USING REGULARIZATION TO IDENTIFY MEASUREMENT BIAS ACROSS MULTIPLE BACKGROUND CHARACTERISTICS: A PENALIZED EXPECTATION-MAXIMIZATION ALGORITHM

William C. M. Belzak

duolingo

Daniel J. Bauer

university of north carolina at chapel hill

Testing for differential item functioning (DIF) has undergone rapid statistical developments recently. Moderated nonlinear factor analysis (MNLFA) allows for simultaneous testing of DIF among multiple categorical and continuous covariates (e.g., sex, age, ethnicity, etc.), and regularization has shown promising results for identifying DIF among many covariates. However, computationally inefficient estimation methods have hampered practical use of the regularized MNFLA method. We develop a penalized expectation-maximization (EM) algorithm with soft- and firm-thresholding to more efficiently estimate regularized MNLFA parameters. Simulation and empirical results show that, compared to previous implementations of regularized MNFLA, the penalized EM algorithm is faster, more flexible, and more statistically principled. This method also yields similar recovery of DIF relative to previous implementations, suggesting that regularized DIF detection remains a preferred approach over traditional methods of identifying DIF.

Key words: differential item functioning, measurement invariance, regularization, expectation-maximization, item response theory

Correspondence regarding this paper should be addressed to William Belzak, wbelzak@gmail.com.
This work was mostly completed at the University of North Carolina at Chapel Hill as the first author's dissertation, and was financially supported by a National Defense Science and Engineering Graduate (NDSEG) Fellowship from the U.S. Department of Defense. We thank JR Lockwood for substantive feedback on an earlier draft of this manuscript, Sophie Wodzak for copyedits that improved readability, and James Sharpnack for expertise on LASSO and non-convex regularization properties. We are also grateful for reviewers' comments which considerably improved the quality of the manuscript.

A basic requirement for making valid inferences in the behavioral sciences is that all observations are measured on the same scale. Failure to meet this requirement implies measurement bias/non-invariance is present. Millsap (2011) defines measurement bias as $f(\mathbf{Y}|\boldsymbol{\theta}, \mathbf{X}) \neq f(\mathbf{Y}|\boldsymbol{\theta})$, such that the multivariate distribution of item responses, $f(\mathbf{Y}|\cdot)$ (e.g., questions on a test), depends not only on one or more target variables, $\boldsymbol{\theta}$ (e.g., latent ability)[1], but also on one or more non-target variables, $\mathbf{X}$ (e.g., sex, ethnicity, age). Measurement bias implies that differential item functioning (DIF) is present in one or more items, defined as $f(\mathbf{y}_j|\boldsymbol{\theta}, \mathbf{X}) \neq f(\mathbf{y}_j|\boldsymbol{\theta})$ for item j.

DIF is often evaluated with multiple-group item response theory (MG-IRT) or confirmatory factor analysis (MG-CFA) by testing whether item parameters significantly vary across a discrete background characteristic (e.g., sex) through the use of likelihood ratio, score, or Wald tests. In contrast, DIF can be evaluated across multiple categorical and continuous characteristics simultaneously with moderated nonlinear factor analysis (MNLFA; Bauer, 2017; Bauer & Hussong, 2009), providing researchers with more flexibility in explaining why DIF occurs.

Despite more flexible modeling capabilities, MNLFA can be challenging to use. Some challenges include identifying DIF among many covariates, as well as efficiently estimating large numbers of model parameters. Recent work on the former has shown promising results for the use of regularization to identify DIF (Bauer, Belzak, & Cole, 2020; Belzak & Bauer, 2020; Chen, Li, & Xu, 2021; Huang, 2018; Liang & Jacobucci, 2020; Magis, Tuerlinckx, & De Boeck, 2015; Robitzsch, 2020; Schauberger & Mair, 2020; Tutz & Schauberger, 2015). The latter problem is only beginning to be addressed (Chen, Li, & Xu, 2021; Wang, Zhu, & Xu, 2022).

Parameter estimation of regularized MNLFA models is challenging because there can be many parameters to estimate (e.g., intercept and slope DIF effects), a variety of background variable types and item response functions to model (e.g., continuous and categorical), and different penalty functions to incorporate (e.g., soft-thresholding and firm-thresholding). Typically, MNLFA models are estimated by directly maximizing the observed data (or marginal) log-likelihood function (Bauer & Hussong, 2009; Schauberger & Mair, 2020).[2] Of course, this approach becomes impractical with more than 10 items or so (Bock & Aitkin, 1981; Bock & Lieberman, 1970) as all model parameters must be optimized simultaneously.

More recently, researchers of regularized IRT models (Sun et al., 2016) and regularized DIF analysis (Belzak, 2023; Huang, 2018; Wang et al., 2022) have adapted the expectation-maximization (EM) algorithm to estimate parameters in penalized latent variable models. This paper extends this work by developing a penalized EM algorithm that allows for the evaluation of DIF across multiple continuous and categorical predictors simultaneously for a variety of 2-parameter IRT models. This algorithm incorporates elements of Bock and Aitken's (1981) EM algorithm (Dempster, Laird, & Rubin, 1977) with coordinate descent to perform

[1]Target variables may not be "latent" or modeled using an latent variable model (e.g., item response theory). The Maentel-Haenszel method (Holland & Thayer, 1986) is one such non-latent-variable approach to evaluating measurement invariance.

[2]MNFLA models may also be fit using an EM algorithm in M*plus* software (Muthén & Muthén, 2017). However, this program does not currently allow for the inclusion of penalties into the likelihood function to perform regularization of DIF.

soft-thresholding (i.e., Least Absolute Shrinkage and Selection Operator or LASSO); Donoho, 1994; Tibshirani, 1996) and firm-thresholding (i.e., Minimax Concave Penalty or MCP; Breheny & Huang, 2011; Friedman, Hastie, & Tibshirani, 2010; Zhang, 2010), making it more flexible, computationally faster, and more statistically principled than previous implementations of regularized MNLFA (Bauer et al., 2020). This method also yields similar recovery of DIF relative to observed/marginal maximum likelihood estimation. The regDIF R package (Belzak, 2023) has also been developed so that this method can be easily used by other researchers.

In what follows, we (1) provide background on the MNLFA model; (2) show how regularization is used to identify DIF; (3) give mathematical details on the penalized EM algorithm; (4-5) evaluate the accuracy of the penalized EM algorithm; (6) use the algorithm with empirical data; and finally, (7) discuss the implications of our results.

1. Model Specification

The MNFLA model parallels the generalized linear model framework (McCullagh & Nelder, 1989) in its use of a linear predictor and link function to relate the expected value of each item response to the latent variable and background characteristics (Bauer & Hussong, 2009). First, the linear predictor, η_{ij}, for person i and item j is defined

$$\eta_{ij} = \nu_{ij} + \lambda_{ij}\theta_i, \tag{1}$$

where person i's latent score, θ_i, influences the linear predictor through the intercept, ν_{ij}, and slope, λ_{ij}, both of which are decomposed as linear combinations of the background characteristics:

$$\begin{aligned} \nu_{ij} &= \nu_{0j} + \mathbf{x}_i^T \boldsymbol{\nu}_{1j}, \\ \lambda_{ij} &= \lambda_{0j} + \mathbf{x}_i^T \boldsymbol{\lambda}_{1j}. \end{aligned} \tag{2}$$

The baseline intercept and slope, ν_{0j} and λ_{0j}, represent the intercept and slope of item j when all background characteristics equal 0. $\boldsymbol{\nu}_{1j}$ and $\boldsymbol{\lambda}_{1j}$ (both $K \times 1$ vectors, with K equal to the number of predictors) represent the effects of person i's background characteristics, $\mathbf{x}_i$, on the item intercept and slope, respectively. In other words, $\boldsymbol{\nu}_{1j}$ and $\boldsymbol{\lambda}_{1j}$ represent the intercept and slope DIF effects of focus here.[3] In practice, it is usually desirable that many items are free of DIF, such that $\boldsymbol{\nu}_{1j}$ and $\boldsymbol{\lambda}_{1j}$ are sparse with many zero elements across the set of items (Millsap, 2011; Chen et al., 2023). Some DIF parameters are typically fixed to identify the MNLFA model. One minimally-identified model requires $\boldsymbol{\nu}_{1j}$ and $\boldsymbol{\lambda}_{1j}$ to be constrained to 0 for each background variable on at least one item.[4] In other words, one or more items for each background variable must be assumed free of DIF, or "anchored" across individuals.

[3] In the classical case of two groups, ν_{1j} and λ_{1j} are scalar values and represent group differences between the intercept and slope (i.e., group DIF).

[4] The null effects can be distributed across items – e.g., background variable x_1 has a null effect on item 1, whereas background variable x_2 has a null effect on item 2, and so on.

Next, to estimate differences between the latent scores as a function of the background characteristics, we assume the latent variable is distributed $\theta_i \sim \mathcal{N}(\alpha_i, \psi_i)$, with the latent mean, α_i, and variance, ψ_i, for person i given by

$$\begin{aligned} \alpha_i &= \alpha_0 + \mathbf{x}_i^T \boldsymbol{\alpha}_1, \\ \psi_i &= \psi_0 \ \exp(\mathbf{x}_i^T \boldsymbol{\psi}_1). \end{aligned} \tag{3}$$

The baseline mean and variance, α_0 and ψ_0, represent the latent mean and variance when all background characteristics are scored 0. For identification purposes, α_0 and ψ_0 are also typically set to equal 0 and 1, respectively.[5] In contrast, $\boldsymbol{\alpha}_1$ and $\boldsymbol{\psi}_1$ are $K \times 1$ vectors of coefficients that account for differences in the latent mean and variance due to the background characteristics, otherwise referred to as impact.

Finally, a link function transforms the values of the linear predictor into expected values for the item responses. In the case of a binary (e.g., yes/no) response, we use a logistic link function:

$$\mu_{ij} = P(y_{ij} = 1 | \theta_i, \mathbf{x}_i) = \frac{1}{1 + \exp(-\eta_{ij})}, \tag{4}$$

such that μ_{ij} is the probability of person i endorsing or correctly answering item j, which is distributed $y_{ij} \sim \text{Bernoulli}(\mu_{ij})$. In the case of an ordinal response, we can instead use a graded response model (Samejima, 1969), and for a continuous response, a normal distribution. The latter case equates to a linear factor analysis model. We focus on binary responses in this paper, but point researchers to work that implements other response functions for Reg-DIF (Belzak, 2023; Schauberger & Mair, 2020).

1.1. Observed/Marginal Log-Likelihood

Previous implementations of MNLFA (Bauer, 2017; Bauer et al., 2020) obtained parameter estimates by maximizing the observed/marginal log-likelihood function:

$$\log L_o(\boldsymbol{\gamma} | \mathbf{y}_i, \mathbf{x}_i) = \log \left(\prod_{i=1}^{N} \int \phi(\theta_i | \mathbf{x}_i; \boldsymbol{\xi}) \prod_{j=1}^{J} f(y_{ij} | \theta_i, \mathbf{x}_i; \boldsymbol{\varpi}_j) d\theta_i \right), \tag{5}$$

where the vector of model parameters is denoted $\boldsymbol{\gamma} = \text{vec}[\boldsymbol{\xi}, \boldsymbol{\varpi}]$, or a combination of the latent variable parameters, $\boldsymbol{\xi} = \text{vec}[\alpha_0, \psi_0, \boldsymbol{\alpha}_1, \boldsymbol{\psi}_1]$, and the item response parameters, $\boldsymbol{\varpi} = [\boldsymbol{\varpi}_1, ..., \boldsymbol{\varpi}_J]$, with $\boldsymbol{\varpi}_J = \text{vec}[\nu_{0j}, \lambda_{0j}, \boldsymbol{\nu}_{1j}, \boldsymbol{\lambda}_{1j}]$. $\boldsymbol{\xi}$ governs the conditional normal distribution of the latent variable, written

$$\phi(\theta_i | \mathbf{x}_i; \boldsymbol{\xi}) = \frac{1}{\sqrt{2\pi\psi_i}} \exp\left(-\frac{[\theta_i - \alpha_i]^2}{2\psi_i} \right), \tag{6}$$

[5]These model constraints are similar to common MG-IRT and MG-CFA constraints, where the latent scale is often identified by fixing the latent mean and variance for one group to equal 0 and 1, respectively, while the latent mean(s) and variance(s) of the other group(s) are estimated.

and $\boldsymbol{\varpi}_j$ governs the conditional item response function for the observed outcome, defined

$$f(y_{ij}|\theta_i, \mathbf{x}_i; \boldsymbol{\varpi}_j) = \mu_{ij}^{y_{ij}}(1-\mu_{ij})^{1-y_{ij}} \tag{7}$$

for a binary item response. The integral in the observed/marginal log-likelihood of Eq. (5) is often approximated using numerical quadrature, where a finite number of grid points over the range of θ_i are weighted by the expected density of θ_i at each point and summed. A root-finding method (e.g., quasi-Newton) is then used to obtain the maximum likelihood estimates for all model parameters.

2. Regularized Differential Item Functioning

Evaluating DIF in the MNLFA model can be an error-prone, cumbersome process when using an iterative testing approach, such as using all items as anchors except for the item being tested (Bauer, 2017; Kopf, Zeileis, & Strobl, 2015; Thissen, Wainer, & Steinberg, 1993). Specifically, when sample sizes are large and DIF is pervasive in the data-generating model, FP rates tend to be unacceptably high, even after controlling for multiple testing. As the set of anchor items becomes increasingly contaminated with DIF, larger sample sizes often exacerbate the effects of DIF contamination (Belzak & Bauer, 2020).

Methods of mitigating contamination exist, including identifying anchor items through preliminary DIF testing (also known as purification or DIF-free-then-DIF; Wang, Shih, & Sun, 2012) or selecting anchor items according to rankings of DIF statistics (Woods, 2009). However, researchers may find it difficult to choose a DIF detection method because the advantages of each approach can depend on the level of DIF contamination (Kopf et al., 2015), which is usually not known in advance. Two-step, iterative anchor selection approaches may also place a non-trivial burden on the researcher, insofar that each item must be separately evaluated for intercept and slope DIF across all background characteristics (Bauer, 2017).

LASSO regularization (Tibshirani, 1996) is an alternative approach adapted to DIF testing which can reduce FP rates and simplify the evaluation of DIF among multiple covariates (Bauer et al., 2020; Magis, Tuerlinckx, & De Boeck, 2015; Tutz & Schauberger, 2015). Regularized differential item functioning (Reg-DIF) works by appending a penalty function onto the observed/marginal log-likelihood function, written

$$\log L_r(\boldsymbol{\gamma}) = \log L_o(\boldsymbol{\gamma}) - \tau(\|\boldsymbol{\nu}_1\|_1 + \|\boldsymbol{\lambda}_1\|_1), \tag{8}$$

where $\log L_r(\boldsymbol{\gamma})$ is the regularized log-likelihood function to be maximized; $\log L_o(\boldsymbol{\gamma}|\mathbf{y}_i; \mathbf{x}_i)$ is the observed log-likelihood shown in Eq. (5); and $\|\cdot\|$ is the L_1 norm, which sums the absolute values of the intercept and slope DIF parameters for all items, namely, $\boldsymbol{\nu}_1$ and $\boldsymbol{\lambda}_1$. Of course, it is not necessary to penalize intercept and slope DIF effects equally; it could even be helpful to account for potential differences in scale between intercept DIF and slope DIF.[6] However, using two penalty parameters may be computationally expensive, as one would then need to conduct a

[6] An extension of Eq. (8) could be to multiply the slope DIF effects by a fixed scalar which depends on τ: e.g., $\tau(\|\boldsymbol{\nu}_1\|_1 + \delta\|\boldsymbol{\lambda}_1\|_1)$, with $\delta > 0$.

two-dimensional grid search to identify the optimal pair of penalty values. For simplicity and computational efficiency, the present work focuses on the single penalty case.

A key advantage of using LASSO for DIF identification is that no ex-ante anchor items need to be explicitly specified by a researcher. This is possible because LASSO constrains the values of the DIF parameters by the magnitude of the tuning parameter, τ, and as τ gradually increases, some DIF parameters shrink towards zero because they become too costly to maintain in the log-likelihood function. In the limit as τ grows large, all DIF parameters shrink to and become zero. Conversely, the regularized log-likelihood simplifies to the MNLFA log-likelihood when $\tau = 0$. LASSO can be considered a continuous form of anchor item selection (Belzak & Bauer, 2020; Liang & Jacobucci, 2020), whereas traditional methods like likelihood ratio testing require categorical specifications of anchor items (e.g., items are anchor items or not). Note that while a researcher need not specify anchor items explicitly, the LASSO approach still requires some anchor items to be automatically specified for the model to be identified.

The main task of performing Reg-DIF is to vary the magnitude of τ across a range of non-negative values and choose the optimal model via some selection criteria. Information criteria and cross-validation may be used to identify the optimal degree of penalization (Hastie, Tibshirani, & Friedman, 2017). The Bayesian Information Criterion (BIC) is one effective approach for selecting τ because of its computational ease and relatively good balance of true and false positives in recovering DIF (Magis et al., 2015).

2.1. Previous Research

Research on regularized DIF analysis (primarily LASSO regularization) has grown rapidly in the last decade. One pervasive finding is that LASSO tends to exhibit fewer false positive DIF effects compared to more traditional methods of DIF testing, while also exhibiting fewer true positives in some cases (Bauer, Belzak, & Cole, 2020; Belzak & Bauer, 2020; Huang, 2018; Liang & Jacobucci, 2020; Magis, Tuerlinckx, & De Boeck, 2015; Tutz & Schauberger, 2015). LASSO may thus be more conservative than other methods (e.g., likelihood ratio tests, Thissen et al., 1993), although this can depend on sample size and proportion of DIF in the measurement scale, among other data conditions.

Research on regularized DIF estimation within latent variable models has also grown, albeit at a slower pace. Different estimation methods developed for regularized DIF analysis include conditional maximum likelihood (Tutz & Schauberger, 2015; Magis et al., 2015), observed/marginal maximum likelihood (Bauer, Belzak, & Cole, 2020; Belzak & Bauer, 2020; Chen et al., 2021; Schauberger & Mair, 2020), and the expectation-maximization algorithm (Belzak, 2021; Huang, 2018; Wang et al., 2022), each of which have advantages and disadvantages.

Conditional maximum likelihood (CML) performs model estimation quickly but only produces consistent estimates for latent variable models where the item slopes are equal (e.g., Rasch models; Rasch, 1961; Andersen, 1970). Marginal/observed maximum likelihood (OML) produces consistent estimates for slope-varying models, although performs model estimation relatively slowly as the number of items grows beyond 10 or so (Bock & Lieberman, 1970). The expectation-maximization (EM) algorithm produces consistent estimates for slope-varying models, like OML, yet it fits models more quickly than OML with larger numbers of model

parameters (Bock & Aitken, 1981).

We focus on the EM algorithm here given its advantages over CML and OML and because recent work has demonstrated it to be a tractable approach for regularized DIF estimation. Namely, Huang (2018) developed a penalized EM algorithm within a limited-information estimation framework (e.g., see Bollen, 1989), and Wang et al. (2022) developed a penalized EM algorithm within a multidimensional 2-parameter logistic IRT framework. These estimation approaches were implemented for multiple-group DIF analyses.

This paper extends this line of estimation work by developing a penalized EM algorithm where multiple categorical and continuous covariates can be evaluated for DIF simultaneously, different item response functions can be modeled together, and a variety of penalty functions can be incorporated into the likelihood function, thereby expanding regularized DIF analysis to the full MNLFA model. In the next section, we describe the mathematical details of this algorithm.

3. Penalized Expectation-Maximization Algorithm

Similar to Bock and Aitken's (1981) approach, we formulate the MNLFA model into a missing data problem by defining the complete data log-likelihood function as

$$\log L_c(\boldsymbol{\gamma}|\mathbf{y}_i, \mathbf{x}_i) = \log \left(\prod_{i=1}^{N} \phi\left(\theta_i|\mathbf{x}_i; \boldsymbol{\xi}\right) \prod_{j=1}^{J} f\left(y_{ij}|\theta_i, \mathbf{x}_i; \boldsymbol{\varpi}_j\right) \right), \tag{9}$$

The complete data log-function is a combination of the missing data or latent scores, θ_i, and the observed data or item responses and background characteristics, y_{ij} and $\mathbf{x}_i$. Because the latent scores are missing from the analysis, maximizing $\log L_c\left(\boldsymbol{\gamma}|\mathbf{y}_i, \mathbf{x}_i\right)$ is not possible. Instead, the EM algorithm makes use of the *expected value* of the complete data log-likelihood, a surrogate function, to simplify the estimation problem.

3.1. Expectation Step

The expected value of the complete data log-likelihood is written

$$\begin{aligned} Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)}) &= E\left[\log L_c\left(\boldsymbol{\gamma}|\mathbf{y}_i, \mathbf{x}_i\right)|\boldsymbol{\gamma}^{(t)}, \mathbf{y}_i, \mathbf{x}_i\right], \\ &= \int \left[\sum_{i=1}^{N} \log\left[\phi(\theta_i|\mathbf{x}_i; \boldsymbol{\xi})\right] P(\theta_i|\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\gamma}^{(t)}) + \right. \\ &\qquad \left. \sum_{i=1}^{N} \sum_{j=1}^{J} \log\left[f(y_{ij}|\theta_i, \mathbf{x}_i; \boldsymbol{\varpi}_j)\right] P(\theta_i|\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\gamma}^{(t)}) \right] d\theta_i, \end{aligned} \tag{10}$$

where $Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)})$ is the surrogate function to maximize; $\boldsymbol{\gamma}^{(t)}$ is the vector of provisional model estimates at iteration t; and $P(\theta_i|\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\gamma}^{(t)})$ is the posterior distribution of the latent scores, θ_i.

Bayes' theorem specifies that the posterior distribution of θ_i is defined

$$\begin{aligned} P(\theta_i|\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\gamma}^{(t)}) &= \frac{P(y_{ij}, \theta_i; \boldsymbol{\gamma}^{(t)})}{P(y_{ij}; \boldsymbol{\gamma}^{(t)})}, \\ &= \frac{\prod_{j=1}^{J} f(y_{ij}|\theta_i, \mathbf{x}_i; \boldsymbol{\varpi}_j^{(t)}) \phi(\theta_i|\mathbf{x}_i; \boldsymbol{\xi}^{(t)})}{\int \prod_{j=1}^{J} f(y_{ij}|\theta_i, \mathbf{x}_i; \boldsymbol{\varpi}_j^{(t)}) \phi(\theta_i|\mathbf{x}_i; \boldsymbol{\xi}^{(t)}) d\theta_i}, \end{aligned} \tag{11}$$

where $P(y_{ij}, \theta_i; \boldsymbol{\gamma}^{(t)})$ is the joint distribution of y_{ij} and θ_i, and $P(y_{ij}; \boldsymbol{\gamma}^{(t)})$ is the marginal distribution of y_{ij}. Quadrature points are used in place of θ_i to compute the posterior density:

$$P(Z_q|\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\gamma}^{(t)}) = \frac{\prod_{j=1}^{J} f(y_{ij}|Z_q, \mathbf{x}_i; \boldsymbol{\varpi}_j^{(t)}) \phi(Z_q|\mathbf{x}_i; \boldsymbol{\xi}^{(t)})}{\sum_{q=1}^{Q} \prod_{j=1}^{J} f(y_{ij}|Z_q, \mathbf{x}_i; \boldsymbol{\varpi}_j^{(t)}) \phi(Z_q|\mathbf{x}_i; \boldsymbol{\xi}^{(t)})}, \tag{12}$$

where Z_q is a real number in the range of the probability density of θ_i (e.g., equally spaced points between -6 and 6). This redefines the surrogate function from Eq. (11) as

$$\begin{aligned} Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)}) = &\sum_{q=1}^{Q} \sum_{i=1}^{N} \log\left[\phi(Z_q|\mathbf{x}_i; \boldsymbol{\xi})\right] P(Z_q|\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\gamma}^{(t)}) + \\ &\sum_{q=1}^{Q} \sum_{i=1}^{N} \sum_{j=1}^{J} \log\left[f(y_{ij}|Z_q, \mathbf{x}_i; \boldsymbol{\varpi}_j)\right] P(Z_q|\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\gamma}^{(t)}), \end{aligned} \tag{13}$$

where the first set of summations represents the contribution of the latent variable distribution, and the second set of summations represents the contributions of the J item response distributions. We show that this function is guaranteed to be concave for all values of $\boldsymbol{\gamma}^{(t)}$ in Appendix A, such that that a set of reasonable starting values will always converge to the set of maximum likelihood estimates under mild regularity conditions (Boyd & Vanderberge, 2004).

Eq. (13) not only shows that we can estimate the latent variable parameters separately from the item parameters, but more importantly that we may also perform maximization independently for each item. That is, we can decompose $Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)})$ as

$$Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)}) = Q(\boldsymbol{\xi}_\theta|\boldsymbol{\gamma}^{(t)}) + \sum_{j=1}^{J} Q(\boldsymbol{\varpi}_j|\boldsymbol{\gamma}^{(t)}), \tag{14}$$

where $\boldsymbol{\xi}_\theta$ is a vector of latent variable parameters, $\boldsymbol{\varpi}_j$ is a vector of parameters for item j, and

$$Q(\boldsymbol{\xi}_\theta|\boldsymbol{\gamma}^{(t)}) = \sum_{q=1}^{Q} \sum_{i=1}^{N} \log\left[\phi(Z_q|\mathbf{x}_i; \boldsymbol{\xi}_\theta)\right] P(Z_q|\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\gamma}^{(t)}), \tag{15}$$

$$Q(\boldsymbol{\varpi}_j|\boldsymbol{\gamma}^{(t)}) = \sum_{q=1}^{Q} \sum_{i=1}^{N} \log\left[f(y_{ij}|Z_q, \mathbf{x}_i; \boldsymbol{\varpi}_j)\right] P(Z_q|\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\gamma}^{(t)}). \tag{16}$$

The main goal of the Expectation step is to compute the posterior density from Eq. (11) using provisional parameter estimates, which either come from reasonable starting values at iteration $t = 0$ or from the previous Maximization step. These posterior values are then used in the next Maximization step.

3.2. Maximization Step

With the posterior values obtained from the Expectation step, we maximize $Q(\boldsymbol{\xi}_\theta|\boldsymbol{\gamma}^{(t)})$ and $Q(\boldsymbol{\varpi}_j|\boldsymbol{\gamma}^{(t)})$ for all j by employing second-order Taylor series expansions of the respective log-likelihood components (Fan & Li, 2001; Sun, Chen, Liu, Ying, & Xin, 2016). First, for a single latent variable parameter denoted ξ_θ, this is defined

$$Q(\xi_\theta|\boldsymbol{\gamma}^{(t)}) = Q(\xi_\theta^{(t)}|\boldsymbol{\gamma}^{(t)}) + Q'(\xi_\theta^{(t)}|\boldsymbol{\gamma}^{(t)})(\xi_\theta - \xi_\theta^{(t)}) + \frac{1}{2}Q''(\xi_\theta^{(t)}|\boldsymbol{\gamma}^{(t)})(\xi_\theta - \xi_\theta^{(t)})^2. \tag{17}$$

We then take the first partial derivative of the quadratic approximation with respect to ξ_θ and set it equal to zero:

$$\frac{\partial Q(\xi_\theta|\boldsymbol{\gamma}^{(t)})}{\partial \xi_\theta} = Q'(\xi_\theta^{(t)}|\boldsymbol{\gamma}^{(t)}) + Q''(\xi_\theta^{(t)}|\boldsymbol{\gamma}^{(t)})(\xi_\theta - \xi_\theta^{(t)}) = 0, \tag{18}$$

where the terms can be rearranged to obtain a Newton-Raphson(-like[7]) update, written

$$\xi_\theta = \xi_\theta^{(t)} - \frac{Q'(\xi_\theta^{(t)}|\boldsymbol{\gamma}^{(t)})}{Q''(\xi_\theta^{(t)}|\boldsymbol{\gamma}^{(t)})}. \tag{19}$$

From Eq (19), we employ coordinate descent (Friedman, Hastie, & Tibshirani, 2010) to maximize the log-likelihood function in Eq. (13). Analytical forms of $Q'(\xi_\theta^{(t)}|\boldsymbol{\gamma}^{(t)})$ and $Q''(\xi_\theta^{(t)}|\boldsymbol{\gamma}^{(t)})$ are provided in Appendix B.

Second, for a single item parameter denoted ϖ_j, we define a quadratic approximation as well, written

$$Q(\varpi_j|\boldsymbol{\gamma}^{(t)}) = Q(\varpi_j^{(t)}|\boldsymbol{\gamma}^{(t)}) + Q'(\varpi_j^{(t)}|\boldsymbol{\gamma}^{(t)})(\varpi_j - \varpi_j^{(t)}) + \frac{1}{2}Q''(\varpi_j^{(t)}|\boldsymbol{\gamma}^{(t)})(\varpi_j - \varpi_j^{(t)}). \tag{20}$$

However, before taking the first derivative of $Q(\varpi_j|\boldsymbol{\gamma}^{(t)})$, we approximate the penalty function and append it to the quadratic approximation above. Zou and Li (2008) show that a linear approximation of the penalty function has better theoretical results compared to a quadratic approximation; therefore, we approximate the penalty function linearly:

$$P(\varpi_j) = P(\varpi_j^{(t)}) + P'(\varpi_j^{(t)})(\varpi_j - \varpi_j^{(t)}), \tag{21}$$

where $P(\varpi_j)$ represents a pre-specified penalty function applied to a DIF effect for item j. This step is repeated for each intercept and slope DIF effect. Note that only the intercept and slope DIF effects are penalized; the baseline item parameters are not.

[7]Using a univariate optimization approach (i.e., coordinate descent) on a multivariate optimization problem is not a true Newton-Raphson method. However, our approach follows a familiar Newton-Raphson form.

We then subtract the linearly-approximated penalty function from Eq. (20) and take the first partial derivative of the regularized function with respect to the item parameter:

$$\frac{\partial Q(\varpi_j|\boldsymbol{\gamma}^{(t)})_r}{\partial \varpi_j} = Q'(\varpi_j^{(t)}|\boldsymbol{\gamma}^{(t)}) + Q''(\varpi_j^{(t)}|\boldsymbol{\gamma}^{(t)})(\varpi_j - \varpi_j^{(t)}) - P'(\varpi_j^{(t)}), \tag{22}$$

such that $Q(\varpi_j|\boldsymbol{\gamma}^{(t)})_r$ denotes the regularized quadratic approximation for item component j. Note that depending on which penalty function, $P(\varpi_j)$, is used for regularization, $P(\varpi_j)$ may not be differentiable at all values of ϖ_j. In fact, the penalties evaluated here (i.e., LASSO and MCP) are not when $\varpi_j = 0$, and thus we must use the subgradient method on $P(\varpi_j)$, as shown in the next section.

Finally, we set Eq. (22) to zero and rearrange to obtain updated item estimates:

$$\begin{aligned} \varpi_j &= \varpi_j^{(t)} - \frac{Q'(\varpi_j^{(t)}|\boldsymbol{\gamma}^{(t)})}{Q''(\varpi_j^{(t)}|\boldsymbol{\gamma}^{(t)})} + \frac{P'(\varpi_j^{(t)})}{Q''(\varpi_j^{(t)}|\boldsymbol{\gamma}^{(t)})}, \\ &= \widetilde{\varpi}_j^{(t)} + \frac{P'(\varpi_j^{(t)})}{Q''(\varpi_j^{(t)}|\boldsymbol{\gamma}^{(t)})}, \end{aligned} \tag{23}$$

where $\widetilde{\varpi}_j^{(t)}$ is the updated estimate before the penalty function has been applied. Eq. (23) shows that the first partial derivative of the penalty function, $P'(\varpi_j^{(t)})$, is scaled by the second partial derivative of the log-likelihood, $Q''(\varpi_j^{(t)}|\boldsymbol{\gamma}^{(t)})$ or Q_j''. This scaling factor implies that the variance of a DIF effect influences the strength of the penalty: larger variance or less information (smaller Q_j'') leads to greater penalization and vice versa for smaller variance or more information (larger Q_j''). In other words, when a DIF effect has a small amount of variance relative to other effects, the penalty function will overwhelm the DIF effect and set it to 0. Conversely, when a DIF effect has a large amount of variance relative to other effects, the penalty function will have little influence on the DIF effect.[8]

3.3. Penalization

We examine two penalty functions in this paper: LASSO and MCP.

3.3.1. LASSO

The LASSO penalty and its subgradient are defined

$$\begin{aligned} P(\varpi_j;\tau) &= \tau|\varpi_j|, \\ P'(\varpi_j;\tau) &= \tau \text{ sign}(\varpi_j). \end{aligned} \tag{24}$$

[8]The effect of a parameter's variance on the magnitude of penalization occurs in relation to other effects. If all DIF effects have equally small or large variance, this will simply scale the tuning parameter to be relatively small or large, respectively.

Inserting the LASSO subgradient into Eq. (23), the intercept DIF estimate shrinks towards zero by a magnitude of τ/Q''_j, either from the left (if $\varpi_j < 0$) or the right (if $\varpi_j > 0$). We solve for each DIF effect using the soft-thresholding operator:

$$\begin{aligned}\frac{S(\widetilde{\varpi}_j, \tau)}{Q''_j} &= \text{sign}(\widetilde{\varpi}_j)(|\widetilde{\varpi}_j| - \tau/Q''_j)_+, \\ &= \begin{cases} \widetilde{\varpi}_j - \tau/Q''_j, & \text{if } \widetilde{\varpi}_j > 0 \text{ and } \tau/Q''_j < |\widetilde{\varpi}_j|; \\ \widetilde{\varpi}_j + \tau/Q''_j, & \text{if } \widetilde{\varpi}_j < 0 \text{ and } \tau/Q''_j < |\widetilde{\varpi}_j|; \\ 0, & \text{if } \tau/Q''_j \geq |\widetilde{\varpi}_j|, \end{cases}\end{aligned} \tag{25}$$

such that the parameter update in Eq. (23) is set to $\varpi_j^{(t+1)} \leftarrow \frac{S(\widetilde{\varpi}_j, \tau)}{Q''_j}$.

3.3.2. MCP

The minimax concave penalty (MCP; Zhang, 2010) uses a second tuning parameter, γ, to de-bias the penalized DIF effects. The MCP function and its subgradient are defined

$$\begin{aligned} P(\varpi_j; \tau, \gamma) &= \begin{cases} \tau|\varpi_j| - \frac{\varpi_j^2}{2\gamma}, & \text{if } |\varpi_j| \leq \gamma\tau; \\ \frac{1}{2}\gamma\tau^2, & \text{if } |\varpi_j| > \gamma\tau, \end{cases} \\ \nabla p(\varpi_j; \tau, \gamma) &= \begin{cases} \tau - \frac{|\varpi_j|}{\gamma} \text{ sign}(\varpi_j), & \text{if } |\varpi_j| \leq \gamma\tau; \\ 0, & \text{if } |\varpi_j| > \gamma\tau, \end{cases} \end{aligned} \tag{26}$$

assuming $\gamma > 1$. Inserting the MCP subgradient into Eq. (23) shows that the MCP function reduces the penalization effect on ϖ_j until $|\varpi_j| > \gamma\tau/Q''_j$, at which point there is no penalization on ϖ_j. The goal of MCP is to remove bias from the penalized effects incurred by the penalty. Similarly for LASSO, we solve for the DIF effects using a thresholding operator. With the additional tuning parameter, γ, this is referred to as the firm-thresholding operator:

$$\frac{F(\widetilde{\varpi_j}, \tau, \gamma)}{Q''_j} = \begin{cases} \frac{\gamma}{\gamma - 1} \frac{S(\widetilde{\varpi_j}, \tau)}{Q''_j}, & \text{if } |\widetilde{\varpi_j}| \leq \gamma\tau/Q''_j; \\ \widetilde{\varpi_j}, & \text{if } |\widetilde{\varpi_j}| > \gamma\tau/Q''_j, \end{cases} \tag{27}$$

such that $\varpi_j^{(t+1)} \leftarrow \frac{F(\widetilde{\varpi_j}, \tau, \gamma)}{Q''_j}$.

The simulation study described next evaluates these regularization approaches and compares them to observed/marginal maximum likelihood estimation (Bauer et al., 2020).

4. Simulation Study

This study uses a subset of previously generated data originally used to evaluate scoring procedures in the MNLFA framework under various conditions of DIF (Curran, Cole, Bauer,

Hussong, & Gottfredson, 2018). Although scoring is not the focus here, these data provide a useful benchmark for identifying DIF as well as a degree of consistency with Bauer et al. (2020), who used the same data. We extend the study factors from Bauer et al. to include a larger number of scale items and exogenous covariates. Table 1 shows the simulated population parameters. Lastly, to limit the scope of the simulation study, we focus on binary item responses only.

4.1. Design Factors

The study design included seven factors:

1. Sample size: 500 or 2000 observations;
2. Number of items: 6, 12, 48*, or 96* items;
3. Number of covariates: 3 or 6* covariates;
4. Proportion of DIF: $\frac{1}{3}$ or $\frac{2}{3}$ of items exhibit DIF;
5. Magnitude of DIF: small or large, as measured by an adaptation of the weighted "area between the curves" metric (Edelen, Stucky, & Chandra, 2015; Hansen et al., 2014);
6. Penalty function: LASSO or MCP;
7. Anchor items known: Yes* or no;

where all simulation factors are crossed, except for the levels denoted with *. Aside from these levels, there are 32 unique cells (conditions) of simulated data.

Originally designed to mimic an integrative data analysis application (Hussong, Curran & Bauer, 2013), the first three background characteristics are generated to reflect age (continuous, ranging from 10 to 17), sex (dichotomous, 0 and 1), and study (dichotomous, 0 and 1), all of which affect both the latent variable distribution and a subset of the item parameters (DIF). We focus on one level of impact (i.e., differences in latent distribution as a function of background variables) from Curran et al. (2018): large mean impact and small variance impact. Note that $\frac{2}{3}$ of items with DIF is a fairly extreme condition. Nevertheless, it provides an upper boundary with which to evaluate the penalized EM algorithm for DIF detection.

The levels denoted with * are new to this study. First, we evaluate LASSO and MCP by crossing the 48 and 96 item levels with the 3 and 6 exogenous covariate levels[9] within a single intersection of the other study factors: a sample size of 2000, $\frac{1}{3}$ proportion of DIF, and large magnitude of DIF. The penalized EM algorithm is likely to perform well with these types of data conditions given the large numbers of parameters to estimate (Bauer et al., 2020; Belzak & Bauer, 2020). Second, we evaluate LASSO and MCP with known anchor items across the 500 and 2000 sample size levels and a single intersection of the other study factors: 6 items, $\frac{1}{3}$ proportion of DIF, and large magnitude of DIF. All DIF effects on items 1-3 and 6 are anchored across individuals (i.e., intercept and slope DIF parameters are set to 0). These additional conditions aim to evaluate the model selection consistency of both penalty methods, or whether true positives increase and false positives decrease as sample sizes increase.

[9]In simulation conditions with 6 exogenous covariates, two covariates were continuous and four were categorical, duplicating the pattern of the first three exogenous covariates.

In total, there are 44 unique conditions of simulated data. The penalty function factor is a "within-subjects" factor, as each replication is tested for all levels of the other factors, whereas the others are "between-subjects" factors. We simulate 500 replications per cell, with each dataset (i.e., replication) including J item responses, N observations, and K covariates.

4.2. Procedure

We use the regDIF R package (Belzak, 2023) to conduct the simulation study. The regDIF package generates 100 tuning parameters by default, starting with the smallest value of τ that penalizes all DIF effects to zero, and increments τ down to zero. By starting the Reg-DIF procedure with a large tuning parameter, no anchor items need to be explicitly specified by the researcher, as DIF effects enter the model until τ approaches zero. The regDIF package has an automatic stopping mechanism, such that at least one DIF parameter per covariate, per parameter (i.e., intercept/slope) is constrained to zero. This stopping mechanism is sufficient to produce a minimally identifiable model in the presence of a weak or zero penalty.[10] If anchor items are specified, the mechanism is bypassed.

All values of τ are fit to each of the 500 simulated replications per cell. For each model, we save information criteria and parameter estimates. Given that previous research has shown BIC to outperform the Akaike Information Criterion (AIC) in model selection (Bauer et al., 2020), we focus on BIC in this study, such that the model associated with the τ value that generates the minimum BIC is selected for further analysis. Furthermore, because the MCP function is nonconcave in certain regions of the regularization path and thus can yield non-unique parameter estimates, the regDIF packages has another automatic stopping rule. Specifically, if the number of EM iterations reaches the maximum limit for a single value of τ, the model-fitting process stops with an abnormal exit status.

For all simulation analyses, we fix the convergence criteria such that the tolerance of parameter changes (i.e., the sum of parameter changes) is no larger than 1×10^{-7}. Additionally, we set the maximum number of EM iterations to be 4000.

[10] In a multiple-group IRT model, this minimally-identifiable constraint is akin to having a single anchor item – that is, fixing the intercept and slope DIF parameter of a single item response to 0.

TABLE 1.

Population parameter values for the Monte Carlo simulation study. Asterisks denote items with DIF in the indicated condition. These data were originally simulated in Curran et al. (2016). The population parameters of the "many items and many covariates" partial design follows the same block pattern as in Table 1. For instance, in the 48-item condition with 6 exogenous covariates, the first 24 items are specified block-wise with DIF (for 1/3 proportion of DIF) and the first 3 exogenous covariates repeat twice with the same population parameter values on the DIF effects (for large magnitude of DIF).

	6 Items		12 Items		Intercept (Small DIF | Large DIF)				Slope (Small DIF | Large DIF)			
Item	DIF 1/3	DIF 2/3	DIF 1/3	DIF 2/3	Baseline	Age	Gender	Study	Baseline	Age	Gender	Study
1					-.5				1			
2		*	*	*	-.9	.125 | .25	-.5 | -1	.5 | 1	1.3	.05 | .075	-.2 | -.3	.2 | .3
3		*	*	*	-1.3	-.125 | -.25	.5 | 1	.5 | 1	1.6	-.05 | -.075	.2 | .3	.2 | .3
4	*	*	*	*	-1.7	.125 | .25			1.9	.05 | .075		
5	*	*	*	*	-2.1		-.5 | -1	.5 | 1	2.2		-.2 | -.3	.2 | .3
6					-2.5				2.5			
7					-.5				1			
8				*	-.9	.125 | .25	-.5 | -1	.5 | 1	1.3	.05 | .075	-.2 | -.3	.2 | .3
9				*	-1.3	-.125 | -.25	.5 | 1	.5 | 1	1.6	-.05 | -.075	.2 | .3	.2 | .3
10				*	-1.7	.125 | .25			1.9	.05 | .075		
11				*	-2.1		-.5 | -1	.5 | 1	2.2		-.2 | -.3	.2 | .3
12					-2.5				2.5			

4.3. Outcomes

We record true positive (TP) and false positive (FP) rates for DIF effects that are identified with the best-BIC model. A true positive refers to a non-zero DIF effect (intercept or slope) that exists in the data-generating model and which also appears in the estimated model chosen for the sample. A false positive refers to a DIF effect that does not exist in the data-generating model but nevertheless appears in the estimated model chosen for the sample.

TPs and FPs are counted at the item-level as opposed to the parameter-level, such that if either the intercept or slope effect exhibits DIF in the best-BIC model, the item (for a particular covariate) is counted as having DIF. This approach is consistent with the motivation of researchers who tend to care more about *whether* an item and/or covariate exhibits DIF, rather than precisely *how* it exhibits DIF. It is worth noting that Bauer et al. (2020) counted DIF effects at the item-level; thus, the results here can be readily compared between the two estimation approaches.

In prior work, Bauer et al. (2020) evaluated TP and FP rates for DIF effects that not only remained in the final model but were also statistically significant when evaluated using "naïve" standard errors (Bauer et al., 2020; Belzak & Bauer, 2020). The standard errors were "naïve" or incorrect because model selection uncertainty was unaccounted for.[11] In the present work, we count DIF effects that remain in the model regardless of statistical significance, which is a more conventional standard in regularization studies (Hastie et al., 2017).

In addition to true and false positive rates, we record abnormal convergence behavior using the early exiting rule described above. All other study outcomes were calculated without any of the replications that resulted in abnormal convergence.

4.4. Proof of Concept

Before running the simulation study, we demonstrate that the penalized EM algorithm yields identical results to observed/marginal maximum likelihood estimation, without penalization, using a single simulated dataset. Details about these data are provided in Appendix C.

We compare the parameter estimates from the EM algorithm against the estimates from observed/marginal maximum likelihood, setting the tuning parameter equal to zero for both estimation methods. This isolates any differences that may be due to the form of the likelihood function rather than the manner in which estimates are set to zero with the penalty. The regDIF R package uses the EM algorithm, whereas the nonlinear mixed effects procedure in SAS (NLMIXED, SAS Institute, 2018) maximizes the observed marginal likelihood. In Appendix C, we provide item and latent variable parameter estimates.

As expected, the results show nearly identical correspondence between both estimation approaches. The largest divergence in parameter estimation is .006, which is likely due to differences in convergence criteria.

[11] Despite being "naïve" standard errors, the significance tests in Bauer et al. (2020) resulted in fewer FPs without much loss in TPs.

5. Results

Table 2 presents proportions of replications with abnormal convergence (i.e., more than 4000 EM iterations during model-fitting) for LASSO and MCP. MCP exhibited higher amounts of abnormal convergence than LASSO, particularly in conditions with the most amount of DIF and smaller sample sizes.

Higher rates of abnormal convergence for MCP are not surprising, as non-convexity can produce multiple likelihood solutions (i.e., different sets of parameter estimates) and multiple solutions may result in longer-than-normal convergence. Conditions with large amounts of DIF and small sample sizes may produce convergence issues because there is less information to identify many DIF effects. Nevertheless, we removed abnormally converged replications before recording true and false positives, but recognize the subset that remains may differ from the full sample of replications. We provide recommendations in the Discussion section on how to deal with abnormal convergence.

TABLE 2.
Proportion of replications exhibiting abnormal convergence, including those that failed to converge within 2000 EM iterations.

Sample Size	DIF		**LASSO**		**MCP**	
			Items		Items	
	Proportion	Magnitude	6	12	6	12
500	1/3	Small	.01	.00	.07	.04
		Large	.01	.00	.07	.04
	2/3	Small	.01	.00	.09	.06
		Large	.04	.02	.13	.09
2000	1/3	Small	.00	.00	.04	.03
		Large	.00	.00	.04	.02
	2/3	Small	.00	.00	.03	.04
		Large	.01	.00	.07	.03

5.1. Main Study Design

Figures 1 and 2 show FPs and TPs at sample sizes of 500 and 2000, respectively. Several expected outcomes validated the simulation study, such as larger sample sizes yielding greater power in detecting DIF (higher TP rates) across all penalty methods, and greater magnitudes and proportions of DIF producing higher FP rates. Other outcomes between the penalty approaches were more informative if less predictable.

In particular, LASSO showed higher FP rates compared to MCP in conditions with larger

sample sizes and more DIF, which was generally offset by higher rates of TPs for LASSO. This pattern of results was reversed for a select few conditions, where MCP showed higher levels of TPs compared to LASSO (e.g., 1/3 of items with small DIF in smaller sample sizes). As before though, this pattern was offset by higher rates of FPs.

5.2. Current Results vs. Bauer et al.'s Results

In Figures 3 and 4, we replot the EM LASSO results against Bauer et al.'s (2020) results, including the observed/marginal likelihood LASSO approach (without naïve standard errors) and IRT-LR-DIF (Thissen, Steinberg, & Wainer, 1993), the latter of which iteratively tested each item for DIF (across all background variables) while all other items were anchored without DIF.

The EM LASSO method tended to be more sensitive in identifying DIF (higher TPs) compared to the observed/marginal likelihood LASSO approach, but also less discerning (higher FPs). However, both LASSO methods were more discerning (fewer FPs) than IRT-LR-DIF, but also less sensitive (higher TPs). This pattern was strongest in conditions with pervasive DIF and larger sample sizes. Furthermore, when the measurement scale contained $\frac{2}{3}$ of items with DIF, no method showed acceptable FP rates.

5.3. Many Items and Covariates

Figure 5 presents results for the 48- and 96-item conditions by the 3- and 6-covariate conditions. Notably, MCP exhibited fewer FPs than LASSO as the number of items and covariates increased. Although both penalty methods showed nearly perfect recovery of true DIF, only MCP held FP rates below the nominal alpha rate (.05) in three of the four large item $\times$ covariate conditions.

5.4. Known Anchor Items

Figures 1-4 show that LASSO exhibited higher rates of FPs as sample sizes grew. This implies that LASSO does not have model selection consistency, at least not in these simulated data conditions. To determine whether the lack of model selection consistency occurs because of contaminated anchor items, we evaluated LASSO and MCP while correctly specifying the anchor items. Table 3 shows that even with known anchor items specified, LASSO exhibits higher FPs rates as sample sizes increase. Conversely, MCP has the same FP rates as sample sizes increase.

6. Empirical Example

Having studied the penalized EM approaches under known data conditions, we provide an empirical demonstration with real data measuring delinquent and violent behaviors in adolescents. Our goals with these empirical data are threefold. First, we wish to demonstrate the relevance and usefulness of the penalized EM algorithm for behavioral and social scientists. Second, these data were also analyzed by Bauer (2017) to illustrate the MNLFA model for DIF testing, where an iterative sequence of likelihood ratio tests was used to evaluate DIF across five predictors. Reanalyzing these data here allows us to compare the Reg-DIF results with Bauer's (2017)

FIGURE 1.
True and false positive rates at sample sizes of 500.

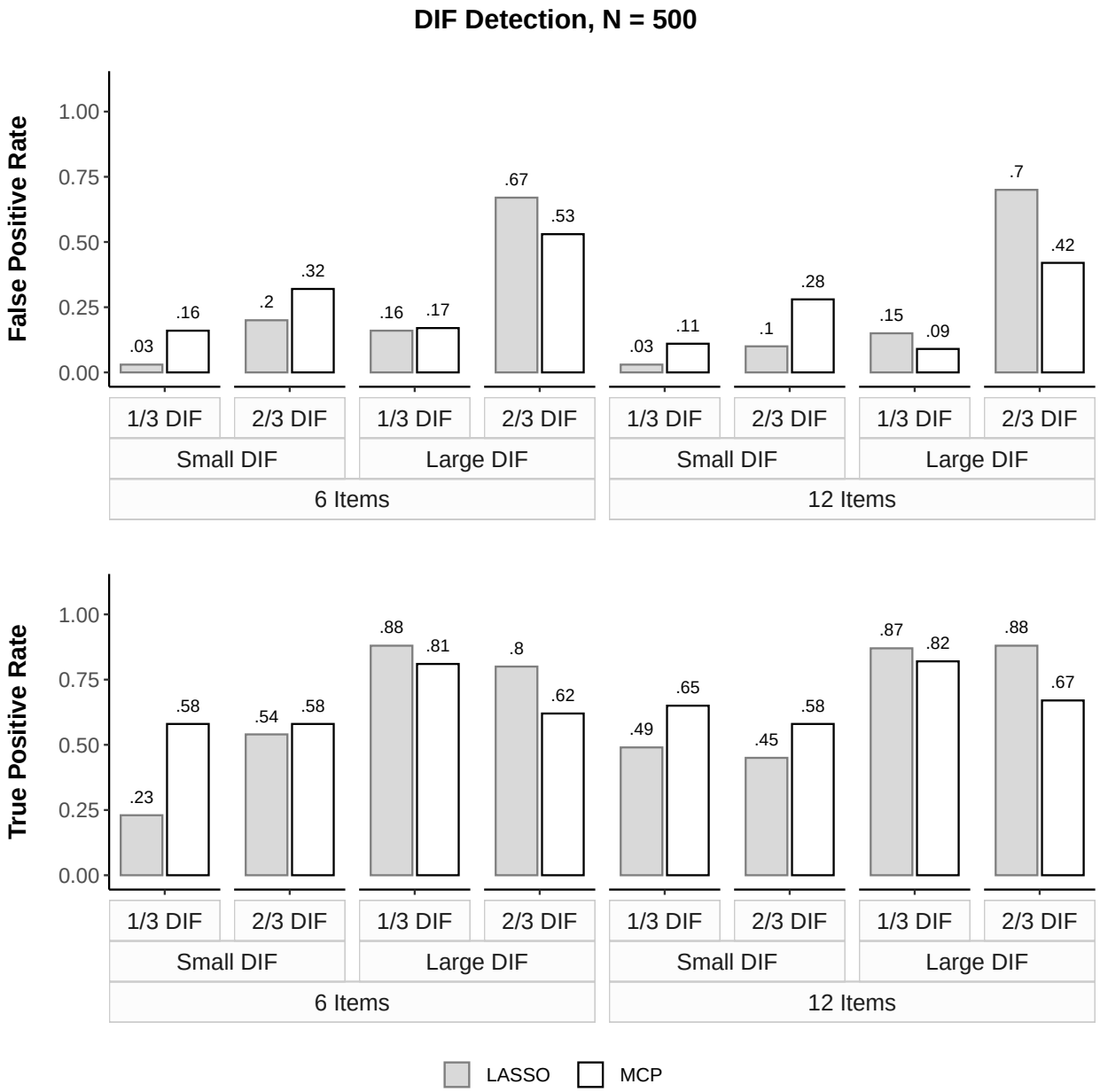

FIGURE 2.
True and false positive rates at sample sizes of 2000.

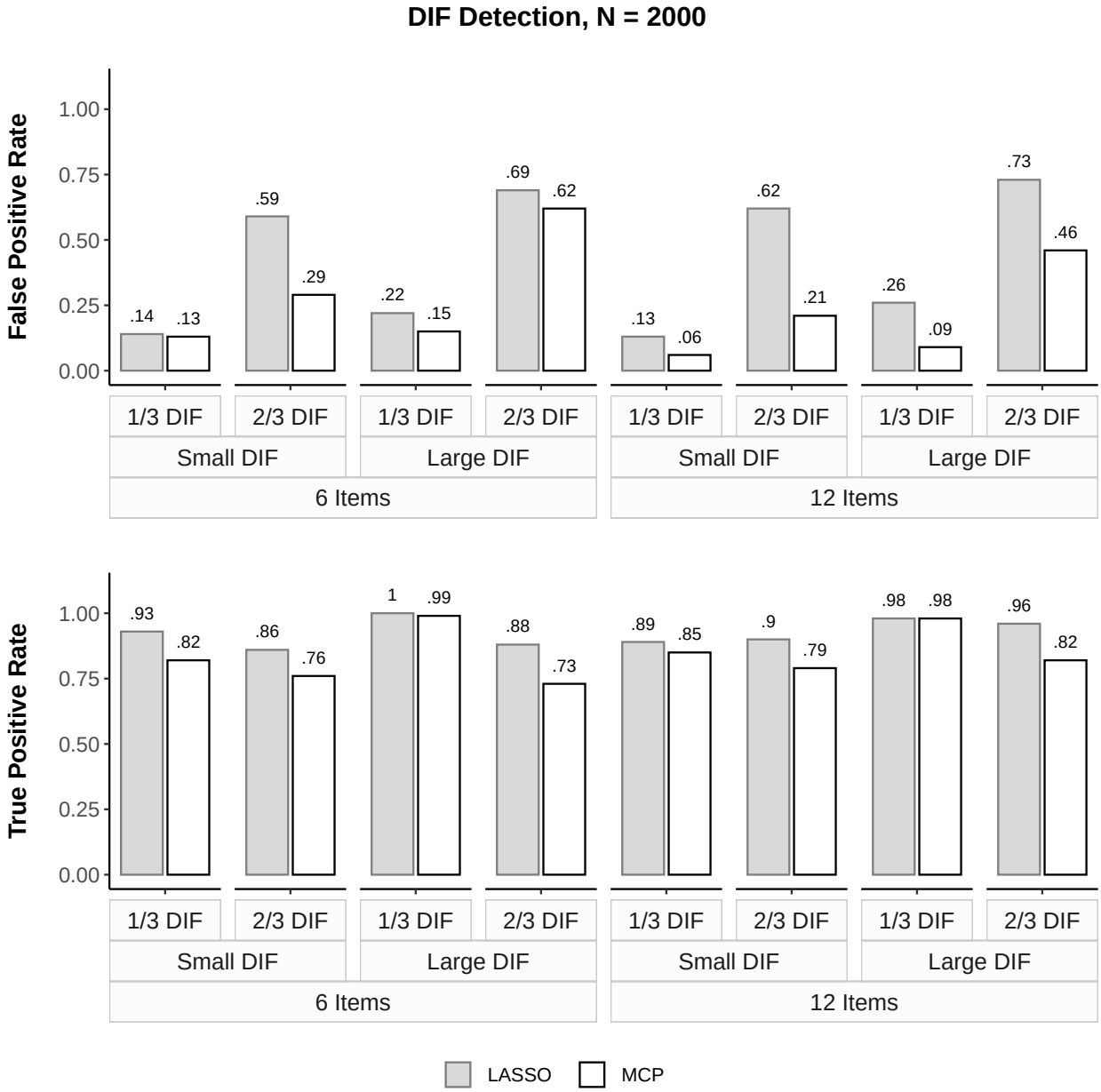

FIGURE 3.
Comparing Bauer et al.'s (2020) true and false positive rates at sample sizes of 500 and 6 items. EM LASSO refers to the penalized EM LASSO method. OML LASSO refers to the observed marginal likelihood LASSO method. IRT-LR-DIF refers to the iterative likelihood ratio test method. The OML LASSO and IRT-LR-DIF results are reprinted from Bauer et al. (2020), whereas the EM LASSO results are reprinted from Figure 1 (referred to as LASSO there). Only the results for the 6-item conditions are presented.

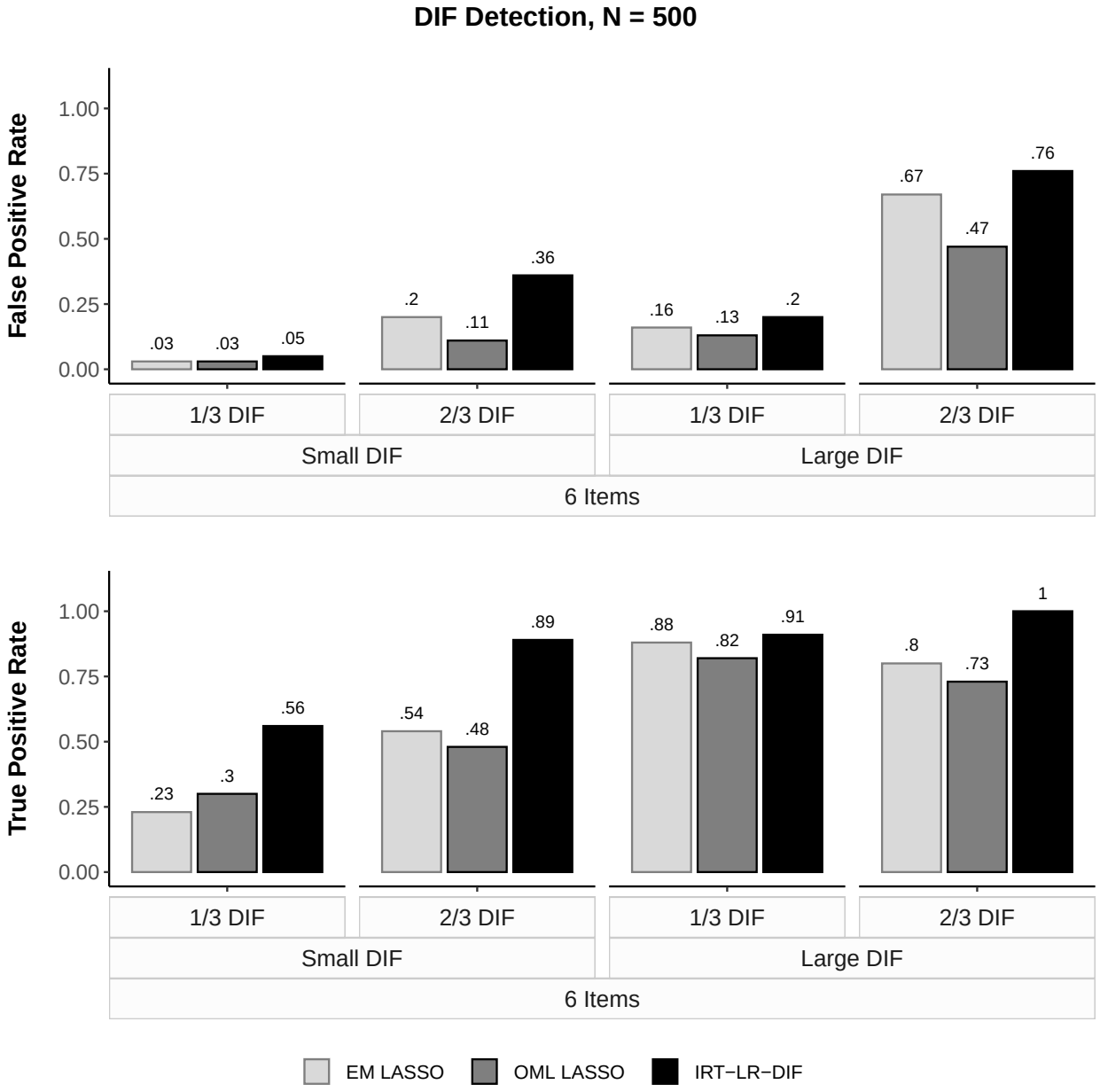

FIGURE 4.
Comparing Bauer et al.'s (2020) true and false positive rates at sample sizes of 2000 and 6 items. EM LASSO refers to the penalized EM LASSO method. OML LASSO refers to the observed marginal likelihood LASSO method. IRT-LR-DIF refers to the iterative likelihood ratio test method. The OML LASSO and IRT-LR-DIF results are reprinted from Bauer et al. (2020), whereas the EM LASSO results are reprinted from Figure 1 (referred to as LASSO there). Only the results for the 6-item conditions are presented.

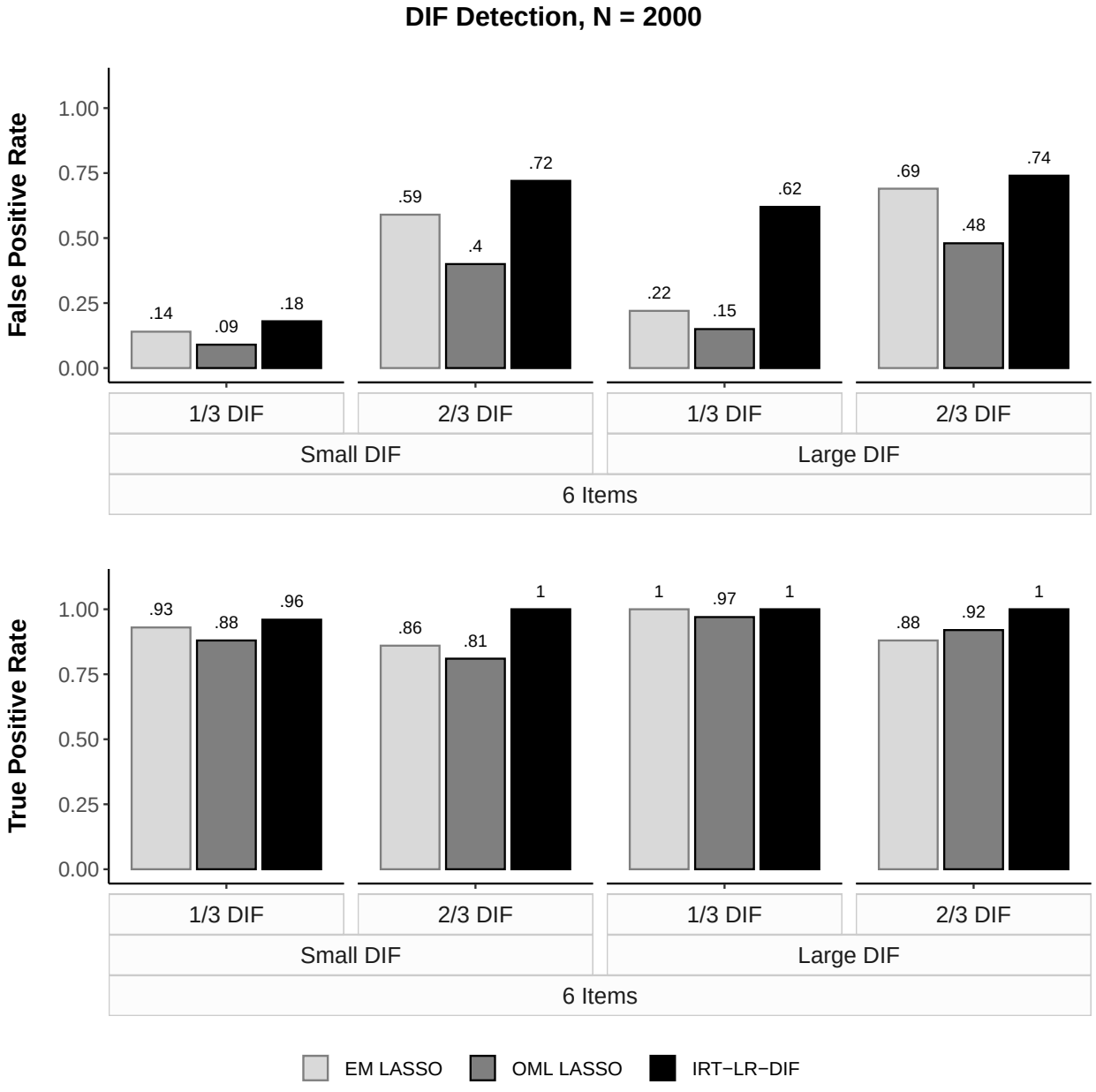

FIGURE 5.
True and false positive rates for many items and covariates. The data generated here included 1/3 items with large DIF and sample sizes of 2000.

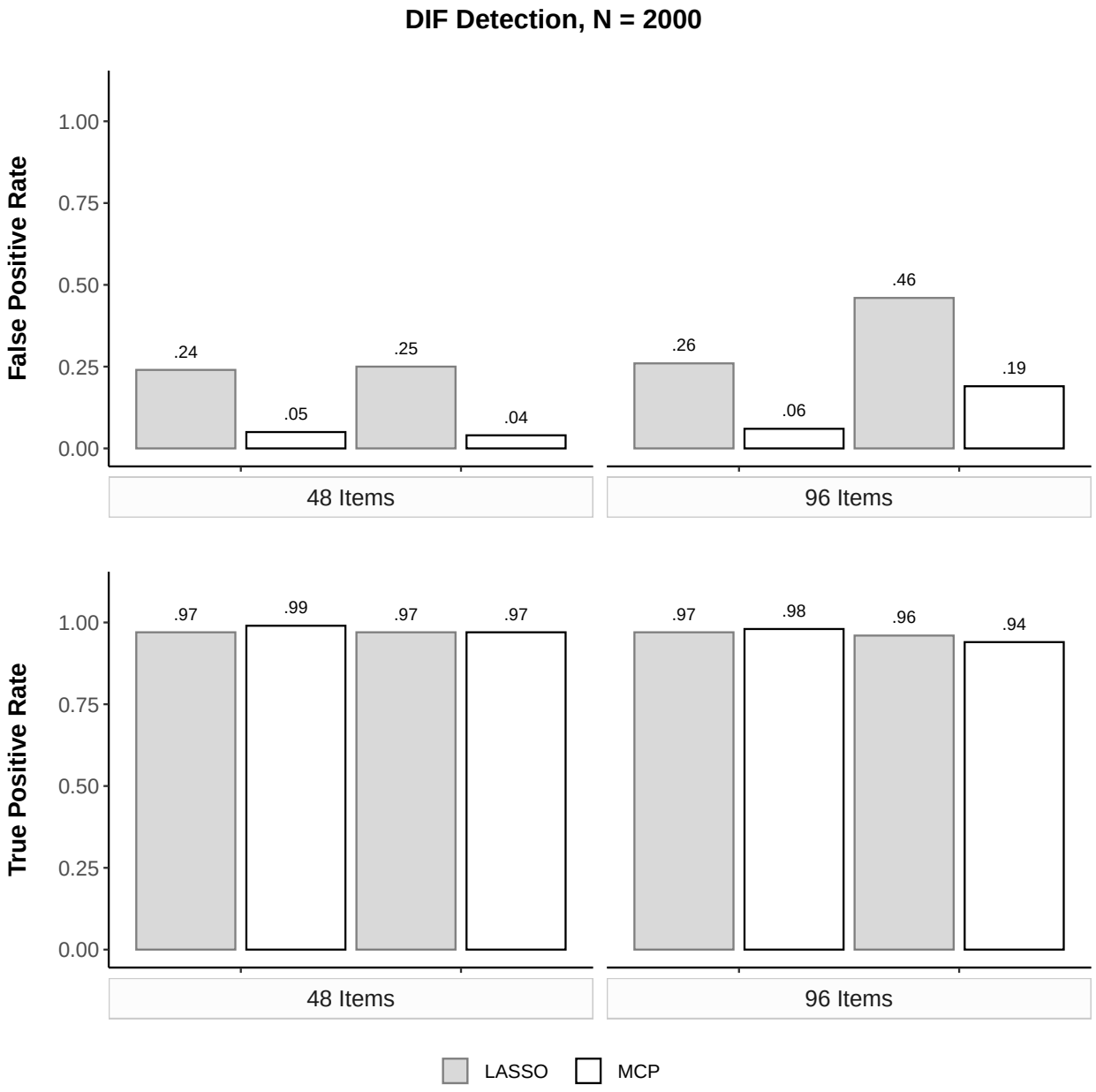

TABLE 3.
True and false positive rates for sample sizes of 500 and 2000 with correctly specified anchor items. All other conditions were fixed to 6 items and 1/3 of items with large DIF.

Outcome	Sample Size	**LASSO**	**MCP**
False Positive	500	.12	.05
	2000	.14	.05
True Positive	500	.97	.88
	2000	1.00	.99

findings. Finally, the dimensions of the data (number of participants, items, and covariates) made the prior implementation of Reg-DIF via observed/marginal maximum likelihood estimation computationally infeasible. This highlights the advantages of the the penalized EM algorithm.

6.1. Data and Items

These data were selected from the National Longitudinal Study of Adolescent to Adult Health and consist of a self-weighting sub-sample of 4,243 adolescents in the United States. Adolescents provided self-reports on two dimensions of problematic behavior, responding to 8 items about (non-violent) delinquent behavior (D) and 10 items about violent behavior (V). Although the items were presented with ordinal response options, Bauer (2017) dichotomized the responses (yes/no) due to sparseness in the upper category levels. Marginal endorsement rates for the item responses were fairly low (ranging from 1% to 53%; see Bauer, 2017). Impact and DIF were evaluated as a function of age (ranging from 12 to 18, $M_{\text{age}} = 14.9$, $SD_{\text{age}} = 1.7$) and sex (47% male). In line with Bauer, five features were constructed from the two covariates: age, age^2, male, male $\times$ age, and male $\times$ age^2. In the original analysis, age was centered, and male was dummy-coded.

6.2. Previous Results

Using the likelihood ratio test method, Bauer (2017) identified significant DIF in half of the items across both factors. The lower-order terms were retained even if non-significant in the event that a higher-order term was significant. The delinquency factor exhibited mean impact for all features except the male $\times$ age^2 interaction, whereas variance impact was identified for age and male. Similarly, the violent factor showed mean impact for all features, as well as variance impact for age and male. The final model allowed for the factor covariance to be a function of both age and male.

In addition to using a different method for DIF detection, the current application of Reg-DIF differ from Bauer's (2017) prior analysis in two important ways. First, we evaluate DIF in separate unidimensional models for each factor, primarily for computational convenience. Second,

we standardize the background variables (age and male) and then use them to compute the higher-level effects, which is consistent with prior research (Bien, Taylor, & Tibshirani, 2013). Because Bauer did not standardize the predictors, the parameter estimates are not directly comparable. Thus, our primary focus is on whether similar patterns of DIF are detected.

6.3. Current Results

The previous results are compared with the current results, obtained using the regDIF R package (Belzak, 2023). We use bold and non-bold text, as well as gray shading, to indicate DIF effects that were (1) identified by both Reg-DIF (LASSO or MCP) and likelihood ratio tests (bold text), (2) identified by Reg-DIF but not by likelihood ratio tests (non-bold text), and (3) identified by likelihood ratio tests but not by Reg-DIF (gray shading).

Table 3 presents parameter estimates from using LASSO and MCP on the delinquency factor. Eight of the same DIF effects were detected by both LASSO and likelihood ratio tests, and six of the same DIF effects for MCP and likelihood ratio tests. These findings provide some measure of confidence that DIF exists in the following three items: 'D3. Lie parents', 'D8. Car w/o permission', and 'D15. Rowdy in public'. For DIF identified with regularization but not by Bauer (2017), we identified five additional DIF effects for LASSO, and one additional DIF effect for MCP. Given that in large sample sizes, MCP showed better control of FPs and LASSO showed better recovery of TPs, it is plausible that DIF exists in the population where both LASSO and MCP, but not likelihood ratio tests, identified DIF. The difference between MCP and LASSO may be explained by higher FPs from LASSO. Finally, for DIF that was identified by likelihood ratio tests but not by Reg-DIF, these may be FPs given that likelihood ratio tests exhibited much higher FPs than LASSO in Bauer et al.'s (2020) simulation.

Table 4 presents parameter estimates from using both penalty methods on the violent factor. We observed less correspondence between the methods than before. Specifically, only five DIF effects were identified by both LASSO and Bauer's (2017) likelihood ratio tests, and 13 DIF effects were identified by both MCP and likelihood ratio tests. As before, these findings suggest that DIF exists in three items – 'V1. Saw shoot/stab', 'V2. Pulled weapon on you', and 'D14. Group fight' – but there were many DIF effects identified by likelihood ratio tests but not by either penalty approach, as well as many DIF effects identified by MCP but not by the other DIF detection methods. That is, 17 DIF effects were identified by likelihood ratio tests but not by LASSO, and nine effects were identified by likelihood ratio tests but not by MCP. One possible reason for this discrepancy is that Bauer (2017) retained lower-order effects when higher-order terms were present, whereas neither MCP nor LASSO implemented this rule. Additionally, MCP identified 24 additional DIF effects that were not identified by either likelihood ratio tests or LASSO. Based on the simulation results, we would expect both penalty methods to show high power in identifying DIF with large sample sizes. Although this suggests that DIF likely exists in the population where both penalty methods identified DIF, the lack of correspondence between LASSO and MCP suggests higher FPs for one of the procedures as well as differentiated power. Unfortunately, we cannot directly answer this given that the population structure is unknown.

One caveat to note for LASSO and MCP is that some higher-order effects were included in the final model without their corresponding lower-order effects. One possible reason for this

outcome was that the higher-order terms were not standardized themselves (consistent with recommendations from Cohen, Cohen, West, & Aiken, 2013), which implies that the interaction and quadratic effects may have had larger or smaller weights in the penalization process, and in turn, may have contributed to a different pattern of results relative to standardizing all effects in the model. If lower-order effects were included with higher-order terms when using LASSO and MCP, we may have also had greater correspondence between the penalty methods and likelihood ratio tests, particularly with the violent factor.

TABLE 4.

Reg-DIF parameter estimates for delinquency factor using BIC. Bold text indicates parameters present in the final MNLFA model estimated by Bauer (2017), who used likelihood ratio tests with pruning of non-significant DIF effects), and also present in the best-fitting Reg-DIF model estimated here with the penalized EM algorithm. Non-bold text indicates parameters present only in the best-fitting Reg-DIF model estimated here. Gray shading indicates parameters present only in the final MNLFA model estimated by Bauer (2017). Finally, Reg-DIF with MCP was implemented with $\gamma = 3$.

	LASSO						MCP					
Reference Parameter	Baseline	Age	Age^2	Male	Male × Age	Male × Age^2	Baseline	Age	Age^2	Male	Male × Age	Male × Age^2
Delinquency Factor												
Mean	0	**−0.02**	**−0.15**	**0.09**	**0.04**	0.05	0	**−0.05**	**−0.13**	**0.15**	**0.06**	0.04
Variance	1	**−0.14**	0.07	**0.21**	0.04	−0.11	1	**−0.12**	0.05	**0.14**	0.00	−0.12
D1. Graffiti												
Intercept	−3.39						−3.40					
Slope	1.82						1.84					
D2. Property damage												
Intercept	−2.61	**−0.01**					−2.67					
Slope	2.51			**0.12**		0.04	2.57					0.22
D3. Lie parents												
Intercept	0.39	**0.28**	**−0.05**	**−0.24**	**−0.12**	−0.04	0.46	**0.38**	**−0.11**	**−0.40**	**−0.21**	
Slope	1.32						1.40					
D8. Car w/o permission												
Intercept	−2.83	**0.06**					−2.84	**0.11**				
Slope	1.29						1.30					
D9. Steal > $50												
Intercept	−4.59						−4.60					
Slope	2.06						2.08					
D10. Steal from house												
Intercept	−4.76						−4.78					
Slope	2.30						2.32					
D13. Steal < $50												
Intercept	−2.11						−2.12					
Slope	1.90						1.90					
D15. Rowdy in public												
Intercept	−0.07		0.01	**−0.09**	0.03	−0.02	−0.07			**−0.22**		
Slope	1.25						1.30					

TABLE 5.

Reg-DIF parameter estimates for violent factor using BIC. Bold text indicates parameters present in the final MNLFA model estimated by Bauer (2017), who used likelihood ratio tests with pruning of non-significant DIF effects), and also present in the best-fitting Reg-DIF model estimated here with the penalized EM algorithm. Non-bold text indicates parameters present only in the best-fitting Reg-DIF model estimated here. Gray shading indicates parameters present only in the final MNLFA model estimated by Bauer (2017). Finally, Reg-DIF with MCP was implemented with $\gamma = 3$.

	LASSO						MCP					
Reference Parameter	Baseline	Age	Age^2	Male	Male × Age	Male × Age^2	Baseline	Age	Age^2	Male	Male × Age	Male × Age^2
Violent Factor												
Mean	0	**−0.07**	**−0.03**	**0.32**	**0.04**	**0.04**	0	**0.01**	**0.05**	**0.28**	**0.00**	**0.00**
Variance	1	**−0.04**	−0.13	**−0.10**	0.07	0.03	1	**−0.05**	−0.27	**−0.09**	0.09	0.04
V1. Saw shoot/stab												
Intercept	−3.13	**0.02**		**−0.12**		−0.09	−3.45	**0.09**			0.22	−0.12
Slope	1.84						2.24	**0.16**		−0.25	−0.29	
V2. Pulled weapon on you												
Intercept	−3.48	**0.24**					−3.75	**0.30**		0.22		
Slope	2.29						2.53	**0.14**	**−0.04**			
V3. Shot you												
Intercept	−5.62						−5.97	**0.47**				
Slope	1.58						1.62				−0.50	0.32
V4. Cut/stabbed you												
Intercept	−5.10						−5.34			−0.66		
Slope	2.26						2.46			0.31		
V5. Physical fight												
Intercept	−1.58	−0.11			−0.05		−1.60	−0.34	−0.22	0.12	−0.08	0.03
Slope	2.51						2.50		0.21			
V6. Were jumped												
Intercept	−3.58						−3.74			0.30		
Slope	2.14						2.25		−0.06			
V7. You pulled weapon												
Intercept	−5.99						−6.23					
Slope	2.99						3.21	0.11				
V8. You shot/stabbed												
Intercept	−9.27						−9.64					
Slope	3.94						4.08		0.23			
D6. Hurt other badly												
Intercept	−2.69						−2.42	**−0.26**	−0.52	**0.34**		
Slope	2.16						1.92		0.38	−0.23	0.13	0.12
D14. Group fight												
Intercept	−2.32	**−0.02**		**−0.07**		−0.04	−2.25		−0.46	**−0.32**	**0.45**	−0.07
Slope	1.91						1.77	**−0.28**	0.55	**0.32**	**−0.39**	

7. Discussion

The primary goals of this paper were to: (1) develop a more flexible, computationally efficient algorithm for using Reg-DIF among many item responses and background variables; (2) evaluate the penalized EM algorithm with a simulation study and compare it to previous implementations of regularized DIF; and (3) illustrate Reg-DIF for applied researchers while using the penalized EM algorithm on empirical data. We discuss the progress made on each of these goals below, while noting the main limitations of this study and considering future directions for research.

7.1. Developing a Flexible and Efficient Algorithm

Regularized DIF analysis has been difficult to implement in applied research settings because the options for model estimation are limited. Most previous research has either focused on simpler IRT models with faster estimation methods (e.g., Rasch with conditional maximum likelihood; Tutz & Schauberger, 2015; Magis et al., 2015) or more complex models with slower estimation (e.g., 2-parameter logistic IRT with observed/marginal maximum likelihood; Bauer et al., 2020; Belzak & Bauer, 2020; Chen et al., 2023; Schauberger & Mair, 2020). More recently, Huang (2018) and Wang et al. (2022) developed penalized EM algorithms for regularized DIF analysis, and Liang and Jacobucci (2022) developed a penalized limited information estimation approach for DIF analysis. Both estimation frameworks are capable of estimating complex IRT models relatively quickly.

In this paper, we developed a penalized EM algorithm which performs regularized DIF analysis across multiple continuous and categorical covariates, a limitation of other recent EM approaches. This extension allows researchers to begin investigating the causes of DIF in latent variable models. The penalized EM method also allows for varying item response functions and different penalty functions (see Belzak, 2023).

Previous implementations of regularized MNFLA employed observed/marginal maximum likelihood for model estimation (Bauer et al., 2020). Although this method allows for the evaluation of multiple DIF covariates in more complex IRT models (Schauberger & Mair, 2020), it becomes computationally intractable as the number of scale items and DIF covariates increase. For instance, in the simulated data condition with 96 item responses and six background characteristics, there were 1152 DIF effects to estimate and select from.[12] Fitting this model with observed/marginal maximum likelihood estimation is impossible, yet it was tractable with the penalized EM method. Furthermore, we calculated that for a simulated dataset with 12 item responses, three background characteristics, and 2000 observations, it took approximately 141 hours to run 100 values of tau (i.e., 100 fitted models) while using observed/marginal maximum likelihood in the SAS NLMIXED procedure. In contrast, fitting these same data using the penalized EM algorithm in the regDIF R package only took about 30 minutes, or roughly .5% of the time. The penalized EM algorithm developed here is thus more flexible and efficient for regularized DIF analysis.

[12]In this condition, there were 96 items × 6 covariates × 2 intercept and slope DIF parameters (1152), 96 items × 2 baseline intercept and slope parameters (192), and 6 covariates × 2 latent mean and variance impact parameters (12), totaling 1356 parameters.

7.2. Comparing Estimation and Penalization Methods

An important finding from the simulation study was that MCP generally exhibited variable selection consistency as sample sizes and the number of scale items increased, although the FP rate increased slightly when $\frac{2}{3}$ of items exhibited large DIF (53% $\rightarrow$ 62% for 6 items, and 42% $\rightarrow$ 46% for 12 items). However, given the extremity of this condition, no method of DIF identification may exhibit model selection consistency in conditions with such pervasive DIF. In contrast to MCP, LASSO exhibited higher FP rates as sample sizes increased, even when anchor items were correctly specified. These findings are not surprising given that non-convex approaches like MCP exhibit the oracle property (Fan, 1997; Zhang, 2010), whereas LASSO does not unless certain restrictive conditions are met (Meinshausen and Bühlmann, 2004; Zhao & Yu, 2006; Zou, 2006). Furthermore, the LASSO results are consistent with Tutz and Schauberger (2015), who also found that LASSO had (marginally) higher FPs as sample sizes increased. On the other hand, Wang et al. (2022) found that larger sample sizes led to fewer FPs for LASSO. This discrepancy may be explained by the number of spurious DIF covariates in the data-generating model, such that if there are other confounding DIF effects, the LASSO method may not have model selection consistency (e.g., see irrepresentability condition[13] in Zhao & Yu, 2006).

Despite MCP showing model selection consistency in conditions with moderate amounts of DIF, a downside of this method is that model convergence may become problematic when many scale items have large DIF effects. One possible solution is that when using MCP, researchers could use (the lack of) model convergence as an indicator of whether the measurement scale is heavily contaminated with DIF. This strategy could be coupled with using different starting values (during initialization of the algorithm) to ensure the model solution is unique, or by increasing the value of γ in the MCP method, as larger values of γ (more LASSO-like in penalization) may not only result in better convergence, but also greater recovery of true DIF. Larger sample sizes may also be required to produce fewer instances of abnormal convergence, particularly when there are fewer scale items which may be heavily contaminated with DIF. Although we cannot offer exact sample size recommendations while using Reg-DIF as there are too many factors to consider in any given dataset, we recommend that researchers should plan for larger sample sizes (e.g., $N >= 2000$) if there are relatively few item responses (e.g., $P <= 6$) and many DIF covariates to evaluate (e.g., $P >= 3$).

Finally, the penalized EM algorithm showed greater power in recovering true DIF compared to observed/marginal maximum likelihood estimation, but also more spurious DIF (see Figures 3 and 4). This outcome may be due to differences in how the penalty function was implemented in Bauer et al. (2020), namely, outside of the optimization routine. Different results may also be due to unstable estimation of the parameter variance ($Q''(\varpi_j^{(t)})$ in Eq. 23), which is used to scale the tuning parameter in the penalized EM algorithm. That is, inaccurate estimation of $Q''(\varpi_j^{(t)})$ may

[13]Irrepresentability states that LASSO will achieve model selection consistency if, by regressing the irrelevant covariates (e.g., confounding DIF effects) onto the the relevant covariates (e.g., true DIF effects), the model coefficients sum to less than 1. That is, the spurious DIF effects cannot be too highly correlated with the true DIF effects or else LASSO will not achieve model selection consistency. Zhao and Yu (2006) proved this condition for linear regression models, but it is unclear whether it holds for nonlinear models with one or more latent variables.

alter which DIF effects receive more or less penalization, thus giving a sub-optimal final result (via model selection). The observed/marginal maximum likelihood method did not directly scale the tuning parameter by the parameter variance, in effect giving all DIF parameters equal penalization.

7.3. Illustrating the Penalized EM Algorithm with an Empirical Example

The empirical example showed that LASSO and MCP largely identified the same items with DIF, but not necessarily the same covariates across both the violent and delinquency scales. For instance, LASSO identified six more DIF effects than MCP on the delinquency scale, whereas MCP identified 31 more DIF effects than LASSO in the violent scale. Based on the simulation results, the additional LASSO effects could be FPs given the large sample size (i.e., N = 4,243). The additional MCP effects are more difficult to explain, as we would expect MCP to have fewer FPs and fewer TPs than LASSO.

In comparison to Bauer's (2017) likelihood ratio results, LASSO and MCP showed some consistency in results. In the delinquency scale, for instance, LASSO identified eight of the same DIF effects as Bauer's likelihood ratio test method, and MCP identified six of the same DIF effects. In the violent scale, LASSO identified five of the same effects and MCP identified 13 of the same effects. There were also some clear discrepancies between regularization and likelihood ratio testing. For instance, LASSO yielded a sparser solution (i.e., fewer DIF effects) than Bauer's (2017) likelihood ratio testing method in the violent scale.[14] This spareness could have been due to the omission of lower-order terms while using regularization. In the delinquency scale, however, LASSO identified DIF in more items than the likelihood ratio test method. In both scales, MCP was more sensitive to identifying DIF. As before, these contradictory results make it difficult to determine where DIF does or does not occur in these scales.

7.4. Limitations and Future Directions

There are a variety of ways that future research could advance the development and evaluation of Reg-DIF. First, to better understand why LASSO exhibited higher rates of FPs as sample sizes increased (i.e., a lack of model selection consistency), more theoretical work could extend Zhao and Yu's (2006) approach to latent variable models and nonlinear regression functions, and determine when irrepresentability holds while using LASSO to evaluate DIF across multiple background covariates.

Future work could also implement the penalized EM algorithm more efficiently. Despite it being considerably faster than observed/marginal maximum likelihood estimation, the penalized EM approach remains somewhat slow for models with many item responses, many DIF covariates, and large sample sizes. Because the algorithm allows multiple covariates to vary across observations/individuals rather than groups, the EM computations must occur across $N \times Q \times J$

[14]Bauer (2017) performed multiple sweeps of testing while progressively relaxing the anchor item set. He also pruned DIF effects that were non-significant in the final model-fitting process. Our simulations evaluated a simpler method of using likelihood ratio tests to detect DIF in the MNLFA model (Bauer et al., 2020, i.e., a single sweep across the scale items).

arrays. In our formulation of the log-likelihood function, there is not a straightforward way to reduce the computations related to N or J. However, the number of quadrature points, Q, could be considerably reduced if a more precise calculation of the latent variable density is achieved (e.g., see Wang et al., 2022). An alternative root-finding method such as proximal gradient descent (Friedman et al., 2010; Lee, Sun, & Saunders, 2012) may also achieve faster optimization solutions compared to coordinate descent with analytical derivatives.

Other penalty functions could also be evaluated for regularized DIF analysis. This includes the elastic net penalty, which selects highly-correlated predictors by combining the ridge (L_2 norm) and LASSO penalties, as well as the group LASSO and MCP penalties, which aim for model sparsity between groups as well as within. Alternatively, researchers may not assume DIF in the measurement model is sparse (i.e., a majority of items are DIF-free). Using a ridge penalty would specify DIF on all item responses and DIF covariates. Not only is this model estimable due to the statistical properties of the L2 norm (Hoerl & Kennard, 1970), it may be more defensible in less established measurement models. Another extension of Reg-DIF would be to apply different penalties to the DIF intercepts (i.e., main effects) and DIF slopes (i.e., interaction effects). Because slope DIF is typically more difficult to identify than intercept DIF, applying a smaller penalty to the slope may result in better recovery of DIF.

Finally, there are alternative approaches to model selection besides information criteria (e.g., BIC), including hypothesis testing and k-fold cross-validation. For example, researchers have developed a LASSO-type approach which allows for unbiased statistical inference via hypothesis testing without the need to select a tuning parameter (Chen et al., 2023), both of which are advantages over the penalized EM method. However, the LASSO-type approach has only been implemented and evaluated for group-based, uniform (intercept) DIF analyses, whereas the penalized EM algorithm works for multiple continuous and categorical DIF covariates, as well as the selection of both intercept and slope DIF. Future work could evaluate the performance of the LASSO-like method with multiple DIF covariates and compare it to the method developed here. Other researchers have also shown k-fold cross-validation to be a reliable method for model selection in identifying DIF (Tutz & Schauberger, 2015). Model selection with the penalized EM algorithm could therefore be extended to cross-validation and compared to information criteria. Yet, a downside to cross-validation is that it is far more computationally intense than calculating information criteria.

Notwithstanding the limitations above and many avenues of potential expansion, the research here demonstrates that regularization can be used to efficiently and effectively identify DIF across multiple background variables in the MNLFA model, providing a new valuable tool for psychological assessment.

References

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society: Series B (Methodological), 32*(2), 283-301.

Bauer, D.J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*(3), 507-526.

Bauer, D.J., Belzak, W.C.M. & Cole, V.T. (2020). Simplifying the assessment of measurement invariance over multiple background variables: using regularized moderated nonlinear factor analysis to detect differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, 27(1), 43-55.

Bauer, D.J. & Hussong, A.M (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods*, *14*(2), 101-125.

Belzak, W. C. M. (2021). *Using regularization to evaluate differential item functioning among multiple covariates: A penalized expectation maximization algorithm via coordinate descent and soft-thresholding.* [Doctoral Dissertation]. University of North Carolina, Chapel Hill.

Belzak, W. C. M. (2023). The regDIF R Package: Evaluating Complex Sources of Measurement Bias Using Regularized Differential Item Functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-11.

Belzak, W. C. M. & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods*, *25*(6), 673-690.

Bien, J., Taylor, J., & Tibshirani, R. (2013). A lasso for hierarchical interactions. *Annals of Statistics*, *41*(3), 1111-1141.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443-459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*(2), 179-197.

Boyd, S. P., & Vandenberghe, L. (2004). Convex optimization. Cambridge, U.K.: Cambridge University Press.

Bollen, K. A. (1989). *Structural equations with latent variables (Vol. 210).* John Wiley & Sons.

Breheny, P., & Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, *5*(1), 232-253.

Breheny, P., & Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, *25*(2), 173-187.

Chen, Y., Li, C., Ouyang, J., & Xu, G. (2023). DIF statistical inference without knowing anchoring items. *Psychometrika*, 1-26.

Cohen, J., Cohen P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences. *Routledge*.

Curran, P. J., Cole, V. T., Bauer, D. J., Rothenberg, W. A., & Hussong, A. M. (2018). Recovering predictor–criterion relations using covariate-informed factor score estimates. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(6), 860-875.

Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, *41*(3), 613-627.

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.

Edelen, M. O., Stucky, B. D., & Chandra, A. (2015). Quantifying 'problematic' DIF within an IRT framework: application to a cancer stigma index. *Quality of Life Research*, *24*(1), 95-103.

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.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1-22.

Hansen, M., Cai, L., Stucky, B. D., Tucker, J. S., Shadel, W. G., & Edelen, M. O. (2014). Methodology for developing and evaluating the PROMIS smoking item banks. *Nicotine & Tobacco Research*, *16*, S175–S189.

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning.* New York, NY: Springer.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In *H. Wainer & H. I. Braun (Eds.), Test validity* (p. 129–145). Lawrence Erlbaum Associates, Inc.

Huang, P. H. (2018). A penalized likelihood method for multi-group structural equation modelling. *British Journal of Mathematical and Statistical Psychology*, *71*(3), 499-522.

Hussong, A. M., Curran, P. J., & Bauer, D. J. (2013). Integrative data analysis in clinical psychology research. *Annual Review of Clinical Psychology*, *9*, 61-89.

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.

Liang, X., & Jacobucci, R. (2020). Regularized structural equation modeling to detect measurement bias: Evaluation of lasso, adaptive lasso, and elastic net. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(5), 722-734.

Lee, J. D., Sun, Y., & Saunders, M. (2012). Proximal Newton-type methods for convex optimization. *Advances in Neural Information Processing Systems, 25*, 1-9.

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

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.

McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models. Vol. 2nd.* Boca Raton, FL: Chapman & Hall, CRCPress.

Meier, L., Van De Geer, S., & Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(1), 53-71.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance.* New York, NY: Routledge.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus User's Guide (8th ed.).* Los Angeles, CA: Muthén & Muthén.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proc. Fourth Berk. Symp., 4*, 321-333.

Robitzsch, A. (2020). Lp loss functions in invariance alignment and Haberman linking with few or many groups. *Stats, 3*(3), 246-283.

Samejima, F. (1997). Graded response model. In *Handbook of modern item response theory* (pp. 85-100). Springer, New York, NY.

SAS Institute. (2018). *Base SAS 9.4 Procedures Guide.* SAS Institute.

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.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, *22*(2), 231-245.

Sun, J., Chen, Y., Liu, J., Ying, Z., & Xin, T. (2016). Latent variable selection for multidimensional item response theory models via regularization. *Psychometrika*, *81*(4), 921–939.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361-370.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. *In P. W. Holland & H. Wainer (Eds.), Differential item functioning* (pp. 67-113). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

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.

Wang, W. C., Shih, C. L., & Sun, G. W. (2012). The DIF-free-then-DIF strategy for the assessment of differential item functioning. *Educational and Psychological Measurement, 72*(4), 687-708.

Wang, C., Zhu, R., & Xu, G. (2021). Using lasso and adaptive lasso to identify DIF in multidimensional 2PL models. *Multivariate Behavioral Research*, 1-21.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49-67.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *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.

Zou, H., & Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, *36*(4), 1509.

APPENDIX A: CONCAVITY EVIDENCE

The log-likelihood function is defined

$$Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)}) = \sum_{q=1}^{Q}\sum_{i=1}^{N} \log\left[\phi(Z_q|\mathbf{x}_i;\boldsymbol{\xi})\right] P(Z_q|\mathbf{y}_i,\mathbf{x}_i;\boldsymbol{\gamma}^{(t)}) + \sum_{q=1}^{Q}\sum_{i=1}^{N}\sum_{j=1}^{J} \log\left[f(y_{ij}|Z_q,\mathbf{x}_i;\boldsymbol{\varpi}_j)\right] P(Z_q|\mathbf{y}_i,\mathbf{x}_i;\boldsymbol{\gamma}^{(t)}), \tag{1A}$$

where $\phi(Z_q|\mathbf{x}_i;\boldsymbol{\xi})$ is the normal density function for the latent variable; $f(y_{ij}|Z_q,\mathbf{x}_i;\boldsymbol{\varpi}_j)$ is the item likelihood function for item j; and $P(Z_q|\mathbf{y}_i,\mathbf{x}_i;\boldsymbol{\gamma}^{(t)})$ is the posterior distribution of the latent variable. We want to maximize $Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)})$ with respect to $\boldsymbol{\gamma}$.

To ensure the optimum $\boldsymbol{\gamma}^*$ is a global optimum rather than a local optimum, we must establish that $Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)})$ is concave for all values of $\boldsymbol{\gamma}$. If the additive components of a function are concave, then the function is also concave (Boyd & Vandenberghe, 2004). We thus establish concavity separately for the normal density of the latent variable and for the item likelihoods.

The latent variable component of $Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)})$ is written

$$Q_\theta(\boldsymbol{\xi}|\boldsymbol{\gamma}^{(t)}) = \sum_{q=1}^{Q}\sum_{i=1}^{N} \log\left[\phi(Z_q|\mathbf{x}_i;\boldsymbol{\xi})\right] P(Z_q|\mathbf{y}_i,\mathbf{x}_i;\boldsymbol{\gamma}^{(t)}), \tag{2A}$$

where $Q_\theta(\boldsymbol{\xi}|\boldsymbol{\gamma}^{(t)})$ represents the log-likelihood function with respect to the latent variable parameters. An easy way to establish concavity is to check whether the second partial derivative of $Q_\theta(\boldsymbol{\xi}|\boldsymbol{\gamma}^{(t)})$ with respect to a parameter in $\boldsymbol{\gamma}_\theta$ is equal to or less than 0 (non-positive) for all possible values of $\boldsymbol{\gamma}_\theta$ (Boyd & Vandenberghe, 2004). Thus, the second partial derivative with respect to a generic mean impact parameter, ζ_α, is

$$\frac{\partial^2 Q_\theta(\boldsymbol{\xi}|\boldsymbol{\gamma}^{(t)})}{\partial \zeta_\alpha^2} = \sum_{q=1}^{Q}\sum_{i=1}^{N} -\frac{\mathbf{x}_i^T\mathbf{x}_i}{\psi} P(Z_q|\mathbf{y}_i,\mathbf{x}_i;\boldsymbol{\gamma}^{(t)}), \tag{3A}$$

which is guaranteed to be non-positive for all values of ζ_α, assuming ψ is always positive and $P(Z_q|\mathbf{y}_i,\mathbf{x}_i;\boldsymbol{\gamma}^{(t)}) \geq 0$. Next, the second partial derivative with respect to a generic variance impact parameter, ζ_ψ, is

$$\frac{\partial^2 Q_\theta(\boldsymbol{\xi}|\boldsymbol{\gamma}^{(t)})}{\partial \zeta_\psi^2} = \sum_{q=1}^{Q}\sum_{i=1}^{N} -\frac{\mathbf{x}_i^T\mathbf{x}_i(Z_q-\alpha)^2}{\psi} P(Z_q|\mathbf{y}_i,\mathbf{x}_i;\boldsymbol{\gamma}^{(t)}), \tag{4A}$$

which is also guaranteed to be non-positive for all values of ζ_α because $\frac{\mathbf{x}_i^T\mathbf{x}_i(Z_q-\alpha)^2}{\psi} \geq 0$. In effect, we conclude that $Q_\theta(\boldsymbol{\xi}|\boldsymbol{\gamma}^{(t)})$ is concave for all values of ζ_α and ζ_ψ.

Finally, the item likelihood component of $Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)})$ is written

$$Q_j(\boldsymbol{\varpi}|\boldsymbol{\gamma}^{(t)}) = \sum_{q=1}^{Q}\sum_{i=1}^{N} \log\left[f(y_{ij}|Z_q, \mathbf{x}_i; \boldsymbol{\varpi}_j)\right] P(Z_q|\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\gamma}^{(t)}), \tag{5A}$$

and we assume

$$f(y_{ij}|\theta_i, \mathbf{x}_i; \boldsymbol{\varpi}_j) = P(y_{ij} = 1|\theta_i, \mathbf{x}_i)^{y_{ij}}[1 - P(y_{ij} = 1|\theta_i, \mathbf{x}_i)]^{1-y_{ij}}, \tag{6A}$$

or that the item response function is distributed Bernoulli. The second derivative with respect to a generic item parameter, ζ_j, is then written

$$\frac{\partial^2 Q_j(\boldsymbol{\varpi}|\boldsymbol{\gamma}^{(t)})}{\partial \zeta_j^2} = \sum_{q=1}^{Q}\sum_{i=1}^{N} -\left(\frac{\partial \eta_{ij}}{\partial \zeta_j}\right)^2 P(y_{ij} = 1|\theta_i, \mathbf{x}_i)\left[1 - P(y_{ij} = 1|\theta_i, \mathbf{x}_i)\right] P(Z_q|\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\gamma}^{(t)}), \tag{7A}$$

and is guaranteed to be non-positive because $P(y_{ij} = 1|\theta_i, \mathbf{x}_i) \geq 0$ and $P(Z_q|\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\gamma}^{(t)}) \geq 0$. We therefore conclude that $Q_j(\boldsymbol{\varpi}|\boldsymbol{\gamma}^{(t)})$ is concave for all values of ζ_j.

APPENDIX B: ANALYTICAL DERIVATIVES

To obtain the quadratic approximation of the expected value of the complete data log-likelihood function for item j, we provide analytical first- and second-order partial derivatives of the expected value of the complete data log-likelihood function for item j. We take the derivatives with respect to a generic item parameter ζ_j holding all other parameters $\boldsymbol{\gamma}_{j\zeta_j}$ at their current value. In the following, we first provide the mathematical steps taken to obtain the first and second partial derivatives, and then summarize these results by noting that the analytical derivatives are identical to numerically approximated derivatives. Ignoring the latent distribution in Equation 19, the first partial derivative is written

$$\frac{\partial Q_j\left(\boldsymbol{\gamma}_j, \zeta_j | \boldsymbol{\gamma}_j{}^{(t)}, \zeta_j{}^{(t)}\right)}{\partial \zeta_j} = \frac{\partial}{\partial \zeta_j}\left[\sum_{i=1}^{N}\sum_{q=1}^{Q} \log\left[f\left(y_{ij}\middle|\ Z_q, \mathbf{x}_i; \$_j{}^{(t)}\right)\right] P\left(Z_q\middle|\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\gamma}_j{}^{(t)}\right)\right], \tag{28}$$

which, for Bernoulli-distributed item responses, expands to

$$= \frac{\partial}{\partial \zeta_j}\left[\sum_{i=1}^{N}\sum_{q=1}^{Q} \left(y_{ij} \log\ \left[P\left(y_{ij}=1|Z_q, \mathbf{x}_i;\ \boldsymbol{\varpi}_j\right)\right] + \left(1-y_{ij}\right)\log\left[1-P\left(y_{ij}=1|Z_q, \mathbf{x}_i; \boldsymbol{\varpi}_j\right)\right]\right) P\left(Z_q\middle|\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\gamma}_j{}^{(t)}\right)\right], \tag{29}$$

$$= \sum_{i=1}^{N}\sum_{q=1}^{Q}\left[\frac{\partial}{\partial \zeta_j}\left(y_{ij} \log\ \left[P\left(y_{ij}=1|Z_q, \mathbf{x}_i;\ \boldsymbol{\varpi}_j\right)\right] + \left(1-y_{ij}\right)\log\left[1-P\left(y_{ij}=1|Z_q, \mathbf{x}_i; \boldsymbol{\varpi}_j\right)\right]\right)\right] P\left(Z_q\middle|\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\gamma}_j{}^{(t)}\right) \tag{30}$$

where y_{ij} is an indicator variable for whether individual i endorsed item j. Note that because the posterior probabilities $P\left(Z_q\middle|\mathbf{y}_i, \mathbf{x}_i; \mathbf{fl}_j{}^{(t)}\right)$ are fixed and known from the E-step, the partial derivative of this density equals zero. Therefore, using the chain rule, we obtain

$$= \sum_{i=1}^{N}\sum_{q=1}^{Q}\left[y_{ij}\frac{\partial \eta_{ij}}{\partial \zeta_j}\left[1-P\left(y_{ij}=1|Z_q, \mathbf{x}_i; \boldsymbol{\varpi}_j\right)\right] + \left(y_{ij}-1\right)\frac{\partial \eta_{ij}}{\partial \zeta_j}\left[P\left(y_{ij}=1|Z_q, \mathbf{x}_i; \boldsymbol{\varpi}_j\right)\right]\right] P\left(Z_q\middle|\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\gamma}_j{}^{(t)}\right), \tag{31}$$

$$= \sum_{i=1}^{N}\sum_{q=1}^{Q}\frac{\partial \eta_{ij}}{\partial \zeta_j}\left[y_{ij} - P\left(y_{ij}=1|Z_q, \mathbf{x}_i; \boldsymbol{\varpi}_j\right)\right] P\left(Z_q\middle|\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\gamma}_j{}^{(t)}\right). \tag{32}$$

The partial derivative of the linear predictor $\eta_{ij} = \left(\nu_{0j} + \mathbf{x}_i\boldsymbol{\nu}_{1j}\right) + \left(\lambda_{0j} + \mathbf{x}_i\boldsymbol{\lambda}_{1j}\right)\theta_i$ is easily computed. If $\zeta_j = \nu_{0j}$, for instance, then $\frac{\partial \eta_{ij}}{\partial \zeta_j} = 1$. This represents the solved first partial derivative of Q_j with respect to ζ_j.

The second partial derivative of $Q_j\left(\boldsymbol{\gamma}_{j(\zeta_j)}, \zeta_j | \boldsymbol{\gamma}_{j(\zeta_j)}{}^{(t)}, \zeta_j{}^{(t)}\right)$ with respect to ζ_j is

$$\frac{\partial^2 Q_j\left(\boldsymbol{\gamma}_j, \zeta_j | \boldsymbol{\gamma}_j{}^{(t)}, \zeta_j{}^{(t)}\right)}{\partial^2 \zeta_j} = \sum_{i=1}^{N}\sum_{q=1}^{Q} \frac{\partial}{\partial \zeta_j}\left[\frac{\partial \eta_{ij}}{\partial \zeta_j}\left[y_{ij} - P\left(y_{ij}=1|Z_q, \mathbf{x}_i; \boldsymbol{\varpi}_j\right)\right]\right] P\left(Z_q \middle| \mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\gamma}_j{}^{(t)}\right), \tag{33}$$

which is solved as

$$= \sum_{i=1}^{N}\sum_{q=1}^{Q} -\left[\frac{\partial \eta_{ij}}{\partial \zeta_j}\right]^2 P\left(y_{ij}=1|Z_q, \mathbf{x}_i; \boldsymbol{\varpi}_j\right)\left[1 - P\left(y_{ij}=1|Z_q, \mathbf{x}_i; \boldsymbol{\varpi}_j\right)\right] \times P\left(Z_q \middle| \mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\gamma}_j{}^{(t)}\right). \tag{34}$$

We use the first and second derivatives to update each DIF parameter estimate at a time via soft-thresholding and each baseline parameter estimate at a time via coordinate descent optimization.

APPENDIX C: PROOF OF CONCEPT

The data consisted of six Bernoulli distributed item responses from 500 individuals that loaded on a single normally distributed latent factor. Three exogenous variables were generated to have large DIF effects on two of the six item responses (items 4 and 5), as well as large impact on the latent mean and small impact on the latent variance. These exogenous variables were also generated to be background characteristics and thus referred to as study, sex, and age. Furthermore, study and sex were effect coded (-.5,.5), whereas age was continuous and standardized. This implied that the variances for all three background variables were equal to 1, ensuring that penalization was unaffected by scaling.

The model was identified by setting the baseline latent mean and variance parameters to 0 and 1, respectively, and constraining all DIF parameters on Item 1 to zero. This made Item 1 an anchor item. Although neither the previous approach nor the current one requires an a priori designation of anchor items, doing so here allowed us to compare the estimators without penalization (i.e., the tuning parameter was set to zero).

TABLE 6.

ν_{0j} and λ_{0j} are baseline intercept and slope estimates for item j. ν_{kj} and λ_{kj} are DIF intercept and slope estimates for covariate k on item j. EM refers to the Expectation-Maximization algorithm. Observed ML refers to the observed marginal likelihood approach. regDIF (EM estimation) used Guass-Hermite quadrature with 51 points, and the convergence criterion was set to 1×10^{-7} (i.e., change in the sum of the absolute values of the point estimates from iteration t to $t+1$). In contrast, NLMIXED (Observed ML approach) used adaptive quadrature with 21 points, and the convergence criterion was set to 1×10^{-11}.

Item	Parameter	EM	Observed ML	Item	Parameter	EM	Observed ML
1	ν_{01}	−.534	−.533	4	ν_{04}	−1.711	−1.717
	λ_{01}	1.023	1.023		λ_{04}	1.656	1.655
	ν_{11}	-	-		ν_{14}	.195	.195
	ν_{21}	-	-		ν_{24}	−.096	−.096
	ν_{31}	-	-		ν_{34}	.056	.056
	λ_{11}	-	-		λ_{14}	−.002	−.003
	λ_{21}	-	-		λ_{24}	.089	.089
	λ_{31}	-	-		λ_{34}	−.791	−.791
2	ν_{02}	−.843	−.843	5	ν_{05}	−1.776	−1.776
	λ_{02}	1.060	1.059		λ_{05}	1.583	1.583
	ν_{12}	.078	.078		ν_{15}	−.325	−.325
	ν_{22}	−.070	−.070		ν_{25}	−.840	−.841
	ν_{32}	−.051	−.051		ν_{35}	.693	.692
	λ_{12}	.022	.022		λ_{15}	−.025	−.026
	λ_{22}	.145	.145		λ_{25}	−.591	−.591
	λ_{32}	−.150	−.151		λ_{35}	−.212	−.211
3	ν_{03}	−1.243	−1.242	6	ν_{06}	−2.026	−2.025
	λ_{03}	1.375	1.374		λ_{06}	1.510	1.509
	ν_{13}	−.237	−.237		ν_{16}	−.020	−.020
	ν_{23}	.301	.301		ν_{26}	.187	.186
	ν_{33}	−.156	−.156		ν_{36}	.120	.120
	λ_{13}	−.066	−.067		λ_{16}	.014	.013
	λ_{23}	−.311	−.311		λ_{26}	−.234	−.234
	λ_{33}	−.636	−.636		λ_{36}	−.484	−.485