stats

MDPI

Article

Comparing Robust Linking and Regularized Estimation for Linking Two Groups in the 1PL and 2PL Models in the Presence of Sparse Uniform Differential Item Functioning

Alexander Robitzsch [1,2]

[1] IPN—Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118 Kiel, Germany; robitzsch@leibniz-ipn.de
[2] Centre for International Student Assessment (ZIB), Olshausenstraße 62, 24118 Kiel, Germany

Abstract: In the social sciences, the performance of two groups is frequently compared based on a cognitive test involving binary items. Item response models are often utilized for comparing the two groups. However, the presence of differential item functioning (DIF) can impact group comparisons. In order to avoid the biased estimation of groups, appropriate statistical methods for handling differential item functioning are required. This article compares the performance-regularized estimation and several robust linking approaches in three simulation studies that address the one-parameter logistic (1PL) and two-parameter logistic (2PL) models, respectively. It turned out that robust linking approaches are at least as effective as the regularized estimation approach in most of the conditions in the simulation studies.

Keywords: item response model; robust linking; regularization; differential item functioning

check for updates

Citation: Robitzsch, A. Comparing Robust Linking and Regularized Estimation for Linking Two Groups in the 1PL and 2PL Models in the Presence of Sparse Uniform Differential Item Functioning. *Stats* **2023**, *6*, 192–208. https://doi.org/10.3390/stats6010012

Academic Editor: Dungang Liu

Received: 8 December 2022
Revised: 18 January 2023
Accepted: 20 January 2023
Published: 25 January 2023

1. Introduction

Item response theory (IRT) models [1,2] are an important class of multivariate statistical models for analyzing dichotomous random variables used to model testing data from social science applications. Of vital importance is the application of item response models in educational large-scale assessment (LSA; [3,4]), such as the programme for international student assessment (PISA; [5]) study.

In this article, we only investigate unidimensional IRT models. Let $\boldsymbol{X} = (X_1, \ldots, X_I)$ be the vector of I dichotomous random variables $X_i \in \{0, 1\}$ (also referred to as test items or items). A unidimensional item response model [6] is a statistical model for the probability distribution $\mathrm{P}(\boldsymbol{X} = \boldsymbol{x})$ for the vector $\boldsymbol{x} = (x_1, \ldots, x_I) \in \{0, 1\}^I$, where

$$\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\delta}, \boldsymbol{\gamma}) = \int_{-\infty}^{\infty} \prod_{i=1}^{I} \left[P_i(\theta; \gamma_i)^{x_i} (1 - P_i(\theta; \gamma_i))^{1-x_i} \right] \phi(\theta; \mu, \sigma) \, \mathrm{d}\theta \,, \tag{1}$$

where ϕ denotes the density of the normal distribution with mean μ and standard deviation σ. The vector $\boldsymbol{\delta} = (\mu, \sigma)$ contains the distribution parameters of the ability variable θ. The vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_I)$ contains all estimated item parameters of item response functions $P_i(\theta; \gamma_i) = \mathrm{P}(X_i = 1 | \theta)$.

Different IRT models emerge by choosing particular item response functions P_i in (1). The one-parameter logistic (1PL) model (also referred to as the Rasch model; [7]) employs the item response function $P_i(\theta) = \Psi(\theta - b_i)$, where Ψ denotes the logistic distribution function, and b_i is the item difficulty of item i (i.e., $\gamma_i = (b_i)$). The two-parameter logistic (2PL) model [8] additionally includes the item discrimination a_i (i.e., $\gamma_i = (a_i, b_i)$), and the item response function is given by $P_i(\theta) = \Psi(a_i(\theta - b_i))$.

Note that distribution parameters $\boldsymbol{\delta}$ and item parameters $\boldsymbol{\gamma}$ cannot be simultaneously identified. For example, in the 2PL model, the mean μ and the standard deviation σ must

be fixed to 0 or 1, respectively, if all item discriminations a_i and item difficulties b_i were estimated. As an alternative, one could fix at least (or all) item parameters to predefined values to enable the estimation of distribution parameters.

In this article, we investigate the estimation of the 1PL and 2PL models in the presence of two groups. In this case, we are primarily interested in comparing the mean μ_2 of the second group with the mean μ_1 of the first group. In order to enable group comparisons, assumptions on item parameters must be imposed in the 1PL and 2PL models. First, one must fix μ_1 to zero. Second, some assumptions on item parameters must be made to enable the identification of the distribution parameters of the second group. A typical assumption is the measurement invariance assumption [9,10], which says that item parameters do not differ across the two groups. However, in practical applications, some item parameters will almost always differ across groups. This situation is also referred to as differential item functioning (DIF; [11,12]). In this article, we assume that there could be DIF in item difficulties in the 1PL and 2PL models (i.e., uniform DIF; [12]). The classes of techniques of robust linking and regularized estimation are compared through three simulation studies regarding the performance of parameter recovery of the group comparison.

The rest of the article is structured as follows. Different statistical methods for group comparisons in the 1PL and the 2PL models in the presence of uniform DIF are discussed in Section 2. Section 3 presents results from Simulation Study 1, which investigates the group comparison in the 1PL model in the presence of DIF. Section 4 more thoroughly investigates the choice of tuning parameters for regularized estimation in a subset of conditions of Simulation Study 1 in the Focused Simulation Study 1A. Section 5 presents results from Simulation Study 2, which investigates the group comparison in the 2PL model in the presence of uniform DIF effects. Finally, the article closes with a discussion in Section 6.

2. Two-Group Comparison under Sparse DIF

We now present IRT estimation in two groups $g = 1, 2$. Let $\boldsymbol{X}_{pg} = (X_{pg1}, \ldots, X_{pgI})$ of person $p = 1, \ldots, N_g$ in group $g = 1, 2$. We now define the log-likelihood function for data $\mathcal{D}_g = (\boldsymbol{X}_{1g}, \ldots, \boldsymbol{X}_{N_g g})$ in group g ($g = 1, 2$) as

$$l(\mu_g, \sigma_g, \boldsymbol{a}_g, \boldsymbol{b}_g; \mathcal{D}_g) = \sum_{p=1}^{N_g} \log \left[\int \prod_{i=1}^{I} P_i(x_{pgi}, \theta; a_{ig}, b_{ig}) \phi(\theta; \mu_g, \sigma_g) \, \mathrm{d}\theta \right] , \tag{2}$$

where $P_i(1, \theta; a_i, b_i) = \Psi(a_i(\theta - b_i))$ and $P_i(0, \theta; a_i, b_i) = 1 - P_i(1, \theta; a_i, b_i)$ is the 2PL model, and the vectors of item parameters are defined as $\boldsymbol{a}_g = (a_{1g}, \ldots, a_{Ig})$ and $\boldsymbol{b}_g = (b_{1g}, \ldots, b_{Ig})$

For reasons of identification, we fix the mean μ_1 in the first group to zero and the standard deviation σ_1 to 1. We assume that the 2PL (or the 1PL model as a restricted version) holds in both groups, and there is uniform DIF in item difficulties (i.e., equal item discriminations $a_{i1} = a_{i2}$ in the two groups are assumed). That is, we model the difference in item difficulties as

$$b_{i2} = b_{i1} + e_i \, . \tag{3}$$

We assume that DIF effects are fixed (see [13] for a random DIF perspective). Throughout this paper, we impose a sparsity assumption on DIF effects e_i. In this case, the majority of items have DIF effects of zero, while only a few DIF effects differ from zero [14,15]. The situation is known as partial invariance [16–19]. Hence, DIF effects can be regarded as outliers that might bias the estimation of group mean differences [14,20–24].

In the next subsections, we discuss alternative methods for two-group comparisons in the 1PL and 2PL models in the presence of uniform DIF. In Section 2.1, concurrent calibration relying on invariant item parameters is discussed. Section 2.2 investigates regularization approaches to handling DIF in two-group comparisons. Robust linking approaches are treated in Section 2.3. Finally, relationships between regularization and robust linking are highlighted in Section 2.4.

2.1. Concurrent Calibration

Concurrent calibration jointly estimates common (group-invariant) item parameters and distribution parameters in one estimation run. This property is particularly convenient for practitioners because no additional steps or postprocessing of model results is required. We now present concurrent calibration separately for the 1PL and 2PL models.

2.1.1. 1PL Model

We now estimate the distribution parameters of the second group (i.e., μ_2 and σ_2) with a concurrent calibration approach in the 1PL model. In the 1PL model, item discriminations are set to one and indicate this using the notation $\boldsymbol{a}_g = \mathbf{1}$ of a vector of ones. The distribution parameters and the vector of common item difficulties are estimated by minimizing the negative of the log-likelihood function (see (2))

$$(\hat{\mu}_2, \hat{\sigma}_2, \hat{\sigma}_1, \hat{\boldsymbol{b}}) = \underset{(\mu_2, \sigma_2, \sigma_1, \boldsymbol{b})}{\arg\min} \left\{ -l(0, \sigma_1, \mathbf{1}, \boldsymbol{b}; \mathcal{D}_1) - l(\mu_2, \sigma_2, \mathbf{1}, \boldsymbol{b}; \mathcal{D}_2) \right\} . \tag{4}$$

Note that the minimization in (4) assumes invariant item difficulties $\boldsymbol{b}$ across the two groups. In the presence of uniform DIF, the log-likelihood function in (4) is misspecified. However, under certain conditions, it is possible that group differences could nevertheless be unbiasedly estimated [25,26].

2.1.2. 2PL Model

We now weaken the assumption of equal item discriminations in the 2PL model. Common item discriminations $\boldsymbol{a}$ and item difficulties $\boldsymbol{b}$ are estimated by minimizing the estimation function

$$(\hat{\mu}_2, \hat{\sigma}_2, \hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}) = \underset{(\mu_2, \sigma_2, \boldsymbol{a}, \boldsymbol{b})}{\arg\min} \left\{ -l(0, 1, \boldsymbol{a}, \boldsymbol{b}; \mathcal{D}_1) - l(\mu_2, \sigma_2, \boldsymbol{a}, \boldsymbol{b}; \mathcal{D}_2) \right\} . \tag{5}$$

Like in the 1PL model, the log-likelihood function in (5) is misspecified in the presence of uniform DIF. Nevertheless, it is interesting to empirically investigate situations in which misspecified concurrent calibration can provide approximately unbiased results.

2.2. Regularization Approaches

In practical applications, all items could be prone to DIF effects. However, modeling all DIF effects without any constraints leads to an unidentified IRT model. To circumvent this issue, regularization techniques have been proposed in statistics to estimate nonidentified models under sparsity assumptions [27,28]. The main idea of using regularization techniques for multiple-group IRT estimation is that by adding an appropriate penalty term to the negative log-likelihood function, some simplified structure on DIF effects is imposed. It can be shown that under sparse DIF effects, regularized estimation provides unbiased group means [20]. Regularized estimation recently became popular in psychometrics, such as item response modeling [29,30], structural equation modeling [31–33], structured latent class analysis [34–36], and mixture models [37,38]. The investigation of regularization approaches of known demographic groups, such as gender or language groups, is an important topic in educational measurement. Moreover, regularization techniques were recently discussed for manifest DIF detection in the 1PL and 2PL models [19,20,39–44].

For a scalar parameter x, lasso penalty is a popular penalty function used in regularization [28], and it is defined as

$$\mathcal{P}_{\text{Lasso}}(x, \lambda) = \lambda |x| , \tag{6}$$

where λ is a non-negative regularization parameter that controls the extent of regularization. It is known that the lasso penalty induces bias in estimated parameters. To circumvent this

issue, the smoothly clipped absolute deviation (SCAD; [45]) penalty has been proposed. It is defined by

$$\mathcal{P}_{\text{SCAD}}(x, \lambda, a) = \begin{cases} \lambda|x| & \text{if } |x| < \lambda \\ -(x^2 - 2a\lambda|x|^2 + \lambda^2)(2(a-1))^{-1} & \text{if } \lambda \leq |x| \leq a\lambda \\ (a+1)\lambda^2 & \text{if } |x| > a\lambda \end{cases} \tag{7}$$

with $a > 2$. In many studies, the recommended value of $a = 3.7$ (see [45]) has been adopted (e.g., [27,36,38,46–48]). However, other studies considered the simultaneous selection of both tuning parameters λ and a [31,49–51].

Figure 1 displays the SCAD penalty function for different values of a with a fixed λ value of 0.2. The SCAD penalty retains the penalization rate and the induced bias of the lasso for model parameters close to zero, but continuously relaxes the rate of penalization as the absolute value of the model parameters increases. Note that $\mathcal{P}_{\text{SCAD}}$ has the property of the lasso penalty around zero, but has zero derivatives for x values that strongly differ from zero. In contrast, the derivative of the lasso penalty is 1 or -1 for positive and negative x values, respectively.

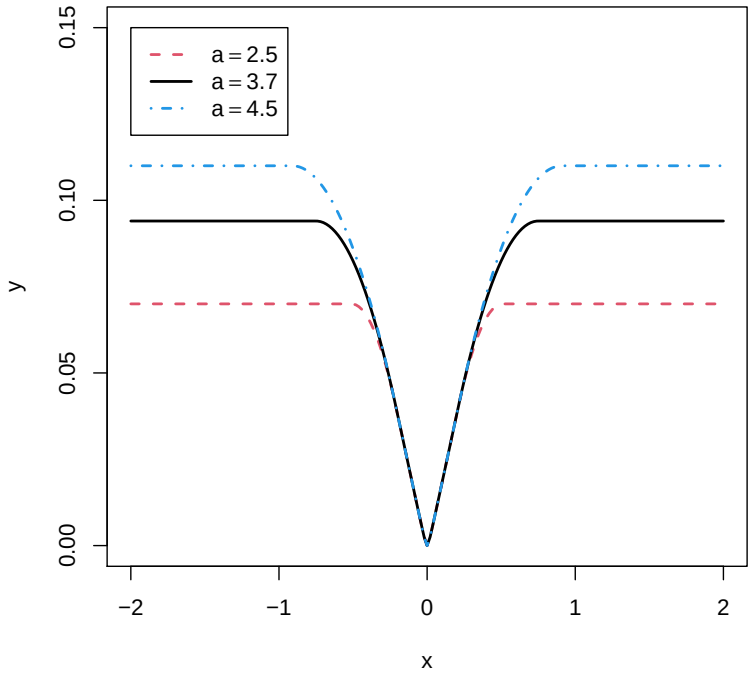

Figure 1. SCAD penalty function $\mathcal{P}_{\text{SCAD}}$ for different values of a for $\lambda = 0.2$.

We now present regularization estimation under uniform DIF for the 1PL and the 2PL model.

2.2.1. 1PL Model

In regularization, we specify an overidentified IRT model but enable the identification of distribution parameters of the second group (i.e., μ_2 and σ_2) by adding a penalty function to the negative log-likelihood function. Compared with (4), we additionally introduce the vector of DIF effects $\boldsymbol{e} = (e_1, \ldots, e_I)$ and consider the minimization problem

$$(\hat{\mu}_2, \hat{\sigma}_2, \hat{\sigma}_1, \hat{\boldsymbol{b}}, \hat{\boldsymbol{e}}) = \underset{(\mu_2, \sigma_2, \sigma_1, \boldsymbol{b}, \boldsymbol{e})}{\arg\min} \left\{ -l(0, \sigma_1, \mathbf{1}, \boldsymbol{b}; \mathcal{D}_1) - l(\mu_2, \sigma_2, \mathbf{1}, \boldsymbol{b} + \boldsymbol{e}; \mathcal{D}_2) + N^* \sum_{i=1}^{I} \mathcal{P}_{\text{SCAD}}(e_i, \lambda, a) \right\}, \tag{8}$$

where $N^* = (N_1 + N_2)/2$ includes the sample sizes in the penalty term.

In practice, the minimization of (8) for a fixed value of λ results in a subset of DIF effects that are different from zero, where the rest of the DIF effects has been set to zero. In essence, the group difference is only based on those items whose DIF effects e_i are estimated equal to zero.

Typically, the regularization parameter λ is an unknown nuisance parameter in (8) that must also be estimated. In practice, the minimization of (8) is carried out on a discrete grid of λ values, and the optimal regularization parameter λ_{opt} is selected that minimizes the Akaike information criterion (AIC) or the Bayesian information criterion (BIC).

The regularized estimation problem (8) can be minimized using marginal maximum likelihood estimation and the expectation maximization (EM) algorithm [30,36,46]. The EM algorithm alternates between the E-step and the M-step. The E-step computation is identical to the estimation in nonregularized item response models. In the M-step, the minimizing of the regularized negative log-likelihood function is carried out using expected counts that are computed in the expected log-likelihood function. The difference between regularized estimation and ordinary maximum likelihood estimation is that the optimization function becomes nondifferentiable, because the SCAD penalty is nondifferentiable. The optimization of nondifferentiable optimization can be performed using gradient descent [28] approaches or by substituting the nondifferentiable optimization functions with differentiable approximating functions [36,52–54]. In our experience, the latter approach performs quite satisfactorily in applications.

2.2.2. 2PL Model

We now turn to regularized estimation in the 2PL model. Like in the 1PL model, the overidentified vector of the DIF effects $\boldsymbol{e}$ is introduced. In addition, the vector of common item discriminations $\boldsymbol{a}$ is estimated. The following estimation function is minimized for determining group differences

$$(\hat{\mu}_2, \hat{\sigma}_2, \hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{e}}) = \underset{(\mu_2, \sigma_2, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{e})}{\arg\min} \left\{ -l(0, 1, \boldsymbol{a}, \boldsymbol{b}; \mathcal{D}_1) - l(\mu_2, \sigma_2, \boldsymbol{a}, \boldsymbol{b} + \boldsymbol{e}; \mathcal{D}_2) + N^* \sum_{i=1}^{I} \mathcal{P}_{\text{SCAD}}(e_i, \lambda, a) \right\}. \tag{9}$$

In principle, regularized estimation in the 2PL model using common item discriminations is not different from regularized estimation in the 1PL model. The principle of regularization can be extended to modeling DIF effects in item discriminations by introducing additional penalty terms for these DIF effects [44].

2.3. Robust Linking Approaches

Linking methods are typically two-step methods that separately estimate IRT models, such as the 1PL or 2PL models in each of the groups, and compute group distribution parameters (i.e., group means and standard deviations) in a second step [25,55–57]. We replace the separate estimation with a simultaneous estimation in a two-group IRT model. We differ from this usual setup in this article for two reasons. First, we want to highlight the notational similarity to regularized estimation. Second, we also want to conduct linking in the 2PL model under the assumption of common item discriminations if there is a strong belief that most of the DIF effects can be attributed to uniform instead of nonuniform DIF effects. Robust linking as a concept refers to the property that group differences can be estimated without bias (or only with small bias) despite the presence of (uniform) DIF [25,58].

2.3.1. 1PL Model

We now present robust linking approaches in the 1PL model. In the first step, we consider the minimization problem

$$(\hat{\sigma}_2, \hat{\sigma}_1, \hat{\boldsymbol{b}}, \hat{\boldsymbol{e}}) = \underset{(\sigma_2, \sigma_1, \boldsymbol{b}, \boldsymbol{e})}{\arg\min} \left\{ -l(0, \sigma_1, \boldsymbol{1}, \boldsymbol{b}; \mathcal{D}_1) - l(0, \sigma_2, \boldsymbol{1}, \boldsymbol{b} + \boldsymbol{e}; \mathcal{D}_2) \right\}. \tag{10}$$

Note that (10) differs from regularized estimation (8) in two aspects. First, it does not involve a penalty function. Second, it does not provide an estimate of the group mean μ_2 of the second group. We want to point out that the estimates from the minimization (10) are equivalent to the separate estimation of the 1PL model with item difficulties $\hat{\boldsymbol{b}}$ in the first group and difficulties $\hat{\boldsymbol{b}} + \hat{\boldsymbol{e}}$ in the second group.

Robust (and nonrobust) linking methods employ estimated DIF effects $\hat{\boldsymbol{e}} = (\hat{e}_1, \ldots, \hat{e}_I)$ to determine a group mean estimate $\hat{\mu}_2$.

Robust Linking Using the L_p Loss Function

The first robust linking method uses the L_p loss function $\rho(x) = |x|^p$ for $p > 0$ (see [25,59,60]) and determines the group mean estimate as

$$\hat{\mu}_2 = \arg\min_{\mu_2} \left\{ \sum_{i=1}^{I} \rho(\hat{e}_i + \mu_2) \right\} . \tag{11}$$

Mean–mean linking (MM; [55,61]) results with $p = 2$. The case $p = 1$ corresponds to median linking [20] and is expected to be more robust than $p = 2$. Finally, the case $p = 0.5$ corresponds to the loss function used in invariance alignment [59,62,63]. In optimization, one can replace the nondifferentiable function ρ with $\tilde{\rho}(x) = (x^2 + \varepsilon)^{p/2}$ using a sufficiently small $\varepsilon > 0$, such as $\varepsilon = 0.01$ [59].

Robust Linking Using the MAD Statistic

Researchers Matthias von Davier and Bezirhan proposed robust outlier detection of items with large DIF effects and removed them from linking [64] (see also [61,65,66]). For each item i, a z statistic is defined by

$$z_i = \frac{\hat{e}_i - \text{Mdn}(\hat{\boldsymbol{e}})}{\text{MAD}(\hat{\boldsymbol{e}})} , \tag{12}$$

where $\text{Mdn}(\hat{\boldsymbol{e}})$ denotes the median of the vector $\hat{\boldsymbol{e}}$ of estimated DIF effects, and MAD is the (scaled) median absolute deviation of the DIF effects. An item is declared an outlier (because of possessing a large DIF effect) if $|z_i|$ exceeds the cutoff of 2.7 [64]. The group mean difference is defined as

$$\hat{\mu}_2 = -\frac{\sum_{i=1}^{I} w_i \hat{e}_i}{\sum_{i=1}^{I} w_i} \text{ with } w_i = \mathbf{1}_{\{|z_i| \leq 2.7\}} . \tag{13}$$

Obviously, items with large DIF effects are removed from the computation of the group mean in (13).

2.3.2. 2PL Model

Linking based on the 2PL model can be applied in two variants. The first approach relies on a first simultaneous estimation step in which common item discriminations $\boldsymbol{a}$ are estimated. It is well known that the 2PL model gets unstable with small samples, which motivates the estimation of a simplified model with joint item discriminations. In the first step of the linking approach, we determine item parameters and the standard deviation in the second group by

$$(\hat{\sigma}_2, \hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{e}}) = \arg\min_{(\sigma_2, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{e})} \left\{ -l(0, 1, \boldsymbol{a}, \boldsymbol{b}; \mathcal{D}_1) - l(0, \sigma_2, \boldsymbol{a}, \boldsymbol{b} + \boldsymbol{e}; \mathcal{D}_2) \right\} . \tag{14}$$

Alternatively, separate estimation of the 2PL model in the two groups can be conducted, which results in estimated item parameters $\hat{\boldsymbol{a}}_1$ and $\hat{\boldsymbol{b}}_1$ and $\hat{\boldsymbol{a}}_2$ and $\hat{\boldsymbol{b}}_2$, respectively.

Robust Linking Using L_p Loss Function or MAD Statistic

The estimation of μ_2 based on the L_p loss function and outlier removal can also be based on the estimated vector $\hat{\boldsymbol{e}}$. No change in formulas in (11), (12), and (13) is required.

Joint Haberman Linking Using Common Item Discriminations

We now show how to perform Haberman linking [67,68] if the first linking step is carried out using common item discriminations (see minimization problem (14). The more general L_p loss function is again used for Haberman linking [60]. In this case, common item difficulties $\tilde{\boldsymbol{b}} = (\tilde{b}_1, \ldots, \tilde{b}_I)$ are estimated by minimizing

$$(\hat{\mu}_2, \hat{\tilde{\boldsymbol{b}}}) = \arg\min_{(\mu_2, \tilde{\boldsymbol{b}})} \Big\{ \sum_{i=1}^{I} \rho(\hat{b}_i - \tilde{b}_i) + \sum_{i=1}^{I} \rho(\hat{b}_i + \hat{e}_i - \tilde{b}_i + \mu_2) \Big\} . \tag{15}$$

Haberman Linking Based on Separate Calibration

We also compare the performance of joint Haberman linking using common item discriminations with the ordinary Haberman linking that is based on separation calibration [60,68]. In the first step of Haberman linking, common logarithmized item discriminations $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_I)$, and the logarithmized standard deviation s_2 of the second group is determined as the minimizer of

$$(\hat{s}_2, \hat{\boldsymbol{\alpha}}) = \arg\min_{(s_2, \boldsymbol{\alpha})} \Big\{ \sum_{i=1}^{I} \rho(\log \hat{a}_{i1} - \alpha_i) + \sum_{i=1}^{I} \rho(\log \hat{a}_{i2} - \alpha_i - s_2) \Big\} . \tag{16}$$

Note that the standard deviation of the second group is given by $\sigma_2 = \exp(s_2)$.

In the second step of Haberman linking, the group means μ_2 of the second group are estimated along with common item difficulties $\boldsymbol{b}_i$

$$(\hat{\mu}_2, \hat{\boldsymbol{b}}) = \arg\min_{(\mu_2, \boldsymbol{b})} \Big\{ \sum_{i=1}^{I} \rho(\hat{b}_{i1} - b_i) + \sum_{i=1}^{I} \rho(\hat{\sigma}_2 \hat{b}_{i2} - b_i + \mu_2) \Big\} . \tag{17}$$

2.4. On the Relation of Robust Linking and Regularized Estimation

It has been argued that robust linking yields very similar results to regularization approaches for linking two groups [20]. We now sketch a heuristic proof of why this is the case. In the regularization approach, one uses the SCAD penalty, which behaves similarly to the lasso penalty $\mathcal{P}_{\text{Lasso}}(x) = \lambda |x|$ for x values close to zero. We now use the notation of ρ for the penalty function in regularized estimation to indicate that a general L_p loss function can be used in regularized estimation. The optimization function in regularized estimation in the 1PL model is then given by

$$(\hat{\mu}_2, \hat{\sigma}_2, \hat{\sigma}_1, \hat{\boldsymbol{b}}, \hat{\boldsymbol{e}}) = \arg\min_{(\mu_2, \sigma_2, \sigma_1, \boldsymbol{b}, \boldsymbol{e})} \Big\{ - l(0, \sigma_1, \mathbf{1}, \boldsymbol{b}; \mathcal{D}_1) - l(\mu_2, \sigma_2, \mathbf{1}, \boldsymbol{b} + \boldsymbol{e}; \mathcal{D}_2) + N^* \lambda \sum_{i=1}^{I} \rho(e_i) \Big\} . \tag{18}$$

We demonstrated that robust linking relies on the first minimization step

$$(\hat{\sigma}_2, \hat{\sigma}_1, \hat{\boldsymbol{b}}, \hat{\boldsymbol{e}}) = \arg\min_{(\sigma_2, \sigma_1, \boldsymbol{b}, \boldsymbol{e})} \Big\{ - l(0, \sigma_1, \mathbf{1}, \boldsymbol{b}; \mathcal{D}_1) - l(0, \sigma_2, \mathbf{1}, \boldsymbol{b} + \boldsymbol{e}; \mathcal{D}_2) \Big\} . \tag{19}$$

Importantly, the log-likelihood in (19) does not change under reparametrization by including the redundant group mean μ_2. We obtain

$$\Big\{ - l(0, \sigma_1, \mathbf{1}, \boldsymbol{b}; \mathcal{D}_1) - l(0, \sigma_2, \mathbf{1}, \boldsymbol{b} + \boldsymbol{e}; \mathcal{D}_2) \Big\} = \Big\{ - l(0, \sigma_1, \mathbf{1}, \boldsymbol{b}; \mathcal{D}_1) - l(\mu_2, \sigma_2, \mathbf{1}, \boldsymbol{b} + \tilde{\boldsymbol{e}}; \mathcal{D}_2) \Big\} \tag{20}$$

for any μ_2 and $\tilde{\boldsymbol{e}} = \mu_2 + \boldsymbol{e}$. Hence, the overidentified right-hand side of (20) can be made identifiable by adding a penalty function to the log-likelihood function. A condition for unbiased estimation of the group mean μ_2 is to define DIF effects e_i in such a way that

$\sum_{i=1}^{I} \rho(e_i) = 0$ is minimal [20]. This condition can also encode sparsity assumptions on DIF effects using the L_p loss function and $p \to 0$. Importantly, robust linking can be defined as

$$\hat{\mu}_2 = \arg\min_{\mu_2} \left\{ \sum_{i=1}^{I} \rho(\hat{e}_i + \hat{\mu}_2) \right\} , \tag{21}$$

which demonstrates the practical equivalence of robust linking and regularized estimation in sufficiently large samples.

3. Simulation Study 1: DIF Effects in the 1PL Model

In this Simulation Study 1, we compare robust linking and regularized estimation in the 1PL model in the presence of DIF effects. We consider the case of two groups.

3.1. Method

In this simulation study, we fixed the number of items to 20 and fixed item difficulties throughout all replications. Item parameters and DIF effects can be found at https://osf.io/tma3f/ (accessed on 8 December 2022). Item difficulties b_i (see column "b") ranged between -1.88 and 1.61, with a mean of 0.00 and a standard deviation of 1.20. Items with DIF effects e_i (column "e") can also be found at https://osf.io/tma3f/ (accessed on 8 December 2022). The DIF effects e_i had values $\{-1, 0, 1\}$. Only 4 out of 20 items were simulated to have DIF effects different from zero.

Item response data were generated according to the 1PL model using item difficulties b_i in the first group and $b_i + \delta e_i$ in the second group. The DIF effect size δ was either 0.5 (i.e., small DIF effects) or 1.0 (i.e., large DIF effects). Furthermore, DIF was chosen to be balanced or unbalanced. In the balanced DIF condition, two items had DIF effects $-\delta$, and two items had DIF effects δ. In the unbalanced DIF condition, all items had DIF effects δ.

The distribution of the ability variable θ was assumed as standard normal (i.e., $\theta \sim \mathrm{N}(0, 1)$) in the first group and had a mean of $\mu_2 = 0.3$ and a standard deviation $\sigma_2 = 1.2$ in the second group. Finally, we varied sample sizes per group N as 500, 1000, 2500, and 5000.

The different linking approaches presented for the 1PL model in Section 2 were evaluated in Simulation Study 1. In the scaling models, we used rectangular integration on a discrete quadrature of 41 equidistant θ points on $[-8.0, 8.0]$. In concurrent calibration, we assumed invariant item parameters across groups. Regularized estimation was carried out using the SCAD penalty, and the regularized model was estimated at a grid of regularization parameters between 1.0 and 0.005 (see replication material on https://osf.io/tma3f/(accessed on 8 December 2022) for specification details). In this simulation study, the tuning parameter a was fixed to 3.7. We chose estimates of the regularization approach using the optimal regularization λ_{opt} based on AIC, BIC, and fixed λ values of 0.05, 0.10, and 0.15 (see [38,69] for a similar approach). Moreover, powers $p = 2$ (L_2; mean–mean linking), $p = 1$ (L_1; median–median linking), and $p = 0.5$ ($L_{0.5}$; invariance alignment) were used in the robust linking approach that utilizes the L_p loss function. Finally, we determined the group mean of the second group on the outlier removal approach using the MAD statistic and a cutoff value of 2.7.

In total, 3000 replications were conducted in each simulation condition. We assessed the bias and root mean square error (RMSE) of the estimated group mean $\hat{\mu}_2$. In each of the R replications in a simulation condition, the group mean $\hat{\mu}_{2r}$ ($r = 1, \ldots, R$) was estimated. The bias was estimated by

$$\text{Bias} = \frac{1}{R} \sum_{r=1}^{R} (\hat{\mu}_{2r} - \mu_2) \, . \tag{22}$$

The RMSE was estimated by

$$\text{RMSE} = \frac{1}{R}\sum_{r=1}^{R}(\hat{\mu}_{2r} - \mu_2)^2 \,. \tag{23}$$

To ease the comparability of the RMSE between different methods across sample sizes, we used a relative RMSE in which we divided the RMSE of a particular method by the RMSE of the best-performing method in a simulation condition. Hence, a relative RMSE of 100 is the reference value for the best-performing method.

The statistical software R [70] was employed for all parts of the simulation. Concurrent calibration in a multiple-group IRT model was estimated using the TAM package [71]. Regularized estimation was carried out using the `xxirt` function in the sirt package [72]. Replication material can be found at https://osf.io/tma3f/ (accessed on 8 December 2022).

3.2. Results

In Table 1, the bias of the estimated group mean $\hat{\mu}_2$ as a function of the size of the DIF effect δ and the sample size N is presented. It turned out that all methods performed well in the case of balanced DIF.

Table 1. Simulation Study 1: Bias of estimated group means for balanced and unbalanced DIF effects as a function of the size of DIF effects δ and sample size N.

			Choice of λ								
δ	N	**MAD**	**AIC**	**BIC**	**0.05**	**0.10**	**0.15**	$L_{0.5}$	L_1	L_2	**CC**
		Balanced DIF									
0.5	500	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	2500	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	5000	0.00	0.00	0.00	0.00	0.00	−0.01	0.00	0.00	0.00	0.00
1.0	500	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	2500	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	5000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		Unbalanced DIF									
0.5	500	**−0.06**	−0.02	**−0.03**	**−0.04**	−0.02	**−0.06**	**−0.03**	**−0.05**	**−0.10**	**−0.09**
	1000	**−0.03**	−0.02	−0.01	−0.01	−0.01	**−0.08**	−0.02	**−0.04**	**−0.10**	**−0.10**
	2500	0.00	−0.02	−0.01	−0.01	−0.01	**−0.09**	−0.01	−0.02	**−0.10**	**−0.10**
	5000	0.00	−0.01	−0.01	−0.01	−0.01	**−0.09**	−0.01	−0.02	**−0.10**	**−0.10**
1.0	500	−0.01	−0.02	0.00	**−0.05**	0.00	0.00	−0.02	**−0.05**	**−0.20**	**−0.18**
	1000	0.00	−0.02	0.00	−0.02	0.00	0.00	−0.01	**−0.04**	**−0.20**	**−0.18**
	2500	0.00	−0.01	0.00	−0.02	0.00	0.00	−0.01	−0.02	**−0.20**	**−0.18**
	5000	0.00	−0.01	0.00	**−0.03**	0.00	0.00	−0.01	−0.02	**−0.20**	**−0.18**

Note. MAD = robust linking using the MAD statistic; Choice of λ = method or value for determining the regularization parameter λ; L_p = linking employing the L_p loss function with $p = 0.5$, 1.0, or 2.0; CC = concurrent calibration assuming invariant item parameters; Absolute biases larger than 0.03 are printed in bold.

In the case of unbalanced DIF, nonrobust linking methods such as concurrent calibration (CC) and mean–mean linking (L_2) were substantially biased. Interestingly, there was also a bias for median–median linking (L_1) with unbalanced DIF. However, the bias decreased in larger samples. This can be expected because median–median linking (i.e., L_1 linking) recovers the true group difference in infinite sample sizes. In finite samples, estimated DIF effects with a true value of 0 will differ from 0, but have an expected mean of 0. As a consequence, the median of all estimated DIF effects (DIF items and non-DIF items) will be negative in the case of unbalanced DIF, because estimated DIF effects for DIF items are positive, resulting in a negative bias of the estimated group mean of the second group.

The robust $L_{0.5}$ linking approach performed well regarding the bias, particularly in larger samples. Overall, the estimated group means tended to be less biased when the BIC instead of the AIC was used in regularized estimation. However, the bias for robust linking and regularization approaches under unbalanced DIF turned out to be larger in conditions with small DIF effects of $\delta = 0.5$ compared with large DIF effects of $\delta = 1.0$.

Table 2 presents the relative RMSE as a function of the size of DIF effects δ and the sample size N. It can be seen that CC is the frontrunner in terms of RMSE in the condition of unbalanced DIF. Notably, regularized estimation was particularly inefficient in larger samples of $N = 2500$ or $N = 5000$ in the condition of a small DIF effect. Furthermore, the robust linking approach based on the MAD statistic that performs outlier removal showed less variability than linking based on the $L_{0.5}$ invariance alignment loss function.

Table 2. Simulation Study 1: Relative root mean sqaure error (RMSE) of estimated group means for balanced and unbalanced DIF effects as a function of the size of DIF effects δ and sample size N.

			Choice of λ								
δ	N	**MAD**	**AIC**	**BIC**	**0.05**	**0.10**	**0.15**	$L_{0.5}$	L_1	L_2	**CC**
		Balanced DIF									
0.5	500	111	115	111	120	110	111	122	110	101	100
	1000	111	112	106	109	108	118	115	108	101	100
	2500	104	**133**	103	122	118	**135**	114	108	101	100
	5000	103	**147**	**129**	**139**	**126**	**153**	111	107	100	100
1.0	500	107	111	105	115	106	104	118	109	102	100
	1000	104	110	103	108	103	103	115	108	102	100
	2500	105	113	104	112	103	103	114	109	102	100
	5000	104	111	103	**126**	103	103	112	108	102	100
		Unbalanced DIF									
0.5	500	120	108	104	120	100	117	113	110	**142**	**138**
	1000	124	**126**	100	122	108	**164**	117	119	**196**	**187**
	2500	102	**203**	**185**	**194**	**161**	**258**	114	120	**288**	**274**
	5000	100	**249**	**240**	**243**	**217**	**374**	114	124	**408**	**383**
1.0	500	108	109	100	**141**	101	100	122	121	**270**	**247**
	1000	100	113	100	125	100	100	117	122	**370**	**336**
	2500	101	113	101	**244**	103	100	113	121	**572**	**519**
	5000	100	115	100	**352**	103	100	111	124	**808**	**730**

Note. MAD = robust linking using the MAD statistic; Choice of λ = method or value for determining the regularization parameter λ; L_p = linking employing the L_p loss function with $p = 0.5$, 1.0, or 2.0; CC = concurrent calibration assuming invariant item parameters; Relative RMSE values larger than 125 are printed in bold.

The situation changed in the conditions of unbalanced DIF. As expected, nonrobust linking approaches CC and L_2 had large RMSE values. In the case of small DIF effects (i.e., $\delta = 0.5$), the performance of the methods depended on the sample size. For moderate sample sizes, $N = 500$ and $N = 1000$, regularized estimation based on the BIC was satisfactory, while the MAD approach would be preferred in larger sample sizes $N = 2500$ and $N = 5000$. Nevertheless, the $L_{0.5}$ robust linking approach was quite competitive across all sample sizes for small DIF effects. For large unbalanced DIF effects (i.e., $\delta = 1.0$), regularized estimation based on BIC and robust linking using the MAD statistic outperformed other methods. Interestingly, regularized estimation using fixed λ values of 0.10 or 0.15 also yielded satisfactory group mean estimates. Notably, the invariance alignment $L_{0.5}$ approach had some efficiency loss but might still be considered a viable alternative to the MAD linking and regularization approach.

4. Focused Simulation Study 1A: Optimal Choice of two Tuning Parameters for the SCAD Penalty

In this Focused Simulation Study 1A, we additionally investigated the impact of different choices of the second tuning parameter a for the SCAD penalty.

4.1. Method

This focused simulation study only considered the impact of the tuning parameter a for the SCAD penalty in regularized estimation. We only investigated selected conditions of Simulation Study 1. We focused on unbalanced DIF and chose sample sizes $N = 500$, 1000, and 2000. Regularized estimation with the SCAD penalty was carried out using different a values of 2.2, 2.5, 3, 3.7, 4.5, 6, and 9. The same grid of λ values as in Simulation Study 1 was employed (see Section 3.1). For each fixed a value, the optimal λ value was determined by means AIC or BIC. In addition, we also determined estimated group means by choosing the optimal pair $(\lambda_{\text{opt}}, a_{\text{opt}})$ for AIC and BIC.

The performance of the different analytical choices was evaluated by using a relative RMSE. The relative RMSE was obtained by dividing the RMSE of a method by the RMSE of the best-performing method. The best-performing method could either be a regularized estimation with a particular choice of λ and a or a regularized estimation with an optimal choice of λ or a using AIC or BIC.

4.2. Results

Table 3 presents the relative RMSE of estimated group means for different choices of a and λ. Like in Simulation Study 1, the RMSE was smaller for methods that relied on BIC than on AIC. For methods based on AIC, it turned out that using the optimal a parameter across a range of a values resulted in the least RMSE for $N = 500$ and $N = 1000$. However, this was not the case for $N = 2500$. Across different conditions, the choice $a = 3.7$ did not result in estimated group means with the least RMSE values. This finding also occurred for regularized estimation based on BIC. However, differences between different choices of a values turned out to be smaller. For $N = 500$ or $N = 1000$, the choice of a for the SCAD penalty does not seem to matter. In contrast, using the optimal a value resulted in a substantial RMSE decrease for small DIF effects (i.e., $\delta = 0.5$). On the other hand, using the optimal a value resulted in an RMSE increase for $N = 2500$ in the presence of large DIF effects (i.e., $\delta = 1$) compared with methods that use a fixed a value of the SCAD penalty. Interestingly, the least RMSE values could also be obtained with proper choices of fixed values of a and λ. In particular, for a large sample size of $N = 2500$, relying on a fixed instead of an optimal λ value can substantially decrease the RMSE.

As a conclusion of this focused simulation study, one could state that the choice of a for the SCAD penalty can have some impact. However, it is more important whether regularized estimation is carried out using AIC or BIC.

Table 3. Focused Simulation Study 1A: Relative root mean square error (RMSE) of estimated group means for unbalanced DIF effects as a function of the size of DIF effects δ and sample size N for different values a of the SCAD penalty.

		Best		**Choice of λ Based on AIC with $a =$**								**Choice of λ Based on BIC with $a =$**							
δ	N	a	λ	**2.2**	**2.5**	**3**	**3.7**	**4.5**	**6**	**9**	a_{opt}	**2.2**	**2.5**	**3**	**3.7**	**4.5**	**6**	**9**	a_{opt}
0.5	500	9	0.04	110.8	111.2	110.9	111.8	110.2	110.3	109.6	108.0	103.4	103.4	103.3	103.4	103.4	103.5	104.0	104.0
	1000	3.7	BIC	128.7	130.1	129.5	127.7	129.4	126.1	122.7	121.4	100.2	100.0	100.1	100.0	100.2	100.1	100.4	100.2
	2500	2.2	0.19	139.2	138.6	137.7	136.3	138.1	137.8	132.3	129.4	126.0	126.6	125.2	122.7	125.1	124.6	117.8	111.3
1	500	3.7	0.13	111.4	110.4	110.5	108.7	108.7	107.9	109.9	108.6	102.5	102.4	102.5	100.3	100.2	100.2	100.4	100.6
	1000	3.7	0.13	113.1	112.5	112.5	112.9	112.1	111.2	108.9	110.3	100.6	100.6	100.7	100.6	100.7	100.6	100.9	100.8
	2500	9	0.08	126.9	119.8	120.4	119.5	120.4	114.4	116.4	122.8	111.3	103.4	103.4	105.8	103.4	103.4	103.3	111.2

Note. Best = pair of (a, λ) values in all estimated models that resulted in the least relative RMSE value of 100; Choice of λ = method or value for determining the regularization parameter λ; and a_{opt} = choice of optimal a parameter based on AIC or BIC with corresponding optimal λ parameter.

5. Simulation Study 2: Uniform DIF Effects in the 2PL Model

In this Simulation Study 2, robust linking and regularization estimation is compared in the 2PL IRT model in the presence of uniform DIF effects.

5.1. Method

The simulation design of Simulation Study 2 closely follows the design of Simulation Study 1 described in Section 3.1. We now only emphasize the differences between this study and the first one.

Notably, item response data were generated based on the 2PL model using the same invariant item discriminations across the two groups (see https://osf.io/tma3f/ (accessed on 8 December 2022)) for data-generating item parameters; see column "a" for item discriminations. Like in Simulation Study 1, item difficulties b_i ranged between -1.88 and 1.61 with a mean of 0.00 and a standard deviation of 1.20. Item discriminations a_i ranged between 0.58 and 1.57 with a mean of 1.00 and a standard deviation of 0.30. Item difficulties and item discriminations were essentially uncorrelated ($r = -0.02$). DIF effects only occurred in item difficulties (i.e., uniform DIF). Like in Simulation Study 1, we set $\mu_2 = 0.3$ and $\sigma_2 = 1.2$ as data-generating distribution parameters in the second group. Moreover, the number of items (i.e., $I = 20$) was also fixed throughout all conditions. DIF effects were either small ($\delta = 0.5$) or large ($\delta = 1.0$) and either balanced or unbalanced. Finally, we also simulated four different sample sizes N of 500, 1000, 2500, and 5000.

For regularized estimation, the same grid of λ values like in Simulation Study 1 and Focused Simulation Study 1A was utilized. In this simulation study, the tuning parameter a was fixed to 3.7.

In the linking approaches, we either relied on a simultaneous first estimation step assuming common item discriminations or separate estimation assuming groupwise item discriminations. In contrast to Simulation Study 1, we also applied joint Haberman linking (JHL) to common item discriminations and ordinary Haberman linking (HL) to groupwise estimated item discriminations with powers $p = 2$, 1, and 0.5.

In total, 3000 replications were simulated in each condition. R software [70] was used throughout the whole simulation. The R packages TAM [71] and sirt [72] were employed for estimating the IRT models. Replication material can again be found at https://osf.io/tma3f/ (accessed on 8 December 2022).

5.2. Results

Table 4 presents the bias for the estimated group mean $\hat{\mu}_2$ as a function of the size of DIF effects δ and the sample size N. It turned out that all methods except concurrent calibration (CC) resulted in unbiased estimates of group means. Consistent with other studies, CC resulted in slightly biased estimates in the case of the 2PL model, even in the situation of balanced DIF effects. The reason might be that the presence of DIF negatively impacted the estimation of common item discriminations and the standard deviation of the second group.

In the case of unbalanced DIF, robust linking approaches based on $L_{0.5}$ or MAD, as well as the regularization approach based on BIC, performed satisfactorily in terms of bias. Joint Haberman linking (JHL) and Haberman linking based on separate calibration (HL) resulted in approximately unbiased estimates with $p = 0.5$.

Table 5 presents the relative RMSE of the estimated group mean as a function of the size of the DIF effect δ and the sample size N. In the case of unbalanced DIF, mean–mean linking L_2 and JHL performed the best across all conditions. It should be emphasized that the RMSE of CC became unacceptable in larger sample sizes due to bias. Moreover, a robust linking approach with powers $p = 0.5$ or $p = 1$ resulted in some efficiency losses.

Table 4. Simulation Study 2: Bias of estimated group means for balanced and unbalanced DIF effects as a function of the size of DIF effects δ and sample size N.

			Choice of λ					JHL with $p =$			HL with $p =$						
δ	N	MAD	AIC	BIC	0.05	0.10	0.15	0.5	1	2	0.5	1	2	$L_{0.5}$	L_1	L_2	CC
		Balanced DIF															
0.5	500	−0.01	−0.01	−0.01	−0.02	−0.01	**−0.04**	−0.01	−0.01	−0.01	0.00	0.00	0.01	−0.01	−0.01	−0.01	**−0.04**
	1000	0.01	0.00	0.01	0.00	0.01	−0.03	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	**−0.04**
	2500	−0.01	−0.01	−0.01	−0.01	0.00	**−0.05**	−0.01	−0.01	−0.01	0.00	0.00	0.00	−0.01	−0.01	−0.01	**−0.04**
	5000	0.00	0.00	0.01	0.01	0.01	**−0.04**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**−0.04**
1.0	500	−0.01	−0.02	−0.01	−0.02	−0.01	−0.01	−0.02	−0.02	−0.01	0.00	0.00	0.01	−0.02	−0.02	−0.01	**−0.06**
	1000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**−0.06**
	2500	−0.01	−0.02	−0.01	−0.01	−0.01	−0.01	−0.02	−0.01	−0.01	0.00	0.00	0.00	−0.02	−0.01	−0.01	**−0.06**
	5000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**−0.06**
		Unbalanced DIF															
0.5	500	**−0.07**	**−0.03**	**−0.04**	**−0.04**	−0.03	**−0.07**	**−0.04**	**−0.06**	**−0.10**	**−0.04**	**−0.06**	**−0.10**	**−0.04**	**−0.05**	**−0.10**	**−0.10**
	1000	**−0.03**	−0.01	0.00	−0.01	−0.01	**−0.07**	−0.02	**−0.04**	**−0.10**	−0.03	**−0.05**	**−0.10**	−0.02	**−0.03**	**−0.10**	**−0.10**
	2500	−0.01	**−0.03**	−0.03	−0.03	−0.02	**−0.10**	−0.02	**−0.04**	**−0.10**	−0.02	**−0.04**	**−0.10**	−0.01	−0.03	**−0.10**	**−0.10**
	5000	0.01	−0.01	0.00	−0.01	0.00	**−0.09**	0.00	−0.02	**−0.10**	−0.01	−0.03	**−0.10**	0.00	−0.01	**−0.10**	**−0.10**
1.0	500	−0.03	**−0.03**	−0.02	**−0.06**	−0.02	−0.02	**−0.03**	**−0.07**	**−0.21**	**−0.03**	**−0.07**	**−0.20**	**−0.04**	**−0.06**	**−0.21**	**−0.17**
	1000	0.00	−0.01	0.00	−0.02	0.00	0.00	−0.01	**−0.04**	**−0.20**	−0.02	**−0.05**	**−0.20**	−0.01	**−0.03**	**−0.20**	**−0.17**
	2500	−0.01	−0.02	−0.02	**−0.05**	−0.02	−0.02	−0.02	**−0.04**	**−0.21**	−0.01	**−0.04**	**−0.20**	−0.02	**−0.03**	**−0.21**	**−0.17**
	5000	0.00	0.00	0.00	−0.02	0.00	0.00	0.00	−0.02	**−0.20**	−0.01	−0.03	**−0.20**	0.00	−0.01	**−0.20**	**−0.17**

Note. MAD = robust linking using the MAD statistic; Choice of λ = method or valuze for determining the regularization parameter λ; JHL = joint Haberman linking using joint item discriminations; HL = Haberman linking using group-specific item discriminations; L_p = linking employing the unweighted L_p loss function with $p = 0.5$, 1.0, or 2.0 using joint item discriminations; CC = concurrent calibration assuming invariant item parameters; Absolute biases larger than 0.03 are printed in bold.

Table 5. Simulation Study 2: Relative root mean square error (RMSE) of estimated group means for balanced and unbalanced DIF effects as a function of the size of DIF effects δ and sample size N.

			Choice of λ					JHL with $p =$			HL with $p =$						
δ	N	MAD	AIC	BIC	0.05	0.10	0.15	0.5	1	2	0.5	1	2	$L_{0.5}$	L_1	L_2	CC
		Balanced DIF															
0.5	500	108	119	109	**130**	111	120	112	103	100	**127**	115	125	115	105	100	113
	1000	109	113	106	111	106	**127**	107	102	100	120	112	118	112	104	100	121
	2500	104	**185**	117	**183**	**139**	**178**	105	103	100	113	110	116	111	105	100	**152**
	5000	103	120	112	115	104	**209**	103	102	100	110	109	113	110	104	100	**194**
1.0	500	105	109	101	117	101	100	109	103	100	**127**	117	**127**	113	105	100	**127**
	1000	104	108	102	107	102	101	109	103	100	124	116	123	112	106	100	**149**
	2500	105	113	100	**127**	100	100	106	104	102	114	111	116	112	107	102	**192**
	5000	103	108	100	108	100	100	102	101	100	113	113	118	110	104	100	**258**
		Unbalanced DIF															
0.5	500	118	118	101	**130**	100	121	103	107	**140**	119	118	**148**	109	106	**140**	**135**
	1000	**126**	**136**	100	**133**	113	**163**	107	118	**190**	**126**	**137**	**201**	116	115	**190**	**192**
	2500	105	**293**	**272**	**299**	**212**	**269**	107	**134**	**288**	118	**146**	**284**	113	123	**288**	**276**
	5000	102	**276**	**279**	**279**	**253**	**356**	100	123	**375**	115	**156**	**391**	109	110	**375**	**384**
1.0	500	109	115	100	**146**	107	105	110	**125**	**265**	**128**	**141**	**271**	120	123	**265**	**231**
	1000	100	114	105	**131**	105	105	105	122	**359**	124	**145**	**366**	113	118	**359**	**315**
	2500	101	**179**	**169**	**308**	**163**	**171**	108	**144**	**536**	113	**146**	**529**	112	**131**	**536**	**459**
	5000	103	117	100	**345**	100	100	104	**135**	**776**	116	**161**	**778**	112	118	**776**	**677**

Note. MAD = robust linking using the MAD statistic; Choice of λ = method or valuze for determining the regularization parameter λ; JHL = joint Haberman linking using joint item discriminations; HL = Haberman linking using group-specific item discriminations; L_p = linking employing the unweighted L_p loss function with $p = 0.5$, 1.0, or 2.0 using joint item discriminations; CC = concurrent calibration assuming invariant item parameters; Relative RMSE values larger than 125 are printed in bold.

The case of unbalanced DIF in the 2PL model was similar to the 1PL model presented in Simulation Study 1 in Section 3.2. For small DIF effects (i.e., $\delta = 0.5$), regularized estimation using BIC was the frontrunner for moderate sample sizes $N = 500$ and $N = 1000$, while robust linking based on the MAD statistic performed well in larger sample sizes $N = 2500$ and $N = 5000$. For large DIF effects (i.e., $\delta = 1.0$), robust linking using the MAD

statistic performed satisfactorily. Moreover, JHL with $p = 0.5$ and the robust $L_{0.5}$ linking also resulted in group mean estimates of acceptable variability.

6. Discussion

In this article, we investigated the performance of robust linking and regularized estimation in the presence of sparse uniform differential item functioning by means of three simulation studies. It turned out that robust linking approaches were competitive or outperformed regularized estimation in most conditions in the simulation studies. In particular, regularized estimation was not quite successful in parameter recovery in conditions with small DIF effects. It was also found that robust linking based on outlier removal using the MAD statistic [64] was often superior to robust linking using the L_p loss function with $p = 0.5$, which corresponds to invariance alignment. Overall, one could generally conclude that there is probably no need for using the computationally much more demanding regularization approaches instead of employing robust linking.

As in any simulation study, our findings are limited to the studied conditions. First, we only considered a fixed test of 20 items. Additional studies could also involve a larger number of items or balanced incomplete block designs for item response data [73]. Second, only 4 out of 20 items (i.e., 20% of the items) showed uniform DIF effects. Other research indicated that higher DIF rates would also be possible in robust linking, as long as the threshold of 50% of DIF items is not exceeded [21,24]. Third, we simulated uniform DIF under a sparsity condition. The 16 non-DIF items had DIF effects of exactly 0. Future research could also assume that there would also be small DIF effects that could add up to zero [26,74]. Fourth, future research could also investigate the case of nonuniform DIF in item discriminations. It could be that larger sample sizes would be required for robust linking in this situation, and regularized estimation might be advantageous.

In this article, we assumed that DIF effects potentially bias group mean differences. Hence, DIF effects should be essentially removed from group comparisons. However, it could be argued that eliminating items from comparisons poses a threat to validity [74–77], and statistical criteria should not determine which items enter group comparisons [78].

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

1PL	one-parameter logistic
2PL	two-parameter logistic
AIC	Akaike information criterion
BIC	Bayesian information criterion
CC	concurrent calibration
DIF	differential item functioning
DWLS	diagonally weighted least squares
FIPC	fixed item parameter calibration
IPD	item parameter drift

IRT	item response theory
JK	jackknife
LE	linking error
LSA	large-scale assessment studies
MAD	median absolute deviation
PISA	programme for international student assessment
RMSE	root mean square error
SCAD	smoothly clipped absolute deviation

References

1. Van der Linden, W.J.; Hambleton, R.K. (Eds.) *Handbook of Modern Item Response Theory*; Springer: New York, NY, USA, 1997. https://doi.org/10.1007/978-1-4757-2691-6.
2. Van der Linden, W.J. Unidimensional logistic response models. In *Handbook of Item Response Theory, Volume 1: Models*; van der Linden, W.J., Ed.; CRC Press: Boca Raton, FL, USA, 2016; pp. 11–30. https://doi.org/10.1201/9781315374512-2.
3. Lietz, P.; Cresswell, J.C.; Rust, K.F.; Adams, R.J. (Eds.) *Implementation of Large-scale Education Assessments*; Wiley: New York, NY, USA, 2017. https://doi.org/10.1002/9781118762462.
4. Rutkowski, L.; von Davier, M.; Rutkowski, D. (Eds.) *A Handbook of International Large-scale Assessment: Background, Technical Issues, and Methods of Data Analysis*; Chapman Hall/CRC Press: London, UK, 2013. https://doi.org/10.1201/b16061.
5. OECD. *PISA 2018. Technical Report*; OECD: Paris, France, 2020. Available online: https://bit.ly/3zWbidA (accessed on 8 December 2022).
6. Yen, W.M.; Fitzpatrick, A.R. Item response theory. In *Educational Measurement*; Brennan, R.L., Ed.; Praeger Publishers: Westport, CT, USA, 2006; pp. 111–154.
7. Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*; Danish Institute for Educational Research: Copenhagen, Denmark, 1960.
8. Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In *Statistical Theories of Mental Test Scores*; Lord, F.M., Novick, M.R., Eds.; MIT Press: Reading, MA, USA, 1968; pp. 397–479.
9. Mellenbergh, G.J. Item bias and item response theory. *Int. J. Educ. Res.* **1989**, *13*, 127–143. https://doi.org/10.1016/0883-0355(89)90002-5.
10. Millsap, R.E. *Statistical Approaches to Measurement Invariance*; Routledge: New York, NY, USA, 2011. https://doi.org/10.4324/9780203821961.
11. Holland, P.W.; Wainer, H. (Eds.) *Differential Item Functioning: Theory and Practice*; Lawrence Erlbaum: Hillsdale, NJ, USA, 1993. https://doi.org/10.4324/9780203357811.
12. Penfield, R.D.; Camilli, G. Differential item functioning and item bias. In *Handbook of Statistics, Vol. 26: Psychometrics*; Rao, C.R., Sinharay, S., Eds.; 2007; Elsevier: Amsterdam, The Netherlands pp. 125–167. https://doi.org/10.1016/S0169-7161(06)26005-X.
13. Robitzsch, A. A comparison of linking methods for two groups for the two-parameter logistic item response model in the presence and absence of random differential item functioning. *Foundations* **2021**, *1*, 116–144. https://doi.org/10.3390/foundations1010009.
14. De Boeck, P. Random item IRT models. *Psychometrika* **2008**, *73*, 533–559. https://doi.org/10.1007/s11336-008-9092-x.
15. Frederickx, S.; Tuerlinckx, F.; De Boeck, P.; Magis, D. RIM: A random item mixture model to detect differential item functioning. *J. Educ. Meas.* **2010**, *47*, 432–457. https://doi.org/10.1111/j.1745-3984.2010.00122.x.
16. Byrne, B.M.; Shavelson, R.J.; Muthén, B. Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychol. Bull.* **1989**, *105*, 456–466. https://doi.org/10.1037/0033-2909.105.3.456.
17. Lee, S.S.; von Davier, M. Improving measurement properties of the PISA home possessions scale through partial invariance modeling. *Psych. Test Assess. Model.* **2020**, *62*, 55–83. https://bit.ly/3FRN6Qf.
18. Magis, D.; Tuerlinckx, F.; De Boeck, P. Detection of differential item functioning using the lasso approach. *J. Educ. Behav. Stat.* **2015**, *40*, 111–135. https://doi.org/10.3102/1076998614559747.
19. Tutz, G.; Schauberger, G. A penalty approach to differential item functioning in Rasch models. *Psychometrika* **2015**, *80*, 21–43. https://doi.org/10.1007/s11336-013-9377-6.
20. Chen, Y.; Li, C.; Xu, G. DIF statistical inference and detection without knowing anchoring items. *arXiv* **2021**, arXiv:2110.11112. https://doi.org/10.48550/arXiv.2110.11112.
21. Halpin, P.F. Differential item functioning via robust scaling. *arXiv* **2022**, arXiv:2207.04598. https://doi.org/10.48550/arXiv.2207.04598.
22. Magis, D.; De Boeck, P. Identification of differential item functioning in multiple-group settings: A multivariate outlier detection approach. *Multivar. Behav. Res.* **2011**, *46*, 733–755. https://doi.org/10.1080/00273171.2011.606757.
23. Magis, D.; De Boeck, P. A robust outlier approach to prevent type I error inflation in differential item functioning. *Educ. Psychol. Meas.* **2012**, *72*, 291–311. https://doi.org/10.1177/0013164411416975.
24. Wang, W.; Liu, Y.; Liu, H. Testing differential item functioning without predefined anchor items using robust regression. *J. Educ. Behav. Stat.* **2022**, *47*, 666–692. https://doi.org/10.3102/10769986221109.
25. Robitzsch, A. Robust and nonrobust linking of two groups for the Rasch model with balanced and unbalanced random DIF: A comparative simulation study and the simultaneous assessment of standard errors and linking errors with resampling techniques. *Symmetry* **2021**, *13*, 2198. https://doi.org/10.3390/sym13112198.

26. Robitzsch, A.; Lüdtke, O. A review of different scaling approaches under full invariance, partial invariance, and noninvariance for cross-sectional country comparisons in large-scale assessments. *Psych. Test Assess. Model.* **2020**, *62*, 233–279. Available online: {h}ttps://bit.ly/3ezBB05 (accessed on 8 December 2022).
27. Fan, J.; Li, R.; Zhang, C.H.; Zou, H. *Statistical Foundations of Data Science*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2020. https://doi.org/10.1201/9780429096280.
28. Hastie, T.; Tibshirani, R.; Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations*; CRC Press: Boca Raton, FL, USA, 2015. https://doi.org/10.1201/b18401.
29. Chen, Y.; Li, X.; Liu, J.; Ying, Z. Robust measurement via a fused latent and graphical item response theory model. *Psychometrika* **2018**, *83*, 538–562. https://doi.org/10.1007/s11336-018-9610-4.
30. Sun, J.; Chen, Y.; Liu, J.; Ying, Z.; Xin, T. Latent variable selection for multidimensional item response theory models via L_1 regularization. *Psychometrika* **2016**, *81*, 921–939. https://doi.org/10.1007/s11336-016-9529-6.
31. Geminiani, E.; Marra, G.; Moustaki, I. Single- and multiple-group penalized factor analysis: A trust-region algorithm approach with integrated automatic multiple tuning parameter selection. *Psychometrika* **2021**, *86*, 65–95. https://doi.org/10.1007/s11336-021-09751-8.
32. Huang, P.H.; Chen, H.; Weng, L.J. A penalized likelihood method for structural equation modeling. *Psychometrika* **2017**, *82*, 329–354. https://doi.org/10.1007/s11336-017-9566-9.
33. Jacobucci, R.; Grimm, K.J.; McArdle, J.J. Regularized structural equation modeling. *Struct. Equ. Modeling* **2016**, *23*, 555–566. https://doi.org/10.1080/10705511.2016.1154793.
34. Chen, Y.; Li, X.; Liu, J.; Ying, Z. Regularized latent class analysis with application in cognitive diagnosis. *Psychometrika* **2017**, *82*, 660–692. https://doi.org/10.1007/s11336-016-9545-6.
35. Robitzsch, A.; George, A.C. The R package CDM for diagnostic modeling. In *Handbook of Diagnostic Classification Models*; von Davier, M., Lee, Y.S., Eds.; Springer: Cham, Switzerland, 2019; pp. 549–572. https://doi.org/10.1007/978-3-030-05584-4_26.
36. Robitzsch, A. Regularized latent class analysis for polytomous item responses: An application to SPM-LS data. *J. Intell.* **2020**, *8*, 30. https://doi.org/10.3390/jintelligence8030030.
37. Fop, M.; Murphy, T.B. Variable selection methods for model-based clustering. *Stat. Surv.* **2018**, *12*, 18–65. https://doi.org/10.1214/18-SS119.
38. Robitzsch, A. Regularized mixture Rasch model. *Information* **2022**, *13*, 534. https://doi.org/10.3390/info13110534.
39. Belzak, W.C. The multidimensionality of measurement bias in high-stakes testing: Using machine learning to evaluate complex sources of differential item functioning. *Educ. Meas.* **2022**, Epub ahead of print, https://doi.org/10.1111/emip.12486.
40. Belzak, W.; Bauer, D.J. Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychol. Methods* **2020**, *25*, 673–690. https://doi.org/10.1037/met0000253.
41. Bauer, D.J.; Belzak, W.C.M.; Cole, V.T. Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. *Struct. Equ. Model.* **2020**, *27*, 43–55. https://doi.org/10.1080/10705511.2019.1642754.
42. Gürer, C.; Draxler, C. Penalization approaches in the conditional maximum likelihood and Rasch modelling context. *Brit. J. Math. Stat. Psychol.* **2022**, Epub ahead of print, https://doi.org/10.1111/bmsp.12287.
43. Liang, X.; Jacobucci, R. Regularized structural equation modeling to detect measurement bias: Evaluation of lasso, adaptive lasso, and elastic net. *Struct. Equ. Model.* **2020**, *27*, 722–734. https://doi.org/10.1080/10705511.2019.1693273.
44. Schauberger, G.; Mair, P. A regularization approach for the detection of differential item functioning in generalized partial credit models. *Behav. Res. Methods* **2020**, *52*, 279–294. https://doi.org/10.3758/s13428-019-01224-2.
45. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. https://doi.org/10.1198/016214501753382273.
46. Chen, Y.; Liu, J.; Xu, G.; Ying, Z. Statistical analysis of Q-matrix based diagnostic classification models. *J. Am. Stat. Assoc.* **2015**, *110*, 850–866. https://doi.org/10.1080/01621459.2014.934827.
47. Umezu, Y.; Shimizu, Y.; Masuda, H.; Ninomiya, Y. AIC for the non-concave penalized likelihood method. *Ann. Inst. Stat. Math.* **2019**, *71*, 247–274. https://doi.org/10.1007/s10463-018-0649-x.
48. Zhang, H.; Li, S.J.; Zhang, H.; Yang, Z.Y.; Ren, Y.Q.; Xia, L.Y.; Liang, Y. Meta-analysis based on nonconvex regularization. *Sci. Rep.* **2020**, *10*, 5755. https://doi.org/10.1038/s41598-020-62473-2.
49. Breheny, P.; Huang, J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.* **2011**, *5*, 232–253. https://doi.org/10.1214/10-AOAS388.
50. Xiao, H.; Sun, Y. On tuning parameter selection in model selection and model averaging: a Monte Carlo study. *J. Risk Financ. Manag.* **2019**, *12*, 109. https://doi.org/10.3390/jrfm12030109.
51. Williams, D.R. Beyond lasso: A survey of nonconvex regularization in Gaussian graphical models. *PsyArXiv* **2020**, 9 November 2020. https://doi.org/10.31234/osf.io/ad57p.
52. Battauz, M. Regularized estimation of the nominal response model. *Multivar. Behav. Res.* **2020**, *55*, 811–824. https://doi.org/10.1080/ 00273171.2019.1681252.
53. Oelker, M.R.; Tutz, G. A uniform framework for the combination of penalties in generalized structured models. *Adv. Data Anal. Classif.* **2017**, *11*, 97–120. https://doi.org/10.1007/s11634-015-0205-y.

54. Tutz, G.; Gertheiss, J. Regularized regression for categorical data. *Stat. Model.* **2016**, *16*, 161–200. https://doi.org/10.1177/1471082X 16642560.
55. Kolen, M.J.; Brennan, R.L. *Test Equating, Scaling, and Linking*; Springer: New York, NY, USA, 2014. https://doi.org/10.1007/978-1-4939-0317-7.
56. Lee, W.C.; Lee, G. IRT linking and equating. In *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test*; Irwing, P., Booth, T., Hughes, D.J., Eds.; Wiley: New York, NY, USA, 2018; pp. 639–673. https://doi.org/10.1002/9781118489772.ch21.
57. Sansivieri, V.; Wiberg, M.; Matteucci, M. A review of test equating methods with a special focus on IRT-based approaches. *Statistica* **2017**, *77*, 329–352. https://doi.org/10.6092/issn.1973-2201/7066.
58. Robitzsch, A. Robust Haebara linking for many groups: Performance in the case of uniform DIF. *Psych* **2020**, *2*, 155–173. https://doi.org/10.3390/psych2030014.
59. Pokropek, A.; Lüdtke, O.; Robitzsch, A. An extension of the invariance alignment method for scale linking. *Psych. Test Assess. Model.* **2020**, *62*, 303–334. Available online: https://bit.ly/2UEp9GH (accessed on 8 December 2022).
60. Robitzsch, A. L_p loss functions in invariance alignment and Haberman linking with few or many groups. *Stats* **2020**, *3*, 246–283. https://doi.org/10.3390/stats3030019.
61. Manna, V.F.; Gu, L. *Different Methods of Adjusting for form Difficulty under the Rasch Model: Impact on Consistency of Assessment Results*; (Research Report No. RR-19-08); Educational Testing Service: Princeton, NJ, USA, 2019. https://doi.org/10.1002/ets2.12244.
62. Asparouhov, T.; Muthén, B. Multiple-group factor analysis alignment. *Struct. Equ. Model.* **2014**, *21*, 495–508. https://doi.org/10.1080/10705511.2014.919210.
63. Muthén, B.; Asparouhov, T. IRT studies of many groups: The alignment method. *Front. Psychol.* **2014**, *5*, 978. https://doi.org/10.3389/fpsyg.2014.00978.
64. von Davier, M.; Bezirhan, U. A robust method for detecting item misfit in large scale assessments. *Educ. Psychol. Meas.* **2022**, Epub ahead of print, https://doi.org/10.1177/00131644221105819.
65. Huynh, H.; Meyer, P. Use of robust z in detecting unstable items in item response theory models. *Pract. Assess. Res. Eval.* **2010**, *15*, 2. https://doi.org/10.7275/ycx6-e864.
66. Liu, C.; Jurich, D. Outlier detection using t-test in Rasch IRT equating under NEAT design. *Appl. Psychol. Meas.* **2022**, Epub ahead of print, https://doi.org/10.1177/01466216221124045.
67. Battauz, M. Multiple equating of separate IRT calibrations. *Psychometrika* **2017**, *82*, 610–636. https://doi.org/10.1007/s11336-016-9517-x.
68. Haberman, S.J. *Linking Parameter Estimates Derived from an Item Response Model through Separate Calibrations*; (Research Report No. RR-09-40); Educational Testing Service: Princeton, NJ, USA, 2009. https://doi.org/10.1002/j.2333-8504.2009.tb02197.x.
69. Liu, X.; Wallin, G.; Chen, Y.; Moustaki, I. Rotation to sparse loadings using L^p losses and related inference problems. *arXiv* **2022**, arXiv:2206.02263. https://doi.org/10.48550/arXiv.2206.02263.
70. R Core Team. *R: A Language and Environment for Statistical Computing*; Vienna, Austria, R Core Team, 2022. Available online: https://www.R-project.org/ (accessed on 11 January 2022).
71. Robitzsch, A.; Kiefer, T.; Wu, M. TAM: Test Analysis Modules, 2022. R Package Version 4.1-4. Available online: https://CRAN.R-project.org/package=TAM (accessed on 28 August 2022).
72. Robitzsch, A. Sirt: Supplementary Item Response Theory Models, 2022. R Package Version 3.12-66. Available online: https://CRAN.R-project.org/package=sirt (accessed on 17 May 2022).
73. Frey, A.; Hartig, J.; Rupp, A.A. An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educ. Meas.* **2009**, *28*, 39–53. https://doi.org/10.1111/j.1745-3992.2009.00154.x.
74. Robitzsch, A.; Lüdtke, O. Mean comparisons of many groups in the presence of DIF: An evaluation of linking and concurrent scaling approaches. *J. Educ. Behav. Stat.* **2022**, *47*, 36–68. https://doi.org/10.3102/10769986211017479.
75. Camilli, G. The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In *Differential Item Functioning: Theory and Practice*; Holland, P.W., Wainer, H., Eds.; Erlbaum: Hillsdale, NJ, USA, 1993; pp. 397–417.
76. El Masri, Y.H.; Andrich, D. The trade-off between model fit, invariance, and validity: The case of PISA science assessments. *Appl. Meas. Educ.* **2020**, *33*, 174–188. https://doi.org/10.1080/08957347.2020.1732384.
77. Robitzsch, A.; Lüdtke, O. Some thoughts on analytical choices in the scaling model for test scores in international large-scale assessment studies. *Meas. Instrum. Soc. Sci.* **2022**, *4*, 9. https://doi.org/10.1186/s42409-022-00039-w.
78. Brennan, R.L. Misconceptions at the intersection of measurement theory and practice. *Educ. Meas.* **1998**, *17*, 5–9. https://doi.org/10.1111/j.1745-3992.1998.tb00615.x.