J Multidisciplinary Scientific Journal

MDPI

Article

Bias-Reduced Haebara and Stocking–Lord Linking

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: Haebara and Stocking–Lord linking methods are frequently used to compare the distributions of two groups. Previous research has demonstrated that Haebara and Stocking–Lord linking can produce bias in estimated standard deviations and, to a smaller extent, in estimated means in the presence of differential item functioning (DIF). This article determines the asymptotic bias of the two linking methods for the 2PL model. A bias-reduced Haebara and bias-reduced Stocking–Lord linking method is proposed to reduce the bias due to uniform DIF effects. The performance of the new linking method is evaluated in a simulation study. In general, it turned out that Stocking–Lord linking had substantial advantages over Haebara linking in the presence of DIF effects. Moreover, bias-reduced Haebara and Stocking–Lord linking substantially reduced the bias in the estimated standard deviation.

Keywords: item response model; linking; 2PL model; Haebara linking; Stocking–Lord linking; bias reduction

check for updates

Citation: Robitzsch, A. Bias-Reduced Haebara and Stocking–Lord Linking. *J* **2024**, *7*, 373–384. https://doi.org/10.3390/j7030021

Academic Editor: Pietro Cipresso

Received: 10 May 2024
Revised: 14 August 2024
Accepted: 21 August 2024
Published: 4 September 2024

Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Item response theory (IRT) models [1,2] are multivariate statistical models for multivariate binary random variables. These kinds of models are frequently used to model cognitive testing data stemming from educational or psychological applications. This article only considers unidimensional IRT models [3]. Let $\boldsymbol{X} = (X_1, \ldots, X_I)$ be the vector of I dichotomous (i.e., binary) random variables $X_i \in \{0, 1\}$ that are typically referred to as items in the psychometric literature. A unidimensional IRT model [4] is a statistical model for the probability distribution $\mathsf{P}(\boldsymbol{X} = \boldsymbol{x})$ for $\boldsymbol{x} = (x_1, \ldots, x_I) \in \{0, 1\}^I$, where

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

where ϕ denotes the density of a normal distribution with mean μ and standard deviation σ. The parameters of the distribution's latent variable θ (also referred to as the factor variable, trait, or ability) are contained in $\delta = (\mu, \sigma)$. The vector $\gamma = (\gamma_1, \ldots, \gamma_I)$ contains all estimated item parameters of item response functions (IRFs) $P_i(\theta; \gamma_i) = \mathsf{P}(X_i = 1 | \theta)$ $(i = 1, \ldots, I)$. The two-parameter logistic (2PL) model [5] is the most popular IRT model for dichotomous items and possesses the IRF

$$P_i(\theta; \gamma_i) = \Psi(a_i(\theta - b_i)) \,, \tag{2}$$

where a_i and b_i are the item discrimination and item difficulty, respectively, and $\Psi(x) = (1 + \exp(-x))^{-1}$ denotes the logistic distribution function. For independent and identically distributed observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ of N persons from the distribution of the random variable $\boldsymbol{X}$, the unknown model parameters in Equation (1) can be estimated by (marginal) maximum likelihood estimation [6–8].

In many educational applications, IRT models are employed to compare the distributions of two groups in a test (i.e., on a set of items) regarding the factor variable θ in

the IRT model in Equation (1). Linking methods [9–11] estimate the 2PL model separately in the two groups in the first step. All group differences are captured in the item parameters. Linking methods transform the estimated item parameters from the first step in a subsequent second step to determine a difference regarding the mean μ and the standard deviation σ between the two groups. The separate application of the 2PL model in each of the two groups has the advantage that it allows items to function differently across groups: a property that is referred to as differential item functioning (DIF) [12,13]. It has been pointed out that the occurrence of DIF causes additional variability in the estimated mean μ and standard deviation σ when applying a linking method [14–19].

However, it has also been shown that the presence of DIF effects produces a bias in the parameters that quantify the group difference [20]. The bias occurs for the popular Haebara [21] and Stocking–Lord [22] linking methods. This article proposes bias-reduced variants of these linking methods that rely on a derivation of the asymptotic by using a second-order Taylor expansion of the optimization function involved in the linking method for the 2PL model. A simulation study is carried out to investigate the usefulness of the proposed modified linking estimator.

The rest of the article is organized as follows. The derivation of the bias-reduced variants of Haebara and Stocking–Lord linking is presented in Section 2. Section 3 reports findings from a simulation study that compares the proposed new estimators with the currently used linking methods in the presence of DIF. Finally, the article closes with a discussion in Section 4.

2. Bias Reduction in Haebara and Stocking–Lord Linking

2.1. Group Comparisons in the 2PL Model

In this section, we explain how a group comparison involving two groups is conducted within the 2PL model. Assume that the 2PL model holds for the first group, and the ability distribution is given as a standard normal distribution; that is, $\theta \sim \mathsf{N}(0,1)$. The item discriminations and item intercepts in the first group are defined as $a_{i1} = a_i$ and $b_{i1} = b_i$, respectively. In the second group, the ability distribution is given as $\theta \sim \mathsf{N}(\mu, \sigma^2)$, where μ and σ are the mean and the standard deviation (SD), respectively, of the second group. By using this parametrization, μ reflects group differences regarding the mean, while σ quantifies group differences in standard deviations. The item discriminations is assumed to be invariant across the two groups such that $a_{i2} = a_i$. A uniform DIF [12] effect e_i is assumed for item difficulties:

$$b_{i2} = b_i + e_i \text{ , where } \mathsf{E}(e_i) = 0 \text{ and } \mathsf{Var}(e_i) = \tau^2 \text{ .} \tag{3}$$

The quantity τ^2 is also called the DIF variance [23,24], and τ is denoted as the DIF SD.

An anonymous reviewer wondered whether the assumption of a mean DIF of zero is tenable in applications. We favor the assumption of a zero mean DIF because we consider DIF effects as construct-relevant [25–27]. In this case, the estimated group differences (i.e., estimates of μ and σ) should be based on all items, and no items should be removed from group comparisons [28], as they are done under a partial invariance assumption [29,30].

A linking method proceeds in two steps. In the first step, the 2PL model is separately fitted within the two groups while assuming a standard normal distribution $\mathsf{N}(0,1)$ for the ability θ. Due to sampling errors (i.e., sampling of persons), the estimated item parameters $\widehat{a}_{i1}$ and $\widehat{b}_{i1}$ will slightly differ from the data-generating parameters $a_{i1} = a_i$ and $b_{i1} = b_i$. In the second group, the original item parameters a_{i2} and b_{i2} are not recovered because the 2PL model is fitted with $\theta \sim \mathsf{N}(0,1)$, but the data-generating model imposes $\theta \sim \mathsf{N}(\mu, \sigma^2)$. However, it holds that

$$a_{i2}(\theta - b_{i2}) = a_{i2}(\sigma\theta^* + \mu - b_{i2}) \text{ for } \theta^* \sim \mathsf{N}(0,1) \text{ ,} \tag{4}$$

and we have that the identified item parameters

$$a_{i2}^* = a_i\sigma \text{ and } b_{i2}^* = \sigma^{-1}(b_i + e_i - \mu) \tag{5}$$

We can now compute the identified item parameters. By construction, we have $\widehat{a}_{i1} = a_{i1}$ and $\widehat{b}_{i1} = b_i$. Note that

$$a_{i2}(\theta - b_{i2}) = a_{i2}(\sigma\theta^* + \mu - b_{i2}) \,, \tag{6}$$

and we have that the identified item parameters

$$\widehat{a}_{i2} = a_i\sigma \text{ and } \widehat{b}_{i2} = \sigma^{-1}(b_i + e_i - \mu) \tag{7}$$

Due to sampling errors, the estimated item parameters $\widehat{a}_{i2}$ and $\widehat{b}_{i2}$ will slightly differ from a_{i2}^* and b_{i2}^*.

If no uniform DIF effects occur (i.e., $\tau = 0$) and there is no sampling error, the group mean μ and the group SD σ can be determined from a single item i using the identified item parameters a_{i2}^* and b_{i2}^*. In the presence of sampling errors and DIF, a linking function H is chosen that estimates μ and σ based on estimated item parameters $\widehat{a}_{ig}$ and $\widehat{b}_{ig}$ for $i = 1, \ldots, I$ and $g = 1, 2$. A linking method is formally described in the following Section 2.2.

2.2. Bias Reduction in a Linking Method Due to DIF

In a linking method, the vector $\boldsymbol{\delta} = (\mu, \sigma)$ is the statistical parameter of interest. The corresponding parameter estimate for $\boldsymbol{\delta}$ is denoted as $\widehat{\boldsymbol{\delta}} = (\widehat{\mu}, \widehat{\sigma})$. The linking function H is defined as a function of $\boldsymbol{\delta}$ and the estimated item parameters $\widehat{\boldsymbol{\gamma}} = (\widehat{\boldsymbol{\gamma}}_1, \ldots, \widehat{\boldsymbol{\gamma}}_I)$ (where $\widehat{\boldsymbol{\gamma}}_i = (\widehat{a}_{i1}, \widehat{b}_{i1}, \widehat{a}_{i2}, \widehat{b}_{i2})$) in the two groups such that

$$\widehat{\boldsymbol{\delta}} = \arg\min_{\boldsymbol{\delta}} H(\boldsymbol{\delta}; \widehat{\boldsymbol{\gamma}}) \,. \tag{8}$$

By assuming differentiability of H, the estimate $\widehat{\boldsymbol{\delta}}$ fulfills the estimating equation

$$\boldsymbol{H}_{\boldsymbol{\delta}}(\widehat{\boldsymbol{\delta}}; \widehat{\boldsymbol{\gamma}}) = \frac{\partial H}{\partial \boldsymbol{\delta}}(\widehat{\boldsymbol{\delta}}; \widehat{\boldsymbol{\gamma}}) = \mathbf{0} \,. \tag{9}$$

A minimal requirement of a linking method is that it should recover the true group mean μ_0 and the true group SD σ_0 if there is no sampling error and no DIF (i.e., $\tau = 0$). Hence, it is assumed that

$$\boldsymbol{H}_{\boldsymbol{\delta}}(\boldsymbol{\delta}_0; \boldsymbol{\gamma}_0) = \mathbf{0} \text{ for } \boldsymbol{\delta}_0 = (\mu_0, \sigma_0) \,, \tag{10}$$

where $\boldsymbol{\gamma}_0$ denotes joint item parameters a_i and b_i, and all DIF effects e_i are assumed to be zero.

We now derive an expression of the bias of the linking method due to the presence of DIF effects e_i. The idea is to utilize a second-order Taylor expansion of $\boldsymbol{H}_{\boldsymbol{\delta}}$ and to derive a bias correction term of the parameter estimate $\widehat{\boldsymbol{\delta}}$. Note that the estimated item parameters $\widehat{\boldsymbol{\gamma}}_i$ for item i are a function of a common item parameter $\boldsymbol{\gamma}_i = (a_i, b_i)$, and the uniform DIF effect e_i. The estimating equation in Equation (9) can be formalized as

$$\boldsymbol{H}_{\boldsymbol{\delta}}(\boldsymbol{\delta}; \boldsymbol{\gamma}, \boldsymbol{e}) = \mathbf{0} \,, \tag{11}$$

where $\boldsymbol{\gamma}$ collects all item parameters, and $\boldsymbol{e} = (e_1, \ldots, e_I)$ denotes the vector of DIF effects. Due to Equation (10), it holds that $\boldsymbol{H}_{\boldsymbol{\delta}}(\boldsymbol{\delta}_0; \boldsymbol{\gamma}, \mathbf{0}) = 0$ for $\boldsymbol{e} = 0$. This property ensures that the application of the linking method provides the true mean μ_0 and the true SD σ_0.

By assuming independence of DIF effects e_i, we have, by using a first-order Taylor expansion with respect to $\boldsymbol{\delta}$ and a second-order Taylor expansion with respect to DIF effects e_i,

$$\mathbf{0} = \boldsymbol{H}_{\boldsymbol{\delta}}(\widehat{\boldsymbol{\delta}}; \boldsymbol{\gamma}, \boldsymbol{e}) = \boldsymbol{H}_{\boldsymbol{\delta}}(\boldsymbol{\delta}_0; \boldsymbol{\gamma}, \mathbf{0}) + \boldsymbol{H}_{\boldsymbol{\delta}\boldsymbol{\delta}}(\boldsymbol{\delta}_0; \boldsymbol{\gamma}, \mathbf{0})(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0) + \sum_{i=1}^{I} \boldsymbol{H}_{\boldsymbol{\delta} e_i}(\boldsymbol{\delta}_0; \boldsymbol{\gamma}, \mathbf{0}) e_i + \frac{1}{2} \sum_{i=1}^{I} \boldsymbol{H}_{\boldsymbol{\delta} e_i e_i}(\boldsymbol{\delta}_0; \boldsymbol{\gamma}, \mathbf{0}) e_i^2 \tag{12}$$

Due to $\boldsymbol{H}_\delta(\delta_0;\gamma,\mathbf{0}) = 0$, we obtain from Equation (12)

$$\widehat{\delta} - \delta_0 = -\boldsymbol{H}_{\delta\delta}(\delta_0;\gamma,\mathbf{0})^{-1}\left(\sum_{i=1}^{I}\boldsymbol{H}_{\delta e_i}(\delta_0;\gamma,\mathbf{0})e_i + \frac{1}{2}\sum_{i=1}^{I}\boldsymbol{H}_{\delta e_i e_i}(\delta_0;\gamma,\mathbf{0})e_i^2\right) \tag{13}$$

Because the DIF effects e_i are random variables with zero expectation $\mathsf{E}(e_i) = 0$, we can determine an approximate (i.e., expected) bias in the estimate $\widehat{\delta}$ as

$$\mathsf{Bias}(\widehat{\delta}) = -\frac{1}{2}\boldsymbol{H}_{\delta\delta}(\delta_0;\gamma,\mathbf{0})^{-1}\sum_{i=1}^{I}\boldsymbol{H}_{\delta e_i e_i}(\delta_0;\gamma,\mathbf{0})\mathsf{E}(e_i^2) \tag{14}$$

Note that the bias due to DIF effects does not vanish in large samples of persons or for a large number of items. Let $\tau^2 = \mathsf{E}(e_i^2)$; the asymptotic bias of $\widehat{\delta}$ can be described as

$$\mathsf{Bias}(\widehat{\delta}) = -\frac{1}{2}\tau^2\boldsymbol{H}_{\delta\delta}(\delta_0;\gamma,\mathbf{0})^{-1}\sum_{i=1}^{I}\boldsymbol{H}_{\delta e_i e_i}(\delta_0;\gamma,\mathbf{0}) \tag{15}$$

Equation (14) gives rise to the idea of using an empirical version of it as a bias correction (or maybe only a bias-reducing) term for $\widehat{\delta}$. We compute the estimated DIF effects $\widehat{e}_i$ by

$$\widehat{e}_i = \widehat{\mu} + \widehat{\sigma}\widehat{b}_{i2} - \widehat{b}_{i1} \tag{16}$$

to obtain a bias-reduced estimate of δ as

$$\widehat{\delta}_{\mathrm{br}} = \widehat{\delta} + \frac{1}{2}\boldsymbol{H}_{\delta\delta}(\widehat{\delta};\widehat{\gamma},\mathbf{0})^{-1}\sum_{i=1}^{I}\boldsymbol{H}_{\delta e_i e_i}(\widehat{\delta};\widehat{\gamma},\mathbf{0})\widehat{e}_i^2 \tag{17}$$

The vector $\boldsymbol{H}_{\delta e_i e_i}$ contains second-order derivatives of $\boldsymbol{H}_\delta$ with respect to e_i and is given by $\boldsymbol{H}_{\delta e_i e_i} = (\boldsymbol{H}_{\mu e_i e_i}, \boldsymbol{H}_{\sigma e_i e_i})$.

The described construction scheme for a bias-reduced linking estimate is applied for Haebara linking in Section 2.3 and Stocking–Lord linking in Section 2.4.

2.3. Bias-Reduced Haebara Linking (brHAE)

The optimization function in Haebara (HAE) [21] linking is defined as

$$H(\delta;\widehat{\gamma}) = \sum_{i=1}^{I}\sum_{t=1}^{T}\omega_t\left[\Psi\left(\widehat{a}_{i1}(\sigma\theta_t + \mu - \widehat{b}_{i1})\right) - \Psi\left(\widehat{a}_{i2}(\theta_t - \widehat{b}_{i2})\right)\right]^2 \text{ for } \delta = (\mu,\sigma)\,. \tag{18}$$

with a priori specification of weights ω_t on a grid of θ values $\theta_1,\ldots,\theta_T$. A convenient choice is to use an equidistant grid of θ values on $[-6,6]$ and weights ω_t that are proportional to the density function of a normal distribution with a mean of 0 and an SD of 2. The corresponding estimating equations to the minimization problem in Equation (18) are given by

$$H_\mu = \frac{\partial H}{\partial \mu} = 2\sum_{i=1}^{I}\sum_{t=1}^{T}\omega_t\left[\Psi\left(\widehat{a}_{i1}(\sigma\theta_t + \mu - \widehat{b}_{i1})\right) - \Psi\left(\widehat{a}_{i2}(\theta_t - \widehat{b}_{i2})\right)\right]\Psi'\left(\widehat{a}_{i1}(\sigma\theta_t + \mu - \widehat{b}_{i1})\right)\widehat{a}_{i1} = 0 \text{ and} \tag{19}$$

$$H_\sigma = \frac{\partial H}{\partial \sigma} = 2\sum_{i=1}^{I}\sum_{t=1}^{T}\omega_t\left[\Psi\left(\widehat{a}_{i1}(\sigma\theta_t + \mu - \widehat{b}_{i1})\right) - \Psi\left(\widehat{a}_{i2}(\theta_t - \widehat{b}_{i2})\right)\right]\Psi'\left(\widehat{a}_{i1}(\sigma\theta_t + \mu - \widehat{b}_{i1})\right)\widehat{a}_{i1}\theta_t = 0\,. \tag{20}$$

In Equation (20), Ψ' denotes the derivative of the logistic function Ψ. The second-order derivative $\boldsymbol{H}_{\delta\delta}$ (i.e., the partial derivatives of H_μ and H_σ with respect to μ and σ) can be simply computed, although the formulas are a bit cumbersome. In a numerical implementation of the bias-reduced linking method, the derivatives can be obtained with numerical approximation.

Using the true parameter vector $\delta_0 = (\mu_0,\sigma_0)$, we can write

$$\Psi\left(\widehat{a}_{i2}(\theta_t - \widehat{b}_{i2})\right) = \Psi\left(a_i\sigma_0(\theta_t - \sigma_0^{-1}(b_i - e_i + \mu_0))\right) = \Psi(a_i(\sigma_0\theta_t - (b_i - e_i + \mu_0))) \quad (21)$$

The unknown DIF effect e_i in Equation (21) can be substituted by $\widehat{e}_i$ such that we obtain

$$\Psi(a_i(\sigma_0\theta_t - (b_i - e_i + \mu_0))) \approx \Psi\left(\widehat{a}_{i1}(\widehat{\sigma}\theta_t - (\widehat{b}_{i1} - \hat{e}_i + \widehat{\mu}))\right) \quad (22)$$

We now compute the required derivatives with respect to e_i in the bias-reduced estimate derived in Equation (17). The derivatives with respect to μ are given by

$$\boldsymbol{H}_{\mu e_i} = 2\sum_{t=1}^{T}\omega_t\Psi'(a_i(\sigma\theta_t - (b_i + e_i) - \mu))\Psi'\left(\widehat{a}_{i1}(\sigma\theta_t + \mu - \widehat{b}_{i1})\right)\widehat{a}_{i1}^2 \text{ and} \quad (23)$$

$$\boldsymbol{H}_{\mu e_i e_i} = -2\sum_{t=1}^{T}\omega_t\Psi''(a_i(\sigma\theta_t - (b_i + e_i) - \mu))\Psi'\left(\widehat{a}_{i1}(\sigma\theta_t + \mu - \widehat{b}_{i1})\right)\widehat{a}_{i1}^3\,, \quad (24)$$

where Ψ'' is the second derivative of Ψ. In analogy, we have for the required second derivative of H_σ with respect to e_i,

$$\boldsymbol{H}_{\sigma e_i e_i} = -2\sum_{t=1}^{T}\omega_t\Psi''(a_i(\sigma\theta_t - (b_i + e_i) - \mu))\Psi'\left(\widehat{a}_{i1}(\sigma\theta_t + \mu - \widehat{b}_{i1})\right)\widehat{a}_{i1}^3\theta_t\,. \quad (25)$$

The term $a_i(\sigma\theta_t - (b_i + e_i) - \mu)$ in Equations (24) and (25) can be estimated by $\widehat{a}_{i2}(\theta_t - \widehat{b}_{i2})$. Hence, the required terms for the bias-reduced Haebara linking (brHAE) can be written as

$$\boldsymbol{H}_{\mu e_i e_i} = -2\sum_{t=1}^{T}\omega_t\Psi''\left(\widehat{a}_{i2}(\theta_t - \widehat{b}_{i2})\right)\Psi'\left(\widehat{a}_{i1}(\sigma\theta_t + \mu - \widehat{b}_{i1})\right)\widehat{a}_{i1}^3 \text{ and} \quad (26)$$

$$\boldsymbol{H}_{\sigma e_i e_i} = -2\sum_{t=1}^{T}\omega_t\Psi''\left(\widehat{a}_{i2}(\theta_t - \widehat{b}_{i2})\right)\Psi'\left(\widehat{a}_{i1}(\sigma\theta_t + \mu - \widehat{b}_{i1})\right)\widehat{a}_{i1}^3\theta_t\,. \quad (27)$$

2.4. Bias-Reduced Stocking–Lord Linking (brSL)

The Stocking–Lord (SL) [22] linking method minimizes the weighted squared distance of test characteristic functions:

$$H(\boldsymbol{\delta};\widehat{\boldsymbol{\gamma}}) = \sum_{t=1}^{T}\omega_t\left[\sum_{i=1}^{I}\Psi\left(\widehat{a}_{i1}(\sigma\theta_t + \mu - \widehat{b}_{i1})\right) - \sum_{i=1}^{I}\Psi\left(\widehat{a}_{i2}(\theta_t - \widehat{b}_{i2})\right)\right]^2 \text{ with } \boldsymbol{\delta} = (\mu,\sigma)\,. \quad (28)$$

SL linking offers more flexibility in allowing differences in the IRFs between groups because the alignment occurs at the level of the test characteristic function. In contrast, HAE linking explicitly defines the squared distance between IRFs. The corresponding estimating equations of SL linking defined in Equation (28) for the parameters μ and σ are given as

$$H_\mu = 2\sum_{t=1}^{T}\omega_t\left[\sum_{i=1}^{I}\Psi\left(\widehat{a}_{i1}(\sigma\theta_t + \mu - \widehat{b}_{i1})\right) - \sum_{i=1}^{I}\Psi\left(\widehat{a}_{i2}(\theta_t - \widehat{b}_{i2})\right)\right]\left[\sum_{i=1}^{I}\Psi'\left(\widehat{a}_{i1}(\sigma\theta_t + \mu - \widehat{b}_{i1})\right)\widehat{a}_{i1}\right] = 0 \text{ and} \quad (29)$$

$$H_\sigma = 2\sum_{t=1}^{T}\omega_t\left[\sum_{i=1}^{I}\Psi\left(\widehat{a}_{i1}(\sigma\theta_t + \mu - \widehat{b}_{i1})\right) - \sum_{i=1}^{I}\Psi\left(\widehat{a}_{i2}(\theta_t - \widehat{b}_{i2})\right)\right]\left[\sum_{i=1}^{I}\Psi'\left(\widehat{a}_{i1}(\sigma\theta_t + \mu - \widehat{b}_{i1})\right)\widehat{a}_{i1}\theta_t\right] = 0\,. \quad (30)$$

The second-order derivatives $\boldsymbol{H}_{\delta\delta}$ can be similarly obtained, although they can be numerically obtained in a practical implementation of the bias-reduced linking method.

The second-order derivatives of H_μ and H_σ with respect to e_i can be computed similarly as for Haebara linking. We obtain the following quantities that are required in the formula for the bias reduction:

$$\boldsymbol{H}_{\mu e_i e_i} = -2 \sum_{t=1}^{T} \omega_t \Psi'' \left(\widehat{a}_{i2} (\theta_t - \widehat{b}_{i2}) \right) \left[\sum_{i=1}^{I} \Psi' \left(\widehat{a}_{i1} (\sigma \theta_t + \mu - \widehat{b}_{i1}) \right) \widehat{a}_{i1} \right] \widehat{a}_{i1}^2 \text{ and} \tag{31}$$

$$\boldsymbol{H}_{\sigma e_i e_i} = -2 \sum_{t=1}^{T} \omega_t \Psi'' \left(\widehat{a}_{i2} (\theta_t - \widehat{b}_{i2}) \right) \left[\sum_{i=1}^{I} \Psi' \left(\widehat{a}_{i1} (\sigma \theta_t + \mu - \widehat{b}_{i1}) \right) \widehat{a}_{i1} \right] \widehat{a}_{i1}^2 \theta_t \, . \tag{32}$$

By using Equations (31) and (32) in Formula (17) for the bias-reduced estimate, we obtain the bias-reduced Stocking–Lord (brSL) linking method.

2.5. Mean-Geometric-Mean Linking (MGM)

The mean-geometric-mean (MGM) linking method [10] is a two-step linking method that computes the SD by a geometric mean of item discriminations in the first step and the mean by an ordinary average of transformed item difficulties in the second step. The estimating equations for μ and σ are given by

$$H_\mu = \sum_{i=1}^{I} \left(\sigma \widehat{b}_{i2} - \widehat{b}_{i1} + \mu \right) = 0 \text{ and} \tag{33}$$

$$H_\sigma = \sum_{i=1}^{I} (\log \widehat{a}_{i2} - \log(\sigma) - \log \widehat{a}_{i1}) = 0 \, . \tag{34}$$

One can easily show that $\boldsymbol{H}_{\mu e_i e_i} = 0$ and $\boldsymbol{H}_{\sigma e_i e_i} = 0$. Hence, no asymptotic bias of the MGM linking method is expected. Therefore, exploring a bias-reducing variant of the MGM linking method is unnecessary.

3. Simulation Study

3.1. Method

Item responses in two groups were simulated according to the 2PL model. For identification reasons, the mean and the standard deviation of the normally distributed ability variable θ in the first group were set to 0 to 1, respectively. In the second group, we chose the mean $\mu = 0.3$ and the SD $\sigma = 1.2$ for the normally distributed ability variable. The parameters μ and σ were held constant in the simulation.

In the simulation, the number of items I was chosen as 10, 20, or 40, indicating short, medium, or long tests, respectively. The group-specific item parameters a_{ig} and b_{ig} for $i = 1, \ldots, I$ and $g = 1, 2$ relied on base item parameters that were fixed in the simulation and a random DIF effect that was simulated in each replication of the simulation study. The common item discriminations a_i in the case of $I = 10$ items were chosen as 0.83, 1.02, 0.88, 0.80, 1.04, 0.95, 1.00, 1.13, 1.32, and 1.11, resulting in a mean $M = 1.01$ and an $SD = 0.16$. 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, resulting in $M = -0.41$ and $SD = 0.86$. For item numbers as multiples of 10, we duplicated the item parameters of the 10 items accordingly. The item parameters in the second group included a uniform DIF effect e_i that was added to the common item difficulty b_i. As described in Section 2.1, the item difficulty b_{i2} in the second group was simulated as

$$b_{i2} = b_i + e_i \, . \tag{35}$$

The DIF effects e_i were independently and identically normally distributed with a mean of zero and a DIF SD τ. Note that DIF effects varied across replications within simulation conditions. The DIF SD τ was chosen as 0, 0.25, or 0.5, indicating no DIF, moderate DIF, or large DIF.

Item responses were simulated according to the 2PL model for sample sizes N per group of 500, 1000, 2000, or 4000. We also investigated an infinite sample size (denoted by $N = \text{Inf}$) in which only item parameters with DIF effects were simulated.

Five different linking methods were applied to estimate the mean μ and the SD σ in the second group: Haebara (HAE) linking, bias-reduced Haebara (brHAE) linking, Stocking–Lord (SL) linking, bias-reduced Stocking–Lord (brSL) linking, and mean-geometric-mean (MGM) linking. In finite sample sizes, all linking methods relied on estimated item discriminations $\widehat{a}_{ig}$ and item difficulties $\widehat{b}_{ig}$ ($i = 1, \ldots, I, g = 1, 2$) obtained from separately fitting the 2PL model to the item response datasets in the two groups. In an infinite sample size, the identified item parameters as defined in Equation (7) were used, and no item responses were simulated.

In each of the 5 (sample size N) $\times$ 3 (DIF standard deviation τ) $\times$ 3 (number of items I) $= 45$ cells of the simulation, 3000 replications were conducted. We computed the empirical bias, the empirical SD, and the root mean square error (RMSE) for the estimated mean $\widehat{\mu}$ and the estimated standard deviation $\widehat{\sigma}$. A relative percentage RMSE was computed as the ratio of the RMSE values of a particular linking method and the RMSE of Haebara linking.

R software (Version 4.3.1, [31]) was used for the entire analysis in this simulation study. The 2PL model was fitted using the `TAM::tam.mml.2pl()` function in the R package TAM [32]. MGM linking was implemented using the function `sirt::linking.haberman()` in the R package sirt [33]. Dedicated R functions were written by the author for HAE and SL linking and their bias-reduced variants. These functions and replication material for this Simulation Study can be found at https://osf.io/8mk49 (accessed on 10 May 2024).

3.2. Results

Table 1 presents the bias, the SD, and the relative RMSE for the estimated mean $\widehat{\mu}$. By construction, all linking methods estimated the true population value $\mu = 0.3$ in an infinite sample size in the no-DIF condition. Furthermore, the five linking methods performed similarly regarding the RMSE in the no-DIF condition $\tau = 0$. However, an efficiency loss due to MGM linking and an efficiency advantage due to SL compared to HAE linking were observed for short tests with a small number of items $I = 10$. A small positive bias was induced by brHAE and brSL in the small sample size $N = 500$.

Table 1. Simulation study: bias, standard deviation (SD), and relative root mean square error (RMSE) for the estimated mean $\widehat{\mu}$ as a function of the uniform DIF standard deviation τ, number of items I, and sample size N.

		Bias					**SD**					**Relative RMSE**				
I	***N***	**HAE**	**brHAE**	**SL**	**brSL**	**MGM**	**HAE**	**brHAE**	**SL**	**brSL**	**MGM**	**HAE**	**brHAE**	**SL**	**brSL**	**MGM**
		No DIF (DIF SD $\tau = 0$)														
10	500	0.008	**0.012**	0.003	**0.010**	0.004	0.094	0.095	0.091	0.092	0.103	100 ‡	101.7	96.7	98.1	**109.9**
	1000	0.004	0.006	0.002	0.005	0.002	0.064	0.064	0.063	0.063	0.069	100 ‡	100.7	97.3	97.9	**108.1**
	2000	0.002	0.003	0.001	0.002	0.001	0.046	0.046	0.045	0.045	0.050	100 ‡	100.3	97.0	97.3	**107.8**
	4000	0.000	0.001	0.000	0.000	0.000	0.032	0.032	0.031	0.031	0.035	100 ‡	100.1	97.1	97.2	**107.3**
	Inf	0	0	0	0	0	0	0	0	0	0	100 ‡	100.0	100.0	100.0	100.0
20	500	0.007	**0.011**	0.003	0.009	0.003	0.082	0.083	0.081	0.082	0.087	100 ‡	101.4	98.6	99.7	106.1
	1000	0.004	0.006	0.002	0.005	0.002	0.058	0.058	0.057	0.057	0.060	100 ‡	100.7	98.3	98.9	104.8
	2000	0.002	0.003	0.001	0.002	0.001	0.041	0.041	0.040	0.040	0.042	100 ‡	100.3	99.2	99.4	104.4
	4000	0.000	0.000	−0.001	0.000	−0.001	0.029	0.029	0.028	0.028	0.030	100 ‡	100.1	98.7	98.8	103.4
	Inf	0	0	0	0	0	0	0	0	0	0	100 ‡	100.0	100.0	100.0	100.0
40	500	0.007	**0.011**	0.003	0.009	0.004	0.077	0.078	0.077	0.077	0.080	100 ‡	101.4	99.4	100.6	102.9
	1000	0.002	0.004	0.000	0.003	0.000	0.055	0.055	0.054	0.054	0.056	100 ‡	100.6	99.3	99.7	102.3
	2000	0.002	0.003	0.001	0.002	0.001	0.039	0.039	0.039	0.039	0.040	100 ‡	100.3	98.9	99.2	102.3
	4000	0.000	0.000	−0.001	0.000	0.000	0.027	0.028	0.027	0.027	0.028	100 ‡	100.1	99.5	99.6	101.8
	Inf	0	0	0	0	0	0	0	0	0	0	100 ‡	100.0	100.0	100.0	100.0

Table 1. *Cont.*

		Bias					SD					Relative RMSE				
I	*N*	HAE	brHAE	SL	brSL	MGM	HAE	brHAE	SL	brSL	MGM	HAE	brHAE	SL	brSL	MGM
		Moderate Uniform DIF (DIF SD $\tau = 0.25$)														
	500	0.006	**0.015**	0.000	**0.012**	0.005	0.127	0.130	0.121	0.123	0.131	100 ‡	103.5	95.5	97.8	103.2
	1000	0.001	0.007	−0.003	0.006	0.003	0.107	0.110	0.102	0.104	0.107	100 ‡	102.2	95.4	96.7	99.9
10	2000	−0.002	0.003	−0.005	0.001	0.000	0.096	0.097	0.091	0.092	0.094	100 ‡	101.9	95.5	96.5	97.9
	4000	−0.003	0.002	−0.006	0.000	−0.001	0.091	0.092	0.086	0.087	0.087	100 ‡	101.7	95.4	96.2	96.0
	Inf	−0.004	0.000	−0.006	−0.001	0.000	0.085	0.086	0.080	0.081	0.079	100 ‡	101.7	95.0	95.7	93.5
	500	0.003	**0.012**	−0.002	**0.010**	0.004	0.100	0.103	0.097	0.098	0.102	100 ‡	103.4	96.6	98.8	101.7
	1000	0.000	0.007	−0.003	0.005	0.003	0.083	0.085	0.080	0.081	0.082	100 ‡	102.3	96.6	97.8	99.4
20	2000	−0.001	0.005	−0.003	0.004	0.003	0.072	0.073	0.069	0.070	0.070	100 ‡	102.0	96.1	97.1	97.6
	4000	−0.003	0.002	−0.005	0.001	0.001	0.066	0.067	0.063	0.064	0.063	100 ‡	101.8	96.0	96.8	95.5
	Inf	−0.005	0.000	−0.006	0.000	−0.001	0.059	0.060	0.056	0.057	0.056	100 ‡	101.5	95.9	96.3	94.3
	500	0.000	0.009	−0.005	0.007	0.002	0.087	0.089	0.086	0.087	0.088	100 ‡	102.8	98.3	99.7	101.0
	1000	−0.001	0.006	−0.004	0.004	0.002	0.068	0.070	0.067	0.067	0.068	100 ‡	102.2	98.0	99.0	100.1
40	2000	−0.004	0.002	−0.005	0.001	0.001	0.057	0.058	0.055	0.056	0.056	100 ‡	101.6	97.1	97.6	97.3
	4000	−0.003	0.002	−0.005	0.001	0.000	0.050	0.051	0.048	0.048	0.048	100 ‡	101.7	96.4	96.9	95.6
	Inf	−0.007	−0.002	−0.008	−0.002	−0.002	0.043	0.043	0.041	0.041	0.040	100 ‡	100.6	95.7	95.2	93.0
		Large Uniform DIF (DIF SD $\tau = 0.5$)														
	500	−0.008	**0.014**	**−0.019**	0.007	0.003	0.188	0.203	0.177	0.186	0.189	100 ‡	**107.8**	94.5	98.4	100.2
	1000	**−0.012**	0.006	**−0.020**	0.001	0.002	0.176	0.187	0.168	0.174	0.174	100 ‡	**106.0**	95.7	98.6	98.6
10	2000	**−0.017**	−0.001	**−0.024**	−0.005	−0.004	0.172	0.182	0.162	0.168	0.167	100 ‡	**105.5**	95.0	97.5	96.6
	4000	**−0.012**	0.003	**−0.019**	−0.001	0.001	0.169	0.179	0.160	0.165	0.162	100 ‡	**105.4**	94.9	97.3	95.5
	Inf	**−0.011**	0.005	**−0.018**	0.000	0.002	0.165	0.175	0.156	0.161	0.156	100 ‡	**105.8**	94.6	97.3	94.4
	500	−0.008	**0.014**	**−0.017**	0.009	0.007	0.144	0.154	0.137	0.142	0.144	100 ‡	**107.0**	95.8	99.0	100.2
	1000	**−0.013**	0.004	**−0.020**	0.001	0.003	0.130	0.138	0.124	0.128	0.127	100 ‡	105.5	96.0	98.1	97.5
20	2000	**−0.015**	0.001	**−0.020**	−0.001	0.003	0.121	0.128	0.116	0.120	0.118	100 ‡	105.0	96.2	98.0	97.0
	4000	**−0.019**	−0.003	**−0.024**	−0.005	−0.003	0.122	0.129	0.115	0.119	0.116	100 ‡	**104.8**	95.7	97.2	94.3
	Inf	**−0.015**	0.000	**−0.020**	−0.002	0.002	0.115	0.122	0.110	0.113	0.111	100 ‡	104.8	95.8	97.5	95.3
	500	**−0.011**	0.009	**−0.019**	0.006	0.006	0.112	0.118	0.108	0.112	0.113	100 ‡	105.9	97.5	99.7	100.7
	1000	**−0.015**	0.002	**−0.021**	0.000	0.003	0.099	0.104	0.095	0.098	0.098	100 ‡	104.3	97.6	98.5	98.7
40	2000	**−0.018**	−0.002	**−0.023**	−0.003	0.000	0.092	0.097	0.088	0.091	0.090	100 ‡	103.8	96.9	97.2	95.9
	4000	**−0.016**	−0.001	**−0.021**	−0.002	0.001	0.087	0.092	0.082	0.085	0.083	100 ‡	**104.1**	96.2	96.7	94.3
	Inf	**−0.019**	−0.004	**−0.023**	−0.005	0.000	0.082	0.087	0.077	0.080	0.078	100 ‡	**103.4**	96.0	95.7	92.8

Note: HAE = Haebara linking; brHAE = bias-reduced Haebara linking; SL = Stocking–Lord linking; brSL = bias-reduced Stocking–Lord linking; MGM = mean-geometric-mean linking; Inf = infinite sample size; ‡ Haebara (HAE) linking was the reference method for computing the relative RMSE. Biases with absolute values larger than 0.010 are printed in bold font. Relative RMSE values larger than 10% more than the relative RMSE of the best-performing method in the respective condition are printed in bold font.

Overall, SL outperformed HAE regarding the RMSE, which was a direct consequence of the lower standard deviation of SL linking compared to HAE linking. Notably, the negative bias of HAE and SL was reduced by the bias-reduced variants brHAE and brSL in the large-DIF condition $\tau = 0.5$. MGM linking was advantageous over the other linking methods in the presence of large DIF and larger sample sizes.

We also computed descriptive statistics for the absolute biases and the relative RMSEs of the different linking methods across all simulation conditions. MGM linking had the lowest average absolute bias with a small SD (MGM: $M = 0.002$, $SD = 0.002$). The bias-reduced variants of HAE and SL slightly outperformed HAE and SL with respect to the average absolute bias (HAE: $M = 0.006$, $SD = 0.006$; brHAE: $M = 0.005$, $SD = 0.004$; SL: $M = 0.009$, $SD = 0.009$; brSL: $M = 0.003$, $SD = 0.003$). SL can be considered the frontrunner among the linking methods regarding the average relative RMSE for the estimated mean $\widehat{\mu}$ (HAE: $M = 100$, $SD = 0$; brHAE: $M = 102.62$, $SD = 2.18$; SL: $M = 96.93$, $SD = 1.60$; brSL: $M = 96.93$, $SD = 1.34$; MGM: $M = 99.48$, $SD = 4.59$).

Table 2 reports the bias, the SD, and the relative RMSE for the estimated SD $\widehat{\sigma}$. The MGM linking method exactly recovers the true SD $\sigma = 1.2$ in an infinite sample size in all DIF conditions because uniform DIF does not affect estimates of the SD. There was noticeable bias in the large-DIF condition for HAE and SL linking that did not vanish in an

infinite sample size. Overall, MGM was the best-performing method regarding bias and relative RMSE for $\widehat{\sigma}$ in DIF conditions and had only minor efficiency losses in the no-DIF conditions. In general, SL outperformed HAE linking regarding the RMSE, which was a consequence of the lower SD. Note that the advantages of SL linking over HAE linking were more pronounced in tests with fewer items (i.e., $I = 10$). The bias-reduced variants of HAE and SL (i.e., brHAE and brSL) effectively reduced a large portion of the bias in the estimated standard deviation $\widehat{\sigma}$. However, positive bias for brHAE and brSL was observed for a sample size $N = 500$.

In summary, descriptive statistics across all simulation conditions for the absolute bias resulted in the MGM linking method as the frontrunner, followed by brHAE and brSL (HAE: $M = 0.013$, $SD = 0.012$; brHAE: $M = 0.006$, $SD = 0.005$; SL: $M = 0.016$, $SD = 0.016$; brSL: $M = 0.005$, $SD = 0.004$; MGM: $M = 0.002$, $SD = 0.002$). The aggregated statistics for the relative RMSEs demonstrated the good performance of MGM linking. SL clearly outperformed the HAE and brHAE methods, but brSL additionally had improvements over SL, in particular for large DIF (HAE: $M = 100$, $SD = 0$; brHAE: $M = 103.29$, $SD = 4.78$; SL: $M = 89.02$, $SD = 10.00$; brSL: $M = 85.87$, $SD = 13.50$; MGM: $M = 69.95$, $SD = 33.90$).

Table 2. Simulation study: bias, standard deviation (SD), and relative root mean square error (RMSE) for the estimated standard deviation $\widehat{\sigma}$ as a function of the uniform DIF standard deviation τ, number of items I, and sample size N.

		Bias					**SD**					**Relative RMSE**				
I	***N***	**HAE**	**brHAE**	**SL**	**brSL**	**MGM**	**HAE**	**brHAE**	**SL**	**brSL**	**MGM**	**HAE**	**brHAE**	**SL**	**brSL**	**MGM**
		No DIF (DIF SD $\tau = 0$)														
	500	**0.013**	**0.018**	0.008	**0.015**	0.008	0.099	0.100	0.097	0.098	0.099	100 ‡	102.4	97.3	98.9	99.5
	1000	0.008	**0.011**	0.005	0.008	0.005	0.068	0.068	0.067	0.067	0.068	100 ‡	101.2	98.0	98.9	100.4
10	2000	0.002	0.004	0.001	0.003	0.001	0.047	0.047	0.046	0.047	0.048	100 ‡	100.5	98.7	99.0	101.4
	4000	0.001	0.002	0.000	0.001	0.000	0.033	0.033	0.033	0.033	0.034	100 ‡	100.2	99.1	99.2	101.7
	Inf	0	0	0	0	0	0	0	0	0	0	100 ‡	100.0	100.0	100.0	100.0
	500	**0.010**	**0.015**	0.005	**0.012**	0.005	0.077	0.078	0.076	0.076	0.077	100 ‡	102.4	97.5	99.0	98.8
	1000	0.005	0.008	0.003	0.006	0.003	0.054	0.054	0.053	0.053	0.054	100 ‡	101.2	98.6	99.4	100.6
20	2000	0.003	0.004	0.001	0.003	0.002	0.037	0.038	0.037	0.037	0.038	100 ‡	100.6	99.0	99.4	100.5
	4000	0.001	0.002	0.000	0.001	0.000	0.027	0.027	0.026	0.026	0.027	100 ‡	100.3	99.0	99.2	100.7
	Inf	0	0	0	0	0	0	0	0	0	0	100 ‡	100.0	100.0	100.0	100.0
	500	0.007	**0.012**	0.003	0.009	0.003	0.066	0.067	0.065	0.066	0.066	100 ‡	102.4	98.1	99.6	99.0
	1000	0.004	0.007	0.002	0.005	0.002	0.046	0.046	0.046	0.046	0.046	100 ‡	101.2	98.6	99.4	99.3
40	2000	0.002	0.003	0.000	0.002	0.000	0.033	0.033	0.032	0.032	0.033	100 ‡	100.5	99.1	99.4	99.8
	4000	0.000	0.001	−0.001	0.000	−0.001	0.022	0.022	0.022	0.022	0.022	100 ‡	100.2	99.4	99.4	100.5
	Inf	0	0	0	0	0	0	0	0	0	0	100 ‡	100.0	100.0	100.0	100.0
		Moderate Uniform DIF (DIF SD $\tau = 0.25$)														
	500	0.007	**0.020**	−0.001	**0.015**	0.009	0.108	0.113	0.099	0.101	0.098	**100** ‡	**105.9**	91.3	94.4	90.6
	1000	0.000	**0.010**	−0.005	0.007	0.005	0.081	0.084	0.072	0.073	0.069	**100** ‡	**104.5**	88.5	90.5	84.8
10	2000	−0.002	0.007	−0.007	0.004	0.002	0.065	0.068	0.053	0.055	0.048	**100** ‡	**104.6**	**82.7**	**84.5**	74.0
	4000	−0.005	0.004	−0.008	0.002	0.001	0.056	0.058	0.042	0.044	0.033	**100** ‡	**104.4**	**76.7**	**77.8**	59.6
	Inf	−0.007	0.001	**−0.010**	0.000	0.000	0.044	0.046	0.026	0.028	0.000	**100** ‡	**103.9**	**63.1**	**62.7**	0.0
	500	0.002	**0.015**	−0.005	**0.012**	0.006	0.085	0.088	0.079	0.081	0.079	100 ‡	**105.6**	93.4	96.2	92.9
	1000	−0.003	0.008	−0.008	0.005	0.003	0.062	0.064	0.056	0.057	0.053	**100** ‡	**104.1**	90.9	92.0	86.6
20	2000	−0.006	0.004	−0.009	0.003	0.002	0.050	0.052	0.042	0.043	0.038	**100** ‡	**103.7**	**85.9**	**86.5**	76.6
	4000	−0.009	0.000	**−0.011**	0.000	0.000	0.041	0.043	0.032	0.033	0.026	**100** ‡	**102.0**	**80.7**	**78.6**	62.8
	Inf	−0.008	0.001	**−0.010**	0.000	0.000	0.031	0.032	0.018	0.020	0.000	**100** ‡	**102.3**	**66.3**	**61.9**	0.0
	500	−0.001	**0.012**	−0.007	0.009	0.004	0.070	0.073	0.066	0.068	0.065	100 ‡	**105.2**	94.9	97.0	93.3
	1000	−0.006	0.005	**−0.010**	0.004	0.001	0.051	0.053	0.047	0.048	0.046	**100** ‡	**103.2**	94.0	94.0	90.2
40	2000	−0.008	0.001	**−0.011**	0.001	0.000	0.039	0.041	0.035	0.035	0.032	**100** ‡	**101.4**	**90.3**	87.9	80.3
	4000	−0.009	0.000	**−0.012**	−0.001	−0.002	0.032	0.033	0.026	0.027	0.023	**100** ‡	**100.1**	**86.4**	**80.8**	68.8
	Inf	**−0.010**	−0.001	**−0.011**	−0.001	0.000	0.022	0.023	0.013	0.014	0.000	**100** ‡	**96.8**	**72.1**	**58.7**	0.0

Table 2. *Cont.*

		Bias					SD					Relative RMSE				
I	*N*	HAE	brHAE	SL	brSL	MGM	HAE	brHAE	SL	brSL	MGM	HAE	brHAE	SL	brSL	MGM
		Large Uniform DIF (DIF SD $\tau = 0.5$)														
	500	**−0.017**	**0.014**	**−0.032**	0.006	0.006	0.130	0.149	0.106	0.115	0.098	**100** ‡	**114.2**	**84.3**	**87.6**	74.6
	1000	**−0.022**	0.007	**−0.033**	0.001	0.005	0.108	0.124	0.082	0.090	0.069	**100** ‡	**112.5**	**80.5**	**82.0**	62.9
10	2000	**−0.025**	0.003	**−0.036**	−0.003	0.002	0.097	0.113	0.067	0.076	0.048	**100** ‡	**112.7**	**75.9**	**76.2**	47.6
	4000	**−0.025**	0.002	**−0.036**	−0.004	0.000	0.092	0.108	0.060	0.070	0.033	**100** ‡	**112.9**	**73.6**	**73.2**	35.0
	Inf	**−0.026**	0.001	**−0.036**	−0.005	0.000	0.085	0.100	0.050	0.060	0.000	**100** ‡	**113.0**	**70.0**	**68.2**	0.0
	500	**−0.023**	0.008	**−0.035**	0.003	0.005	0.096	0.109	0.081	0.087	0.077	**100** ‡	**110.9**	**89.3**	**88.4**	77.9
	1000	**−0.029**	−0.001	**−0.039**	−0.004	0.002	0.080	0.091	0.063	0.068	0.054	**100** ‡	**107.1**	**86.3**	**80.2**	62.7
20	2000	**−0.033**	−0.005	**−0.041**	−0.007	0.000	0.070	0.080	0.051	0.057	0.038	**100** ‡	**104.5**	**84.3**	**74.2**	48.9
	4000	**−0.030**	−0.002	**−0.039**	−0.005	0.001	0.064	0.075	0.043	0.050	0.027	**100** ‡	**105.7**	**82.1**	**70.7**	37.9
	Inf	**−0.032**	−0.005	**−0.040**	−0.007	0.000	0.059	0.070	0.036	0.043	0.000	**100** ‡	**104.0**	**79.4**	**64.6**	0.0
	500	**−0.026**	0.004	**−0.037**	0.002	0.005	0.080	0.089	0.070	0.075	0.068	**100** ‡	**106.4**	**94.5**	89.1	81.0
	1000	**−0.033**	−0.005	**−0.041**	−0.006	0.001	0.062	0.070	0.051	0.056	0.046	**100** ‡	**100.1**	**93.5**	**79.2**	65.2
40	2000	**−0.034**	−0.006	**−0.042**	−0.007	−0.001	0.053	0.060	0.041	0.045	0.033	**100** ‡	**96.3**	**92.9**	**72.0**	52.3
	4000	**−0.034**	−0.007	**−0.041**	−0.007	0.000	0.048	0.055	0.033	0.038	0.023	**100** ‡	**95.0**	**90.9**	**66.2**	39.0
	Inf	**−0.035**	−0.008	**−0.042**	−0.008	0.000	0.042	0.050	0.026	0.031	0.000	**100** ‡	**91.9**	**89.0**	**58.1**	0.0

Note: HAE = Haebara linking; brHAE = bias-reduced Haebara linking; SL = Stocking–Lord linking; brSL = bias-reduced Stocking–Lord linking; MGM = mean-geometric-mean linking; Inf = infinite sample size; ‡ Haebara (HAE) linking was the reference method for computing the relative RMSE. Biases with absolute values larger than 0.010 are printed in bold font. Relative RMSE values larger than 10% more than the relative RMSE of the best-performing method in the respective condition are printed in bold font.

4. Discussion

Previous research has highlighted that the HAE and SL linking methods can produce substantial bias for SDs and for the mean (to a smaller extent) in the presence of DIF effects. Importantly, the bias does not vanish for large sample sizes of persons or a large number of items. To this end, bias-reduced variants of HAE and SL linking were proposed in this article to reduce large portions of the bias in the estimated SD. However, using the bias-reduced variants of the HAE and SL linking methods comes with the price of an increased SD. Whether it is beneficial regarding the RMSE to apply the bias-reduced linking method depends on the size of the DIF variance. In general, bias-reducing methods might be preferred in the presence of large DIF effects, but they could potentially hurt in situations with small DIF effects, because in those situations, the biases of the original HAE and SL linking methods are relatively small. Overall, SL linking outperformed HAE linking (see also [34–36]).

In future research, the methodology could be further improved by considering alternative bias-reducing linking estimators. Moreover, the methodology could be extended to linking polytomous item responses using the generalized partial credit model [37]. Also, this article only treated the impact of uniform DIF effects (i.e., DIF effects in item difficulties) in HAE and SL linking. It might also be interesting to study bias-reducing linking methods that can also handle nonuniform DIF effects (i.e., DIF effects in item discriminations).

Our simulation study relied on marginal maximum likelihood estimation of item response models. However, the item parameters in separate scalings could also be obtained with limited information estimation [38,39]. Stocking–Lord and Haebara linking remain unchanged when using a different estimation procedure of item parameters. Hence, our proposed methodology also applies to alternative IRT estimation methods.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Replication material for the simulation study can be found at https://osf.io/8mk49.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

2PL	two-parameter logistic
brHAE	bias-reduced Haebara
brSL	bias-reduced Stocking–Lord
HAE	Haebara
IRF	item response function
IRT	item response theory
MGM	mean-geometric-mean
RMSE	root mean square error
SD	standard deviation
SL	Stocking–Lord

References

1. Bock, R.D.; Moustaki, I. Item response theory in a general framework. In *Handbook of Statistics, Vol. 26: Psychometrics*; Rao, C.R., Sinharay, S., Eds.; Elsevier: Amsterdam, The Netherlands, 2007; pp. 469–513. [CrossRef]
2. Chen, Y.; Li, X.; Liu, J.; Ying, Z. Item response theory—A statistical framework for educational and psychological measurement. *Stat. Sci.* 2024, *Epub ahead of print*. Available online: https://rb.gy/1yic0e (accessed on 10 May 2024).
3. 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. [CrossRef]
4. 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.
5. 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.
6. Aitkin, M. Expectation maximization algorithm and extensions. In *Handbook of Item Response Theory, Vol. 2: Statistical Tools*; van der Linden, W.J., Ed.; CRC Press: Boca Raton, FL, USA, 2016; pp. 217–236. [CrossRef]
7. Bock, R.D.; Aitkin, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **1981**, *46*, 443–459. [CrossRef]
8. Glas, C.A.W. Maximum-likelihood estimation. In *Handbook of Item Response Theory, Vol. 2: Statistical Tools*; van der Linden, W.J., Ed.; CRC Press: Boca Raton, FL, USA, 2016; pp. 197–216. [CrossRef]
9. 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. [CrossRef]
10. Kolen, M.J.; Brennan, R.L. *Test Equating, Scaling, and Linking*; Springer: New York, NY, USA, 2014. [CrossRef]
11. 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. [CrossRef]
12. Millsap, R.E. *Statistical Approaches to Measurement Invariance*; Routledge: New York, NY, USA, 2011. [CrossRef]
13. Penfield, R.D.; Camilli, G. Differential item functioning and item bias. In *Handbook of Statistics, Vol. 26: Psychometrics*; Rao, C.R., Sinharay, S., Eds.; Elsevier: Amsterdam, The Netherlands, 2007; pp. 125–167. [CrossRef]
14. Monseur, C.; Berezner, A. The computation of equating errors in international surveys in education. *J. Appl. Meas.* **2007**, *8*, 323–335. Available online: https://bit.ly/2WDPeqD (accessed on 10 May 2024). [PubMed]
15. 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. [CrossRef]
16. Robitzsch, A. Linking error in the 2PL model. *J* **2023**, *6*, 58–84. [CrossRef]
17. Sachse, K.A.; 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*, 152–171. [CrossRef]
18. Sachse, K.A.; 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*, 102–116. [CrossRef]
19. Wu, M. Measurement, sampling, and equating errors in large-scale assessments. *Educ. Meas.* **2010**, *29*, 15–27. [CrossRef]
20. 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. [CrossRef]
21. Haebara, T. Equating logistic ability scales by a weighted least squares method. *Jpn. Psychol. Res.* **1980**, *22*, 144–149. [CrossRef]
22. Stocking, M.L.; Lord, F.M. Developing a common metric in item response theory. *Appl. Psychol. Meas.* **1983**, *7*, 201–210. [CrossRef]
23. Longford, N.T.; Holland, P.W.; Thayer, D.T. Stability of the MH D-DIF statistics across populations. In *Differential Item Functioning*; Holland, P.W., Wainer, H., Eds.; Routledge: New York, NY, USA, 1993; pp. 171–196. [CrossRef]

24. Penfield, R.D.; Algina, J. A generalized DIF effect variance estimator for measuring unsigned differential test functioning in mixed format tests. *J. Educ. Meas.* **2006**, *43*, 295–312. [CrossRef]
25. 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.
26. 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. [CrossRef]
27. Shealy, R.; Stout, W. A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika* **1993**, *58*, 159–194. [CrossRef]
28. Robitzsch, A.; Lüdtke, O. Why full, partial, or approximate measurement invariance are not a prerequisite for meaningful and valid group comparisons. *Struct. Equ. Model.* **2023**, *30*, 859–870. [CrossRef]
29. 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. [CrossRef]
30. von Davier, M.; Yamamoto, K.; Shin, H.J.; Chen, H.; Khorramdel, L.; Weeks, J.; Davis, S.; Kong, N.; Kandathil, M. Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assess. Educ.* **2019**, *26*, 466–488. [CrossRef]
31. R Core Team. R: A Language and Environment for Statistical Computing. 2023. Available online: https://www.R-project.org (accessed on 15 March 2023).
32. Robitzsch, A.; Kiefer, T.; Wu, M. *TAM: Test Analysis Modules*, R Package Version 4.2-21. 2024. Available online: https://cran.r-project.org/web/packages/TAM/index.html (accessed on 20 April 2024).
33. Robitzsch, A. *sirt: Supplementary Item Response Theory Models*, R Package Version 4.2-57. 2024. Available online: https://github.com/alexanderrobitzsch/sirt (accessed on 20 April 2024).
34. Kang, T.; Petersen, N.S. Linking item parameters to a base scale. *Asia Pacific Educ. Rev.* **2012**, *13*, 311–321 [CrossRef]
35. Kilmen, S.; Demirtasli, N. Comparison of test equating methods based on item response theory according to the sample size and ability distribution. *Procedia Soc. Behav. Sci.* **2012**, *46*, 130–134. [CrossRef]
36. Lee, W.C.; Ban, J.C. A comparison of IRT linking procedures. *Appl. Meas. Educ.* **2009**, *23*, 23–48. [CrossRef]
37. Muraki, E. A generalized partial credit model: Application of an EM algorithm. *Appl. Psychol. Meas.* **1992**, *16*, 159–176. [CrossRef]
38. Forero, C.G.; Maydeu-Olivares, A.; Gallardo-Pujol, D. Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Struct. Equ. Model.* **2009**, *16*, 625–641. [CrossRef]
39. Robitzsch, A. A comparison of limited information estimation methods for the two-parameter normal-ogive model with locally dependent items. *Stats* **2024**, *7*, 576–591. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.