

Article

Estimation of Standard Error, Linking Error, and Total Error for Robust and Nonrobust Linking Methods in the Two-Parameter Logistic Model

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: The two-parameter logistic (2PL) item response theory model is a statistical model for analyzing multivariate binary data. In this article, two groups are brought onto a common metric using the 2PL model using linking methods. The linking methods of mean–mean linking, mean–geometric–mean linking, and Haebara linking are investigated in nonrobust and robust specifications in the presence of differential item functioning (DIF). M-estimation theory is applied to derive linking errors for the studied linking methods. However, estimated linking errors are prone to sampling error in estimated item parameters, thus resulting in artificially increased the linking error estimates in finite samples. For this reason, a bias-corrected linking error estimate is proposed. The usefulness of the modified linking error estimate is demonstrated in a simulation study. It is shown that a simultaneous assessment of the standard error and linking error in a total error must be conducted to obtain valid statistical inference. In the computation of the total error, using the bias-corrected linking error estimate instead of the usually employed linking error provides more accurate coverage rates.

Keywords: item response model; linking; 2PL model; M-estimation; linking error; standard error; total error; robust linking; nonrobust linking



Citation: Robitzsch, A. Estimation of Standard Error, Linking Error, and Total Error for Robust and Nonrobust Linking Methods in the Two-Parameter Logistic Model. *Stats* **2024**, *7*, 592–612. <https://doi.org/10.3390/stats7030036>

Academic Editor: Marco Riani

Received: 29 April 2024

Revised: 17 June 2024

Accepted: 19 June 2024

Published: 21 June 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 [1–3] are multivariate statistical models for multivariate binary random variables. These models are frequently used to model cognitive testing data from educational or psychological applications. For example, IRT models are operationally utilized in educational large-scale assessments [4,5], like the programme for international student assessment (PISA; [6]) study.

In this article, we only treat unidimensional IRT models [7]. Let $\mathbf{X} = (X_1, \dots, X_I)$ be the vector of I dichotomous random variables $X_i \in \{0, 1\}$ (also referred to as items or (scored) item responses). A unidimensional IRT model [8] is a statistical model for the probability distribution $P(\mathbf{X} = \mathbf{x})$ for $\mathbf{x} = (x_1, \dots, x_I) \in \{0, 1\}^I$, where

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

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 the factor variable, trait, or ability) are contained in the vector $\boldsymbol{\delta} = (\mu, \sigma)$. The vector $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_I)$ contains all the estimated item parameters of item response functions (IRFs); where $P_i(\theta; \boldsymbol{\gamma}_i) = P(X_i = 1 | \theta)$ ($i = 1, \dots, I$). The two-parameter logistic (2PL) model [9] possesses the following IRF:

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

Using the item discrimination a_i and item difficulty b_i , $\Psi(x) = (1 + \exp(-x))^{-1}$ denotes the logistic distribution function. The 2PL model could also be estimated for non-normal distributions [10–15]. In this case, an identification constraint is typically applied to an item X_{i_0} such that $a_{i_0} = 1$ and $b_{i_0} = 0$.

The Rasch model [16] is obtained from the 2PL model as a particular case in which all item discriminations equal 1. Some researchers believe that the Rasch model offers particular measurement (i.e., metrological) properties in contrast to the 2PL model (e.g., [17–19]). However, in our view, the Rasch model has only one advantage over the 2PL model, which is that a conditional maximum likelihood estimation is applicable [20]. Moreover, the Rasch model possesses the unweighted sum score as a sufficient statistic for θ , which offers many interpretational advantages [21–26]. There is a disbelief that group comparisons can only be conducted with the Rasch model because it has a so-called property of separability, which entails specific objective comparisons [27,28]. However, this reasoning is incorrect and can be disproved by empirical data [29]. In fact, any IRT model with invariant item parameters across groups allows for invariant group comparisons [7,30], although proponents of the Rasch model frequently claim otherwise [31,32].

If independent and identically distributed observations $x_1, \dots, x_N$ of N persons from the distribution of the random variable X are available, and the unknown model parameters of the IRT model in (1) can be estimated using the marginal maximum likelihood (MML) using an expectation–maximization algorithm [33,34].

IRT models are frequently used to compare the performance of two groups in a test (i.e., on a set of items) regarding the factor variable θ in the IRT model (1). In the following, we only discuss the 2PL model. Two primary approaches can be distinguished [35]. First, concurrent calibration can be applied in which a joint IRT model is estimated in the two groups by assuming common (i.e., invariant) item discriminations and item difficulties in the two groups. While the mean and the standard deviation of θ are fixed in the first group for identification reasons, the mean μ and the standard deviation σ can be identified for the second group. Hence, these two parameters summarize the group difference regarding the factor variable θ . Second, the 2PL model can be separately estimated in each of the two groups. This approach allows items to function differently across groups, which is a property that is referred to as differential item functioning (DIF; see [36–38]). In the second step, the differences in item parameters are used to determine the group difference regarding the θ variable by means of a linking method [39–41]. The occurrence of DIF causes additional variability in the estimated mean μ and standard deviation σ [42–46]. Therefore, the estimated distribution parameters (μ, σ) depend on the choice of selected items, even for infinite sample sizes of persons. This variability is quantified in the linking error [47–52].

An anonymous reviewer pointed out that a simultaneous estimation, while allowing for group-specific DIF effects, would also be possible [53–56]. In fact, the two-step procedure of separate scaling with subsequent linking can be equivalently formulated as a one-step simultaneous estimation (i.e., concurrent calibration) with nonlinear constraints on group-specific item parameters [57].

This article investigates the computation of the total uncertainty of linking methods in a general treatment based on M-estimation theory [58]. A new bias-corrected linking error is derived, which results in the better performance of the coverage rates of constructed confidence intervals for linking estimates. In particular, it turned out that the newly proposed bias-corrected linking error estimate has a smaller bias for the linking error than the estimators currently employed in the literature.

The rest of the article is organized as follows. Section 2 formalizes the linking methods in the statistical language of estimating equations (i.e., M-estimation). Section 3 presents examples of linking methods that are subsequently investigated in two simulation studies. In Section 4, M-estimation theory is applied to compute the linking error, standard error, total error, and the newly proposed bias-corrected linking error. Then, Sections 5 and 6

present the findings from two simulation studies. Finally, the article closes with a discussion in Section 7.

2. Linking Method

In this section, we formally define linking methods as a particular M-estimation [58,59] problem. We refer to Section 3 for examples of linking methods. Throughout this paper, it is assumed that $\hat{\gamma}_i = (\hat{a}_{i1}, \hat{b}_{i1}, \hat{a}_{i2}, \hat{b}_{i2})$ contains the estimated item parameters from the 2PL model for item i in both groups. One can assume that there exists a true item parameter $\gamma_i = (a_{i1}, b_{i1}, a_{i2}, b_{i2})$ in the population (i.e., for an infinite sample size). The quantity $\hat{\gamma}_i$ is a consistent estimate for γ_i (under the scheme $N \rightarrow \infty$). The goal of a linking method consists of estimating the mean μ and the standard deviation σ in the second group based on all the estimated item parameters $\hat{\gamma}_i$.

2.1. One-Step Linking Method

In a one-step method linking method, the parameter estimate $\hat{\delta} = (\hat{\mu}, \hat{\sigma})$ is obtained as the minimizer of a nonlinear optimization function H that is a sum of terms, in which each term refers to a single item. Formally, we define

$$\hat{\delta} = \arg \min_{\delta} H(\delta) \text{ with } H(\delta) = \sum_{i=1}^I h(\delta; \hat{\gamma}_i). \quad (3)$$

Now, we define the partial derivatives $H_{\mu} = (\partial H)/(\partial \mu)$, $H_{\sigma} = (\partial H)/(\partial \sigma)$, $h_{\mu} = (\partial h)/(\partial \mu)$, $h_{\sigma} = (\partial h)/(\partial \sigma)$ and $\mathbf{H}_{\delta} = (H_{\mu}, H_{\sigma})$, and $\mathbf{h}_{\delta} = (h_{\mu}, h_{\sigma})$ for brevity. The parameter estimate $\hat{\delta}$ solves the following estimating equation:

$$\mathbf{H}_{\delta}(\delta) = \sum_{i=1}^I \mathbf{h}_{\delta}(\delta; \hat{\gamma}_i) = \mathbf{0}. \quad (4)$$

M-estimation theory is applied for computing a variance estimate for $\hat{\delta}$ under the assumption that $I \rightarrow \infty$ and $\{\hat{\gamma}_i\}_{i=1, \dots, I}$ are independent realizations from a distribution [58]. In this sense, the uncertainty of $\hat{\delta}$ regarding the selected set of items is somehow quantified.

2.2. Two-Step Linking Method

In a two-step linking method, the standard deviation σ is estimated in the first step. Afterward, the mean μ is estimated in the second step. The estimate $\hat{\sigma}$ of the standard deviation σ is determined as a root of the nonlinear equation (that is, it is additive in items).

$$H_{\sigma}(\sigma) = \sum_{i=1}^I h_{\sigma}(\sigma; \hat{\gamma}_i) = 0. \quad (5)$$

In the second step, the estimate $\hat{\mu}$ is obtained as the root of the estimating equation

$$H_{\mu}(\mu, \hat{\sigma}) = \sum_{i=1}^I h_{\mu}(\mu, \hat{\sigma}; \hat{\gamma}_i) = 0. \quad (6)$$

Note that the two-step method can be alternatively considered as a one-step linking method, because the estimate $\hat{\delta} = (\hat{\mu}, \hat{\sigma})$ solves the stacked estimating equation (see [58]).

$$\mathbf{H}_{\delta}(\delta) = \mathbf{0}, \text{ where } \mathbf{H}_{\delta}(\delta) = \begin{pmatrix} H_{\mu}(\mu, \sigma) \\ H_{\sigma}(\sigma) \end{pmatrix} \text{ for } \delta = (\mu, \sigma). \quad (7)$$

Hence, one-step and two-step linking methods can be simultaneously analyzed using M-estimation theory.

2.3. Statistical Inference

This article is concerned with estimating the uncertainty of the estimated mean and standard deviation that are contained in the vector δ . The uncertainty is due to the sampling (or selection) of persons (i.e., under the scheme $N \rightarrow \infty$) and the selection (or sampling/modeling the randomness in group comparisons) of items (i.e., under the scheme $I \rightarrow \infty$).

Most of the linking literature that treats the uncertainty in $\hat{\gamma}$ by computing a standard error (SE) of $\hat{\delta}$ for a fixed number of items [60–62]. In this case, the variability in $\hat{\gamma}$ exists because there is sampling variability in the estimated item parameters $\hat{\gamma}_i$.

We have argued that the estimated item parameters $\hat{\gamma}_i$ have a population analogue γ_i in an infinite sample size. If there is a model error (i.e., DIF), the estimated linking parameter $\hat{\delta}$ depends on the chosen set of items, even for an infinite sample size. This variability due to item selection, which is referred to as the linking error (LE; see [48,63]) and the variance estimation, operates under the scheme where $I \rightarrow \infty$.

The total error (TE) includes both sources of uncertainty: the standard error due to randomness due to persons and the linking error due to randomness in items [47,50,51,64]. However, it has been argued that the ordinary linking error estimate is partly affected by the sampling error [65]. In this article, a bias-corrected linking error resulting in a bias-corrected total error estimate is examined to try to reduce the portion of the estimated linking error variance that is due to the sampling error. We outline the statistical underpinnings of the estimators in Section 4.

3. Robust and Nonrobust Linking Methods

In this section, we discuss the most frequently employed linking methods for the 2PL model.

3.1. Mean–Mean Linking (MM)

The mean–mean (MM) linking method is a two-step linking method [40]. The standard deviation σ is estimated in the first step as the ratio of the means of the item discriminations in the two groups.

$$\hat{\sigma} = \frac{\frac{1}{I} \sum_{i=1}^I \hat{a}_{2i}}{\frac{1}{I} \sum_{i=1}^I \hat{a}_{1i}}. \quad (8)$$

In the second step, the mean μ is estimated by

$$\hat{\mu} = -\hat{\sigma} \frac{1}{I} \sum_{i=1}^I \hat{b}_{i2} + \frac{1}{I} \sum_{i=1}^I \hat{b}_{i1}. \quad (9)$$

The estimate $\hat{\delta} = (\hat{\mu}, \hat{\sigma})$ can be written as the solution $\delta = (\mu, \sigma)$ of the estimating equations

$$\sum_{i=1}^I (\sigma \hat{b}_{i2} - \hat{b}_{i1} + \mu) = 0 \text{ and } \sum_{i=1}^I (\sigma \hat{a}_{i1} - \hat{a}_{i2}) = 0. \quad (10)$$

The MM linking method is considered nonrobust to outliers in the item parameter differences between groups, because the mean is an outlier-sensitive location measure.

3.2. Mean–Geometric–Mean Linking (MGM and RMGM)

The mean–geometric–mean (MGM) linking [40] (also referred to as log–mean–mean linking in [66]) is another two-step linking method that estimates σ as the ratio of the geometric means of the item discriminations. Using a general loss function ρ , the standard deviation σ is estimated in the first step as

$$\hat{\sigma} = \arg \min_{\sigma} \sum_{i=1}^I \rho(\log \hat{a}_{i2} - \log(\sigma) - \log \hat{a}_{i1}), \quad (11)$$

where $\rho(x) = (x^2 + \varepsilon)^{p/2}$ is a differentiable approximation of the L_p loss function $x \mapsto |x|^p$ for $p \in (0, 2]$ and a sufficiently small $\varepsilon > 0$ (see [67]). In this article, we use $\varepsilon = 0.01$ in the simulation studies [68]. The ordinary mean is obtained using the L_p loss function for $p = 2$. The median as a location measure can be approximately obtained for $p = 1$. The choice $p = 0.5$ is advantageous for asymmetrically distributed error distributions for DIF effects [69] and appears in invariance alignment [68,70]. In the second step, the mean μ is estimated by

$$\hat{\mu} = \arg \min_{\mu} \sum_{i=1}^I \rho(\hat{\sigma} \hat{b}_{i2} - \hat{b}_{i1} + \mu). \quad (12)$$

By defining $\rho_1 = d\rho/dx$, the solution in MGM linking is given as the root of the estimating equations as follows:

$$\sum_{i=1}^I \rho_1(\log \hat{a}_{i2} - \log(\sigma) - \log \hat{a}_{i1}) = 0 \text{ and } \sum_{i=1}^I \rho_1(\hat{\sigma} \hat{b}_{i2} - \hat{b}_{i1} + \mu) = 0. \quad (13)$$

MGM linking in a nonrobust variant is defined with $p = 2$, while robust MGM (abbreviated as RMGM) linking is defined by choosing $p = 0.5$ in the loss function ρ .

3.3. Symmetric Haebara Linking (HAE and RHAE)

Haebara linking is a one-step linking method [71] that relies on aligning IRFs to determine μ and σ . Using a discrete grid of θ points θ_t ($t = 1, \dots, T$) and weights ω_t , the mean μ and the standard deviation σ are estimated by minimizing a weighted distance between the IRFs, which is shown as follows:

$$H(\mu, \sigma) = \sum_{i=1}^I \sum_{t=1}^T \omega_t \rho(P(\sigma \theta_t + \mu; \hat{a}_{i1}, \hat{b}_{i1}) - P(\theta_t; \hat{a}_{i2}, \hat{b}_{i2})), \quad (14)$$

where $P(\theta; a, b) = \Psi(a(\theta - b))$, and ρ again define a differentiable approximation of the L_p loss function. For example, the θ grid can be equidistantly chosen between -6 and 6 , and ω_t could be proportional to the values of the density of a normal distribution with a standard deviation of 2. Some researchers alternatively prefer to set all weights ω_t equal to 1. The nonrobust loss function $\rho(x) = x^2$ has been originally proposed by Haebara [71] and is widely used. The robust loss function with $p = 1$ has been proposed in Refs. [72,73]. The general robust L_p loss function in Haebara linking for $p \leq 1$ was utilized in [74]. Corresponding estimating equations to (14) are obtained by computing the partial derivatives of H with respect to μ and σ .

The linking function H in (14) is referred to as asymmetric Haebara linking, because it aligns the IRF of the first group with the IRF of the second group. This kind of asymmetry induces non-negligibly biased estimates for the standard deviation and for the mean to a smaller extent [66]. To this end, symmetric Haebara linking [75] has been proposed to simultaneously align the IRFs of both groups and to reduce the bias [66]. The linking function of symmetric Haebara linking is defined by

$$H(\mu, \sigma) = \sum_{i=1}^I \sum_{t=1}^T \omega_t \rho(P(\sigma \theta_t + \mu; a_{i1}, b_{i1}) - P(\theta_t; a_{i2}, b_{i2})) + \sum_{i=1}^I \sum_{t=1}^T \omega_t \rho(P(\theta_t; a_{i1}, b_{i1}) - P(\sigma^{-1} \theta_t - \sigma^{-1} \mu; a_{i2}, b_{i2})). \quad (15)$$

The estimate $\hat{\delta} = (\hat{\mu}, \hat{\sigma})$ is obtained by minimizing H in (15) with respect to $\delta = (\mu, \sigma)$. Nonrobust symmetric Haebara (HAE) linking is obtained with the choice $p = 2$ of the L_p loss function, while robust symmetric Haebara (RHAE) linking is obtained with the choice $p = 0.5$.

4. Estimation of Standard Error, Linking Error, and Total Error

In this section, we derive the standard error, linking error, and total error for linking estimates in the framework of M-estimation theory [58,59,76]. The treatment in this section is an extension of the material presented in [65]. Assume the item parameter estimate $\hat{\gamma}_i = (\hat{a}_{i1}, \hat{b}_{i1}, \hat{a}_{i2}, \hat{b}_{i2})$ with an estimated variance matrix $V_{\hat{\gamma}_i}$. Moreover, let $\hat{\gamma} = (\hat{\gamma}_1, \dots, \hat{\gamma}_I)$ be the vector of all the item parameters with an estimated variance matrix $V_{\hat{\gamma}}$. The corresponding population analogs of the estimators are denoted by γ_i and γ and are effectively the estimates for an infinite sample size. As described in Section 2, the linking method provides an estimate $\hat{\delta} = (\hat{\mu}, \hat{\sigma})$ for the population parameter $\delta = (\mu, \sigma)$ as a root of the estimating equation.

$$H_{\delta}(\hat{\delta}) = \sum_{i=1}^I h_{\delta}(\hat{\delta}; \hat{\gamma}_i) = 0. \quad (16)$$

4.1. Linking Error

First, we derive the linking error of the estimate $\hat{\delta}$ of the estimating equation (16) that quantifies the uncertainty in the estimate due to the selection (or randomness) of items. As is usual in M-estimation theory, we carry out a Taylor approximation of H_{δ} around the true parameter δ (sometimes also referred to as the pseudotrue parameter; [58])

$$H_{\delta}(\hat{\delta}) = H_{\delta}(\delta) + H_{\delta\delta}(\delta)(\hat{\delta} - \delta) = 0, \quad (17)$$

where $H_{\delta\delta}$ denotes the vector of partial derivatives of H_{δ} with respect to δ . Moreover, it holds that $H_{\delta}(\delta) = 0$ because of the definition of the true parameter [58]. Hence, we obtain from (17) the following:

$$\hat{\delta} - \delta = -H_{\delta\delta}(\delta)^{-1} H_{\delta}(\hat{\delta}). \quad (18)$$

M-estimation theory provides the variance matrix of $\hat{\delta}$ as the sandwich variance estimate:

$$V_{LE} = \text{Var}(\hat{\delta}) = A^{-1} B A^{-\top}, \text{ where} \quad (19)$$

$$A = H_{\delta\delta}(\delta) = \sum_{i=1}^I h_{\delta\delta}(\delta; \hat{\gamma}_i) \text{ and} \quad (20)$$

$$B = \text{Var}(H_{\delta}(\hat{\delta})) = \sum_{i=1}^I \text{Var}(h_{\delta}(\delta; \hat{\gamma}_i)). \quad (21)$$

In (21), we used the approximate independence of the item parameters across items. In M-estimation, the matrix A is called the bread matrix, and B is the meat matrix. The unknown quantities in (19) can be estimated by

$$\hat{A} = \sum_{i=1}^I h_{\delta\delta}(\hat{\delta}; \hat{\gamma}_i) \text{ and} \quad (22)$$

$$\hat{B} = \sum_{i=1}^I h_{\delta}(\hat{\delta}; \hat{\gamma}_i) h_{\delta}(\hat{\delta}; \hat{\gamma}_i)^{\top}. \quad (23)$$

Hence, an estimate of the variance matrix V_{LE} is given by

$$\hat{V}_{LE} = \frac{I}{I-1} \cdot \hat{A}^{-1} \hat{B} \hat{A}^{-\top}. \quad (24)$$

The factor $I/(I-1)$ in (24) is included to correct for finite-sample bias [65,77–79].

4.2. Standard Error

We now compute the standard error of $\hat{\delta}$ due to the sampling of persons (see [80–85]). A Taylor approximation of h_{δ} around (δ, γ_i) is carried out and results in

$$h_{\delta}(\hat{\delta}; \hat{\gamma}_i) = h_{\delta}(\delta, \gamma_i) + h_{\delta\gamma}(\delta, \gamma_i)(\hat{\gamma}_i - \gamma_i) + h_{\delta\delta}(\delta, \gamma_i)(\hat{\delta} - \delta). \quad (25)$$

We can now use

$$\sum_{i=1}^I h_{\delta}(\hat{\delta}; \hat{\gamma}_i) = \mathbf{0} \text{ and } \sum_{i=1}^I h_{\delta}(\delta, \gamma_i) = \mathbf{0}, \text{ and arrive at} \quad (26)$$

$$\hat{\delta} - \delta = \left(\sum_{i=1}^I h_{\delta\delta}(\delta, \gamma_i) \right)^{-1} \sum_{i=1}^I h_{\delta\gamma}(\delta, \gamma_i)(\hat{\gamma}_i - \gamma_i) = A^{-1}C(\hat{\gamma} - \gamma), \text{ where} \quad (27)$$

$$C = \begin{pmatrix} h_{\delta\gamma}(\delta; \gamma_1) & \cdots & h_{\delta\gamma}(\delta; \gamma_I) \end{pmatrix}. \quad (28)$$

This allows us to compute the variance matrix in $\hat{\delta}$ due to the sampling error as follows:

$$V_{SE} = A^{-1}DA^{-T} \text{ with } D = CV_{\hat{\gamma}}C^T. \quad (29)$$

The unknown quantities in (29) can be estimated using $\hat{A}$ in (22), and we obtain the following:

$$\hat{D} = \hat{C}V_{\hat{\gamma}}\hat{C}^T \text{ with } \hat{C} = \begin{pmatrix} h_{\delta\gamma}(\hat{\delta}, \hat{\gamma}_1) & \cdots & h_{\delta\gamma}(\hat{\delta}, \hat{\gamma}_I) \end{pmatrix}. \quad (30)$$

4.3. Total Error and Bias-Corrected Linking Error

We now compute the total uncertainty in $\hat{\delta}$ (i.e., the total error). The variance as the total error has been defined as the sum of the variances due to the sampling error and linking error, and it is written as follows (see [50,65]):

$$V_{TE} = V_{SE} + V_{LE} \text{ and } \hat{V}_{TE} = \hat{V}_{SE} + \hat{V}_{LE}. \quad (31)$$

We now derive a bias-corrected estimate of the linking error variance matrix V_{LE} , which, in turn, allows us to compute a bias-corrected variance matrix for the linking error. The estimated meat matrix in the variance matrix for the linking error is given as

$$\hat{B} = \sum_{i=1}^I h_{\delta}(\hat{\delta}; \hat{\gamma}_i)h_{\delta}(\hat{\delta}; \hat{\gamma}_i)^T. \quad (32)$$

However, the linking error should only be computed based on the true item parameters γ_i instead of the estimated item parameters $\hat{\gamma}_i$, which appear in (32). A Taylor approximation provides

$$h_{\delta}(\hat{\delta}; \hat{\gamma}_i) = h_{\delta}(\hat{\delta}; \gamma_i) + h_{\delta\gamma}(\hat{\delta}, \gamma_i)(\hat{\gamma}_i - \gamma_i). \quad (33)$$

Hence, the inflated variance contribution in $\hat{B}$ due to the sampling error can be determined as

$$\text{Var} \left(\sum_{i=1}^I h_{\delta\gamma}(\hat{\delta}, \gamma_i)(\hat{\gamma}_i - \gamma_i) \right) = \sum_{i=1}^I h_{\delta\gamma}(\hat{\delta}, \gamma_i)V_{\hat{\gamma}_i}h_{\delta\gamma}(\hat{\delta}, \gamma_i)^T, \quad (34)$$

where we used the approximate independence of the item parameters across items. As a result, we compute a bias-corrected meat matrix as follows:

$$\hat{B}_{bc} = \hat{B} - \tilde{D} \text{ with } \tilde{D} = \sum_{i=1}^I h_{\delta\gamma}(\hat{\delta}, \hat{\gamma}_i)V_{\hat{\gamma}_i}h_{\delta\gamma}(\hat{\delta}, \hat{\gamma}_i)^T \quad (35)$$

Note that the correction term $\tilde{D}$ in (35) corresponds to the matrix $\hat{D}$ in (29) in the variance due to standard errors if the item parameters $\hat{\gamma}_i$ were uncorrelated across items. Next, a bias-corrected variance matrix due to linking error is given as

$$\hat{V}_{LE,bc} = \frac{I}{I-1} \cdot \hat{A}^{-1} \hat{B}_{bc} \hat{A}^{-\top} \quad (36)$$

and the variance matrix for the total error is given by

$$\hat{V}_{TE,bc} = \hat{V}_{SE} + \hat{V}_{LE,bc} . \quad (37)$$

To sum up, the variance matrix referring to the total error can be written as

$$\hat{V}_{TE} = \hat{A}^{-1} \left(\frac{I}{I-1} \hat{B} + \hat{D} \right) \hat{A}^{-\top} \text{ and } \hat{V}_{TE,bc} = \hat{A}^{-1} \left(\frac{I}{I-1} \hat{B} + \hat{D} - \frac{I}{I-1} \tilde{D} \right) \hat{A}^{-\top} . \quad (38)$$

To obtain standard errors, linking errors, bias-corrected linking errors, total errors and bias-corrected total errors, the square root of the diagonal elements of the corresponding matrices can be taken. In cases of negative variances for bias-corrected estimates, a corresponding linking error estimate is set to zero.

5. Simulation Study 1: Assessing Linking Error in Infinite Sample Size

In the first Simulation Study 1, the validity of the variance estimates based on M-estimation (see Section 4) for the estimated mean $\hat{\mu}$ and the estimated standard deviation $\hat{\sigma}$ was investigated for an infinite sample size of persons.

5.1. Method

In this study, only item parameters were simulated in each replication. No item responses were simulated, because the case of an infinite sample size N was investigated. The 2PL model was used as the IRF in the IRT model. There were two groups. For identification reasons, the mean and the standard deviation of the factor variable θ in the first group were set to 0 to 1, respectively. The mean μ and the standard deviation σ of the second group parametrized the group differences. Throughout the simulation, the choices $\mu = 0.3$ and $\sigma = 1.2$ were made.

We simulated item parameters for $I = 10, 20$, and 40 items. The group-specific item parameters a_{ig} and b_{ig} for $i = 1, \dots, I$ 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 base item discriminations a_{i0} in the case of $I = 10$ items were chosen as 0.73, 1.25, 1.20, 1.47, 0.97, 1.38, 1.05, 1.14, 1.15, and 0.67. The base item discriminations b_{i0} were chosen as $-1.31, 1.44, -1.20, 0.10, 0.10, -0.74, 1.48, -0.61, 0.82$, and -0.07 . The same item parameters were also chosen in [65]. For item numbers as multiples of 10, we duplicated the item parameters of the 10 items accordingly. The group-specific item difficulties b_{ig} ($g = 1, 2$) were simulated as

$$b_{i1} = b_{i0} - e_i/2 \text{ and } b_{i2} = b_{i0} + e_i/2 , \quad (39)$$

where e_i is a random DIF effect. Note that $e_i = b_{i2} - b_{i1}$ parametrizes a uniform DIF effect as the difference in group-specific item difficulties. Group-specific item discriminations a_{ig} ($g = 1, 2$) were simulated as

$$a_{i1} = a_{i0} \exp(-f_i/2) \text{ and } a_{i2} = a_{i0} \exp(f_i/2) , \quad (40)$$

where f_i is another random DIF effect. The nonuniform DIF effect f_i can be computed as the difference of logarithms of item discriminations (i.e., $f_i = \log a_{i2} - \log a_{i1}$). In the simulation, we assumed that e_i and f_i were uncorrelated. Both DIF effects had zero means and had standard deviation τ for e_i and $0.3 \times \tau$ for f_i . Two distributions of e_i and f_i were specified: a normal distribution or a scaled t distribution with three degrees of freedom

(denoted as t_3). In this simulation study, we varied the DIF standard deviation for DIF effects for e_i item difficulties as 0.25 and 0.50. According to the definition, the respective DIF standard deviation for DIF effects f_i for (logarithmized) item discriminations were 0.075, and 0.15, respectively.

Five different linking methods were utilized to compute to estimate the mean difference $\hat{\mu}$ and the standard deviation $\hat{\sigma}$: mean–mean (MM) linking, mean–geometric–mean (MGM) linking, robust mean–geometric–mean (RGM) linking, symmetric Haebara (HAE) linking, and robust symmetric Haebara (RHAE) linking. The linking methods rely on estimated item discriminations $\hat{a}_{ig}$ and item difficulties $\hat{b}_{ig}$ ($i = 1, \dots, I, g = 1, 2$). In an infinite sample size, these identified item parameters are given as

$$\hat{a}_{i1} = a_{i1}, \hat{b}_{i1} = b_{i1}, \hat{a}_{i2} = \sigma^{-1}a_{i2}, \text{ and } \hat{b}_{i2} = \sigma^{-1}(b_{i2} - \mu). \quad (41)$$

In each of the 2 (type of distribution) $\times$ 2 (DIF standard deviation τ) $\times$ 3 (number of items I) = 12 cells of the simulation, 5000 replications were conducted. We computed bias and root mean square error (RMSE) for the estimated mean $\hat{\mu}$ and the estimated standard deviation $\hat{\sigma}$. A relative percentage RMSE was computed as the ratio of the RMSE values of a particular linking method and the chosen reference method of MGM linking. We also assessed the coverage rate for $\hat{\mu}$ and $\hat{\sigma}$ at the 95% confidence level based on the normal distribution as the percentage of events in which an estimated confidence interval contained the true value $\mu = 0.3$ or $\sigma = 1.2$, respectively.

The R software (Version 4.3.0; [86]) was used for the entire analysis in this simulation study. We wrote an R function `linking_2groups_dich()` that allows for the computation of estimates and their standard errors of any user-defined linking method. The function and replication material for this Simulation Study 1 can be found at <https://osf.io/6bp3t> (accessed on 29 April 2024).

5.2. Results

It turned out that all five linking methods resulted in approximately unbiased estimates for the mean $\hat{\mu}$ and the standard deviation $\hat{\sigma}$. Table 1 displays the relative RMSE for the estimates as a function of the DIF effect distribution type, the DIF standard deviation τ , and the number of items I . The MM linking method performed similarly to the MGM method with respect to the mean μ (i.e., the MM had relative RMSE values close to 100; that is, those of the MGM) but was slightly less efficient with respect to the standard deviation σ , particularly for DIF effects that followed the t_3 distribution. The RMGM linking method had substantial efficiency losses for normally distributed DIF effects (i.e., it had RMSE values much larger than 100). On the other hand, the RMGM method provided large efficiency gains compared to the MGM method for the heavy-tailed t_3 distribution for μ and σ (i.e., it had RMSE values much smaller than 100). Note that the RHAE method outperformed the HAE method for DIF effects with a t_3 distribution when analyzing the RMSE for the estimated mean $\hat{\mu}$. However, RHAE (or HAE) should not be preferred over RMGM (or MGM) with respect to the RMSE of the estimated standard deviation $\hat{\sigma}$, because it had RMSE values much larger than 100.

Table 2 reports the coverage rate for $\hat{\mu}$ and $\hat{\sigma}$ for the five linking methods. Coverage rates that are within the interval [92.5,97.5] indicate acceptable performance. Overall, the coverage rates based on M-estimation theory performed satisfactorily for at least 20 items, except for the RMGM linking method, which resulted in undercoverage for $\hat{\mu}$ (i.e., it had coverage rates much lower than 92.5). In this case, the estimated linking errors were, on average, too small compared to the standard deviation of the $\hat{\mu}$ estimates across replications in this simulation study. As expected, the coverage rates improved with an increasing number of items I .

Table 1. Simulation Study 1: Relative RMSE for the estimated mean $\hat{\mu}$ and for the estimated standard deviation $\hat{\sigma}$ as a function of different DIF distributions, DIF standard deviation τ , and number of items I for an infinite sample size.

Dist	τ	I	$\hat{\mu}$					$\hat{\sigma}$				
			MM	MGM	RMGM	HAE	RHAE	MM	MGM	RMGM	HAE	RHAE
norm	0.25	10	100.0	100 ‡	125.8	104.6	105.2	102.8	100 ‡	104.7	189.3	175.5
		20	100.0	100 ‡	126.0	104.4	105.7	102.9	100 ‡	104.4	189.0	175.6
		40	100.0	100 ‡	125.5	103.9	105.4	102.6	100 ‡	104.9	185.0	172.0
	0.5	10	100.0	100 ‡	141.1	104.5	110.8	102.0	100 ‡	114.3	186.6	166.9
		20	100.1	100 ‡	141.4	104.3	109.9	102.9	100 ‡	113.7	186.6	168.2
		40	100.0	100 ‡	143.5	104.5	110.9	102.8	100 ‡	114.8	187.6	166.1
t_3	0.25	10	100.1	100 ‡	76.5	99.3	79.9	110.6	100 ‡	69.8	163.9	129.9
		20	100.1	100 ‡	74.9	96.7	78.4	102.8	100 ‡	74.3	175.2	137.6
		40	100.2	100 ‡	75.4	97.5	78.9	108.8	100 ‡	73.3	170.7	132.9
	0.5	10	100.1	100 ‡	86.5	94.3	77.4	105.1	100 ‡	74.8	169.5	127.9
		20	100.2	100 ‡	82.5	93.3	76.1	114.8	100 ‡	71.9	167.4	125.2
		40	100.1	100 ‡	74.6	87.0	69.7	114.3	100 ‡	70.9	161.7	123.5

Note. Dist = type of DIF distribution; norm = normal distribution; t_3 = t distribution with three degrees of freedom; MM = mean-mean linking; MGM = mean-geometric-mean linking; RMGM = robust mean-geometric-mean linking; HAE = symmetric Haebara linking; RHAE = robust symmetric Haebara linking; ‡ = MGM was the reference method for computing the relative RMSE. Relative RMSE values larger than 110.0 are printed in bold font.

Table 2. Simulation Study 1: Coverage rate at 95% confidence level for the estimated mean $\hat{\mu}$ and for the estimated standard deviation $\hat{\sigma}$ as a function of different DIF distributions, DIF standard deviation τ , and number of items I for an infinite sample size.

Dist	τ	I	$\hat{\mu}$					$\hat{\sigma}$				
			MM	MGM	RMGM	HAE	RHAE	MM	MGM	RMGM	HAE	RHAE
norm	0.25	10	91.8	91.7	84.3	89.8	91.1	92.2	93.5	93.5	88.4	90.3
		20	93.4	93.3	88.3	92.9	93.5	93.6	94.4	94.2	92.1	93.1
		40	94.0	94.0	91.0	93.9	94.2	95.5	96.0	95.3	93.7	94.2
	0.5	10	91.2	91.2	68.9	89.9	89.7	92.7	93.7	90.9	89.3	91.2
		20	92.9	92.8	78.0	92.5	92.9	93.6	94.9	93.4	92.2	93.2
		40	93.9	93.8	84.5	93.9	93.9	94.7	95.6	94.1	93.6	94.5
t_3	0.25	10	92.2	92.1	89.8	91.2	92.6	93.7	95.0	94.6	91.0	92.2
		20	93.7	93.7	92.2	93.5	94.1	95.4	95.7	95.5	92.6	93.1
		40	94.0	93.8	93.3	93.8	93.9	95.9	96.6	95.8	94.4	94.6
	0.5	10	93.2	92.9	80.9	92.5	92.8	94.2	95.1	93.8	90.1	92.0
		20	93.8	93.7	86.8	93.8	94.2	95.4	96.2	95.0	93.4	93.8
		40	94.1	94.0	90.2	93.7	94.2	95.8	96.4	95.6	93.6	93.9

Note. Dist = type of DIF distribution; norm = normal distribution; t_3 = t distribution with three degrees of freedom; MM = mean-mean linking; MGM = mean-geometric-mean linking; RMGM = robust mean-geometric-mean linking; HAE = symmetric Haebara linking; RHAE = robust symmetric Haebara linking; Coverage rates smaller than 92.5 and larger than 97.5 are printed in bold font.

6. Simulation Study 2: Assessing Total Error in Finite Sample Sizes

In the second Simulation Study 2, we investigated the statistical performance of the linking estimation methods in finite samples. In this case, there was uncertainty regarding the sampling of persons, thus resulting in standard error results and randomness in the DIF effects, which resulted in linking error results. Both sources of errors can be summarized as the total error.

6.1. Method

The item responses were simulated according to the 2PL model for a test with $I = 10$, 20, or 40 items. The same item parameters as in Simulation Study 1 (see Section 5.1) were

used. The factor variable θ was assumed to be normally distributed in both groups. Like in Simulation Study 1, the mean and the standard deviation for the factor variable θ in the first group were fixed at 0 and 1, respectively. The variable θ in the second group had a mean $\mu = 0.3$ and a standard deviation $\sigma = 1.2$. We used the normal distribution and the scaled t_3 distribution for DIF effects and varied the DIF standard deviations τ as 0.25 and 0.5. Moreover, we simulated a condition of no DIF effects (i.e., $\tau = 0$). The sample sizes $N = 500, 1000$, and 2000 were chosen in order to mimic sample sizes that are typically in the applications of the 2PL model [8].

In contrast to Simulation Study 1, the item parameters of the 2PL model were separately estimated for the two groups in the first step using MML estimation. In the second step, the performance of the five linking methods MM, MGM, RMGM, HAE, and RHAE was studied. The estimated mean $\hat{\mu}$ and the estimated standard deviation $\hat{\sigma}$ for the five methods were compared regarding the bias, RMSE, and relative RMSE. As in Simulation Study 1, MGM linking was used as the reference method for computing the RMSE.

In total, 5000 replications were conducted in each of the 5 (type of distribution for DIF effects combined with DIF standard deviation σ) $\times$ 3 (number of items I) $\times$ 3 (sample size N) = 45 cells of the simulation.

In the analysis, we computed the median of the linking error estimate LE based on (24) and the median of the bias-corrected linking error estimate LE_{bc} based on (36). Moreover, we compared the coverage rates for the estimates $\hat{\mu}$ and $\hat{\sigma}$ based on the standard error SE, as well as the (uncorrected) total error TE, based on (31), and the bias-corrected total error TE_{bc} was based on (37).

The R software [86] was used for the entire analysis in this simulation study. The 2PL model was estimated with the `sirt::xxirt()` function in the R package `sirt` [87]. As in Simulation Study 1, the R function `linking_2groups_dich()` was used for computing the estimates $\hat{\mu}$ and $\hat{\sigma}$ and their standard errors for the five linking methods. Replication material for this Simulation Study 2 can be found at <https://osf.io/6bp3t> (accessed on 29 April 2024).

6.2. Results

All five linking methods were approximately unbiased in all conditions of the simulation study. Table 3 presents the relative RMSE as a function of the different DIF distribution types, the DIF effect standard deviation τ , the number of items I , and the sample size N . As expected from the literature, the HAE method was the most efficient linking method in the condition of no DIF (i.e., $\tau = 0$). Across all conditions, the MM method had comparable performance to the MGM method regarding $\hat{\mu}$, but it was slightly more efficient for $\hat{\sigma}$. Efficiency gains of the RMGM method were only realized for the heavy-tailed t_3 distribution in large sample sizes. In these situations, the RHAE method outperformed the RMGM method for the estimated mean but not for the estimated standard deviation.

Table 4 presents the median of the estimated linking error LE and bias-corrected linking error LE_{bc} for the estimated mean $\hat{\mu}$. The results for the DIF effects with a t_3 distribution and a standard deviation of $\tau = 0.5$ were omitted for space reasons. In Table 4, we have also reported the estimated linking error for an infinite sample size (i.e., $N = \text{Inf}$) that was obtained from Simulation Study 1. It can be seen that the estimated linking errors LE and LE_{bc} (almost always) converged to the LE for an infinite sample size with an increasing sample size.

It turned out that the estimated linking error LE was positively biased, while the bias-corrected linking error LE_{bc} was negatively biased (to a lesser extent). In particular, the median estimated linking error of 0.061 for the MM method for 10 items and a small sample size $N = 500$ was substantially larger than the true value of 0 in the condition of no DIF (i.e., $\tau = 0$). On the other hand, the median of the estimated bias-corrected linking error LE_{bc} was 0 in all situations in which no DIF was simulated in the item parameters. Overall, one could conclude that the bias in both linking error types can be reduced with an increasing sample size and an increasing number of items. For all linking methods except

for the RMGM method, the uncorrected linking error LE had worse performance compared to the bias-corrected linking error LE_{bc} . Hence, LE_{bc} could be the preferred choice for a reported linking error.

Table 3. Simulation Study 2: Relative RMSE for the estimated mean $\hat{\mu}$ and for the estimated standard deviation $\hat{\sigma}$ as a function of different DIF distributions, DIF standard deviation τ , number of items I , and sample size N .

Dist	τ	I	N	$\hat{\mu}$					$\hat{\sigma}$				
				MM	MGM	RMGM	HAE	RHAE	MM	MGM	RMGM	HAE	RHAE
-	0	10	500	99.6	100 [‡]	100.7	92.0	89.9	96.9	100 [‡]	118.5	96.2	96.5
			1000	99.9	100 [‡]	100.1	92.6	91.0	98.2	100 [‡]	112.8	95.8	96.0
			2000	99.9	100 [‡]	99.4	94.0	92.9	99.1	100 [‡]	107.9	95.7	95.2
		20	500	100.1	100 [‡]	100.4	94.2	93.6	97.7	100 [‡]	108.7	97.7	97.7
			1000	100.2	100 [‡]	99.5	95.1	94.3	98.3	100 [‡]	105.4	98.1	97.8
			2000	100.1	100 [‡]	98.8	95.8	95.3	98.4	100 [‡]	103.7	98.1	97.9
		40	500	100.0	100 [‡]	99.2	96.3	95.8	98.4	100 [‡]	102.7	99.2	98.6
			1000	100.1	100 [‡]	100.1	97.5	97.4	98.8	100 [‡]	101.7	99.4	99.3
			2000	100.1	100 [‡]	99.8	97.9	97.7	99.4	100 [‡]	100.9	99.3	99.0
norm	0.25	10	500	99.8	100 [‡]	118.0	95.9	97.0	97.6	100 [‡]	118.5	106.4	107.5
			1000	99.9	100 [‡]	122.4	99.3	100.6	99.1	100 [‡]	114.4	118.1	116.0
			2000	100.0	100 [‡]	123.6	100.6	102.0	99.2	100 [‡]	110.8	132.1	128.0
		20	500	100.0	100 [‡]	112.3	96.5	97.0	98.7	100 [‡]	110.6	104.3	104.3
			1000	100.0	100 [‡]	116.5	99.2	100.0	98.5	100 [‡]	107.8	115.0	112.6
			2000	100.0	100 [‡]	119.5	100.7	101.1	99.1	100 [‡]	105.9	124.4	120.7
		40	500	100.0	100 [‡]	107.6	97.7	98.1	98.4	100 [‡]	105.0	104.3	103.5
			1000	100.1	100 [‡]	111.5	99.4	100.0	99.4	100 [‡]	105.4	110.4	109.0
			2000	100.0	100 [‡]	116.3	101.4	102.3	99.3	100 [‡]	103.6	116.9	114.6
norm	0.5	10	500	99.8	100 [‡]	130.9	99.6	104.7	98.0	100 [‡]	122.3	129.4	127.0
			1000	99.9	100 [‡]	133.6	101.8	107.4	100.1	100 [‡]	118.8	143.2	134.0
			2000	99.9	100 [‡]	135.6	102.5	109.0	99.8	100 [‡]	117.6	156.5	144.8
		20	500	99.9	100 [‡]	130.5	100.5	105.0	98.2	100 [‡]	115.1	125.6	122.2
			1000	100.0	100 [‡]	138.1	101.6	108.5	99.8	100 [‡]	114.6	139.0	132.5
			2000	100.0	100 [‡]	138.5	102.5	108.6	100.5	100 [‡]	113.5	155.1	144.7
		40	500	99.9	100 [‡]	126.5	100.4	104.6	98.7	100 [‡]	108.2	115.9	113.3
			1000	100.0	100 [‡]	131.8	102.2	106.5	99.4	100 [‡]	108.5	129.4	125.1
			2000	100.1	100 [‡]	139.8	103.3	108.9	100.9	100 [‡]	111.6	140.9	132.1
t_3	0.25	10	500	99.8	100 [‡]	99.2	94.6	87.0	97.8	100 [‡]	116.6	106.7	103.9
			1000	99.8	100 [‡]	94.5	94.1	84.6	98.8	100 [‡]	112.8	113.9	106.4
			2000	100.0	100 [‡]	90.4	97.8	85.2	100.8	100 [‡]	102.8	125.8	113.3
		20	500	99.9	100 [‡]	100.7	95.4	90.9	98.1	100 [‡]	108.4	106.0	102.7
			1000	100.0	100 [‡]	94.9	95.0	87.7	98.8	100 [‡]	105.8	110.6	104.2
			2000	100.1	100 [‡]	89.1	97.9	86.8	100.1	100 [‡]	100.8	120.8	109.0
		40	500	100.0	100 [‡]	99.2	97.1	94.1	98.5	100 [‡]	103.9	103.7	101.5
			1000	100.0	100 [‡]	96.6	97.4	92.1	99.1	100 [‡]	102.0	108.2	103.6
			2000	100.1	100 [‡]	91.7	97.5	88.7	99.8	100 [‡]	100.1	115.9	107.3
t_3	0.5	10	500	99.8	100 [‡]	93.4	91.9	81.0	99.5	100 [‡]	115.6	122.6	112.4
			1000	99.9	100 [‡]	92.5	92.2	80.4	100.9	100 [‡]	107.2	134.9	116.5
			2000	100.0	100 [‡]	91.3	94.5	80.2	101.6	100 [‡]	99.3	147.3	121.9
		20	500	100.0	100 [‡]	97.0	93.4	84.1	98.7	100 [‡]	107.8	116.0	108.5
			1000	100.0	100 [‡]	92.7	93.6	82.4	101.0	100 [‡]	101.9	127.6	112.7
			2000	100.0	100 [‡]	90.7	93.9	80.0	102.4	100 [‡]	93.1	135.5	112.8
		40	500	100.0	100 [‡]	97.5	95.5	88.7	98.8	100 [‡]	104.5	112.2	106.4
			1000	100.0	100 [‡]	92.0	94.7	83.9	100.3	100 [‡]	100.3	121.0	109.6
			2000	100.0	100 [‡]	87.1	93.3	80.0	101.9	100 [‡]	96.0	132.0	111.6

Note. Dist = type of DIF distribution; norm = normal distribution; $t_3 = t$ distribution with three degrees of freedom; MM = mean-mean linking; MGM = mean-geometric-mean linking; RMGM = robust mean-geometric-mean linking; HAE = symmetric Haebara linking; RHAE = robust symmetric Haebara linking; [‡] = MGM was the reference method for computing the relative RMSE. Relative RMSE values larger than 110.0 are printed in bold font.

Table 4. Simulation Study 2: Median of the estimated linking error LE and the estimate bias-corrected linking error LE_{bc} for the estimated mean $\hat{\mu}$ as a function of different DIF distributions, DIF standard deviation τ , number of items I , and sample size N .

Dist	τ	I	N	MM		MGM		RMGM		HAE		RHAЕ	
				LE	LE _{bc}	LE	LE _{bc}	LE	LE _{bc}	LE	LE _{bc}	LE	LE _{bc}
-	0	10	500	0.061	0.000	0.061	0.000	0.066	0.000	0.051	0.000	0.052	0.000
			1000	0.042	0.000	0.042	0.000	0.045	0.000	0.036	0.000	0.037	0.000
			2000	0.030	0.000	0.030	0.000	0.032	0.000	0.025	0.000	0.025	0.000
			Inf	0.000		0.000		0.000		0.000		0.000	
		20	500	0.043	0.000	0.043	0.000	0.045	0.000	0.036	0.000	0.036	0.000
			1000	0.030	0.000	0.030	0.000	0.031	0.000	0.026	0.000	0.026	0.000
			2000	0.021	0.000	0.021	0.000	0.022	0.000	0.018	0.000	0.018	0.000
			Inf	0.000		0.000		0.000		0.000		0.000	
		40	500	0.030	0.000	0.030	0.000	0.032	0.000	0.025	0.000	0.025	0.000
			1000	0.021	0.000	0.021	0.000	0.022	0.000	0.018	0.000	0.018	0.000
			2000	0.015	0.000	0.015	0.000	0.015	0.000	0.013	0.000	0.013	0.000
			Inf	0.000		0.000		0.000		0.000		0.000	
norm	0.25	10	500	0.102	0.068	0.102	0.068	0.100	0.000	0.094	0.066	0.102	0.066
			1000	0.091	0.074	0.091	0.074	0.090	0.000	0.085	0.071	0.091	0.075
			2000	0.086	0.077	0.086	0.077	0.086	0.039	0.081	0.074	0.088	0.079
			Inf	0.076		0.075		0.077		0.075		0.080	
		20	500	0.072	0.050	0.071	0.050	0.078	0.000	0.068	0.052	0.070	0.051
			1000	0.064	0.053	0.064	0.053	0.071	0.000	0.062	0.054	0.065	0.056
			2000	0.060	0.055	0.060	0.055	0.066	0.040	0.060	0.056	0.062	0.057
			Inf	0.054		0.054		0.061		0.056		0.058	
		40	500	0.050	0.036	0.050	0.036	0.059	0.000	0.048	0.038	0.050	0.037
			1000	0.045	0.038	0.045	0.038	0.054	0.000	0.045	0.039	0.046	0.040
			2000	0.042	0.039	0.042	0.039	0.050	0.033	0.043	0.040	0.044	0.041
			Inf	0.039		0.039		0.046		0.041		0.041	
norm	0.5	10	500	0.174	0.155	0.174	0.154	0.129	0.000	0.166	0.152	0.181	0.156
			1000	0.167	0.157	0.167	0.157	0.122	0.000	0.161	0.153	0.175	0.162
			2000	0.163	0.159	0.163	0.159	0.119	0.065	0.160	0.156	0.171	0.165
			Inf	0.151		0.150		0.116		0.151		0.162	
		20	500	0.122	0.110	0.122	0.110	0.116	0.000	0.121	0.112	0.129	0.115
			1000	0.117	0.112	0.117	0.111	0.111	0.021	0.118	0.114	0.127	0.120
			2000	0.115	0.112	0.114	0.112	0.107	0.074	0.116	0.114	0.124	0.121
			Inf	0.109		0.109		0.105		0.113		0.119	
		40	500	0.085	0.078	0.085	0.077	0.096	0.000	0.086	0.081	0.092	0.083
			1000	0.082	0.078	0.082	0.078	0.093	0.038	0.084	0.082	0.090	0.085
			2000	0.080	0.078	0.080	0.078	0.091	0.069	0.083	0.082	0.088	0.086
			Inf	0.077		0.077		0.087		0.081		0.086	
t_3	0.25	10	500	0.092	0.051	0.092	0.051	0.091	0.000	0.083	0.048	0.086	0.041
			1000	0.079	0.058	0.078	0.057	0.077	0.000	0.073	0.055	0.074	0.054
			2000	0.072	0.061	0.071	0.061	0.067	0.018	0.067	0.058	0.068	0.059
			Inf	0.061		0.061		0.054		0.060		0.059	
		20	500	0.066	0.041	0.065	0.041	0.068	0.000	0.061	0.042	0.059	0.035
			1000	0.057	0.045	0.057	0.045	0.055	0.000	0.054	0.045	0.052	0.040
			2000	0.053	0.047	0.053	0.047	0.049	0.021	0.052	0.047	0.048	0.042
			Inf	0.046		0.046		0.039		0.046		0.042	
		40	500	0.047	0.031	0.047	0.031	0.048	0.000	0.044	0.033	0.042	0.026
			1000	0.041	0.033	0.041	0.033	0.040	0.000	0.040	0.034	0.037	0.029
			2000	0.038	0.034	0.038	0.034	0.035	0.017	0.038	0.035	0.034	0.030
			Inf	0.034		0.034		0.028		0.035		0.030	

Note. Dist = type of DIF distribution; norm = normal distribution; t_3 = t distribution with three degrees of freedom; Inf = infinite sample size; MM = mean-mean linking; MGM = mean-geometric-mean linking; RMGM = robust mean-geometric-mean linking; HAE = symmetric Haebara linking; RHAЕ = robust symmetric Haebara linking; Linking error estimates with an absolute bias larger than 0.01 are printed in bold font.

Table 5 reports the median values of the estimated linking error LE and bias-corrected linking error LE_{bc} for the estimated standard deviation $\hat{\sigma}$. Overall, we observed a similar pattern of findings as in the case of the estimated mean $\hat{\mu}$. Again, the bias-corrected linking error estimates for the RMGM method were unsatisfactory. The bias in the uncorrected linking error estimates was slightly larger for $\hat{\sigma}$ than for $\hat{\mu}$. A simple idea might be to use

the mean of both of the linking error estimates LE and LE_{bc} as another linking error to improve the performance of the linking error estimate.

Table 5. Simulation Study 2: Median of the estimated linking error LE and the estimate bias-corrected linking error LE_{bc} for the estimated standard deviation $\hat{\sigma}$ as a function of different DIF distributions, DIF standard deviation τ , number of items I , and sample size N .

Dist	τ	I	N	MM		MGM		RMGM		HAE		RHAЕ	
				LE	LE_{bc}	LE	LE_{bc}	LE	LE_{bc}	LE	LE_{bc}	LE	LE_{bc}
-	0	10	500	0.086	0.000	0.087	0.000	0.093	0.000	0.074	0.000	0.079	0.000
			1000	0.061	0.000	0.061	0.000	0.068	0.000	0.052	0.000	0.055	0.000
			2000	0.042	0.000	0.043	0.000	0.047	0.000	0.036	0.000	0.038	0.000
			Inf	0.000		0.000		0.000		0.000		0.000	
		20	500	0.052	0.000	0.053	0.000	0.060	0.000	0.048	0.000	0.050	0.000
			1000	0.037	0.000	0.037	0.000	0.042	0.000	0.034	0.000	0.034	0.000
			2000	0.026	0.000	0.026	0.000	0.028	0.000	0.024	0.000	0.024	0.000
			Inf	0.000		0.000		0.000		0.000		0.000	
		40	500	0.034	0.000	0.035	0.000	0.040	0.000	0.032	0.000	0.033	0.000
			1000	0.024	0.000	0.025	0.000	0.027	0.000	0.023	0.000	0.023	0.000
			2000	0.017	0.000	0.017	0.000	0.019	0.000	0.016	0.000	0.016	0.000
			Inf	0.000		0.000		0.000		0.000		0.000	
norm	0.25	10	500	0.092	0.006	0.093	0.009	0.102	0.000	0.091	0.017	0.098	0.016
			1000	0.066	0.007	0.067	0.010	0.076	0.000	0.071	0.033	0.077	0.033
			2000	0.051	0.019	0.052	0.021	0.059	0.006	0.061	0.042	0.065	0.044
			Inf	0.029		0.029		0.032		0.046		0.048	
		20	500	0.056	0.008	0.057	0.009	0.066	0.000	0.060	0.026	0.063	0.015
			1000	0.042	0.015	0.043	0.014	0.049	0.000	0.050	0.032	0.051	0.029
			2000	0.033	0.018	0.034	0.019	0.038	0.005	0.044	0.034	0.044	0.033
			Inf	0.021		0.021		0.022		0.035		0.034	
		40	500	0.037	0.009	0.038	0.009	0.045	0.000	0.042	0.022	0.042	0.013
			1000	0.028	0.013	0.029	0.013	0.033	0.000	0.035	0.024	0.035	0.021
			2000	0.023	0.014	0.023	0.014	0.025	0.004	0.031	0.025	0.030	0.023
			Inf	0.015		0.015		0.015		0.026		0.024	
norm	0.5	10	500	0.104	0.037	0.106	0.041	0.113	0.000	0.125	0.087	0.135	0.072
			1000	0.085	0.051	0.086	0.052	0.098	0.000	0.112	0.092	0.118	0.089
			2000	0.073	0.055	0.074	0.057	0.085	0.019	0.104	0.094	0.107	0.093
			Inf	0.057		0.058		0.065		0.091		0.091	
		20	500	0.068	0.038	0.069	0.038	0.081	0.000	0.089	0.070	0.090	0.058
			1000	0.056	0.039	0.057	0.040	0.066	0.003	0.081	0.071	0.080	0.064
			2000	0.049	0.041	0.050	0.041	0.058	0.027	0.077	0.072	0.074	0.066
			Inf	0.041		0.042		0.047		0.070		0.067	
		40	500	0.045	0.028	0.047	0.028	0.056	0.000	0.062	0.050	0.061	0.041
			1000	0.038	0.029	0.039	0.029	0.046	0.004	0.057	0.051	0.055	0.045
			2000	0.034	0.029	0.035	0.030	0.040	0.023	0.054	0.051	0.052	0.047
			Inf	0.030		0.030		0.033		0.051		0.047	
t_3	0.25	10	500	0.091	0.005	0.092	0.007	0.098	0.000	0.086	0.011	0.092	0.008
			1000	0.066	0.007	0.067	0.010	0.075	0.000	0.066	0.020	0.070	0.018
			2000	0.050	0.013	0.050	0.015	0.056	0.003	0.054	0.031	0.057	0.031
			Inf	0.024		0.024		0.022		0.038		0.037	
		20	500	0.056	0.007	0.057	0.009	0.064	0.000	0.058	0.020	0.059	0.009
			1000	0.041	0.013	0.042	0.013	0.047	0.000	0.046	0.026	0.046	0.019
			2000	0.032	0.016	0.032	0.016	0.034	0.002	0.039	0.029	0.038	0.025
			Inf	0.018		0.018		0.015		0.029		0.026	
		40	500	0.037	0.008	0.038	0.008	0.043	0.000	0.040	0.018	0.039	0.007
			1000	0.028	0.011	0.028	0.011	0.031	0.000	0.032	0.020	0.031	0.015
			2000	0.022	0.012	0.022	0.012	0.023	0.002	0.028	0.022	0.026	0.018
			Inf	0.013		0.013		0.011		0.022		0.019	

Note. Dist = type of DIF distribution; norm = normal distribution; t_3 = t distribution with three degrees of freedom; Inf = infinite sample size; MM = mean-mean linking; MGM = mean-geometric-mean linking; RMGM = robust mean-geometric-mean linking; HAE = symmetric Haebara linking; RHAЕ = robust symmetric Haebara linking; Linking error estimates with an absolute bias larger than 0.01 are printed in bold font.

In Table 6, the coverage rate for the estimated mean $\hat{\mu}$ is displayed. In the no DIF condition $\tau = 0$, the coverage rates based on the standard error SE performed satisfactorily. In the presence of DIF, the uncertainty in the estimated means of $\hat{\mu}$ was underestimated when the standard error was used when computing confidence intervals, thus resulting in substantial undercoverage. The confidence intervals based on the total error TE tended to have slightly increased coverage rates. In such situations, the coverage rates for the confidence intervals based on the bias-corrected linking error TE_{bc} were slightly better.

Generally, the RMGM linking method did not have adequate coverage rates in many situations.

Table 6. Simulation Study 2: Coverage rate at 95% confidence level based on the standard error SE, total error TE, and bias-corrected total error TE_{bc} for the estimated mean $\hat{\mu}$ as a function of different DIF distributions, DIF standard deviation τ , number of items I , and sample size N .

Dist	τ	I	N	SE					TE					TE _{bc}				
				MM	MGM	RMGM	HAE	RHAE	MM	MGM	RMGM	HAE	RHAE	MM	MGM	RMGM	HAE	RHAE
-	0	10	500	94.6	94.6	98.7	94.9	95.2	97.7	97.6	99.4	97.4	97.9	95.0	95.0	98.8	95.1	95.4
			1000	94.4	94.5	97.7	94.9	95.0	97.9	97.9	99.1	97.7	97.8	94.9	94.9	97.7	95.1	95.2
			2000	94.6	94.5	97.0	94.9	95.0	98.0	98.0	98.7	97.7	97.8	95.0	95.0	97.2	95.1	95.3
		20	500	94.5	94.3	98.2	94.6	94.9	97.0	97.0	99.1	96.9	97.1	94.6	94.5	98.2	94.7	95.0
			1000	94.5	94.5	97.3	94.9	95.2	96.8	96.7	98.5	96.7	97.0	94.7	94.7	97.3	95.0	95.2
			2000	94.5	94.5	96.7	95.0	95.0	97.2	97.2	98.4	96.9	97.1	94.8	94.7	96.7	95.1	95.0
		40	500	94.6	94.6	98.0	95.3	95.6	96.7	96.6	98.7	96.5	96.8	94.8	94.7	98.0	95.3	95.6
			1000	95.3	95.2	96.9	95.8	95.7	96.7	96.6	98.0	96.6	96.6	95.4	95.3	96.9	95.8	95.7
			2000	94.6	94.6	96.4	95.3	95.3	96.3	96.1	97.3	96.4	96.4	94.7	94.8	96.4	95.3	95.4
norm	0.25	10	500	86.0	86.0	95.3	84.4	86.3	96.3	96.4	97.5	95.8	96.4	92.0	92.0	95.5	91.7	92.9
			1000	78.7	78.7	88.5	76.7	77.3	95.4	95.4	94.8	94.6	95.4	92.0	92.0	90.1	91.5	92.2
			2000	66.1	65.9	78.6	64.0	64.7	94.5	94.4	91.7	93.0	93.6	92.0	92.0	84.4	90.6	91.3
		20	500	88.3	88.0	96.4	87.9	89.1	95.7	95.6	98.3	95.9	95.9	93.4	93.3	96.4	93.8	93.9
			1000	82.3	82.2	91.8	81.8	82.5	95.6	95.5	96.8	95.3	95.8	93.0	92.8	92.8	93.4	93.7
			2000	75.1	75.1	83.6	72.7	74.1	95.3	95.4	95.6	95.0	95.5	93.7	93.8	89.9	93.5	94.1
		40	500	90.7	90.6	96.7	91.0	91.6	95.1	95.0	98.3	95.1	95.5	93.3	93.2	96.8	93.8	94.2
			1000	87.4	87.4	93.3	87.3	88.0	95.4	95.3	97.2	95.4	95.5	93.8	93.8	94.1	94.3	94.2
			2000	81.4	80.9	86.1	80.5	80.5	95.3	95.2	96.0	95.4	95.4	94.6	94.4	92.3	94.7	94.7
norm	0.5	10	500	66.4	66.4	85.0	63.6	67.4	93.9	93.7	89.9	92.5	92.9	91.0	91.0	85.7	90.2	89.6
			1000	54.2	54.2	73.5	50.5	55.0	93.0	92.9	84.9	91.4	92.4	91.4	91.2	77.6	89.7	90.0
			2000	40.3	40.5	58.3	38.2	41.2	92.2	92.2	78.8	90.9	90.6	91.2	91.2	69.9	90.1	89.3
		20	500	74.6	74.6	89.7	72.9	74.0	95.1	95.0	94.4	95.1	95.1	93.6	93.5	90.2	93.8	93.0
			1000	64.0	63.6	78.2	61.8	62.3	94.6	94.5	90.0	94.5	94.0	93.3	93.3	82.1	93.7	92.8
			2000	48.8	48.8	64.5	47.2	48.9	94.5	94.5	84.7	93.3	92.9	93.9	93.8	78.0	92.9	92.1
		40	500	82.3	82.2	92.8	81.7	82.5	95.5	95.4	96.4	95.7	95.5	94.4	94.4	93.1	94.8	94.5
			1000	72.9	72.5	84.4	71.7	72.5	95.0	94.9	94.2	95.0	95.1	94.5	94.3	87.8	94.5	94.6
			2000	60.7	60.4	71.1	58.7	58.6	94.8	94.7	90.9	94.7	94.6	94.3	94.3	85.6	94.4	94.0
t_3	0.25	10	500	87.0	86.9	97.3	85.8	90.0	97.1	97.3	98.6	96.7	97.6	93.1	93.0	97.4	92.6	93.8
			1000	80.7	80.6	94.1	78.9	84.0	95.9	95.9	97.3	95.2	96.2	92.4	92.3	94.4	91.8	92.5
			2000	70.7	70.5	88.3	69.3	74.5	96.0	95.9	96.2	94.7	95.5	92.5	92.4	91.8	91.6	92.9
		20	500	89.6	89.2	97.6	89.2	91.4	96.3	96.2	98.7	96.1	96.6	93.4	93.4	97.6	93.6	93.9
			1000	84.2	84.1	94.9	84.3	87.9	95.9	95.9	98.0	95.6	96.3	93.7	93.5	95.3	93.7	94.0
			2000	75.0	75.1	90.0	74.5	80.3	95.1	95.0	96.9	94.8	95.1	93.0	92.9	92.9	93.1	93.3
		40	500	91.5	91.5	97.5	92.4	93.4	96.2	96.0	98.9	96.0	96.4	94.1	94.0	97.5	95.0	94.9
			1000	87.8	87.6	95.4	88.5	90.6	96.1	96.1	98.0	96.1	96.2	94.1	94.0	95.7	94.7	94.7
			2000	82.5	82.2	92.5	83.5	87.1	95.6	95.6	97.0	95.6	95.9	94.2	94.2	94.5	94.7	94.5
t_3	0.5	10	500	70.7	70.8	92.3	69.6	78.5	95.2	95.1	95.8	94.2	96.0	92.1	91.9	92.7	91.4	91.9
			1000	59.8	59.9	84.9	58.4	66.9	94.4	94.3	91.8	93.0	94.7	92.1	92.0	86.7	90.9	91.9
			2000	48.1	48.1	73.3	46.1	55.5	94.4	94.2	89.7	93.1	94.0	93.0	93.0	81.8	91.8	92.4
		20	500	77.7	77.8	95.5	77.9	84.9	95.7	95.7	97.8	95.9	96.8	94.2	94.2	95.6	94.3	94.8
			1000	65.7	65.5	89.0	65.6	73.5	95.3	95.1	95.8	94.6	95.6	93.8	93.7	90.7	93.7	93.7
			2000	53.0	52.8	78.7	52.1	62.5	94.8	94.7	93.0	94.5	95.0	93.9	93.9	86.7	93.9	94.1
		40	500	82.7	82.3	96.4	83.9	88.4	96.1	96.0	98.4	95.8	96.1	94.7	94.6	96.6	95.1	94.8
			1000	73.6	73.5	91.3	74.8	81.2	95.0	95.0	96.8	95.3	95.9	94.1	94.1	93.1	94.5	94.8
			2000	62.3	61.9	84.2	63.9	71.3	95.3	95.0	95.3	95.0	95.8	94.5	94.4	90.9	94.4	94.7

Note. Dist = type of DIF distribution; norm = normal distribution; t_3 = t distribution with three degrees of freedom; MM = mean-mean linking; MGM = mean-geometric-mean linking; RMGM = robust mean-geometric-mean linking; HAE = symmetric Haebara linking; RHAE = robust symmetric Haebara linking; Coverage rates smaller than 92.5 and larger than 97.5 are printed in bold font.

Finally, Table 7 reports the coverage rates for the estimated standard deviation $\hat{\sigma}$. The bias-corrected total error outperformed the uncorrected total error regarding coverage rates. In many situations, the coverage rates based on the total error were too high. However, the RMGM method had substantial overcoverage in many conditions for confidence intervals based on TE and TE_{bc}, particularly for fewer items or smaller sample sizes.

Table 7. Simulation Study 2: Coverage rate at 95% confidence level based on the standard error SE, total error TE, and bias-corrected total error TE_{bc} for the estimated standard deviation $\hat{\sigma}$ as a function of different DIF distributions, DIF standard deviation τ , number of items I , and sample size N .

Dist	τ	I	N	SE					TE					TE _{bc}				
				MM	MGM	RMGM	HAE	RHAE	MM	MGM	RMGM	HAE	RHAE	MM	MGM	RMGM	HAE	RHAE
-	0	10	500	95.1	95.3	99.2	94.8	95.9	98.9	98.9	99.5	98.4	98.8	96.1	96.1	99.2	95.3	96.3
			1000	94.8	95.1	99.0	95.3	96.0	99.1	99.0	99.7	98.8	99.0	95.9	95.9	99.0	96.0	96.4
			2000	94.7	94.8	98.0	95.2	95.3	99.0	99.0	99.5	99.0	99.1	95.6	95.6	98.1	95.6	95.8
		20	500	95.0	95.0	99.1	95.1	95.7	98.3	98.4	99.5	98.0	98.4	95.4	95.3	99.1	95.5	96.0
			1000	94.9	94.8	98.3	94.5	95.0	98.0	98.1	99.2	97.7	98.1	95.2	95.2	98.3	94.8	95.3
			2000	95.7	95.5	97.2	95.3	95.4	98.3	98.1	99.0	98.0	98.1	96.1	95.9	97.4	95.5	95.6
		40	500	94.4	94.8	98.3	94.6	95.2	97.2	97.4	98.8	96.8	97.2	94.5	95.0	98.3	94.8	95.4
			1000	95.4	95.4	97.7	95.1	95.2	97.7	97.5	98.9	97.1	97.3	95.6	95.5	97.7	95.2	95.3
			2000	95.6	95.7	97.2	95.7	95.7	97.7	97.7	98.5	97.7	97.6	95.8	95.9	97.2	95.8	95.9
norm	0.25	10	500	93.7	93.8	99.1	91.1	93.4	98.7	98.7	99.5	97.9	98.7	95.1	94.9	99.1	93.2	95.2
			1000	92.7	92.7	98.6	85.9	88.9	98.6	98.8	99.4	97.2	97.9	94.6	94.5	98.6	91.4	93.4
			2000	89.8	90.0	96.6	77.3	80.6	98.4	98.4	98.8	95.3	96.5	93.7	93.9	96.8	88.8	91.0
		20	500	93.1	93.2	98.9	91.2	93.3	98.0	98.1	99.4	97.3	97.8	94.1	94.5	98.9	93.7	94.5
			1000	93.0	93.2	97.8	88.1	90.3	97.9	97.8	99.1	97.0	97.7	94.8	94.5	97.8	92.7	94.2
			2000	91.0	91.2	95.4	81.4	84.2	97.3	97.5	98.8	95.9	96.9	94.0	94.2	96.0	93.0	93.7
		40	500	94.5	94.4	97.9	93.1	94.0	97.0	97.2	99.0	97.0	97.7	95.0	94.9	97.9	94.7	94.8
			1000	94.1	94.2	97.0	90.5	91.9	97.1	97.4	98.6	96.6	97.2	94.9	95.1	97.1	94.4	94.7
			2000	92.9	92.7	94.8	86.2	88.4	97.2	97.4	98.3	96.3	96.6	94.8	95.1	95.4	94.3	94.7
norm	0.5	10	500	89.8	89.9	98.3	78.2	86.4	98.1	97.8	99.1	95.4	97.1	93.2	93.5	98.3	90.1	92.5
			1000	84.0	84.8	96.1	66.5	77.4	97.5	97.7	98.4	94.0	96.0	91.5	92.2	96.3	88.9	91.2
			2000	76.4	76.5	91.5	54.8	64.2	96.3	96.5	96.9	91.8	94.4	91.3	92.1	92.6	88.7	90.3
		20	500	91.2	91.3	98.5	82.6	88.2	97.9	97.8	99.4	95.9	97.4	94.2	94.7	98.6	92.9	93.9
			1000	87.1	87.6	96.0	73.8	80.1	97.4	97.5	98.7	94.8	96.3	93.9	94.5	96.3	92.5	93.3
			2000	80.2	81.2	91.1	59.3	67.2	96.6	97.0	97.6	93.6	95.1	94.3	94.7	94.0	91.9	92.8
		40	500	92.2	92.3	97.5	86.5	90.1	96.6	96.9	98.8	96.1	97.2	94.3	94.5	97.6	94.1	94.5
			1000	90.0	90.7	95.9	79.2	84.3	97.0	97.2	98.5	96.2	96.8	94.7	95.0	96.3	94.4	94.4
			2000	84.2	85.4	90.8	68.7	75.1	96.4	96.7	98.1	95.6	96.3	94.6	95.0	94.4	94.4	95.1
t_3	0.25	10	500	93.9	94.1	99.3	91.7	94.4	98.9	99.0	99.6	98.0	98.7	95.3	95.5	99.3	94.2	95.5
			1000	92.5	92.7	98.7	87.6	91.2	98.6	98.5	99.3	97.4	98.1	94.6	94.6	98.7	92.2	94.0
			2000	89.9	90.4	97.4	80.7	86.0	98.2	98.3	99.1	96.3	97.3	93.9	94.3	97.6	90.8	92.8
		20	500	93.4	93.9	98.7	91.3	93.6	97.8	98.0	99.4	97.3	97.8	94.6	94.7	98.7	93.5	94.7
			1000	93.3	93.5	97.8	89.5	92.4	98.0	98.3	99.3	97.5	98.2	95.0	95.2	97.8	94.1	94.6
			2000	91.1	91.2	96.3	83.3	88.4	98.0	98.2	99.0	96.5	97.2	94.7	94.7	96.6	93.3	93.8
		40	500	94.1	94.2	98.2	93.2	94.4	96.9	97.0	99.1	97.0	97.5	94.6	94.8	98.2	94.7	95.0
			1000	93.4	93.6	97.0	90.6	92.2	97.1	97.0	98.7	96.7	97.2	94.6	94.8	97.0	94.1	94.1
			2000	92.3	92.5	95.1	87.4	90.8	97.0	97.3	98.1	95.9	96.5	94.4	94.8	95.4	93.7	94.4
t_3	0.5	10	500	91.1	91.3	98.8	82.1	90.4	98.4	98.6	99.2	96.8	98.0	94.6	94.6	98.8	91.8	93.6
			1000	86.1	86.6	97.5	71.2	82.6	97.7	97.8	98.8	95.6	97.0	93.3	93.8	97.5	89.9	91.8
			2000	79.7	80.0	94.6	61.1	73.4	97.0	97.2	98.2	93.7	95.1	92.1	92.3	94.9	90.0	91.2
		20	500	91.6	92.0	98.3	84.6	90.4	97.5	97.9	99.4	96.6	97.6	94.7	94.9	98.3	93.3	93.7
			1000	87.8	87.9	96.7	77.6	85.5	97.6	97.7	98.7	95.9	97.0	94.3	94.4	96.9	93.2	93.4
			2000	81.0	81.9	94.1	66.4	78.5	97.3	97.5	98.4	94.8	96.3	94.7	94.8	95.4	92.8	93.6
		40	500	91.8	92.6	98.1	88.2	92.0	97.1	97.4	99.1	96.3	96.8	94.5	94.8	98.1	94.0	94.2
			1000	90.0	90.3	95.8	82.7	88.4	96.9	97.0	98.4	95.9	96.5	94.8	94.9	96.0	93.9	94.4
			2000	86.8	87.6	93.8	73.9	82.3	96.9	97.2	98.3	95.6	96.4	95.5	95.8	95.2	94.4	94.9

Note. Dist = type of DIF distribution; norm = normal distribution; t_3 = t distribution with three degrees of freedom; MM = mean-mean linking; MGM = mean-geometric-mean linking; RMGM = robust mean-geometric-mean linking; HAE = symmetric Haebara linking; RHAE = robust symmetric Haebara linking; Coverage rates smaller than 92.5 and larger than 97.5 are printed in bold font.

7. Discussion

In this article, we simultaneously treated standard errors and linking errors for linking methods in the 2PL model. We proposed a bias-corrected linking error estimate, which delivers a bias-corrected total error estimate. This bias-corrected total error outperformed the usually employed total error that is given as the simple variances due to standard error and the usual uncorrected linking error. In a simulation study, it turned out that the confidence intervals for the linking parameters based on the bias-corrected total error outperformed those based on the usual total error regarding coverage rates. Moreover, the bias-corrected linking error estimate was less biased than the uncorrected linking error estimate.

As with any simulation study, our study had several limitations. First, our study only treated the 2PL model for dichotomous item responses. However, the performance of the linking estimators and their variance estimates could also be investigated for the simpler Rasch model for dichotomous item responses [16] or the generalized partial credit model for polytomous item responses [88]. Furthermore, the theory in this article could also be adapted to the chain linking of multiple groups [83,84]. In addition, the distribution types of the DIF effects in the simulation studies were restricted to the symmetric normal distribution and the t distributions with three degrees of freedom. Future research could focus on alternative and asymmetric distributions such as mixture, uniform, or discrete distributions. Moreover, the factor variable θ was assumed to be normally distributed in both simulation studies. The 2PL model could also be estimated with non-normal θ distributions [10,13], which could be investigated in future studies. Next, followup research could focus on linking with smaller sample sizes, such as $N = 250$, as well as the case of unbalanced group sizes. Furthermore, we only employed 10, 20, or 40 items in the two simulation studies. Future research could also investigate a larger number of items. We do not think that linking should be conducted with an even smaller number of items, such as $I = 5$ items, because the group comparisons will likely become unstable in the presence of DIF, and the representativity of the link items might be questioned (but, see [89]). Also, the extent of nonuniform DIF was not independently manipulated from the extent of uniform DIF in the two simulation studies. Finally, the performance of our proposed error estimates could also be applied to mis-specified IRT models. For example, the 2PL model could be employed for linking if the item response data would be generated from the logistic positive exponential model [90,91] or the monotonic polynomial IRT model [92,93]. All of these limitations could be addressed in extensive future research.

As a final side note, I would like to add that a comparison of two groups regarding the distribution of the factor variable θ could also be conducted using concurrent calibration by assuming invariant (i.e., the same) item parameters across the groups. Some researchers argue that linking uncertainty is reduced by assuming invariant item parameters (see [94,95]). I think that this belief is unjustified. The fact that there could be variability due to item selection does not disappear, because the variability in the model parameters is not represented in the statistical model. The computation of the linking errors under the assumption of invariant item parameters in the statistical model has been carved out in Ref. [96].

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript:

2PL	two-parameter logistic
HAE	Haebara
IRF	item response function
IRT	item response theory
LE	linking error
MGM	mean-geometric-mean
MM	mean-mean
MML	marginal maximum likelihood
RHAE	robust Haebara
RMGM	robust mean-geometric-mean
SE	standard error
TE	total error

References

- Chen, Y.; Li, X.; Liu, J.; Ying, Z. Item response theory—A statistical framework for educational and psychological measurement. *Stat. Sci.* **2023**, *epub ahead of print*. Available online: <https://imstat.org/journals-and-publications/statistical-science/statistical-science-future-papers/> (accessed on 29 April 2024).
- 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. [\[CrossRef\]](#)
- Bock, R.D.; Gibbons, R.D. *Item Response Theory*; Wiley: Hoboken, NJ, USA, 2021. [\[CrossRef\]](#)
- Rutkowski, L.; von Davier, M.; Rutkowski, D. (Eds.) *A Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*; Chapman Hall/CRC Press: New York, NY, USA, 2013. [\[CrossRef\]](#)
- Berezner, A.; Adams, R.J. Why large-scale assessments use scaling and item response theory. In *Implementation of Large-Scale Education Assessments*; Lietz, P., Cresswell, J.C., Rust, K.F., Adams, R.J., Eds.; Wiley: New York, NY, USA, 2017; pp. 323–356. [\[CrossRef\]](#)
- OECD. *PISA 2018. Technical Report*; OECD: Paris, France, 2020. Available online: <https://bit.ly/3zWbidA> (accessed on 29 April 2024).
- 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. [\[CrossRef\]](#)
- 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.
- 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.
- Bartolucci, F. A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika* **2007**, *72*, 141–157. [\[CrossRef\]](#)
- Casabianca, J.M.; Lewis, C. IRT item parameter recovery with marginal maximum likelihood estimation using loglinear smoothing models. *J. Educ. Behav. Stat.* **2015**, *40*, 547–578. [\[CrossRef\]](#)
- von Davier, M. A general diagnostic model applied to language testing data. *Br. J. Math. Stat. Psychol.* **2008**, *61*, 287–307. [\[CrossRef\]](#) [\[PubMed\]](#)
- Xu, X.; von Davier, M. *Fitting the Structured General Diagnostic Model to NAEP Data*; Research Report No. RR-08-28; Educational Testing Service: Princeton, NJ, USA, 2008. [\[CrossRef\]](#)
- Woods, C.M.; Lin, N. Item response theory with estimation of the latent density using Davidian curves. *Appl. Psychol. Meas.* **2009**, *33*, 102–117. [\[CrossRef\]](#)
- Woods, C.M. Estimating the latent density in unidimensional IRT to permit non-normality. In *Handbook of Item Response Theory Modeling*; Reise, S.P., Revicki, D.A., Eds.; Routledge: New York, NY, USA, 2014; pp. 78–102. [\[CrossRef\]](#)
- Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*; Danish Institute for Educational Research: Copenhagen, Denmark, 1960.
- Bond, T.; Yan, Z.; Heene, M. *Applying the Rasch Model*; Routledge: New York, NY, USA, 2020. [\[CrossRef\]](#)
- Linacre, J.M. Understanding Rasch measurement: Estimation methods for Rasch measures. *J. Outcome Meas.* **1999**, *3*, 382–405. Available online: <https://bit.ly/2UV6Eht> (accessed on 29 April 2024).
- Salzberger, T. The illusion of measurement: Rasch versus 2-PL. *Rasch Meas. Trans.* **2002**, *16*, 882. Available online: <https://tinyurl.com/25wzmzb5> (accessed on 29 April 2024).
- van der Linden, W.J. Fundamental measurement and the fundamentals of Rasch measurement. In *Objective Measurement: Theory Into Practice. Vol. 2*; Wilson, M., Ed.; Ablex Publishing Corporation: Hillsdale, NJ, USA, 1994; pp. 3–24.
- Camilli, G. IRT scoring and test blueprint fidelity. *Appl. Psychol. Meas.* **2018**, *42*, 393–400. [\[CrossRef\]](#)
- Edelsbrunner, P.A. A model and its fit lie in the eye of the beholder: Long live the sum score. *Front. Psychol.* **2022**, *13*, 986767. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hemker, B.T. To a or not to a : On the use of the total score. In *Essays on Contemporary Psychometrics*; van der Ark, L.A., Emons, W.H.M., Meijer, R.R., Eds.; Springer: Cham, Switzerland, 2023; pp. 251–270. [\[CrossRef\]](#)
- Robitzsch, A. On the choice of the item response model for scaling PISA data: Model selection based on information criteria and quantifying model uncertainty. *Entropy* **2022**, *24*, 760. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#)
- White, M. A peculiarity in educational measurement practices. *PsyArXiv* **2024**. [\[CrossRef\]](#)
- Engelhard, G. *Invariant Measurement*; Routledge: New York, NY, USA, 2012. [\[CrossRef\]](#)
- Wind, S.A.; Engelhard, G. How invariant and accurate are domain ratings in writing assessment? *Assess. Writ.* **2013**, *18*, 278–299. [\[CrossRef\]](#)
- Heene, M.; Bollmann, S.; Bühner, M. Much ado about nothing, or much to do about something? *J. Individ. Differ.* **2014**, *35*, 245–249. [\[CrossRef\]](#)
- Ballou, D. Test scaling and value-added measurement. *Educ. Financ. Policy* **2009**, *4*, 351–383. [\[CrossRef\]](#)
- Briggs, D.; Maul, A.; McGrane, J. On the nature of measurement. *PsyArXiv* **2023**. [\[CrossRef\]](#)

32. Heine, J.H.; Heene, M. Measurement and mind: Unveiling the self-delusion of metrification in psychology. *Meas. Interdiscip. Res. Persp.* **2024**, *epub ahead of print*. [\[CrossRef\]](#)
33. 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. [\[CrossRef\]](#)
34. Bock, R.D.; Aitkin, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **1981**, *46*, 443–459. [\[CrossRef\]](#)
35. 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*, 233–279. <https://bit.ly/3ezBB05>.
36. Holland, P.W.; Wainer, H., Eds. *Differential Item Functioning: Theory and Practice*; Lawrence Erlbaum: Hillsdale, NJ, USA, 1993. [\[CrossRef\]](#)
37. Millsap, R.E. *Statistical Approaches to Measurement Invariance*; Routledge: New York, NY, USA, 2011. [\[CrossRef\]](#)
38. 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. [\[CrossRef\]](#)
39. 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. [\[CrossRef\]](#)
40. Kolen, M.J.; Brennan, R.L. *Test Equating, Scaling, and Linking*; Springer: New York, NY, USA, 2014. [\[CrossRef\]](#)
41. 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. [\[CrossRef\]](#)
42. Brennan, R.L. *Generalizability Theory*; Springer: New York, NY, USA, 2001. [\[CrossRef\]](#)
43. Michaelides, M.P. 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. [\[CrossRef\]](#)
44. Michaelides, M.P.; Haertel, E.H. 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*, 46–57. [\[CrossRef\]](#)
45. 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. [\[CrossRef\]](#)
46. 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. [\[CrossRef\]](#)
47. Battauz, M. Multiple equating of separate IRT calibrations. *Psychometrika* **2017**, *82*, 610–636. [\[CrossRef\]](#)
48. Monseur, C.; Berezner, A. The computation of equating errors in international surveys in education. *J. Appl. Meas.* **2007**, *8*, 323–335.
49. OECD. *PISA 2012. Technical Report*; OECD: Paris, France, 2014. Available online: <https://bit.ly/2YLG24g> (accessed on 29 April 2024).
50. Robitzsch, A.; Lüdtke, O. Linking errors in international large-scale assessments: Calculation of standard errors for trend estimation. *Assess. Educ.* **2019**, *26*, 444–465. [\[CrossRef\]](#)
51. 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. [\[CrossRef\]](#)
52. Wu, M. Measurement, sampling, and equating errors in large-scale assessments. *Educ. Meas.* **2010**, *29*, 15–27. [\[CrossRef\]](#)
53. Melin, J.; Cano, S.; Pendrill, L. The role of entropy in construct specification equations (CSE) to improve the validity of memory tests. *Entropy* **2021**, *23*, 212. [\[CrossRef\]](#)
54. Melin, J.; Cano, S.; Flöel, A.; Göschel, L.; Pendrill, L. The role of entropy in construct specification equations (CSE) to improve the validity of memory tests: Extension to word lists. *Entropy* **2022**, *24*, 934. [\[CrossRef\]](#)
55. Melin, J.; Fridberg, H.; Hansson, E.E.; Smedberg, D.; Pendrill, L. Exploring a new application of construct specification equations (CSEs) and entropy: A pilot study with balance measurements. *Entropy* **2023**, *25*, 940. [\[CrossRef\]](#)
56. Tennant, A.; Pallant, J.F. DIF matters: A practical approach to test if differential item functioning makes a difference. *Rasch Meas. Trans.* **2007**, *20*, 1082–1084.
57. von Davier, M.; von Davier, A.A. A unified approach to IRT scale linking and scale transformations. *Methodology* **2007**, *3*, 115–124. [\[CrossRef\]](#)
58. Boos, D.D.; Stefanski, L.A. *Essential Statistical Inference*; Springer: New York, NY, USA, 2013. [\[CrossRef\]](#)
59. Stefanski, L.A.; Boos, D.D. The calculus of M-estimation. *Am. Stat.* **2002**, *56*, 29–38. [\[CrossRef\]](#)
60. Andersson, B. Asymptotic variance of linking coefficient estimators for polytomous IRT models. *Appl. Psychol. Meas.* **2018**, *42*, 192–205. [\[CrossRef\]](#)
61. Battauz, M. equateIRT: An R package for IRT test equating. *J. Stat. Softw.* **2015**, *68*, 1–22. [\[CrossRef\]](#)
62. Jewsbury, P.A. *Error Variance in Common Population Linking Bridge Studies*; Research Report No. RR-19-42; Educational Testing Service: Princeton, NJ, USA, 2019. [\[CrossRef\]](#)

63. Monseur, C.; Sibberns, H.; Hastedt, D. Linking errors in trend estimation for international surveys in education. *IERI Monogr. Ser.* **2008**, *1*, 113–122.
64. Haberman, S.J.; Lee, Y.H.; Qian, J. *Jackknifing Techniques for Evaluation of Equating Accuracy*; Research Report No. RR-09-02; Educational Testing Service: Princeton, NJ, USA, 2009. [\[CrossRef\]](#)
65. Robitzsch, A. Linking error in the 2PL model. *J* **2023**, *6*, 58–84. [\[CrossRef\]](#)
66. 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. [\[CrossRef\]](#)
67. Robitzsch, A. L_p loss functions in invariance alignment and Haberman linking with few or many groups. *Stats* **2020**, *3*, 246–283. [\[CrossRef\]](#)
68. Asparouhov, T.; Muthén, B. Multiple-group factor analysis alignment. *Struct. Equ. Model.* **2014**, *21*, 495–508. [\[CrossRef\]](#)
69. Robitzsch, A. Comparing robust linking and regularized estimation for linking two groups in the 1PL and 2PL models in the presence of sparse uniform differential item functioning. *Stats* **2023**, *6*, 192–208. [\[CrossRef\]](#)
70. Robitzsch, A. Examining differences of invariance alignment in the Mplus software and the R package sirt. *Mathematics* **2024**, *12*, 770. [\[CrossRef\]](#)
71. Haebara, T. Equating logistic ability scales by a weighted least squares method. *Jpn. Psychol. Res.* **1980**, *22*, 144–149. [\[CrossRef\]](#)
72. He, Y.; Cui, Z.; Osterlind, S.J. New robust scale transformation methods in the presence of outlying common items. *Appl. Psychol. Meas.* **2015**, *39*, 613–626. [\[CrossRef\]](#)
73. He, Y.; Cui, Z. Evaluating robust scale transformation methods with multiple outlying common items under IRT true score equating. *Appl. Psychol. Meas.* **2020**, *44*, 296–310. [\[CrossRef\]](#)
74. Robitzsch, A. Robust Haebara linking for many groups: Performance in the case of uniform DIF. *Psych* **2020**, *2*, 155–173. [\[CrossRef\]](#)
75. Weeks, J.P. plink: An R package for linking mixed-format tests using IRT-based methods. *J. Stat. Softw.* **2010**, *35*, 1–33. [\[CrossRef\]](#)
76. Zeileis, A. Object-oriented computation of sandwich estimators. *J. Stat. Softw.* **2006**, *16*, 1–16. [\[CrossRef\]](#)
77. Fay, M.P.; Graubard, B.I. Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics* **2001**, *57*, 1198–1206. [\[CrossRef\]](#)
78. Li, P.; Redden, D.T. Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. *Stat. Med.* **2015**, *34*, 281–296. [\[CrossRef\]](#) [\[PubMed\]](#)
79. Zeileis, A.; Köll, S.; Graham, N. Various versatile variances: An object-oriented implementation of clustered covariances in R. *J. Stat. Softw.* **2020**, *95*, 1–36. [\[CrossRef\]](#)
80. Ogasawara, H. Standard errors of item response theory equating/linking by response function methods. *Appl. Psychol. Meas.* **2001**, *25*, 53–67. [\[CrossRef\]](#)
81. Ogasawara, H. Item response theory true score equatings and their standard errors. *J. Educ. Behav. Stat.* **2001**, *26*, 31–50. [\[CrossRef\]](#)
82. Ogasawara, H. Applications of asymptotic expansion in item response theory linking. In *Statistical Models for Test Equating, Scaling, and Linking*; von Davier, A., Ed.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 261–280. [\[CrossRef\]](#)
83. Battauz, M. IRT test equating in complex linkage plans. *Psychometrika* **2013**, *78*, 464–480. [\[CrossRef\]](#) [\[PubMed\]](#)
84. Battauz, M. Factors affecting the variability of IRT equating coefficients. *Stat. Neerl.* **2015**, *69*, 85–101. [\[CrossRef\]](#)
85. Zhang, Z. Asymptotic standard errors of generalized partial credit model true score equating using characteristic curve methods. *Appl. Psychol. Meas.* **2021**, *45*, 331–345. [\[CrossRef\]](#) [\[PubMed\]](#)
86. R Core Team. *R: A Language and Environment for Statistical Computing*; R Core Team: Vienna, Austria, 2023. Available online: <https://www.R-project.org/> (accessed on 15 March 2023).
87. Robitzsch, A. sirt: Supplementary Item Response Theory Models. R Package Version 4.1-15. 2024. Available online: (accessed on 29 April 2024). [\[CrossRef\]](#)
88. Muraki, E. A generalized partial credit model: Application of an EM algorithm. *Appl. Psychol. Meas.* **1992**, *16*, 159–176. [\[CrossRef\]](#)
89. Pibál, F.; Cesnik, H.S. Evaluating the quantity-quality trade-off in the selection of anchor items: A vertical scaling approach. *Pract. Assess. Res. Eval.* **2011**, *16*, 6.
90. Samejima, F. Logistic positive exponent family of models: Virtue of asymmetric item characteristic curves. *Psychometrika* **2000**, *65*, 319–335. [\[CrossRef\]](#)
91. Huang, Q.; Bolt, D.M.; Lyu, W. Investigating item complexity as a source of cross-national DIF in TIMSS math and science. *Large-scale Assess. Educ.* **2024**, *12*, 12. [\[CrossRef\]](#)
92. Falk, C.F.; Cai, L. Semiparametric item response functions in the context of guessing. *J. Educ. Meas.* **2016**, *53*, 229–247. [\[CrossRef\]](#)
93. Feuerstahler, L. Flexible item response modeling in R with the flexmet package. *Psych* **2021**, *3*, 447–478. [\[CrossRef\]](#)
94. OECD. *PISA 2015; Technical Report*; OECD: Paris, France, 2017. Available online: <https://bit.ly/32buWnZ> (accessed on 29 April 2024).

-
95. Robitzsch, A.; Lüdtke, O. An examination of the linking error currently used in PISA. *Meas. Interdiscip. Res. Persp.* **2024**, *22*, 61–77. [[CrossRef](#)]
 96. Robitzsch, A. Analytical approximation of the jackknife linking error in item response models utilizing a Taylor expansion of the log-likelihood function. *AppliedMath* **2023**, *3*, 49–59. [[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.