Article | Received 1 August 2024; Accepted 9 October 2024; Published 18 October 2024
https://doi.org/10.55092/asymmetry20240005

Does random differential item functioning occur in one or two groups? Implications for bias and variance in asymmetric and symmetric Haebara and Stocking-Lord linking

Alexander Robitzsch[1,2,*]

[1] IPN – Leibniz Institute for Science and Mathematics Education, Kiel, Germany

[2] Centre for International Student Assessment (ZIB), Kiel, Germany

* Correspondence author; E-mail: robitzsch@leibniz-ipn.de.

Abstract: Linking methods are frequently applied to analyze the performance of two groups on a set of items. This article shows analytically and by simulation that the occurrence of differential item functioning (DIF) can induce bias and additional variance in parameter estimates of the linking method. Interestingly, the bias of the parameter estimates of a linking method depends on whether random DIF occurs in one or two groups (*i.e.*, the type of DIF effects). The findings are shown utilizing asymmetric and symmetric Haebara and Stocking-Lord linking. Moreover, the latter linking methods were compared with recently proposed corresponding SIMEX-based variants of linking in a simulation study. It turned out that SIMEX-based linking provided unbiased estimates of whether the correct assumption of the type of DIF effects has been implemented.

Keywords: linking; 2PL model; random differential item functioning; Stocking-Lord linking; Haebara linking; item response model

1. Introduction

Item response theory (IRT) models [1–3] are multivariate statistical models to analyze multivariate binary random variables. These models have wide applications in education psychology. For example, IRT models are operationally utilized in educational large-scale assessment studies [4].

In this paper, unidimensional IRT models [5, 6] are only investigated. Let $X = (X_1, \ldots, X_I)$ be the vector of I dichotomous random variables $X_i \in \{0, 1\}$ (also referred to as items or item responses). A unidimensional IRT model is a statistical model for the probability distribution $\mathsf{P}(X = x)$ for $x = (x_1, \ldots, x_I) \in \{0, 1\}^I$ with the parametrized probability distribution

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

where ϕ denotes the density of the normal distribution with the mean μ and the standard deviation σ. The distribution parameters of the latent variable θ (also referred to as a trait or ability) are contained in the vector $\boldsymbol{\delta} = (\mu, \sigma)$. The vector $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_I)$ contains all estimated item parameters of item response functions (IRF) $P_i(\theta; \boldsymbol{\gamma}_i) = \mathsf{P}(X_i = 1|\theta)$ $(i = 1, \ldots, I)$. The two-parameter logistic (2PL) model [7] has the IRF

$$P_i(\theta; \boldsymbol{\gamma}_i) = \Psi(a_i(\theta - b_i)) \tag{2}$$

using the item discrimination a_i and item difficulty b_i, and $\Psi(x) = (1 + \exp(-x))^{-1}$ denotes the logistic distribution function. For independently and identically distributed observations $x_1, \ldots, x_N$ of N subjects from the distribution of the random variable X, unknown model parameters of the IRT model (1) can be estimated by marginal maximum likelihood (MML) estimation using an expectation-maximization algorithm [8–10].

IRT models are frequently used to compare the distribution of X in two groups, summarized by the parameters of the distribution of the ability variable θ in the IRT model (1). In the following, we confine ourselves to the 2PL model. In this paper, we consider linking methods [11] for group comparisons. Linking methods are two-step procedures in which the 2PL model is separately estimated in each of the two groups in the first step. In the second step, differences in estimated item parameters are used to determine the group difference regarding distributional differences of the θ variable by means of a linking method [11–13]. In our article, we focus on the impact of differential item functioning (DIF; [14–18]) on the bias [19] and the variance [20–28] in estimated group means and standard deviations.

This article investigates whether the bias and variance of the Haebara [29] and Stocking-Lord [30] linking methods for linking in two groups depend on whether random DIF [31, 32] occurs in one or two groups. If random DIF occurs in only one group, the presence of DIF is asymmetric, while it is symmetric if DIF occurs in both groups. It will turn out that the findings substantially differ for these two kinds of DIF effects. Moreover, Haebara and Stocking-Lord have been originally proposed to align the IRFs of the first group onto the IRFs of the second group, which entails an asymmetric treatment in handling errors. The original versions of these linking methods can be labeled as asymmetric Haebara and Stocking-Lord linking, respectively. In contrast, we also study symmetric formulations of Haebara and Stocking-Lord linking [33] in this article for the two kinds of random DIF effects. Finally, bias-corrected variants of Haebara and Stocking-Lord linking based on the simulation extrapolation (SIMEX; [34]) measurement error correction method [35] are evaluated in a simulation study. SIMEX requires the specification of whether DIF occurs in one or two groups, and the simulation study presented in this article investigates whether an incorrect choice of the kind of random DIF results in biased estimates.

The rest of the article is organized as follows. Section 2 distinguishes fixed and random DIF. In Section 3, linking methods are generally introduced, and Haebara and Stocking-Lord linking are discussed as particular examples. Section 4 presents an analytical derivation of the bias and the variance of parameter estimates of a general linking method. The findings of Section 4 are illustrated for Haebara linking in Section 5. Section 6 presents findings from a

simulation study. Finally, the article closes with conclusions in Section 7.

2. Fixed and random differential item functioning

In this section, we discuss the concept of DIF and how the determination of the group mean and the group standard deviation is related to identification constraints on DIF effects.

Let us assume that group-specific item parameters a_{ig} and b_{ig} ($i = 1,\ldots,I$; $g = 1,2$) in the 2PL model. In the first group, the mean and the standard deviation are fixed at 0 and 1, respectively. The mean and the standard deviation in the second group are denoted as μ and σ, respectively. For example, the two groups could be two genders (*i.e.*, female and male) or two countries (e.g., China and Germany). Alternatively, researchers could also define a first reference group that involves all subjects (e.g., students of all countries) and a second focal group that involves only subjects from a subpopulation (e.g., students from China).

In the first step, the 2PL model is separately estimated in the two groups, resulting in identified item parameters $\widehat{a}_{ig}$ and $\widehat{b}_{ig}$ for $g = 1,2$. In the two scalings, the mean and the standard deviation of the ability variable are fixed at 0 and 1 for identification reasons. Because we are only interested in identification issues in this section, we can ignore sampling errors and only consider population-level data. Hence, we have $\widehat{a}_{i1} = a_{i1}$ and $\widehat{b}_{i1} = b_{i1}$ for the item parameters in the first group. Furthermore, the identified item parameters for the second group are given as

$$\widehat{a}_{i2} = a_{i2}\sigma \text{ and } \widehat{b}_{i2} = \sigma^{-1}(b_{i2} - \mu) \,. \tag{3}$$

Uniform DIF is present if $b_{i1} \neq b_{i2}$ holds for at least one item $i \in \mathscr{I} = \{1,\ldots,I\}$ and $a_{i1} = a_{i2}$ for all $i \in \mathscr{I}$. Nonuniform DIF is present if $a_{i1} \neq a_{i2}$ holds for at least one item $i \in \mathscr{I}$. In this article, we only consider uniform DIF. The uniform DIF effects are defined by $e_i = b_{i2} - b_{i1}$ and we use the notation $e = (e_1,\ldots,e_I)$.

Two important kinds of DIF must be distinguished: fixed and random DIF [31, 36]. In the case of fixed DIF, DIF effects e_i are treated as fixed parameters, while they are considered as random variables in the case of random DIF.

We first discuss the case of fixed DIF. To disentangle DIF effects from group differences, additional identification constraints on DIF effects must be imposed to identify μ. In a general treatment, there is a nonlinear function g such that $g(e) = 0$, which is referred to as an identification constraint of DIF effects and μ. For example, the DIF effects could average to zero, resulting in the constraint $g(e) = \sum_{i=1}^{I} e_i = 0$. This situation is often referred to as balanced DIF [37, 38]. Alternatively, group differences could only rely on those items that do not exceed a certain threshold κ, resulting in the identification constraint $g(e) = \sum_{i=1}^{I} e_i 1(|e_i| \leq \kappa) = 0$, where 1 denotes the indicator function. Furthermore, researchers could select or detect an anchor set of items $\mathscr{A} \subset \mathscr{I}$ [39, 40] such that the identification constraint is given by $g(e) = \sum_{i=1}^{I} e_i 1(i \in \mathscr{A}) = 0$. In the latter case, group differences represented by μ are only determined by the anchor items.

Now assume that the identified item parameters are given in (3). We set $a_{i1} = a_i$ and $b_{i2} = b_i$ and define for the second group $a_{i2} = a_i$ and $b_{i2} = b_i + e_i$. We collect all identified

item parameters $\widehat{b}_{i2}$ ($i = 1,\ldots,I$) in the vector $\widehat{b}_2$ and obtain

$$\widehat{b}_2 = \sigma^{-1}(b_2 - \mu\mathbf{j}) \,, \tag{4}$$

where $\mathbf{j}$ denotes a vector of ones with length I. The group standard deviation σ can be identified from data by computing $\sigma = \widehat{a}_{i2}/\widehat{a}_{i1}$ for some item $i \in \mathscr{I}$. The group mean μ can be determined as the root of

$$h(\mu) = g(b_2 - b_1) = g\left(\sigma\widehat{b}_2 + \mu\mathbf{j} - \widehat{b}_1\right) = 0 \tag{5}$$

if σ is already identified. Interestingly, the true group mean μ can only be obtained when using the correct identification constraint g for computing μ. As a consequence, researchers must impose assumptions on the structure of DIF effects in order to identify group differences. We reiterate that identified item parameters are given by

$$\widehat{a}_{i1} = a_i \,,\ \widehat{a}_{i2} = a_i\sigma \,,\ \widehat{b}_{i1} = b_i \text{ and } \widehat{b}_{i2} = \sigma^{-1}\left(b_i + e_i - \mu\right) \,. \tag{6}$$

It is evident that DIF effects e_i only occur in item parameters of the second group. Hence, DIF effects seem only to be present in one group (*i.e.*, in the second group). Now define $\widetilde{b}_i = b_i + \frac{1}{2}e_i$ and $\widetilde{e}_i = e_i$. Then, we get an equivalent parametrization of the identified item parameters as

$$\widehat{a}_{i1} = a_i \,,\ \widehat{a}_{i2} = a_i\sigma \,,\ \widehat{b}_{i1} = \widetilde{b}_i - \frac{1}{2}\widetilde{e}_i \text{ and } \widehat{b}_{i2} = \sigma^{-1}\left(\widetilde{b}_i + \frac{1}{2}\widetilde{e}_i - \mu\right) \,. \tag{7}$$

In this parametrization, uniform DIF effects e_i appear in item intercepts of both groups. Importantly, because of equivalent parametrizations, the two cases do not need a different treatment in the case of fixed DIF.

In contrast, random (uniform) DIF assumes that e_i (that equals $\widetilde{e}_i$) is a random variable [31, 41, 42]. Note that randomness in DIF effects is not a consequence of sampling error but a mathematical device for describing the heterogeneity of group differences in items. Items can be considered fixed entities, but deviations in item parameters between two groups operate in some stochastic way that is reflected in a random variable. The variability e_i (*i.e.*, the random DIF effect) can be similarly interpreted like the regression residual ϵ_i in a linear regression $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. The regression provides analytical inference in the case of a fixed finite population of subjects $i = 1,\ldots,N$, and variability (*i.e.*, uncertainty) in regression coefficients only stems from imposing a distribution for the random residual variable ϵ_i (see [43, 44]).

In the case of random DIF, the two parametrizations (6) and (7) are no longer equivalent and require a different treatment. In the rest of the article, the two situations in which random DIF occurs in one or two groups are investigated in in terms of their consequences for linking procedures.

3. Linking methods

In this section, we discuss several linking methods (see [12, 19]). Let $H(\mu,\sigma,\widehat{a}_1,\widehat{b}_1,\widehat{a}_2,\widehat{b}_2)$ be a linking function with input item parameters $\widehat{a}_g = (\widehat{a}_{1g},\ldots,\widehat{a}_{Ig})$ and $\widehat{b}_g = (\widehat{b}_{1g},\ldots,\widehat{b}_{Ig})$ for

$g = 1, 2$. The group mean μ and standard deviation σ can be estimated by

$$(\widehat{\mu}, \widehat{\sigma}) = \arg\min_{(\mu,\sigma)} H(\mu, \sigma, \widehat{a}_1, \widehat{b}_1, \widehat{a}_2, \widehat{b}_2) \,. \tag{8}$$

For a differentiable function H, estimating equations for μ and σ can be obtained by taking partial derivatives of H with respect to μ and σ.

In practical applications, there will frequently be unique items that are only administered in the first or the second group. For example, such a situation might occur when groups of different average abilities are linked and the tests reflect different average item difficulties. However, only the common items (*i.e.*, anchor items) that are administered in both groups are used in the linking methods, and only these items contribute to group differences (at least in correctly specified IRT models).

3.1. Haebara linking

3.1.1 Asymmetric Haebara linking

Haebara linking has been originally proposed in [29] and uses the linking function

$$H(\mu, \sigma, \widehat{a}_1, \widehat{b}_1, \widehat{a}_2, \widehat{b}_2) = \sum_{i=1}^{I} \sum_{t=1}^{T} \omega_t \left[\Psi(\widehat{a}_{i1}(\sigma\theta_t + \mu - \widehat{b}_{i1})) - \Psi(\widehat{a}_{i2}(\theta_t - \widehat{b}_{i2})) \right]^2 , \tag{9}$$

where $\theta_1, \ldots, \theta_T$ is a discrete set of θ points (e.g., $T = 101$ equidistant points on the interval $[-6, 6]$) and ω_t are user-defined weights. For example, the weights could be chosen equal to 1 or proportional to the normal density function with mean 0 and standard deviation σ_ω (e.g., $\sigma_\omega = 2$). The linking function H defined in (9) is referred to as asymmetric Haebara linking because the IRFs in the first group are aligned onto the IRFs in the second group.

3.1.2 Symmetric Haebara linking

Symmetric Haebara linking [33, 42, 45–47] simultaneously aligns the IRFs of the first group onto the IRFs in the second group and the other way around by including a second term in the linking function:

$$\begin{aligned} H(\mu, \sigma, \widehat{a}_1, \widehat{b}_1, \widehat{a}_2, \widehat{b}_2) = & \sum_{i=1}^{I} \sum_{t=1}^{T} \omega_t \left[\Psi(\widehat{a}_{i1}(\sigma\theta_t + \mu - \widehat{b}_{i1})) - \Psi(\widehat{a}_{i2}(\theta_t - \widehat{b}_{i2})) \right]^2 \\ & + \sum_{i=1}^{I} \sum_{t=1}^{T} \omega_t \left[\Psi(\widehat{a}_{i1}(\theta_t - \widehat{b}_{i1})) - \Psi(\widehat{a}_{i2}(\sigma^{-1}(\theta_t - \mu) - \widehat{b}_{i2})) \right]^2 . \end{aligned} \tag{10}$$

While the order of the definition of groups affects the result in asymmetric Haebara linking, it does not have consequences in symmetric Haebara linking, which might be seen as advantageous.

3.1.3 SIMEX-based Haebara linking

It has been demonstrated that Haebara linking provides biased estimates in the presence of random uniform DIF [19]. To this end, a general estimation approach to linking methods based on simulation extrapolation (SIMEX; [34, 48]) has been proposed for bias correction. The core idea of applying SIMEX to linking is to regard DIF as measurement error [35]. The variance of DIF effects has to be known for this method, but it can be computed based on empirical data (see below). We now describe the application of SIMEX to a general linking function H. Therefore, the description is independent of a particular choice of the function H.

Assume that a preliminary estimate of μ and σ is available as

$$(\mu^*, \sigma^*) = \arg\min_{(\mu,\sigma)} H(\mu, \sigma, \widehat{a}_1, \widehat{b}_1, \widehat{a}_2, \widehat{b}_2) \; . \tag{11}$$

Uniform DIF effects e_i can be estimated by the equation

$$\widehat{e}_i = \mu^* + \sigma^* \widehat{b}_{i2} - \widehat{b}_{i1} \; . \tag{12}$$

Then, the DIF variance τ^2 can be estimated as

$$\hat{\tau}^2 = \frac{1}{I} \sum_{i=1}^{I} (\widehat{e}_i - \bar{e})^2 - \frac{1}{I} \sum_{i=1}^{I} v_{\widehat{e}_i} \; , \tag{13}$$

where $v_{\widehat{e}_i}$ is the variance estimate of the DIF estimate $\widehat{e}_i$ and $\bar{e} = I^{-1} \sum_{i=1}^{I} \widehat{e}_i$. Afterwards, adapted DIF effects $e_i^*(\lambda) = \widehat{e}_i + u_i(\lambda)$ with a larger variance $(1+\lambda)\hat{\tau}^2$ with $\lambda = 0.5, 1.0, 1.5$, and 2.0 are calculated, where $u_i(\lambda)$ is a random draw from a normal distribution with zero mean and a variance $\lambda \hat{\tau}^2$.

The application of SIMEX to linking must distinguish the case in which DIF occurs in one group or two groups. In the case of DIF in one group, adapted item difficulties are computed as

$$b_{i2}^*(\lambda) = \frac{1}{\sigma^*} \left(\widehat{b}_{i1} + e_i^*(\lambda) - \mu^* \right) = \widehat{b}_{i2} + \frac{1}{\sigma^*} u_i(\lambda) \; . \tag{14}$$

We collect all pseudo item parameters $b_{i2}^*(\lambda)$ in vector the $b_2^*(\lambda)$. The estimate of the mean μ and the standard deviation σ do now depend on λ and are defined as

$$(\widehat{\mu}(\lambda), \widehat{\sigma}(\lambda)) = \arg\min_{(\mu,\sigma)} H(\mu, \sigma, \widehat{a}_1, \widehat{b}_1, \widehat{a}_2, b_2^*(\lambda)) \; . \tag{15}$$

These estimates are obtained based on item parameters that possess more DIF variance than what is found in the original data. Next, compute a regression function of $\widehat{\mu}(\lambda)$ and $\widehat{\sigma}(\lambda)$ as a quadratic function of λ. For the mean μ, one obtains the regression function

$$\widehat{\mu}(\lambda) \simeq \alpha_0 + \alpha_1 \lambda + \alpha_2 \lambda^2 \; . \tag{16}$$

Finally, the parameter estimate $\widehat{\mu}$ and $\widehat{\sigma}$ is obtained by inserting the value $\lambda = -1$ in the regression function that corresponds to an estimate with DIF variance of zero. In more detail, the final parameter estimate based on (16) is obtained as $\widehat{\mu} = \widehat{\mu}(-1) = \alpha_0 - \alpha_1 + \alpha_2$. The estimate for $\widehat{\sigma}$ is obtained in the same way.

We now describe the application of SIMEX to linking if DIF occurs in in two groups. The

same DIF effects e_i^* as defined above are used. In contrast to the case that DIF is induced in only one group, the item parameters for the first group are also adapted. The modified item parameters are given by

$$b_{i1}^*(\lambda) = \widehat{b}_{i1} + \frac{1}{2}\widehat{e}_i - \frac{1}{2}e_i^*(\lambda) = \widehat{b}_{i1} - \frac{1}{2}u_i(\lambda) \text{ and} \tag{17}$$

$$b_{i2}^*(\lambda) = \frac{1}{\sigma^*}\left(\widehat{b}_{i1} + \frac{1}{2}\widehat{e}_i + \frac{1}{2}e_i^*(\lambda) - \mu^*\right) = \widehat{b}_{i2} + \frac{1}{2\sigma^*}u_i(\lambda)\,. \tag{18}$$

We collect all pseudo item parameters $b_{ig}^*(\lambda)$ in the vectors $b_g^*(\lambda)$ for $g = 1,2$. The estimates for μ and σ, depending on λ, are defined as

$$(\widehat{\mu}(\lambda), \widehat{\sigma}(\lambda)) = \arg\min_{(\mu,\sigma)} H(\mu, \sigma, \widehat{a}_1, b_1^*(\lambda), \widehat{a}_2, b_2^*(\lambda))\,. \tag{19}$$

Again, SIMEX-based estimates of μ and σ are obtained by the extrapolation of the quadratic regression functions at the value $\lambda = -1$.

In a practical implementation of SIMEX-based linking, the Monte Carlo simulation uncertainty can be reduced by using quasi-random methods and systematic perturbation of DIF effects instead of pure simulation-based techniques (see [35]).

3.2. Stocking-Lord linking

3.2.1 Asymmetric Stocking-Lord linking

Stocking-Lord linking has been originally proposed in [30] and uses the linking function

$$H(\mu, \sigma, \widehat{a}_1, \widehat{b}_1, \widehat{a}_2, \widehat{b}_2) = \sum_{t=1}^{T} \omega_t \left[\sum_{i=1}^{I} \Psi(\widehat{a}_{i1}(\sigma\theta_t + \mu - \widehat{b}_{i1})) - \sum_{i=1}^{I} \Psi(\widehat{a}_{i2}(\theta_t - \widehat{b}_{i2}))\right]^2 . \tag{20}$$

Note that this function is referred to as asymmetric Stocking-Lord linking because it aligns the test characteristic function (TCF) in the first group onto the TCF in the second group.

3.2.2 Symmetric Stocking-Lord linking

Stocking-Lord linking has also been proposed for a symmetric link function [33, 47] using the linking function

$$\begin{aligned} H(\mu, \sigma, \widehat{a}_1, \widehat{b}_1, \widehat{a}_2, \widehat{b}_2) = &\sum_{t=1}^{T} \omega_t \left[\sum_{i=1}^{I} \Psi(\widehat{a}_{i1}(\sigma\theta_t + \mu - \widehat{b}_{i1})) - \sum_{i=1}^{I} \Psi(\widehat{a}_{i2}(\theta_t - \widehat{b}_{i2}))\right]^2 \\ &+ \sum_{t=1}^{T} \omega_t \left[\sum_{i=1}^{I} \Psi(\widehat{a}_{i1}(\theta_t - \widehat{b}_{i1})) - \sum_{i=1}^{I} \Psi(\widehat{a}_{i2}(\sigma^{-1}(\theta_t - \mu) - \widehat{b}_{i2}))\right]^2 . \end{aligned} \tag{21}$$

The resulting symmetric Stocking-Lord linking has the advantage that it does not depend on the order of the definition of groups.

3.2.3 SIMEX-based Stocking-Lord linking

SIMEX-based linking as described in Section 3.1.3 can be applied to asymmetric and symmetric Stocking-Lord linking. It has been shown that applying SIMEX to Stocking-Lord linking substantially reduced the bias in the estimate of the group mean and the group standard deviation [35].

4. Bias and variance in linking methods under random DIF

In this section, the bias and the variance of the estimate of the mean and the standard deviation are derived for a general linking method H if random uniform DIF occurs in one or two groups. The bias has already been derived for the case of one group in Ref. [19].

Let $\widehat{\boldsymbol{\delta}} = (\widehat{\mu}, \widehat{\sigma})$ denotes the vector of parameter estimates that solves the nonlinear equation

$$H_{\boldsymbol{\delta}}(\widehat{\boldsymbol{\delta}}, a_1, b_1, a_2, b_2) = 0 \,, \tag{22}$$

where $H_{\boldsymbol{\delta}}$ denotes the vector of partial derivatives of H with respect to $\boldsymbol{\delta}$. We would like to emphasize that the notation in (22) includes (original) item parameters a_g and b_g, not identified item parameters $\widehat{a}_g$ and $\widehat{b}_g$ ($g = 1, 2$). Now assume uniform DIF effects e_i that fulfill $\mathsf{E}(e_i) = 0$ and $\mathsf{E}(e_i^2) = \tau^2$. Then, we get $a_1 = a_2 \equiv a$, and it follows from (22)

$$H_{\boldsymbol{\delta}}(\widehat{\boldsymbol{\delta}}, a, b_1, a, b_2) = 0 \,. \tag{23}$$

4.1. Random DIF in one group

We now derive the bias and variance in the case that random uniform DIF $e = (e_1, \ldots, e_I)$ occurs in one group. In this case, we have $b_1 = b$ and $b_2 = b + e$ with a vector of common item difficulties b (see Section 2). We apply a second-order Taylor expansion of $H_{\boldsymbol{\delta}}$ around $\boldsymbol{\delta}$ and b_2

$$\begin{aligned} 0 &= H_{\boldsymbol{\delta}}(\widehat{\boldsymbol{\delta}}, a, b_1, a, b_2) \\ &= H_{\boldsymbol{\delta}}(\boldsymbol{\delta}, a, b, a, b) + H_{\boldsymbol{\delta}\boldsymbol{\delta}}(\boldsymbol{\delta}, a, b, a, b) \left(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}\right) \\ &\quad + \sum_{i=1}^{I} H_{\boldsymbol{\delta} b_{i2}}(\boldsymbol{\delta}, a, b, a, b) e_i + \frac{1}{2} \sum_{i=1}^{I} H_{\boldsymbol{\delta} b_{i2} b_{i2}}(\boldsymbol{\delta}, a, b, a, b) e_i^2 \,. \end{aligned} \tag{24}$$

Due to $H_{\boldsymbol{\delta}}(\boldsymbol{\delta}, a, b, a, b) = 0$, we arrive at

$$\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta} = -H_{\boldsymbol{\delta}\boldsymbol{\delta}}^{-1} \left[\sum_{i=1}^{I} H_{\boldsymbol{\delta} b_{i2}} e_i + \frac{1}{2} \sum_{i=1}^{I} H_{\boldsymbol{\delta} b_{i2} b_{i2}} e_i^2 \right] \,. \tag{25}$$

Note that we suppress arguments in the derivatives in (25). For example, we write $H_{\boldsymbol{\delta}\boldsymbol{\delta}} = H_{\boldsymbol{\delta}\boldsymbol{\delta}}(\boldsymbol{\delta}, a, b, a, b)$. Then, we can compute the bias from (25) by relying on $\mathsf{E}(e_i) = 0$ as

$$\mathsf{Bias}(\widehat{\boldsymbol{\delta}}) = -\frac{1}{2} H_{\boldsymbol{\delta}\boldsymbol{\delta}}^{-1} \left[\sum_{i=1}^{I} H_{\boldsymbol{\delta} b_{i2} b_{i2}} \right] \tau^2 \,. \tag{26}$$

Furthermore, the variance can be computed from (25) by only considering linear terms e_i and ignoring quadratic terms e_i^2:

$$\mathrm{Var}(\widehat{\boldsymbol{\delta}}) = H_{\boldsymbol{\delta}\boldsymbol{\delta}}^{-1} \left[\sum_{i=1}^{I} H_{\boldsymbol{\delta} b_{i2}} H_{\boldsymbol{\delta} b_{i2}}^{\top} \right] H_{\boldsymbol{\delta}\boldsymbol{\delta}}^{-\top} \tau^2 \, . \tag{27}$$

4.2. Random DIF in two groups

We now derive the bias and the variance of the estimate $\widehat{\boldsymbol{\delta}}$ in the case that random uniform DIF occurs in two groups. In this case, we have $b_1 = b - e/2$ and $b_2 = b + e/2$ (see Section 2). Then, a second-order Taylor expansion with respect to $\boldsymbol{\delta}$, b_1 and b_2 is applied, which provides

$$\begin{aligned} 0 &= H_{\boldsymbol{\delta}}(\widehat{\boldsymbol{\delta}}, a, b_1, a, b_2) \\ &= H_{\boldsymbol{\delta}}(\boldsymbol{\delta}, a, b, a, b) + H_{\boldsymbol{\delta}\boldsymbol{\delta}} \left(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta} \right) \\ &\quad + \frac{1}{2} \sum_{i=1}^{I} \left(H_{\boldsymbol{\delta} b_{i2}} - H_{\boldsymbol{\delta} b_{i1}} \right) e_i + \frac{1}{8} \sum_{i=1}^{I} \left(H_{\boldsymbol{\delta} b_{i2} b_{i2}} + H_{\boldsymbol{\delta} b_{i1} b_{i1}} - 2 H_{\boldsymbol{\delta} b_{i1} b_{i2}} \right) e_i^2 \, . \end{aligned} \tag{28}$$

The bias and the variance can be derived as

$$\mathrm{Bias}(\widehat{\boldsymbol{\delta}}) = -\frac{1}{8} H_{\boldsymbol{\delta}\boldsymbol{\delta}}^{-1} \left[\sum_{i=1}^{I} \left(H_{\boldsymbol{\delta} b_{i2} b_{i2}} + H_{\boldsymbol{\delta} b_{i1} b_{i1}} - 2 H_{\boldsymbol{\delta} b_{i1} b_{i2}} \right) \right] \tau^2 \text{ and} \tag{29}$$

$$\mathrm{Var}(\widehat{\boldsymbol{\delta}}) = \frac{1}{4} H_{\boldsymbol{\delta}\boldsymbol{\delta}}^{-1} \left[\sum_{i=1}^{I} \mathbf{U}_i \mathbf{U}_i^{\top} \right] H_{\boldsymbol{\delta}\boldsymbol{\delta}}^{-\top} \tau^2 \text{ , where } \mathbf{U}_i = H_{\boldsymbol{\delta} b_{i2}} - H_{\boldsymbol{\delta} b_{i1}} \, . \tag{30}$$

5. Analytical derivation of the bias for Haebara linking

In this section, we illustrate the bias formulas from Section 4 for asymmetric and symmetric Haebara linking.

5.1. Asymmetric Haebara linking

The linking function for asymmetric Haebara linking (see Section 3.1.1) is given by

$$H(\mu, \sigma, \widehat{a}_1, \widehat{b}_1, \widehat{a}_2, \widehat{b}_2) = \sum_{i=1}^{I} \sum_{t=1}^{T} \omega_t \Lambda_{1it}^2 \text{ , where} \tag{31}$$

$$\Lambda_{1it} = \Psi(\widehat{a}_{i1}(\sigma\theta_t + \mu - \widehat{b}_{i1})) - \Psi(\widehat{a}_{i2}(\theta_t - \widehat{b}_{i2})) \, . \tag{32}$$

One can simplify (32) in the presence of uniform DIF effects using (3) and $a_{i1} = a_{i2} = a_i$ based on identified item parameters

$$\Lambda_{1it} = \Psi(a_i(\sigma\theta_t + \mu - b_{i1})) - \Psi(a_i(\sigma\theta_t + \mu - b_{i2})) \, . \tag{33}$$

The following derivations use the general parameter κ, denoting $\kappa = \mu$ or $\kappa = \sigma$. For $\kappa = \mu$, we define $f_t = 1$, and we define $f_t = \theta_t$ for $\kappa = \sigma$. The estimating equations for $\kappa = \mu$ and $\kappa = \sigma$ are given as

$$H_{\kappa} = 2 \sum_{i=1}^{I} \sum_{t=1}^{T} \omega_t a_i f_t \Lambda_{1it} \Psi'(a_i(\sigma\theta_t + \mu - b_{i1})) \, . \tag{34}$$

The first-order and second-order derivatives of H_κ with respect to b_{i2} and b_{i1} are calculated as

$$H_{\kappa b_{i2}} = 2\sum_{t=1}^{T} \omega_t a_i^2 f_t \Psi'(a_i(\sigma\theta_t + \mu - b_{i2}))\Psi'(a_i(\sigma\theta_t + \mu - b_{i1}))\,, \tag{35}$$

$$H_{\kappa b_{i2} b_{i2}} = -2\sum_{t=1}^{T} \omega_t a_i^3 f_t \Psi''(a_i(\sigma\theta_t + \mu - b_{i2}))\Psi'(a_i(\sigma\theta_t + \mu - b_{i1}))\,, \tag{36}$$

$$\begin{aligned} H_{\kappa b_{i1}} \;=\; & -2\sum_{t=1}^{T} \omega_t a_i^2 f_t \left[\Psi'(a_i(\sigma\theta_t + \mu - b_{i1}))\right]^2 \\ & -2\sum_{t=1}^{T} \omega_t a_i^2 f_t \Lambda_{1it} \Psi''(a_i(\sigma\theta_t + \mu - b_{i1}))\,, \end{aligned} \tag{37}$$

$$\begin{aligned} H_{\kappa b_{i1} b_{i1}} \;=\; & 4\sum_{t=1}^{T} \omega_t a_i^3 f_t \Psi'(a_i(\sigma\theta_t + \mu - b_{i1}))\Psi''(a_i(\sigma\theta_t + \mu - b_{i1})) \\ & +2\sum_{t=1}^{T} \omega_t a_i^3 f_t \Psi'(a_i(\sigma\theta_t + \mu - b_{i1}))\Psi''(a_i(\sigma\theta_t + \mu - b_{i1})) \\ & +2\sum_{t=1}^{T} \omega_t a_i^3 f_t \Lambda_{1it} \Psi'''(a_i(\sigma\theta_t + \mu - b_{i1})) \text{ and} \end{aligned} \tag{38}$$

$$H_{\kappa b_{i1} b_{i2}} = -2\sum_{t=1}^{T} \omega_t a_i^3 f_t \Psi'(a_i(\sigma\theta_t + \mu - b_{i2}))\Psi''(a_i(\sigma\theta_t + \mu - b_{i1}))\,. \tag{39}$$

We now evaluate the derivatives at $b_{i1} = b_i$ and $b_{i2} = b_i$. Using the abbreviation $\eta_{it} = a_i(\sigma\theta_t + \mu - b_i)$, we obtain

$$H_{\kappa b_{i2} b_{i2}} = -2\sum_{t=1}^{T} \omega_t a_i^3 f_t \Psi'(\eta_{it})\Psi''(\eta_{it})\,, \tag{40}$$

$$H_{\kappa b_{i1} b_{i1}} = 6\sum_{t=1}^{T} \omega_t a_i^3 f_t \Psi'(\eta_{it})\Psi''(\eta_{it}) \text{ and} \tag{41}$$

$$H_{\kappa b_{i1} b_{i2}} = -2\sum_{t=1}^{T} \omega_t a_i^3 f_t \Psi'(\eta_{it})\Psi''(\eta_{it}) \tag{42}$$

Finally, we obtain the bias-determining term corresponding to the case that DIF occurs in two groups (see (29)) as

$$H_{\kappa b_{i2} b_{i2}} + H_{\kappa b_{i1} b_{i1}} - 2H_{\kappa b_{i1} b_{i2}} = 8\sum_{t=1}^{T} \omega_t a_i^3 f_t \Psi'(\eta_{it})\Psi''(\eta_{it})\,. \tag{43}$$

We now illustrate the bias-determining terms for μ and σ as a function of a common item difficulty b_i. We choose $\mu = 0.3$ and $\sigma = 1.2$ as in the Simulation Study presented in the next Section 6. Note that the Hessian matrix $H_{\boldsymbol{\delta}\boldsymbol{\delta}}$ is positive definite for a linking function H that attains the parameter estimate $\widehat{\boldsymbol{\delta}}$ as the minimizer.

Figure 1 presents the bias-determining terms $H_{\kappa b_2 b_2}$ for $\kappa = \mu, \sigma$ for asymmetric Haebara linking if DIF occurs in one group. The derivative with respect to μ can be positive or negative values, depending on the difference $\mu - b$ is negative or positive. The bias-determining term

for σ is positive for a broad range of b values for which $\mu - b$ does not deviate too much from 0. Hence, a negative bias for $\widehat{\sigma}$ can be expected because of the negative sign of the bias in (26).

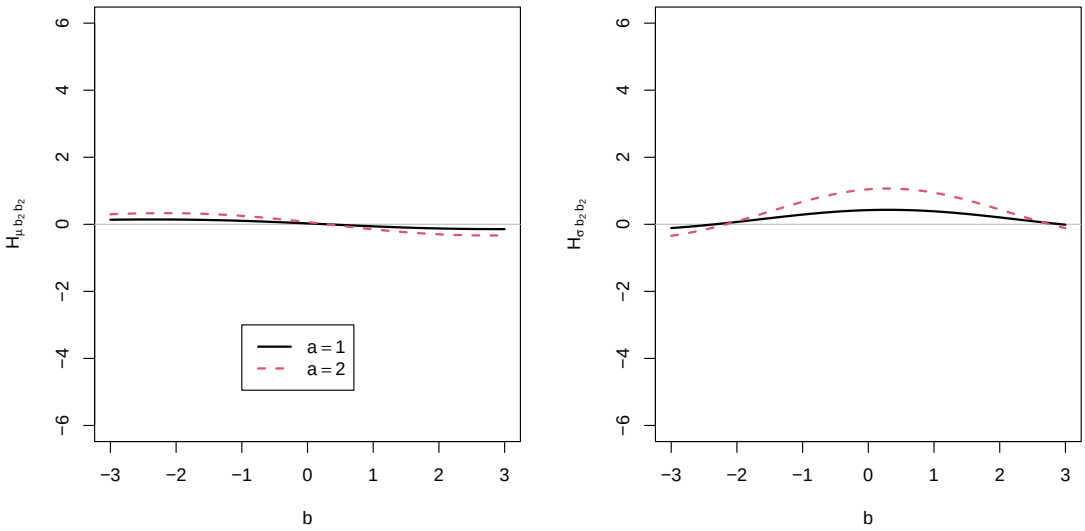

Figure 1. Bias-determining term $H_{\kappa b_2 b_2}$ for $\kappa = \mu$ (left panel) and $\kappa = \sigma$ (right panel) (see (26)) for asymmetric Haebara linking in a test with one item with item discriminations $a = 1$ or $a = 2$ as a function of common item difficulty b with group mean $\mu = 0.3$ and group standard deviation $\sigma = 1.2$.

Figure 2 illustrates the bias-determining term $H_{\kappa b_2 b_2} + H_{\kappa b_1 b_1} - 2H_{\kappa b_1 b_2}$ for asymmetric Haebara linking of random DIF occurs in two groups. It is evident that the bias term for σ now has a negative sign, which subsequently will result in a positive expected bias of $\widehat{\sigma}$ according to (29).

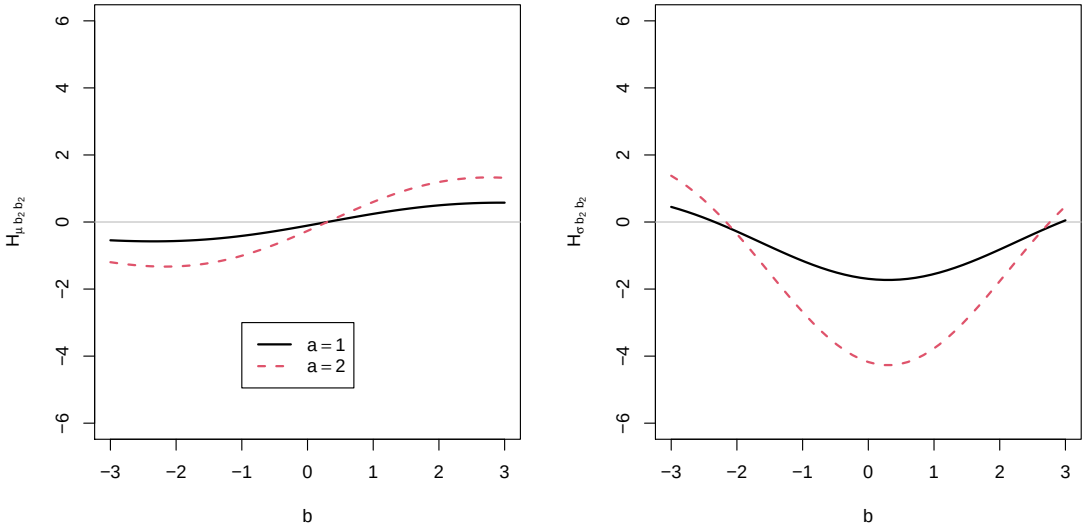

Figure 2. Bias-determining term $H_{\kappa b_2 b_2} + H_{\kappa b_1 b_1} - 2H_{\kappa b_1 b_2}$ for $\kappa = \mu$ (left panel) and $\kappa = \sigma$ (right panel) (see (29)) for asymmetric Haebara linking in a test with one item with item discriminations $a = 1$ or $a = 2$ as a function of common item difficulty b with group mean $\mu = 0.3$ and group standard deviation $\sigma = 1.2$.

5.2. Symmetric Haebara linking

The linking function for symmetric Haebara linking (see Section 3.1.2) is given by

$$H(\mu, \sigma, \widehat{a}_1, \widehat{b}_1, \widehat{a}_2, \widehat{b}_2) = \sum_{i=1}^{I} \sum_{t=1}^{T} \omega_t \Lambda_{1it}^2 + \sum_{i=1}^{I} \sum_{t=1}^{T} \omega_t \Lambda_{2it}^2 \text{ , where} \tag{44}$$

$$\Lambda_{1it} = \Psi(\widehat{a}_{i1}(\sigma\theta_t + \mu - \widehat{b}_{i1})) - \Psi(\widehat{a}_{i2}(\theta_t - \widehat{b}_{i2})) \text{ and} \tag{45}$$

$$\Lambda_{2it} = \Psi(\widehat{a}_{i1}(\theta_t - \widehat{b}_{i1})) - \Psi(\widehat{a}_{i2}(\sigma^{-1}(\theta_t - \mu) - \widehat{b}_{i2})) \,. \tag{46}$$

We can simplify (45) and (46) with $a_{i1} = a_{i2} = a_i$ and using identified item parameters (see (3)) and arrive at

$$\Lambda_{1it} = \Psi(a_i(\sigma\theta_t + \mu - b_{i1})) - \Psi(a_i(\sigma\theta_t + \mu - b_{i2})) \text{ and} \tag{47}$$

$$\Lambda_{2it} = \Psi(a_i(\theta_t - b_{i1})) - \Psi(a_i(\theta_t - b_{i2})) \,. \tag{48}$$

We set $g_t = \sigma^{-1}$ for $\kappa = \mu$ and $g_t = \sigma^{-2}(\theta_t - \mu)$ for $\kappa = \sigma$ and obtain the estimating equations

$$H_\kappa = 2\sum_{i=1}^{I}\sum_{t=1}^{T} \omega_t a_i f_t \Lambda_{1it} \Psi'(a_i(\sigma\theta_t + \mu - b_{i1})) + 2\sum_{i=1}^{I}\sum_{t=1}^{T} \omega_t a_i g_t \Lambda_{2it} \Psi'(a_i(\theta_t - b_{i2})) \tag{49}$$

The first-order and second-order derivatives are given by

$$\begin{aligned} H_{\kappa b_{i2}} = {} & 2\sum_{t=1}^{T} \omega_t a_i^2 f_t \Psi'(a_i(\sigma\theta_t + \mu - b_{i2}))\Psi'(a_i(\sigma\theta_t + \mu - b_{i1})) \\ & + 2\sum_{i=1}^{I}\sum_{t=1}^{T} \omega_t a_i^2 g_t \left[\Psi'(a_i(\theta_t - b_{i2}))\right]^2 \\ & - 2\sum_{i=1}^{I}\sum_{t=1}^{T} \omega_t a_i^2 g_t \Lambda_{2it} \Psi''(a_i(\theta_t - b_{i2})) \,, \end{aligned} \tag{50}$$

$$\begin{aligned} H_{\kappa b_{i2} b_{i2}} = {} & -2\sum_{t=1}^{T} \omega_t a_i^3 f_t \Psi''(a_i(\sigma\theta_t + \mu - b_{i2}))\Psi'(a_i(\sigma\theta_t + \mu - b_{i1})) \\ & - 6\sum_{i=1}^{I}\sum_{t=1}^{T} \omega_t a_i^3 g_t \Psi'(a_i(\theta_t - b_{i2}))\Psi''(a_i(\theta_t - b_{i2})) \\ & + 2\sum_{i=1}^{I}\sum_{t=1}^{T} \omega_t a_i^3 g_t \Lambda_{2it} \Psi'''(a_i(\theta_t - b_{i2})) \,, \end{aligned} \tag{51}$$

$$\begin{aligned} H_{\kappa b_{i1}} = {} & -2\sum_{i=1}^{I}\sum_{t=1}^{T} \omega_t a_i^2 f_t \left[\Psi'(a_i(\sigma\theta_t + \mu - b_{i1}))\right]^2 \\ & - 2\sum_{i=1}^{I}\sum_{t=1}^{T} \omega_t a_i^2 f_t \Lambda_{1it} \Psi''(a_i(\sigma\theta_t + \mu - b_{i1})) \\ & - 2\sum_{i=1}^{I}\sum_{t=1}^{T} \omega_t a_i^2 g_t \Psi'(a_i(\theta_t - b_{i1}))\Psi'(a_i(\theta_t - b_{i2})) \,, \end{aligned} \tag{52}$$

$$
\begin{aligned}
H_{\kappa b_{i1} b_{i1}} &= 6\sum_{i=1}^{I}\sum_{t=1}^{T} \omega_t a_i^3 f_t \Psi'(a_i(\sigma\theta_t + \mu - b_{i1}))\Psi''(a_i(\sigma\theta_t + \mu - b_{i1})) \\
&\quad + 2\sum_{i=1}^{I}\sum_{t=1}^{T} \omega_t a_i^3 f_t \Lambda_{1it} \Psi'''(a_i(\sigma\theta_t + \mu - b_{i1})) \\
&\quad + 2\sum_{i=1}^{I}\sum_{t=1}^{T} \omega_t a_i^3 g_t \Psi''(a_i(\theta_t - b_{i1}))\Psi'(a_i(\theta_t - b_{i2})) \text{ and} \\
H_{\kappa b_{i1} b_{i2}} &= -2\sum_{i=1}^{I}\sum_{t=1}^{T} \omega_t a_i^3 f_t \Psi'(a_i(\sigma\theta_t + \mu - b_{i2}))\Psi''(a_i(\sigma\theta_t + \mu - b_{i1})) \\
&\quad + 2\sum_{i=1}^{I}\sum_{t=1}^{T} \omega_t a_i^3 g_t \Psi'(a_i(\theta_t - b_{i1}))\Psi''(a_i(\theta_t - b_{i2})) \ .
\end{aligned}
\tag{53, 54}
$$

We can now evaluate the derivatives at $b_{i1} = b_{i2} = b_i$ and obtain by using the abbreviations $\eta_{it} = a_i(\sigma\theta_t + \mu - b_i)$ and $\gamma_{it} = a_i(\theta_t - b_i)$

$$
H_{\kappa b_{i2} b_{i2}} = -2\sum_{t=1}^{T} \omega_t a_i^3 f_t \Psi'(\eta_{it})\Psi''(\eta_{it}) - 6\sum_{i=1}^{I}\sum_{t=1}^{T} \omega_t a_i^3 g_t \Psi'(\gamma_{it})\Psi''(\gamma_{it}) \ , \tag{55}
$$

$$
H_{\kappa b_{i1} b_{i1}} = 6\sum_{i=1}^{I}\sum_{t=1}^{T} \omega_t a_i^3 f_t \Psi'(\eta_{it})\Psi''(\eta_{it}) + 2\sum_{i=1}^{I}\sum_{t=1}^{T} \omega_t a_i^3 g_t \Psi'(\gamma_{it})\Psi''(\gamma_{it}) \text{ and} \tag{56}
$$

$$
H_{\kappa b_{i1} b_{i2}} = -2\sum_{i=1}^{I}\sum_{t=1}^{T} \omega_t a_i^3 f_t \Psi'(\eta_{it})\Psi''(\eta_{it}) + 2\sum_{i=1}^{I}\sum_{t=1}^{T} \omega_t a_i^3 g_t \Psi'(\gamma_{it})\Psi''(\gamma_{it}) \ . \tag{57}
$$

Finally, the bias-determining term in the case that DIF occurs in two groups is given as

$$
\begin{aligned}
&H_{\kappa b_{i1} b_{i2}} + H_{\kappa b_{i1} b_{i1}} - 2H_{\kappa b_{i1} b_{i2}} \\
&= 8\sum_{i=1}^{I}\sum_{t=1}^{T} \omega_t a_i^3 f_t \Psi'(\eta_{it})\Psi''(\eta_{it}) - 8\sum_{i=1}^{I}\sum_{t=1}^{T} \omega_t a_i^3 g_t \Psi'(\gamma_{it})\Psi''(\gamma_{it}) \ .
\end{aligned}
\tag{58}
$$

Figure 3 illustrates the bias-determining terms $H_{\kappa b_2 b_2}$ for symmetric Haebara linking if DIF occurs in one group. The term is positive for σ, which implies a negative expected bias for $\widehat{\sigma}$.

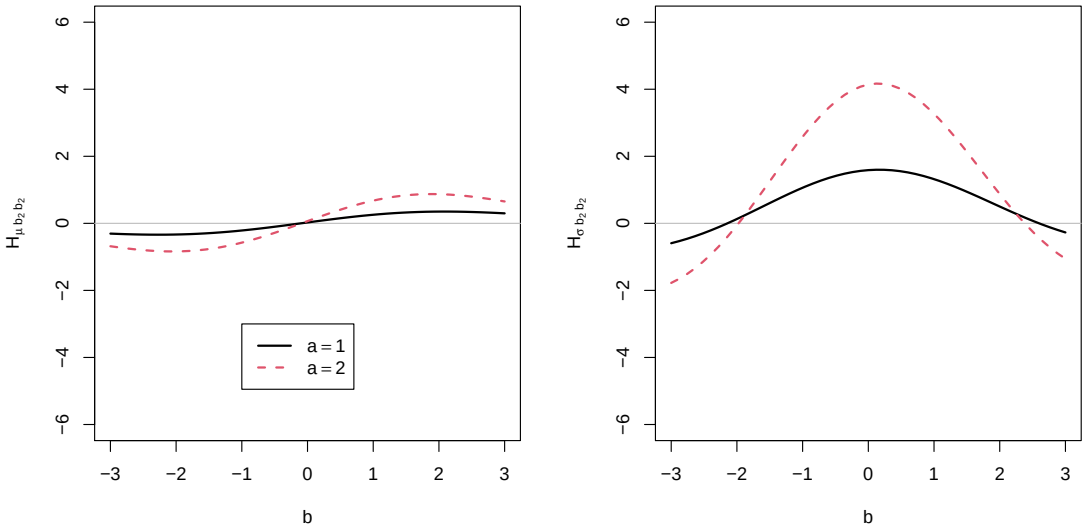

Figure 3. Bias-determining term $H_{\kappa b_2 b_2}$ for $\kappa = \mu$ (left panel) and $\kappa = \sigma$ (right panel) (see (26)) for symmetric Haebara linking in a test with one item with item discriminations $a = 1$ or $a = 2$ as a function of common item difficulty b with group mean $\mu = 0.3$ and group standard deviation $\sigma = 1.2$.

Figure 4 illustrates the bias-determining term $H_{\kappa b_2 b_2} + H_{\kappa b_1 b_1} - 2H_{\kappa b_1 b_2}$ for symmetric Haebara linking if DIF occurs in two groups. It can be seen that the term has only small negative values for σ. Hence, only small biases for $\widehat{\sigma}$ can be expected for symmetric Haebara linking if DIF occurs in two groups.

We would like to emphasize that the bias-determining term $H_{\kappa b_{i1} b_{i2}} + H_{\kappa b_{i1} b_{i1}} - 2H_{\kappa b_{i1} b_{i2}}$ in (58) equals zero if $\mu = 0$ and $\sigma = 1$ (*i.e.*, there are no differences in the distributions of the two groups). In this case, we have $f_t = g_t$ and $\eta_{it} = \gamma_{it}$, which implies the finding.

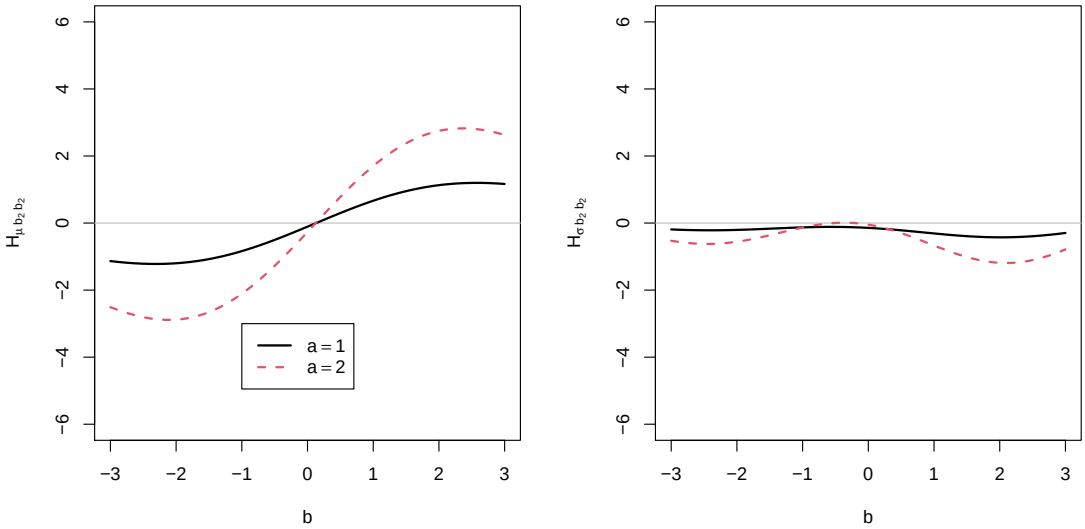

Figure 4. Bias-determining term $H_{\kappa b_2 b_2} + H_{\kappa b_1 b_1} - 2H_{\kappa b_1 b_2}$ for $\kappa = \mu$ (left panel) and $\kappa = \sigma$ (right panel) (see (29)) for symmetric Haebara linking in a test with one item with item discriminations $a = 1$ or $a = 2$ as a function of common item difficulty b with group mean $\mu = 0.3$ and group standard deviation $\sigma = 1.2$.

6. Simulation study

6.1. Method

This simulation study used the 2PL model to simulate item responses in two groups. The mean and standard deviation of the normally distributed ability variable θ in the first group were set to 0 to 1, and the mean and SD for the normally distributed ability variable θ in the second group were set to $\mu = 0.3$ and $\sigma = 1.2$, respectively.

The number of items I in the simulation was varied as 10, 20, and 40. The group-specific item parameters a_{ig} and b_{ig} for $i = 1,\ldots,I$ and $g = 1,2$ relied on common item parameters a_i and b_i that were held fixed in the simulation and a random uniform and normally distributed DIF effect. Uniform DIF effects were simulated in each replication of the simulation study. The same item parameters as in Ref. [19] were used. In the case of $I = 10$ items, the common item discriminations a_i were chosen as 0.83, 1.02, 0.88, 0.80, 1.04, 0.95, 1.00, 1.13, 1.32, and 1.11. The mean of the a_i parameters was $M = 1.008$ and the standard deviation was $SD = 0.156$, referring to a situation of a test with items that had medium item discrimination. The common item difficulties b_i were chosen as -1.74, -1.22, -0.22, 0.54, -0.04, -0.39, -0.73, 0.30, 0.83, and -1.39. The b_i parameters had a mean $M = -0.406$ and a standard deviation $SD = 0.857$, resulting in a test with items that were slightly easier compared to the average ability in the population (*i.e.*, resulting in marginal item response probabilities slightly larger than 0.50). The item parameters of the 10 items were duplicated for item numbers as multiples of 10 (*i.e.*, in the cases $I = 20$ and $I = 40$).

Two types of DIF effects were simulated. DIF could occur in one group (*i.e.*, in the second group, denoted by 1G) or in the two groups (denoted by 2G). In both cases, the group-specific item discriminations a_{ig} were chosen to be equal (*i.e.*, $a_{i1} = a_{i2} = a_i$). In the case that DIF occurs in one group (*i.e.*, the case 1G), we defined $b_{i1} = b_i$ and $b_{i2} = b_i + e_i$. In the case that DIF occurs in two groups (*i.e.*, the case 2G), we defined $b_{i1} = b_i - e_i/2$ and $b_{i2} = b_i + e_i/2$. The uniform DIF effects e_i were simulated from a normal distribution with zero mean and a DIF standard deviation $\tau = 0.50$.

Item responses from the 2PL model were simulated for finite sample sizes $N = 500$, 1000, and 2000. Moreover, we investigated an infinite sample size in which only computed identified item parameters $\widehat{a}_{ig}$ and $\widehat{b}_{ig}$ ($i = 1,\ldots,I$; $g = 1,2$) without simulating item responses.

Different linking methods were compared in this simulation study. First, we investigated mean-geometric-mean (MGM) linking [11, 25, 42]. Moreover, we applied asymmetric and symmetric Haebara linking (AHA and SHA, respectively) as well as asymmetric and symmetric Stocking-Lord linking (ASL and SSL, respectively). Moreover, we applied SIMEX-based linking for the four latter linking methods AHA, SHA, ASL, and SSL. SIMEX-based linking was applied under the assumption that DIF was induced in SIMEX estimation either in only one group (*i.e.*, method SI1G) or in both groups (*i.e.*, method SI2G), as described in Section 3.1.3. Therefore, $4 \times 2 = 8$ SIMEX-based linking methods were utilized in this study. SIMEX estimation was carried out for λ values of 0.5, 1.0, 1.5, and 2.0. Moreover, a quasi Monte Carlo method was used instead of a fully simulation-based approach of SIMEX

as described in Ref. [35]. For $I = 10$ and $I = 20$ items, original item parameters were combined with 50 replications from an (approximately) exact normal distribution resulting in $10 \times 50 = 500$ or $20 \times 50 = 1000$ data points as input for SIMEX. In contrast, 30 replications were used in the simulation conditions involving $I = 40$ items, resulting in $40 \times 30 = 1200$ data points as input for SIMEX. Overall, we compared $1 + 4 + 8 = 13$ different linking methods in this simulation study.

In each of the 4 (sample size N) $\times$ 2 (DIF types 1G and 2G) $\times$ 3 (number of items I) $= 24$ cells of the simulation, 1,500 replications were conducted. We computed the empirical bias and the root mean square error (RMSE) for the estimated mean $\widehat{\mu}$ and the estimated standard deviation $\widehat{\sigma}$. For finite sample sizes, a relative (percentage) RMSE was computed as the ratio of the RMSE values of a particular linking method and the best-performing linking method in a condition.

The R software [49] was used for the entire analysis in this simulation study. The 2PL model was fitted using the sirt::xxirt() function in the R package sirt [50]. The author of this article wrote dedicated R functions for the different linking methods. These functions and replication material for this Simulation Study can be found at https://osf.io/6btr9/ (accessed on 1 August 2024).

6.2. Results

Table 1 presents the bias of the estimated group mean μ and the estimated standard deviation σ in the case of an infinite sample size. The MGM linking method was unbiased in all conditions. Overall, the bias was more pronounced for the standard deviation σ than for the mean μ.

If DIF occurred in only one group (*i.e.*, case 1G), biased estimates were obtained for ASL, SSL, AHA and SHA linking methods. Surprisingly, SHA was more biased than AHA. However, these findings were also predicted from the analytical derivations in Section 5. If the SIMEX-based linking methods SI1G were applied for these four linking methods, unbiased estimates were obtained. However, if the SIMEX-based method SI2G was applied, the bias was not decreased and even increased in some situations. This finding can be explained by the fact that SIMEX would incorrectly assume that DIF effects would occur in two groups, but it occurred in only one group in the data-generating model.

If DIF occurred in two groups, ASL, SSL, and SHAE were approximately unbiased. However, AHA resulted in biased estimates. These findings confirmed our predictions from Section 5 that AHA would result in a bias, while the bias would be small for SHA. For all four linking methods, SIMEX-based linking SI2G with correctly chosen DIF induced in two groups was unbiased, while bias increased if DIF was incorrectly assumed to occur in only one group in SIMEX-based linking SI1G.

Table 1. Simulation Study: Bias of estimated group mean μ and estimated group standard deviation σ for a DIF SD $\tau = 0.5$, uniform DIF in one (1G) or two (2G) groups, and an infinite sample size as a function of the number of items I.

DIF	Par	I	MGM	ASL	ASL-SI1G	ASL-SI2G	SSL	SSL-SI1G	SSL-SI2G	AHA	AHA-SI1G	AHA-SI2G	SHA	SHA-SI1G	SHA-SI2G
1G	μ	10	0.00	**−0.02**	0.01	−0.01	−0.02	0.01	−0.01	−0.01	0.01	**−0.02**	**−0.03**	0.01	**−0.02**
		20	0.00	**−0.02**	0.00	**−0.02**	**−0.02**	0.00	**−0.02**	−0.02	0.01	**−0.03**	**−0.03**	0.00	**−0.03**
		40	0.00	**−0.02**	−0.01	**−0.02**	**−0.02**	−0.01	**−0.02**	**−0.02**	0.00	**−0.03**	**−0.04**	0.00	**−0.04**
	σ	10	0.00	**−0.04**	0.01	**−0.03**	**−0.04**	0.01	**−0.03**	**−0.03**	0.01	**−0.05**	**−0.06**	0.01	**−0.05**
		20	0.00	**−0.04**	0.00	**−0.04**	**−0.04**	0.00	**−0.04**	**−0.03**	0.01	**−0.06**	**−0.07**	0.00	**−0.06**
		40	0.00	**−0.04**	0.00	**−0.04**	**−0.04**	0.00	**−0.04**	**−0.04**	0.00	**−0.07**	**−0.07**	0.00	**−0.07**
2G	μ	10	0.00	0.00	**0.03**	0.01	0.00	**0.03**	0.01	**0.02**	**0.05**	0.01	0.01	**0.05**	0.01
		20	0.00	0.00	**0.03**	0.00	0.00	**0.03**	0.00	**0.02**	**0.05**	0.01	0.00	**0.04**	0.01
		40	0.00	0.00	**0.02**	0.00	0.00	**0.02**	0.00	**0.02**	**0.04**	0.00	0.00	**0.04**	0.00
	σ	10	0.00	0.00	**0.05**	0.01	0.00	**0.05**	0.01	**0.04**	**0.08**	0.01	0.00	**0.08**	0.01
		20	0.00	0.00	**0.05**	0.01	0.00	**0.05**	0.01	**0.04**	**0.09**	0.01	0.00	**0.08**	0.01
		40	0.00	0.00	**0.05**	0.00	0.00	**0.05**	0.00	**0.04**	**0.08**	0.00	0.00	**0.08**	0.00

Note. DIF = differential item functioning; Par = parameter; MGM = mean-geometric mean linking; ASL = asymmetric Stocking-Lord linking; SSL = symmetric Stocking-Lord linking; AHA = asymmetric Haebara linking; SHA = symmetric Haebara linking; SI1G = SIMEX-based linking with assumed DIF in one group; SI2G = SIMEX-based linking with assumed DIF in two groups; Biases with absolute values of at least 0.02 are printed in bold font.

Table 2 presents the RMSE for the estimated mean and standard deviation in an infinite sample size. Generally, the RMSE decreased with an increasing number of items. In line with other studies, MGM had zero variance for the standard deviation σ in infinite samples [25]. Notably, MGM had the least RMSE across all conditions. Generally, SL had smaller RMSE values than HA linking. Interestingly, SSL had a similar RMSE to ASL, while the RMSE of SHA was smaller than AHA. It should also be emphasized that the application of the adequate SIMEX-based linking method (*i.e.*, SI1G in the DIF 1G case and SI2G in the DIF 2G case) did not substantially increase the RMSE.

Table 2. Simulation Study: Root mean square error (RMSE) of estimated group mean μ and estimated group standard deviation σ for a DIF SD $\tau = 0.5$, uniform DIF in one (1G) or two (2G) groups, and an infinite sample size as a function of the number of items I.

DIF	Par	I	MGM	ASL	ASL-SI1G	ASL-SI2G	SSL	SSL-SI1G	SSL-SI2G	AHA	AHA-SI1G	AHA-SI2G	SHA	SHA-SI1G	SHA-SI2G
1G	μ	10	0.159	0.159	0.164	0.159	0.159	0.164	0.159	0.166	0.173	0.163	0.164	0.173	0.164
		20	0.113	0.113	0.116	0.113	0.113	0.116	0.113	0.118	0.122	0.118	0.119	0.122	0.118
		40	0.079	0.081	0.081	0.081	0.081	0.081	0.081	0.084	0.085	0.086	0.088	0.085	0.087
	σ	10	0	0.062	0.055	0.061	0.063	0.056	0.061	0.089	0.093	0.096	0.100	0.093	0.096
		20	0	0.054	0.039	0.052	0.054	0.040	0.052	0.068	0.065	0.082	0.088	0.065	0.083
		40	0	0.048	0.026	0.049	0.048	0.027	0.049	0.054	0.043	0.080	0.080	0.044	0.080
2G	μ	10	0.159	0.164	0.170	0.164	0.164	0.170	0.165	0.178	0.187	0.174	0.174	0.186	0.174
		20	0.113	0.115	0.121	0.116	0.116	0.121	0.116	0.125	0.135	0.122	0.121	0.134	0.122
		40	0.079	0.082	0.085	0.081	0.082	0.085	0.082	0.089	0.098	0.086	0.086	0.096	0.086
	σ	10	0	0.052	0.075	0.053	0.053	0.076	0.054	0.099	0.129	0.090	0.088	0.130	0.090
		20	0	0.036	0.062	0.036	0.036	0.063	0.037	0.074	0.109	0.063	0.061	0.107	0.063
		40	0	0.027	0.054	0.027	0.027	0.054	0.027	0.062	0.093	0.045	0.045	0.093	0.045

Note. DIF = differential item functioning; Par = parameter; MGM = mean-geometric mean linking; ASL = asymmetric Stocking-Lord linking; SSL = symmetric Stocking-Lord linking; AHA = asymmetric Haebara linking; SHA = symmetric Haebara linking; SI1G = SIMEX-based linking with assumed DIF in one group; SI2G = SIMEX-based linking with assumed DIF in two groups.

Table 3 displays the bias of the estimates for μ and σ in finite sample sizes. Overall, the bias in finite sample sizes was very similar to the case of an infinite sample size. Moreover, the size of the bias was relatively independent of sample size N and the number of items I.

Table 3. Simulation Study: Bias of estimated group mean μ and estimated group standard deviation σ for a DIF SD $\tau = 0.5$, and uniform DIF in one (1G) or two (2G) groups as a function of sample size N and of the number of items I.

DIF	Par	I	N	MGM	ASL	ASL-SI1G	ASL-SI2G	SSL	SSL-SI1G	SSL-SI2G	AHA	AHA-SI1G	AHA-SI2G	SHA	SHA-SI1G	SHA-SI2G
1G	μ	10	500	0.01	−0.01	0.01	−0.01	−0.01	0.01	−0.01	0.00	**0.02**	**−0.02**	**−0.02**	0.01	**−0.02**
			1000	0.00	**−0.02**	0.00	**−0.02**	**−0.02**	0.00	**−0.02**	−0.01	0.01	**−0.02**	**−0.03**	0.01	**−0.03**
			2000	0.00	**−0.02**	0.01	−0.02	**−0.02**	0.01	−0.02	−0.01	0.01	**−0.02**	**−0.03**	0.01	**−0.02**
		20	500	0.01	−0.01	0.01	−0.01	−0.01	0.01	−0.01	0.00	**0.02**	**−0.02**	**−0.03**	0.01	**−0.02**
			1000	0.00	**−0.02**	0.00	**−0.02**	**−0.02**	0.01	**−0.02**	−0.01	0.01	**−0.03**	**−0.03**	0.01	**−0.03**
			2000	0.00	**−0.02**	0.00	**−0.02**	**−0.02**	0.00	**−0.02**	**−0.02**	0.01	**−0.03**	**−0.03**	0.00	**−0.03**
		40	500	0.00	**−0.02**	−0.01	**−0.02**	**−0.02**	−0.01	**−0.02**	−0.02	0.00	**−0.03**	**−0.04**	−0.01	**−0.04**
			1000	0.00	**−0.02**	−0.01	**−0.02**	**−0.02**	−0.01	**−0.02**	**−0.02**	0.00	**−0.03**	**−0.04**	0.00	**−0.04**
			2000	0.00	**−0.02**	−0.01	**−0.02**	**−0.02**	−0.01	**−0.02**	**−0.02**	0.00	**−0.03**	**−0.04**	0.00	**−0.04**
	σ	10	500	0.01	**−0.03**	0.01	**−0.03**	**−0.03**	0.01	**−0.03**	**−0.02**	**0.02**	**−0.04**	**−0.05**	0.02	**−0.05**
			1000	0.00	**−0.03**	0.01	**−0.03**	**−0.03**	0.01	**−0.03**	**−0.02**	**0.02**	**−0.05**	**−0.05**	**0.02**	**−0.05**
			2000	0.00	**−0.04**	0.01	**−0.03**	**−0.04**	0.01	**−0.03**	**−0.02**	**0.02**	**−0.05**	**−0.06**	0.01	**−0.05**
		20	500	0.01	**−0.03**	0.01	**−0.03**	**−0.03**	0.01	**−0.03**	**−0.02**	**0.02**	**−0.05**	**−0.06**	0.01	**−0.05**
			1000	0.00	**−0.04**	0.01	**−0.04**	**−0.04**	0.01	**−0.04**	**−0.03**	0.01	**−0.05**	**−0.06**	0.01	**−0.06**
			2000	0.00	**−0.04**	0.00	**−0.04**	**−0.04**	0.00	**−0.04**	**−0.03**	0.01	**−0.06**	**−0.07**	0.01	**−0.06**
		40	500	0.00	**−0.04**	0.00	**−0.04**	**−0.04**	0.00	**−0.04**	**−0.03**	0.01	**−0.06**	**−0.07**	0.00	**−0.07**
			1000	0.00	**−0.04**	0.00	**−0.04**	**−0.04**	0.00	**−0.04**	**−0.03**	0.01	**−0.07**	**−0.07**	0.00	**−0.07**
			2000	0.00	**−0.04**	0.00	**−0.04**	**−0.04**	0.00	**−0.04**	**−0.03**	0.00	**−0.07**	**−0.07**	0.00	**−0.07**
2G	μ	10	500	0.01	0.01	**0.03**	0.01	0.01	**0.03**	0.01	**0.03**	**0.06**	**0.02**	0.01	**0.05**	0.01
			1000	0.00	0.00	**0.03**	0.00	0.00	**0.03**	0.00	**0.02**	**0.05**	0.01	0.00	**0.04**	0.01
			2000	0.00	0.00	**0.03**	0.01	0.00	**0.03**	0.01	**0.03**	**0.05**	0.01	0.01	**0.05**	0.01
		20	500	0.01	0.01	**0.03**	0.01	0.01	**0.03**	0.01	**0.03**	**0.06**	0.01	0.01	**0.05**	0.01
			1000	0.00	0.00	**0.03**	0.01	0.00	**0.03**	0.01	**0.03**	**0.05**	0.01	0.00	**0.05**	0.01
			2000	0.00	0.00	**0.03**	0.01	0.00	**0.03**	0.01	**0.03**	**0.05**	0.01	0.00	**0.05**	0.01
		40	500	0.00	0.00	**0.02**	0.00	0.00	**0.02**	0.00	**0.03**	**0.05**	0.01	0.00	**0.04**	0.01
			1000	0.00	0.00	**0.02**	0.00	0.00	**0.02**	0.00	**0.03**	**0.05**	0.01	0.00	**0.04**	0.00
			2000	0.00	0.00	**0.02**	0.00	0.00	**0.02**	0.00	**0.02**	**0.04**	0.00	0.00	**0.04**	0.00
	σ	10	500	0.01	0.01	**0.05**	0.01	0.01	**0.05**	0.01	**0.05**	**0.10**	**0.02**	0.01	**0.09**	**0.02**
			1000	0.01	0.01	**0.05**	0.01	0.01	**0.05**	0.01	**0.04**	**0.09**	0.02	0.01	**0.09**	0.02
			2000	0.00	0.00	**0.05**	0.01	0.00	**0.05**	0.01	**0.04**	**0.09**	0.01	0.01	**0.09**	0.02
		20	500	0.01	0.01	**0.05**	0.01	0.01	**0.05**	0.01	**0.05**	**0.10**	**0.02**	0.01	**0.09**	0.01
			1000	0.00	0.00	**0.05**	0.01	0.00	**0.05**	0.01	**0.04**	**0.09**	0.01	0.00	**0.09**	0.01
			2000	0.00	0.00	**0.05**	0.01	0.00	**0.05**	0.01	**0.04**	**0.09**	0.01	0.00	**0.09**	0.01
		40	500	0.01	0.01	**0.05**	0.00	0.01	**0.05**	0.00	**0.05**	**0.09**	0.01	0.01	**0.09**	0.01
			1000	0.00	0.00	**0.05**	0.00	0.00	**0.05**	0.00	**0.05**	**0.09**	0.01	0.00	**0.08**	0.01
			2000	0.00	0.00	**0.05**	0.00	0.00	**0.05**	0.00	**0.04**	**0.08**	0.00	0.00	**0.08**	0.00

Note. DIF = differential item functioning; Par = parameter; MGM = mean-geometric mean linking; ASL = asymmetric Stocking-Lord linking; SSL = symmetric Stocking-Lord linking; AHA = asymmetric Haebara linking; SHA = symmetric Haebara linking; SI1G = SIMEX-based linking with assumed DIF in one group; SI2G = SIMEX-based linking with assumed DIF in two groups; Biases with absolute values of at least 0.02 are printed in bold font.

Table 4 presents the relative RMSE for the estimated group mean and standard deviation in finite sample sizes. The different linking methods did not substantially differ for the RMSE regarding the group mean μ. However, the notable difference between the linking methods were obtained for the group standard deviation σ. There were significant efficiency gains when using MGM compared to SL and HA linking methods, particularly for large sample sizes. As for infinite sample sizes, SL outperformed HA linking methods. Again, SIMEX-based linking procedures that assumed the correct DIF model did not substantially increase the RMSE compared to the original linking procedure.

Table 4. Simulation Study: Relative root mean square error (RMSE) of estimated group mean μ and estimated group standard deviation σ for a DIF SD $\tau = 0.5$, and uniform DIF in one (1G) or two (2G) groups as a function of sample size N and of the number of items I.

DIF	Par	I	N	MGM	ASL	ASL-SI1G	ASL-SI2G	SSL	SSL-SI1G	SSL-SI2G	AHA	AHA-SI1G	AHA-SI2G	SHA	SHA-SI1G	SHA-SI2G
1G	μ	10	500	105	100	104	100	100	104	100	106	111	104	104	109	104
			1000	102	100	103	100	100	103	100	106	110	105	104	109	104
			2000	101	100	103	100	100	103	100	106	111	105	105	110	105
		20	500	105	100	104	100	100	104	100	106	111	104	104	109	104
			1000	102	100	103	100	100	103	100	105	109	104	105	108	104
			2000	101	100	102	100	100	102	100	105	108	104	105	107	105
		40	500	102	100	101	100	100	101	100	102	104	102	104	103	104
			1000	100	101	101	101	101	101	101	103	105	104	106	105	106
			2000	100	100	101	100	100	101	100	103	105	105	107	104	106
	σ	10	500	100	112	114	112	112	114	112	**135**	**144**	**137**	**139**	**142**	**137**
			1000	100	**132**	**131**	**130**	**132**	**131**	**130**	**166**	**178**	**172**	**175**	**175**	**171**
			2000	100	**159**	**149**	**156**	**159**	**150**	**157**	**206**	**216**	**221**	**226**	**214**	**220**
		20	500	100	114	112	113	114	113	113	**128**	**136**	**137**	**143**	**134**	**139**
			1000	100	**140**	**126**	**136**	**140**	**127**	**137**	**163**	**166**	**183**	**193**	**164**	**184**
			2000	100	**168**	**142**	**163**	**169**	**143**	**164**	**202**	**199**	**232**	**247**	**198**	**235**
		40	500	100	117	106	118	117	106	118	122	121	**147**	**149**	120	**149**
			1000	100	**137**	115	**139**	**137**	116	**139**	**148**	**141**	**189**	**191**	**140**	**191**
			2000	100	**172**	**129**	**174**	**173**	**130**	**175**	**187**	**170**	**255**	**257**	**170**	**257**
2G	μ	10	500	102	100	104	100	100	104	100	108	114	106	104	112	105
			1000	101	100	103	100	100	103	100	108	113	106	106	112	106
			2000	100	100	104	101	101	104	101	109	114	107	106	114	107
		20	500	102	100	105	100	100	105	100	109	116	105	104	113	105
			1000	100	100	105	101	100	105	101	109	117	106	105	116	106
			2000	100	100	106	101	100	106	101	109	118	106	105	116	106
		40	500	101	100	105	100	100	104	100	109	118	104	104	115	104
			1000	100	100	105	100	100	105	100	109	118	104	104	116	104
			2000	100	101	105	100	101	105	101	109	120	105	105	117	105
	σ	10	500	100	113	**131**	113	113	**132**	114	**152**	**181**	**141**	**137**	**178**	**139**
			1000	100	125	**153**	**126**	**126**	**154**	**127**	**182**	**225**	**169**	**165**	**222**	**168**
			2000	100	**147**	**189**	**149**	**148**	**191**	**150**	**236**	**299**	**217**	**211**	**296**	**215**
		20	500	100	109	**133**	109	109	**134**	110	**147**	**188**	**132**	**128**	**182**	**130**
			1000	100	122	**162**	123	123	**163**	124	**183**	**245**	**161**	**157**	**239**	**161**
			2000	100	**140**	**200**	**141**	**141**	**202**	**143**	**230**	**320**	**199**	**193**	**313**	**198**
		40	500	100	107	**136**	107	107	**137**	107	**147**	**192**	122	120	**186**	121
			1000	100	114	**160**	114	115	**161**	114	**174**	**239**	**138**	**138**	**235**	**138**
			2000	100	**126**	**195**	**126**	**127**	**196**	**127**	**213**	**308**	**164**	**165**	**304**	**166**

Note. DIF = differential item functioning; Par = parameter; MGM = mean-geometric mean linking; ASL = asymmetric Stocking-Lord linking; SSL = symmetric Stocking-Lord linking; AHA = asymmetric Haebara linking; SHA = symmetric Haebara linking; SI1G = SIMEX-based linking with assumed DIF in one group; SI2G = SIMEX-based linking with assumed DIF in two groups; Relative RMSE values larger than 125 are printed in bold font.

7. Conclusion

In this article, we investigated the bias and the variance of linking methods if random uniform DIF occurs in one or two groups. The analytically obtained bias formulas were specialized for asymmetric and symmetric Haebara linking. A simulation study was carried out to study the performance of Haebara and Stocking-Lord linking as well as bias-correction methods that relied on the SIMEX approach. Importantly, the biases qualitatively for the different linking methods differ depending on whether DIF occurs in one or two groups. In the case that DIF occurs in only one group, asymmetric and symmetric Haebara and Stocking-Lord linking resulted in biased estimates. In contrast, asymmetric and symmetric Stocking-Lord linking resulted in nearly unbiased estimates if DIF occurs in two groups. In this situation, asymmetric Haebara linking again resulted in biased estimates, while symmetric Haebara linking was approximately unbiased as predicted from the analytical derivations.

The SIMEX-based linking methods were effective in reducing (or removing) the bias. However, it is important that the type of occurrence of DIF effects (*i.e.*, in one or two groups) must be correctly implemented in SIMEX estimation. If the incorrect type of DIF effects is applied, the SIMEX-based linking method could even introduce bias.

If researchers can ensure in practical applications that random DIF symmetrically occurs in both groups, using asymmetric or symmetric Stocking-Lord linking is advised. If researchers are unsure about whether random DIF occurs in one or two groups and the sample size per group is not overly small, we recommend using mean-geometric mean linking instead of Haebara or Stocking-Lord linking because it results in unbiased estimates and induces only minor efficiency losses if there is no random DIF in the item response data.

Our findings also have implications for how simulation studies about linking are conducted. The way random DIF is simulated will impact the performance of the linking methods under study.

As with any simulation study, our study also possesses some limitations. First, it assumes that the item response model was correctly specified as the 2PL model. Second, we only simulated normally distributed DIF effects. Third, we only considered uniform DIF effects, and nonuniform DIF was assumed to be absent. However, previous research highlighted that uniform DIF is more likely to be found in practical applications than nonuniform DIF [51]. Moreover, the general derivations of Section 4 can be extended to accommodate random nonuniform DIF. Fourth, random DIF was assumed to be independent across items, but DIF effects could be correlated within testlets (*i.e.*, groups of items; see [52, 53]). Fifth, our findings could be generalized to polytomous linking methods [54, 55]. These limitations could be addressed in future research.

An anonymous reviewer suggested using a concurrent calibration (CC) method as a linking method that employs a joint IRT model that involves the two groups. Previous research demonstrated that CC assuming invariant or partially invariant; see [56]) items provides biased parameter estimates in the presence of random DIF [42]. For this reason, we did not include the CC method as a competitor in our simulation study.

Acknowledgment

The author has received no funds.

Conflicts of Interests

The author declares no conflict of interest.

References

[1] Chen Y, Li X, Liu J, Ying Z. Item response theory—A statistical framework for educational and psychological measurement. *arXiv* 2021, arXiv:2108.08604.

[2] Bock RD, Moustaki I. Item response theory in a general framework. In *Handbook of Statistics*, Rao CR, Sinharay S, eds. Amsterdam: Elsevier, 2006, pp. 469–513.

[3] Bock RD, Gibbons RD. *Item Response Theory*. New Jersey: Wiley, 2021.

[4] Rutkowski L, von Davier M, Rutkowski D. *Handbook of International Large-scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, 1st ed. New York: Chapman and Hall/CRC, 2013.

[5] van der Linden WJ. Unidimensional logistic response models. In *Handbook of Item Response Theory, Volume 1: Models*, 1st ed. New York: Chapman and Hall/CRC, 2016, pp. 11–30.

[6] Yen WM, Fitzpatrick AR. Item response theory. In *Educational Measurement*, Brennan RL, ed. Westport: Praeger Publishers, 2006, pp. 111–153.

[7] Birnbaum A. Some latent trait models and their use in inferring an examinee's ability. In *Statistical Theories of Mental Test Scores*, Lord FM, Novick MR, eds. Reading MA: Addison-Wesley, 1968, pp. 397–479.

[8] Aitkin M. Expectation maximization algorithm and extensions. In *Handbook of Item Response Theory, Volume 2: Statistical Tools*, van der Linden WJ, ed. New York: Chapman and Hall/CRC, 2016, pp. 217–236.

[9] Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* 1981, 46(4):443–459.

[10] Glas CAW. Maximum-likelihood estimation. In *Handbook of Item Response Theory, Volume 2: Statistical Tools*, van der Linden WJ, ed. New York: Chapman and Hall/CRC, 2016, pp. 197–216.

[11] Kolen MJ, Brennan RL. *Test Equating, Scaling, and Linking*, 3rd ed. New York: Springer, 2014.

[12] Lee WC, 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 DJ, eds. New York: Wiley, 2018, pp. 639–673.

[13] Sansivieri V, Wiberg M, Matteucci M. A review of test equating methods with a special focus on IRT-based approaches. *Statistica* 2017, 77(4):329–352.

[14] Bauer DJ. Enhancing measurement validity in diverse populations: Modern approaches to evaluating differential item functioning. *Brit. J. Math. Stat. Psychol.* 2023, 76(3):435–461.

[15] Holland PW, Wainer H. *Differential Item Functioning: Theory and Practice*, 1st ed. New York: Routledge, 1993.

[16] Millsap RE. *Statistical Approaches to Measurement Invariance*, 1st ed. New York: Routledge, 2011.

[17] Penfield RD, Camilli G. Differential item functioning and item bias. In *Handbook of Statistics, Vol. 26: Psychometrics*, Rao CR, Sinharay S, Sinharay S, eds. Amsterdam: Elsevier, 2006, pp. 125–167.

[18] Thissen D. A review of some of the history of factorial invariance and differential item functioning. *Multivar. Behav. Res.* 2024, 1–25.

[19] Robitzsch A. Bias-reduced Haebara and Stocking-Lord linking. *J* 2024, 7(3):373–384.

[20] Brennan RL. *Generalizabilty Theory*, 1st ed. New York: Springer 2001.

[21] Michaelides MP. A review of the effects on IRT item parameter estimates with a focus on misbehaving common items in test equating. *Front. Psychol.* 2010, 1:167.

[22] Michaelides MP, Haertel EH. Selection of common items as an unrecognized source of variability in test equating: A bootstrap approximation assuming random sampling of common items. *Appl. Meas. Educ.* 2014, 27(1):46–57.

[23] Monseur C, Berezner A. The computation of equating errors in international surveys in education. *J. Appl. Meas.* 2007, 8(3):323–335.

[24] Robitzsch A. Linking error in the 2PL model. *J* 2023, 6(1):58–84.

[25] Robitzsch A. Estimation of standard error, linking error, and total error for robust and non-robust linking methods in the two-parameter logistic model. *Stats* 2024, 7(3):592–612.

[26] Sachse KA, Roppelt A, Haag N. A comparison of linking methods for estimating national trends in international comparative large-scale assessments in the presence of cross-national DIF. *J. Educ. Meas.* 2016, 53(2):152–171.

[27] Sachse KA, Haag N. Standard errors for national trends in international large-scale assessments in the case of cross-national differential item functioning. *Appl. Meas. Educ.* 2017, 30(2):102–116.

[28] Wu M. Measurement, sampling, and equating errors in large-scale assessments. *Educ. Meas.* 2010, 29(4):15–27.

[29] Haebara T. Equating logistic ability scales by a weighted least squares method. *Jpn. Psychol. Res.* 1980, 22(3):144–149.

[30] Stocking ML, Lord FM. Developing a common metric in item response theory. *Appl. Psychol. Meas.* 1983, 7(2):201–210.

[31] De Boeck P. Random item IRT models. *Psychometrika* 2008, 73(4):533–559.

[32] Fox JP, Verhagen AJ. Random item effects modeling for cross-national survey data. In *Cross-cultural Analysis: Methods and Applications*, Davidov E, Schmidt P, Billiet J, eds. New York: Routledge, 2018, pp. 529–550.

[33] Arai S, Mayekawa Si. A comparison of equating methods and linking designs for developing an item pool under item response theory. *Behaviormetrika* 2011, 38:1–16.

[34] Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd ed. New York: Chapman and Hall/CRC, 2006.

[35] Robitzsch A. SIMEX-based and analytical bias corrections in Stocking-Lord linking. *Analytics* 2024, 3(3):368–388.

[36] 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(1):36–68.

[37] Pohl S, Schulze D. Assessing group comparisons or change over time under measurement non-invariance: The cluster approach for nonuniform DIF. *Psychol. Test Assess. Model.* 2020, 62(2):281–303.

[38] 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. *Psychol. Test Assess. Model.* 2020, 62(2):233–279.

[39] Kopf J, Zeileis A, Strobl C. Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educ. Psychol. Meas.* 2015, 75(1):22–56.

[40] Wang WC, Shih CL, Sun GW. The DIF-free-then-DIF strategy for the assessment of differential item functioning. *Educ. Psychol. Meas.* 2012, 72(4):687–708.

[41] De Boeck P. Wondering and mind-wandering thoughts on IRT. *Stat. Sci.* 2024 (in press).

[42] 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(1):116–144.

[43] Kass RE. Statistical inference: The big picture. *Stat. Sci.* 2011, 26(1):1–9.

[44] Breidt FJ, Opsomer JD. Model-assisted survey estimation with modern prediction techniques. *Stat. Sci.* 2017, 32(2):190–205.

[45] Kim S, Kolen MJ. Effects on scale linking of different definitions of criterion functions for the IRT characteristic curve methods. *J. Educ. Behav. Stat.* 2007, 32(4):371–397.

[46] LeBeau B. Ability and prior distribution mismatch: An exploration of common-item linking methods. *Appl. Psychol. Meas.* 2017, 41(7):545–560.

[47] Weeks JP. plink: An R package for linking mixed-format tests using IRT-based methods. *J. Stat. Softw.* 2010, 35(12):1–33.

[48] Buonaccorsi JP. *Measurement Error: Models, Methods, and Applications*, 1st ed. New York: CRC Press, 2010.

[49] R Core Team. *R: A language and environment for statistical computing*, 2023. Available: https://www.R-project.org (accessed on 15 March 2023).

[50] Robitzsch A. *sirt: Supplementary item response theory models*. R package version 4.2-64. 2024. Available: https://alexanderrobitzsch.r-universe.dev/sirt (accessed on 15 July 2024).

[51] Rutkowski L, Svetina D. Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educ. Psychol. Meas.* 2014, 74(1):31–57.

[52] Kim S, Kolen MJ. Scale linking for the testlet item response theory model. *Appl. Psychol. Meas.* 2022, 46(2):79–97.

[53] Li Y, Bolt DM, Fu J. A test characteristic curve linking method for the testlet model. *Appl. Psychol. Meas.* 2005, 29(5):340–356.

[54] Andersson B. Asymptotic variance of linking coefficient estimators for polytomous IRT models. *Appl. Psychol. Meas.* 2018, 42(3):192–205.

[55] Zhang Z. Asymptotic standard errors of generalized partial credit model true score equating using characteristic curve methods. *Appl. Psychol. Meas.* 2021, 45(5):331–345.

[56] von Davier M, Yamamoto K, Shin HJ, Chen H, Khorramdel L, *et al.* Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assess. Educ.* 2019, 26(4):466–488.