory step size,

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

the algorithm returns the current step size. The implemented line search algorithm is given in the following Line Search Algorithm.

Algorithm 3
Line Search Algorithm

Input: $Y, a^{(0)}, d^{(0)}, \delta^{(0)}, \pi^{(0)}, \mu^{(0)}, \sigma^{(0)}, \nabla_a, \nabla_d, \nabla_\delta, \nabla_\mu, \nabla_\sigma, \lambda, \gamma^{(0)} = 1.$

Initialize the objective function

$$O^{(0)} = -l(\Delta^{(0)}) + \lambda \sum_{k=0}^{K} \sum_{j=1}^{J} |\delta_{jk}^{(0)}|$$

Set the tolerance tol to 10^{-6}
Set maximum number of iterations T to 100

For $t = 1, \ldots, T$ **do**

Update the parameters using the current step size:

$a^{(t)} \leftarrow a^{(t-1)} - \gamma^{(t-1)} \nabla_a$
$d^{(t)} \leftarrow d^{(t-1)} - \gamma^{(t-1)} \nabla_d$
$\delta^{(t)} \leftarrow \text{Prox}(\delta^{(t-1)}, \nabla_\delta, \lambda, \gamma^{(t-1)})$
$\sigma^{(t)} \leftarrow \sigma^{(t-1)} - \gamma^{(t-1)} \nabla_\sigma$
$\mu^{(t)} \leftarrow \mu^{(t-1)} - \gamma^{(t-1)} \nabla_\mu$

Calculate new objective function value

$$O^{(t)} \leftarrow -l(\Delta^{(t)}) + \lambda \sum_{k=0}^{K} \sum_{j=1}^{J} |\delta_{jk}^{(t)}|$$

If $|O^{(t)} - O^{(t-1)}| < tol$ **then**

return $\gamma^{(t)}$

else

Adjust the step size and update the value of the objective function:

$\gamma^{(t)} \leftarrow \gamma^{(t-1)}/2$

end if

end for

$\gamma^* \leftarrow \gamma^{(t)}$

return Optimal step size γ^*

C. The Soft-Thresholding Procedure

The soft threshold function and the proximal gradient function are used to identify the DIF-free items, i.e., the anchor items, from the data. The soft threshold function takes a vector x and a scalar λ as inputs and applies element-wise thresholding to x. It sets elements with absolute values less than or equal to λ to zero, subtracts λ from elements greater than λ, and adds λ to elements

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

less than $-\lambda$. The proximal gradient function takes the (estimated) DIF effect δ, its gradient ∇_x, a regularisation parameter λ, and a step size γ as inputs. It calls the soft threshold function with the updated vector $\delta - \gamma\nabla_\delta$ and the product $\lambda\gamma$. The output of the proximal gradient function is the updated DIF effect parameter estimate after applying the soft threshold function. These algorithms are summarised below.

Algorithm 4
Soft Threshold Function

Input: x, λ

Create a temporary variable $temp \leftarrow x$
Set elements of $temp$ to 0 if their absolute value is less than or equal to λ:

$temp[|x| \leq \lambda] \leftarrow 0$

Update elements of $temp$ where $x > \lambda$:

$temp[x > \lambda] \leftarrow temp[x > \lambda] - \lambda$

Update elements of $temp$ where $x < -\lambda$:

$temp[x < -\lambda] \leftarrow temp[x < -\lambda] + \lambda$

Output: $temp$

Algorithm 5
Proximal Gradient Function

Input: δ, ∇_δ, λ, γ

Call the soft threshold function with input $(\delta - \gamma\nabla_\delta, \lambda\gamma)$:

Prox $\leftarrow soft_thre(\delta - \gamma\nabla_\delta, \lambda\gamma)$

Output: Prox

D. The Closed-Form Solution of the Latent Class Proportion

Using the Lagrange multiplier method, we obtain

$$\nu_k^{(t+1)} = \frac{\sum_{i=1}^{n} \gamma_{ik}^{(t)}}{n},$$

where $\gamma_{ik}^{(t)} = P(\xi_i = k|\mathbf{y}_i)$
We can see this from the following argument. Given that

$$D_t(\Delta_1) = -\sum_{i=1}^{N} \mathbb{E}\left[\log\left(\nu_{\xi_i}|\mathbf{Y}_i, \Delta^{(t)}\right)\right].$$

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

The expectation can be expressed as

$$\mathbb{E}\left[\log(\nu_{\xi_i})\middle|\mathbf{Y}_i, \Delta^{(t)}\right] = \sum_{k=0}^{K} P(\xi_i = k\middle|\mathbf{Y}_i, \Delta^{(t)}) \log(\nu_k).$$

Thus, we have

$$D_t(\Delta_1) = -\sum_{i=1}^{N}\sum_{k=0}^{K} \gamma_{ik}^{(t)} \log(\nu_k),$$

where $\gamma_{ik}^{(t)} = P(\xi_i = k\big|\mathbf{Y}_i, \Delta^{(t)})$.
Construct the Lagrangian function with the constraints:

$$\mathcal{L}(\Delta_1, \boldsymbol{\alpha}, \beta) = D_t(\Delta_1) + \sum_{k=0}^{K} \alpha_k(\nu_k - \epsilon_k) + \beta\left(\sum_{k=0}^{K} \nu_k - 1\right),$$

where $\alpha_k \geq 0$ for $k = 0, 1, \ldots, K$.
Now, take the partial derivative of the Lagrangian function with respect to ν_k:

$$\frac{\partial \mathcal{L}}{\partial \nu_k} = -\sum_{i=1}^{n} \gamma_{ik}^{(t)} \frac{1}{\nu_k} + \alpha_k + \beta, \quad k = 0, 1, \ldots, K.$$

Setting the partial derivatives to zero, we get

$$-\sum_{i=1}^{n} \gamma_{ik}^{(t)} \frac{1}{\nu_k} + \alpha_k + \beta = 0, \quad k = 0, 1, \ldots, K.$$

From the above equation, we have:

$$\nu_k^{(t+1)} = \frac{\sum_{i=1}^{n} \gamma_{ik}^{(t)}}{-\alpha_k - \beta}, \quad k = 0, 1, \ldots, K.$$

To find the values of α_k and β, we need to apply the constraint $\sum_{k=0}^{K} \nu_k = 1$. Plugging in the expression for $\nu_k^{(t+1)}$, we have

$$\sum_{k=0}^{K} \frac{\sum_{i=1}^{n} \gamma_{ik}^{(t)}}{-\alpha_k - \beta} = 1.$$

Since the constraint only involves the sum of the $\nu_k^{(t+1)}$, we can eliminate the Lagrange multipliers α_k by normalizing the solution:

$$\nu_k^{(t+1)} = \frac{\sum_{i=1}^{n} \gamma_{ik}^{(t)}}{\sum_{k=0}^{K}\sum_{i=1}^{n} \gamma_{ik}^{(t)}}, \quad k = 0, 1, \ldots, K.$$

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

However, since $\sum_{k=0}^{K} \gamma_{ik}^{(t)} = 1$ for all $i = 1, 2, \ldots, n$, the denominator simplifies to the total number of observations N. Therefore, we have the closed-form solution:

$$\nu_k^{(t+1)} = \frac{\sum_{i=1}^{n} \gamma_{ik}^{(t)}}{n}, \quad k = 0, 1, \ldots, K.$$

References

Bauer, D. J., Belzak, W. C., & 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.

Bechger, T. M., & Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika, 80*(2), 317–340.

Belzak, W., & 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.

Bennink, M., Croon, M. A., Keuning, J., & Vermunt, J. K. (2014). Measuring student ability, classifying schools, and detecting item bias at school level, based on student-level dichotomous items. *Journal of Educational and Behavioral Statistics, 39*(3), 180–202.

Bhattacharya, S., & McNicholas, P. D. (2014). A lasso-penalized bic for mixture model selection. *Advances in Data Analysis and Classification, 8*(1), 45–61.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–472). Addison-Wesley.

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., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*(4), 331–348.

Bouveyron, C., & Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis, 71*, 52–78.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12*(3), 253–260.

Cao, M., Tay, L., & Liu, Y. (2017). A monte carlo study of an iterative wald test procedure for dif analysis. *Educational and Psychological Measurement, 77*(1), 104–118.

Chen, Y., Li, C., Ouyang, J., & Xu, G. (2023). DIF statistical inference and detection without knowing anchoring items [To appear]. *Psychometrika*.

Chen, Y., Lu, Y., & Moustaki, I. (2022). Detection of two-way outliers in multivariate data and application to cheating detection in educational tests. *The Annals of Applied Statistics, 16*(3), 1718–1746.

Cho, S.-J., & Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics, 35*(3), 336–370.

Cho, S.-J., Suh, Y., & Lee, W.-Y. (2016). An NCME instructional module on latent DIF analysis using mixture item response models. *Educational Measurement: Issues and Practice, 35*(1), 48–61.

Cizek, G. J., & Wollack, J. A. (2017). *Handbook of quantitative methods for detecting cheating on tests*. NY: Routledge New York.

Clauser, B., Mazor, K., & Hambleton, R. K. (1993). The effects of purification of matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6*(4), 269–279.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42*(2), 133–148.

De Boeck, P., Cho, S.-J., & Wilson, M. (2011). Explanatory secondary dimension modeling of latent differential item functioning. *Applied Psychological Measurement, 35*(8), 583–603.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1–22.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement, 23*(4), 355–368.

Drabinová, A., & Martinková, P. (2017). Detection of differential item functioning with nonlinear regression: A Non-IRT approach accounting for guessing. *Journal of Educational Measurement, 54*(4), 498–517.

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.

Fidalgo, A., Mellenbergh, G. J., & Muñiz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online, 5*(3), 43–53.

Finch, W. H., & Hernández Finch, M. E. (2013). Investigation of specific learning disability and testing accommodations based differential item functioning using a multilevel multidimensional mixture item response theory model. *Educational and Psychological Measurement, 73*(6), 973–993.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

Holland, P. W., & Thayer, D. T. (1986). Differential item functioning and the Mantel-Haenszel procedure. *ETS Research Report Series, 1986*(2), i–24.
Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Psychology Press.
Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9*(1–2), 152–194.
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*(351a), 631–639.
Kim, S.-H., Cohen, A. S., & Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement, 32*(3), 261–276.
Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement, 75*(1), 22–56.
Kopf, J., Zeileis, A., & Strobl, C. (2015). A framework for anchor methods and an iterative forward approach for DIF detection. *Applied Psychological Measurement, 39*(2), 83–103.
Kuha, J., & Moustaki, I. (2015). Nonequivalence of measurement in latent variable modeling of multigroup data: A sensitivity analysis. *Psychological Methods, 20*(4), 523–536.
Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19–29). Swets & Zeitlinger Publishers.
Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.
Luo, R., Tsai, C.-L., & Wang, H. (2008). On mixture regression shrinkage and selection via the mr-lasso. *International Journal of Pure and Applied Mathematics, 46*, 403–414.
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.
Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. Routledge.
Muthen, B., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics, 10*(2), 133–142.
Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*(4), 557–585.
O'Leary, L. S., & Smith, R. W. (2016). Detecting candidate preknowledge and compromised content using differential person and item functioning. In *Handbook of quantitative methods for detecting cheating on tests* (pp. 151–163). Routledge.
Parikh, N., & Boyd, S. (2014). Proximal algorithms. *Foundations and Trends®in Optimization*, *1*(3), 127–239.
Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61*(1), 153–172.
Rasch, G. (1960). *Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests.* Nielsen & Lydiche.
Redner, R. A., & Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM review, 26*(2), 195–239.
Reeve, B. B., & Teresi, J. A. (2016). Overview to the two-part series: Measurement equivalence of the Patient Reported Outcomes Measurement Information System®(PROMIS®) short forms. *Psychological Test and Assessment Modeling, 58*(1), 31–35.
Robitzsch, A. (2022). Regularized mixture Rasch model. *Information, 13*(11), 534.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.
Schauberger, G., & Mair, P. (2020). A regularization approach for the detection of differential item functioning in generalized partial credit models. *Behavior Research Methods, 52*(1), 279–294.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.
Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica, 7*(2), 221–242.
Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*(2), 159–194.
Steenkamp, J.-B.E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*(1), 78–90.
Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 62*(4), 795–809.
Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361–370.
Tay, L., Huang, Q., & Vermunt, J. K. (2016). Item response theory with covariates (IRT-C) assessing item recovery and differential item functioning for the three-parameter logistic model. *Educational and Psychological Measurement, 76*(1), 22–42.
Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to irt measurement equivalence analysis. *Organizational Research Methods, 18*(1), 3–46.
Teresi, J. A., & Reeve, B. B. (2016). Epilogue to the two-part series: Measurement equivalence of the Patient Reported Outcomes Measurement Information System®(PROMIS®) short forms. *Psychological Test and Assessment Modeling, 58*(2), 423–433.
Teresi, J. A., Wang, C., Kleinman, M., Jones, R. N., & Weiss, D. J. (2021). Differential item functioning analyses of the Patient-Reported Outcomes Measurement Information System (PROMIS®) measures: Methods, challenges, advances, and future directions. *Psychometrika, 86*(3), 674–711.

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147–172). Lawrence Erlbaum Associates Inc.

Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin, 104*(3), 385–395.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288.

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika, 80*(1), 21–43.

van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics, 36*(2), 614–645.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis, 18*(4), 450–469.

Vermunt, J. K., & Magidson, J. (2021). How to perform three-step latent class analysis in the presence of measurement non-invariance or differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal, 28*(3), 356–364.

Von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika, 76*(2), 318–336.

Wainer, H. (2012). An item response theory model for test bias and differential test functioning. In *Differential item functioning* (pp. 202–244). Routledge.

Wang, C., Zhu, R., & Xu, G. (2023). Using lasso and adaptive lasso to identify dif in multidimensional 2pl models. *Multivariate Behavioral Research, 58*(2), 387–407.

Wang, W.-C., Shih, C.-L., & Yang, C.-C. (2009). The mimic method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement, 69*(5), 713–731.

Wang, W.-C., & Su, Y.-H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education, 17*(2), 113–144.

Wang, W.-C., & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27*(6), 479–498.

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., Cai, L., & Wang, M. (2013). The langer-improved wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement, 73*(3), 532–547.

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, K.-H., Liu, H., & Han, Y. (2021). Differential item functioning analysis without a priori information on anchor items: QQ plots and graphical test. *Psychometrika, 86*(2), 345–377.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics, 38*(2), 894–942.

Zhao, P., & Yu, B. (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research, 7*, 2541–2563.

Zhao, S., Witten, D., & Shojaie, A. (2021). In defense of the indefensible: A very naive approach to high-dimensional inference. *Statistical Science, 36*(4), 562–577.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101*(476), 1418–1429.

Zwick, R., Thayer, D. T., & Lewis, C. (2000). Using loss functions for DIF detection: An empirical Bayes approach. *Journal of Educational and Behavioral Statistics, 25*(2), 225–247.

Manuscript Received: 27 APR 2023
Published Online Date: 21 FEB 2024

https://doi.org/10.1007/s11336-024-09948-7 Published online by Cambridge University Press