stats

MDPI

Article

Bias-Corrected Fixed Item Parameter Calibration, with an Application to PISA Data

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: Fixed item parameter calibration (FIPC) is commonly used to compare groups or countries using an item response theory model with a common set of fixed item parameters. However, FIPC has been shown to produce biased estimates of group means and standard deviations in the presence of random differential item functioning (DIF). To address this, a bias-corrected variant of FIPC, called BCFIPC, is introduced in this article. BCFIPC eliminated the bias of FIPC with only minor efficiency losses in certain simulation conditions, but substantial precision gains in many others, particularly for estimating group standard deviations. Finally, a comparison of both methods using the PISA 2006 dataset revealed relatively large differences in country means and standard deviations.

Keywords: item response model; 2PL model; fixed item parameter calibration; bias correction; differential item functioning; large-scale assessment

check for updates

Academic Editor: Wei Zhu

Received: 19 March 2025
Revised: 18 April 2025
Accepted: 22 April 2025
Published: 24 April 2025

Citation: Robitzsch, A. Bias-Corrected Fixed Item Parameter Calibration, with an Application to PISA Data. *Stats* **2025**, *8*, 29. https://doi.org/10.3390/stats8020029

Correction Statement: This article has been republished with a minor change. The change does not affect the scientific content of the article and further details are available within the backmatter of the website version of this article.

Copyright: © 2025 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 [1–4] provide a statistical framework for analyzing multivariate binary data. Let $\boldsymbol{X} = (X_1, \ldots, X_I)$ represent a vector of I binary random variables ($X_i \in \{0,1\}$ for $i = 1, \ldots, I$), typically referred to as items or (scored) item responses in the psychometric literature. A unidimensional IRT model [5,6] specifies the probability distribution $\mathsf{P}(\boldsymbol{X} = \boldsymbol{x})$ for $\boldsymbol{x} = (x_1, \ldots, x_I) \in \{0,1\}^I$ as

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

where ϕ is the normal density function with mean μ and standard deviation (SD) σ. The latent variable θ is often denoted as ability, trait, or factor in the literature. The vector $\boldsymbol{\delta} = (\mu, \sigma)$ represents distribution parameters, while $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_I)$ contains the estimated item parameters, which define the item response functions (IRF) $P_i(\theta; \boldsymbol{\gamma}_i) = \mathsf{P}(X_i = 1 | \theta)$.

Likely the most popular IRT model for dichotomous responses is the two-parameter logistic (2PL) model [7], which specifies the IRF as

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

where a_i and b_i represent the item discrimination and difficulty parameters, respectively, and $\Psi(x) = (1 + \exp(-x))^{-1}$ is the logistic link function.

If N independent and identically distributed observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ from the distribution of $\boldsymbol{X}$ are available, the model parameters $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ of the IRT model (1) are estimated

using marginal maximum likelihood (MML) estimation [8–10]. Since no closed-form solutions exist, numerical techniques are typically required for estimation. It is important to note that the approach taken in this article is entirely frequentist.

In educational large-scale assessment (LSA; [11,12]) studies, such as programme for international student assessment (PISA; [13]), IRT models are used to compare the distribution of specific groups, such as countries, in a test (i.e., across a set of items) with respect to the latent variable θ in the IRT model (1). Fixed item parameter calibration (FIPC; [14,15]) is commonly applied, assuming that item parameters are known (or estimated) from the total sample comprising all groups (or countries), while only the distribution parameters of a particular group are estimated. The use of FIPC is motivated by the fact that common item parameters across all countries establish a consistent measurement scale, as the same item parameters are applied in the scaling model in every country.

The application of the FIPC method using the 2PL model provides consistent distribution parameter estimates if the item parameters also hold for the group under study (i.e., they are invariant across groups). However, item parameters may be group-specific, a property referred to as differential item functioning (DIF; [16–19]). Research has shown that the presence of DIF introduces additional variability in the estimated mean μ and standard deviation σ of country means and country SDs in LSA studies [20–28].

In this article, the performance of FIPC is investigated in the presence of random DIF [29–31]. In what follows, we assume that these random DIF effects have zero means and a particular SD of DIF effects. The extent of variance in the FIPC estimates due to random DIF is quantified in [32–34]. Moreover, it has also been demonstrated that random DIF causes bias in estimated country means and SDs in the FIPC method [33]. This article addresses these advantages and proposes a bias-corrected variant of FIPC, namely, bias-corrected fixed item parameter calibration (BCFIPC), that aims to remove the bias in the FIPC estimates. A simulation study is carried out in order to demonstrate the satisfactory performance of BCFIPC compared to FIPC. Moreover, it is shown that applying either of the two scaling approaches FIPC or BCFIPC in the PISA 2006 data has a moderate to substantial impact on country means and country SDs.

This article examines the performance of FIPC in the presence of random DIF [29–31]. It is assumed that these random DIF effects have zero means and a positive DIF SD. The variance in FIPC estimates due to random DIF is carved out in [32–34]. Additionally, research has shown that random DIF introduces bias in estimated country means and SDs when using FIPC [33].

This article addresses these limitations by proposing a bias-corrected variant of FIPC, called bias-corrected fixed item parameter calibration (BCFIPC), designed to eliminate bias in FIPC estimates. A simulation study evaluates the performance of BCFIPC relative to FIPC. Additionally, applying FIPC or BCFIPC to PISA 2006 data has a moderate to substantial impact on estimated country means and SDs.

The core idea of the proposed BCFIPC method is to estimate the bias in FIPC as a function of the random DIF variance and apply a correction term to obtain an adjusted estimate with reduced bias. In empirical settings such as large-scale educational studies, the BCFIPC approach is expected to exhibit more desirable statistical properties than the commonly used FIPC method.

The remainder of the article is structured as follows. Section 2 reviews FIPC and introduces the proposed BCFIPC scaling method. Section 3 presents findings from a simulation study comparing these two methods. An empirical comparison of FIPC and BCFIPC using PISA 2006 data is provided in Section 4. Finally, Section 5 concludes with a discussion.

2. Bias Correction for DIF in Fixed Item Parameter Calibration

This section introduces the BCFIPC method by incorporating a bias correction term in FIPC. The section concludes with comments on the implementation of the proposed approach. The bias derivation of the FIPC method closely follows [33].

2.1. Maximum Likelihood Estimation in FIPC

Let $\delta = (\mu, \sigma)$ denote the parameter vector containing the mean μ and the SD σ. Let $\gamma = (\gamma_1, \ldots, \gamma_I)$ collect all item parameters γ_i for $i = 1, \ldots, I$. The true distribution parameter and item parameters in the group are denoted by δ_0 and γ, respectively.

In the scaling model, the estimation of $\widehat{\gamma}$ assumes fixed item parameters $\gamma^* = (\gamma_1^*, \ldots, \gamma_I^*)$. The difference $e = \gamma - \gamma^* = (e_1, \ldots, e_I)$ represents DIF effects in item parameters (i.e., in item discriminations and item difficulties) and reflects misspecification of the IRT model. When the scaling model includes item responses from students in a specific country, the vector γ^* typically consists of international item parameters, and the DIF effects e quantify country-specific DIF [35,36]. This article assumes random DIF effects with zero means, given by

$$\mathsf{E}(e_i) = \mathsf{E}(\gamma_i - \gamma_i^*) = 0. \tag{3}$$

The condition (3) ensures that DIF effects average out at the population level. The following sections demonstrate that, despite this cancellation on average, DIF effects introduce bias in the estimated $\widehat{\mu}$ and $\widehat{\sigma}$ under the FIPC method.

Assume that item response vectors $x_n = (x_{n1}, \ldots, x_{nI})$ are available for subjects $n = 1, \ldots, N$. The log-likelihood function l is defined as (see the IRT model (1))

$$l(\delta, \gamma) = \sum_{n=1}^{N} \log \left\{ \int \prod_{i=1}^{I} \left[P_i(\theta; \gamma_i)^{x_{ni}} (1 - P_i(\theta; \gamma_i))^{1 - x_{ni}} \right] \phi(\theta; \delta) \, \mathrm{d}\theta \right\}. \tag{4}$$

Let $l(\delta, \gamma)$ denote the log-likelihood function. Define the partial derivatives as $l_\delta = (\partial l)/(\partial \delta)$, $l_{\delta\delta} = (\partial^2 l)/(\partial \delta \partial \delta^\top)$, $l_{\delta\gamma_i} = (\partial^2 l)/(\partial \delta \partial \gamma_i^\top)$, and $l_{\delta\gamma_i\gamma_i} = (\partial^3 l)/(\partial \delta \partial \gamma_i \partial \gamma_i^\top)$ for items $i = 1, \ldots, I$.

In FIPC, the parameter estimate $\widehat{\delta} = (\widehat{\mu}, \widehat{\sigma})$ for the distribution parameter δ is obtained as

$$\widehat{\delta} = \arg\min_{\delta} l(\delta, \gamma^*), \tag{5}$$

where γ^* represents the fixed item parameters. Moreover, the estimate $\widehat{\delta}$ satisfies the estimating equation

$$l_\delta(\widehat{\delta}, \gamma^*) = \mathbf{0}. \tag{6}$$

Equation (6) typically requires numerical techniques for its solution.

2.2. Derivation of the Bias in FIPC

The derivation of the bias in the FIPC scaling method involves a second-order Taylor expansion of l_δ around (δ_0, γ) (see [33]). A Taylor expansion provides

$$l_\delta(\widehat{\delta}, \gamma^*) = l_\delta(\delta_0, \gamma) + l_{\delta\delta}(\delta_0, \gamma)(\widehat{\delta} - \delta) + \sum_{i=1}^{I} l_{\delta\gamma_i}(\delta_0, \gamma)(\gamma_i^* - \gamma_i) + \frac{1}{2} \sum_{i=1}^{I} (\gamma_i^* - \gamma_i)^\top l_{\delta\gamma_i\gamma_i}(\delta_0, \gamma)(\gamma_i^* - \gamma_i). \tag{7}$$

In (7), it is assumed that item parameters $\widehat{\gamma}_i$ are approximately independent across items $i = 1, \ldots, I$. This assumption tends to hold more reliably as the number of items increases [37]. The identity $l_\delta(\delta_0, \gamma) = \mathbf{0}$ holds because δ_0 and γ are defined as the true parameters in the data-generating model. Using this identity, along with (3) and (6), the bias of $\widehat{\delta}$ is given by

$$\mathsf{Bias}(\widehat{\delta}) = -\frac{1}{2} I_{\delta\delta}(\delta_0, \gamma)^{-1} \sum_{i=1}^{I} \mathsf{E}\Big[(\gamma_i^* - \gamma_i)^\top I_{\delta\gamma_i\gamma_i}(\delta_0, \gamma)(\gamma_i^* - \gamma_i)\Big]. \tag{8}$$

Equation (8) shows that the variance due to DIF, $e_i = \gamma_i^* - \gamma_i$, induces a bias in the FIPC estimation. It has been demonstrated in [33] that the SD estimate $\widehat{\sigma}$ is typically more biased than the mean estimate $\widehat{\mu}$.

In the following, it is assumed that only uniform DIF is present, meaning there are DIF effects only in item difficulties, not in item discriminations. The unidimensional DIF effects e_i can be written as $e_i = b_i - b_i^*$. Let τ denote the SD of the DIF effects e_i. The bias of the FIPC, as obtained in (8), can be simplified to

$$\mathsf{Bias}(\widehat{\delta}) = -\frac{1}{2}\tau^2 I_{\delta\delta}(\delta_0, \gamma)^{-1} \sum_{i=1}^{I} I_{\delta b_i b_i}(\delta_0, \gamma). \tag{9}$$

Note that expression (9) involves second-order derivatives with respect to item difficulties b_i; that is, only partial derivatives with respect to item difficulty are considered, while derivatives with respect to item discrimination are omitted. In the simplification of (8) leading to (9), DIF effects are treated as stochastic quantities, whereas the terms involving partial derivatives are regarded as fixed.

2.3. Bias-Corrected FIPC

The bias of the FIPC method in (9) can be used to estimate the bias-corrected variant of FIPC (BCFIPC) by subtracting an empirical version of the bias from the FIPC parameter estimate $\widehat{\delta}$.

The derivatives of the likelihood function are modified as follows: $I_{\delta\delta}(\delta_0, \gamma)$ is replaced with $I_{\delta\delta}(\widehat{\delta}, \gamma^*)$ and $I_{\delta b_i b_i}(\delta_0, \gamma)$ with $I_{\delta b_i b_i}(\widehat{\delta}, \gamma^*)$.

Additionally, the SD τ of the DIF effects e_i must be estimated. To achieve this, item difficulties b_i were estimated in a scaling model with fixed item discriminations a_i^* and distribution parameters $\widehat{\mu}$ and $\widehat{\sigma}$. This scaling model also estimates the variances $\mathsf{Var}(\widehat{b}_i)$ of the estimated item difficulties. The variance matrix of the item parameters are computed using the inverse of the observed information matrix [37]. The DIF SD τ can then be estimated as

$$\widehat{\tau} = \mathrm{sqrt}_+\left(\frac{1}{I}\sum_{i=1}^{I}(\widehat{b}_i - b_i^*)^2 - \frac{1}{I}\sum_{i=1}^{I}\mathsf{Var}(\widehat{b}_i)\right), \tag{10}$$

where $\mathrm{sqrt}_+(x) = \sqrt{\max(x, 0)}$. Note that sqrt_+ returns a value of 0 for negative arguments. With these steps, the bias-corrected FIPC estimate $\widehat{\delta}_{\mathrm{bc}}$ is obtained as

$$\widehat{\delta}_{\mathrm{bc}} = \widehat{\delta} + \frac{1}{2}\widehat{\tau}^2 I_{\delta\delta}(\widehat{\delta}, \gamma^*)^{-1} \sum_{i=1}^{I} I_{\delta b_i b_i}(\widehat{\delta}, \gamma^*). \tag{11}$$

2.4. Theoretical Results

When FIPC is applied to data generated under random DIF in item difficulties, the estimation of the distribution parameters μ and σ is based on a misspecified likelihood function. As a result, the statistical properties of the estimator can be derived using M-estimation theory [38,39]. The population parameter $\tilde{\delta}$ minimizes the Kullback-Leibler information, and the estimator $\widehat{\delta} = (\widehat{\mu}, \widehat{\sigma})$ is asymptotically normally distributed with mean $\tilde{\delta}$ and a variance matrix given by the sandwich estimator [38].

The population parameter δ_0 under BCFIPC is defined as the expected result of applying FIPC in the limit as the DIF SD τ approaches zero. The bias correction method is designed to eliminate the bias in FIPC for sufficiently small τ by applying a Taylor

expansion to the log-likelihood function. Although a formal proof of consistency is not provided in this article, the method's validity is supported by simulation results.

Notably, M-estimation theory does not require the data to be generated from a unidimensional IRT model. Even the assumption of local independence is not necessary. The target parameter is defined as the population value that minimizes the Kullback-Leibler information and may not conform to the assumed IRT model (e.g., it may violate local independence). The assumption of independent DIF effects appears to be crucial for the validity of the Taylor expansion. If DIF is clustered within testlets and item independence is violated, the DIF variance correction term must be appropriately adapted to account for this structure.

2.5. Further Adaptations of FIPC and BCFIPC

The FIPC and BCFIPC methods can be easily adapted for the requirements of LSA data applications.

First, the log-likelihood function in (4) can be modified to accommodate missing item responses resulting from planned missingness in balanced incomplete block designs [40,41]. To account for this, define response indicators r_{ni} for subject n and item i, where $r_{ni} = 1$ if item i is observed for subject n and $r_{ni} = 0$ if the item response is missing. The corresponding factor in the log-likelihood function is omitted from the multiplication if item i is missing (i.e., if $r_{ni} = 0$).

Second, in LSA studies, subject-specific sampling weights w_n are typically used. The case-wise likelihood terms should be multiplied by these sampling weights to compute parameter estimates for a defined population of interest.

Combining these two adjustments, the log-likelihood function (4) can be rephrased as

$$l(\delta, \gamma) = \sum_{n=1}^{N} w_n \log \left\{ \int \prod_{i=1}^{I} \left[P_i(\theta; \gamma_i)^{r_{ni} x_{ni}} (1 - P_i(\theta; \gamma_i))^{r_{ni}(1-x_{ni})} \right] \phi(\theta; \delta) \, \mathrm{d}\theta \right\}. \tag{12}$$

2.6. Computation of Derivatives in BCFIPC

The BCFIPC method involves the computation of derivatives in the log-likelihood function. To minimize numerical errors, especially for derivatives of order greater than one, it is crucial not to to rely on numerical differentiation fully.

The integral in the log-likelihood function (12) is approximated using a fixed grid of θ points, $\theta_1, \ldots, \theta_T$. In practice, an equidistant grid of θ values between -6 and 6 is used, typically with $T = 21$ or $T = 41$ points. For larger numbers of items, a denser integration grid is recommended. The log-likelihood function, using shorthand notation, can be written as

$$l = \sum_{n=1}^{N} w_n \log L_n \quad \text{with} \quad L_n = \sum_{t=1}^{T} f_t \prod_{i=1}^{I} g_{nit} = \sum_{t=1}^{T} f_t h_{nt}, \tag{13}$$

where f_t denotes values for the distribution, and g_{nit} represents item-wise likelihood contributions for subject n at grid point t. For missing item responses, we define $g_{nit} = 1$. The term h_{nt} evaluates the individual likelihood at grid point θ_t for $t = 1, \ldots, T$.

The derivative of l with respect to μ (i.e., l_μ) is given by

$$l_\mu = \sum_{n=1}^{N} w_n \frac{L_{n,\mu}}{L_n} = \sum_{n=1}^{N} w_n \frac{\sum_{t=1}^{T} f_{t,\mu} h_{nt}}{L_n}. \tag{14}$$

The derivative with respect to σ can be obtained by replacing μ with σ in (14). The second derivative $l_{\mu\mu}$ can be computed as

$$l_{\mu\mu} = \sum_{n=1}^{N} w_n \frac{L_{n,\mu\mu} L_n - L_{n,\mu}^2}{L_n^2}. \tag{15}$$

The derivatives with respect to b_i can be determined as

$$l_{\mu b_i} = \sum_{n=1}^{N} w_n \left[\frac{\sum_{t=1}^{T} f_{t,\mu} \frac{g_{nti,b_i}}{g_{nti}} h_{nt}}{L_n} \frac{\left(\sum_{t=1}^{T} f_{t,\mu} h_{nt} \right) \left(\sum_{t=1}^{T} f_t \frac{g_{nti,b_i}}{g_{nti}} h_{nt} \right)}{L_n^2} \right]. \tag{16}$$

The remaining necessary derivatives can be derived similarly, though they may involve more tedious algebraic calculations.

3. Simulation Study

This Simulation Study compares the performance of the FIPC and BCFIPC methods. The goal is to determine under which simulation conditions the bias correction in the FIPC method leads to more efficient parameter estimates.

3.1. Method

In this Simulation Study, item responses were simulated under the 2PL model for a single group of interest. Item parameters were generated for $I = 15$, $I = 30$ and $I = 45$ items, representing short and long tests in empirical applications. For the $I = 15$ condition, the fixed item discriminations a_i^* and item difficulties b_i^* used in FIPC and BCFIPC for scaling were specified as follows. The item discriminations a_i^* were 1.14, 1.20, 1.05, 1.11, 0.71, 1.26, 1.38, 1.41, 1.05, 0.58, 1.13, 0.56, 0.63, 1.19, and 0.60, yielding a mean of $M = 1.00$ and an SD of $SD = 0.30$. The item difficulties b_i^* were 0.18, 1.40, -0.86, 0.09, 0.56, 1.11, 0.09, 0.34, -1.80, -1.57, -0.72, -0.19, -0.84, 0.76, and 1.44, resulting in $M = 0.00$ and $SD = 1.00$. For $I = 30$ or $I = 45$ items, the parameters were doubled or tripled accordingly.

In the simulation of item responses under the 2PL model, item discriminations a_i were set equal to a_i^*, while item difficulties were generated as

$$b_i = b_i^* + e_i \text{ with } e_i \sim \mathsf{N}(0, \tau^2). \tag{17}$$

Group-specific item difficulties were modeled with normally distributed random DIF effects with zero mean (see (3)) and DIF SD τ. The SD τ was varied as 0, 0.2, and 0.4, representing no DIF, small DIF, and large DIF, respectively. Because DIF effects were randomly drawn, item difficulties were newly simulated in each replication of a simulation condition.

The θ variable in the 2PL model was assumed to follow a normal distribution. The SD σ of θ was varied as 0.8 and 1.2, while the mean μ was varied as -0.3, 0, 0.3, and 0.6. The condition $\mu = -0.3$ represents a test with slightly too difficult items for the group, while $\mu = 0$ corresponds to a situation where the mean item difficulty matches the mean of the θ distribution. The conditions $\mu = 0.3$ and $\mu = 0.6$ represent tests that are slightly too easy items for the group.

To reflect the application of FIPC and BCFIPC in both small-scale and large-scale assessment studies, the sample size N was varied as 125, 250, 500, 1000, 2000, and 4000.

In each of the 6 (sample size N) $\times$ 3 (number of items I) $\times$ 3 (DIF SD τ) $\times$ 4 (mean μ) $\times$ 2 (SD σ) = 432 simulation conditions, 3000 replications were conducted. Note that all factors were crossed in this simulation study.

To prevent non-convergence in model estimation for small samples—caused by item proportion correct values of 0 or 1—MML estimation was stabilized by augmenting the item response dataset with two additional pseudo-subjects. The first pseudo-subject was assigned the response vector **0** (all zeroes), and the second the response vector **1** (all ones). The response vector **0** received a weight of $\varepsilon_0(1 - M)$, and **1** received a weight of $\varepsilon_0 M$, with $\varepsilon_0 = 0.5$ and M denoting the average of the mean score across all items. The original subjects were assigned a weight of 1. This weighting scheme ensured that the average item mean score remained unchanged between the original and augmented datasets.

Bias and root mean square error (RMSE) were computed for the estimated mean $\widehat{\mu}$ and standard deviation $\widehat{\sigma}$ under the FIPC and BCFIPC methods. A relative RMSE for the BCFIPC method was calculated as the ratio of the RMSE values of BCFIPC to FIPC, multiplied by 100. A relative RMSE exceeding 100 indicates an efficiency loss for BCFIPC, whereas values below 100 signify precision gains. Hence, the relative RMSE reflects the square root of the variance ratio of the parameter estimates from the BCFIPC method to those from the FIPC method, multiplied by 100.

This simulation study was conducted using the statistical software R (Version 4.4.1; [42]). The 2PL model was estimated with the `sirt::xxirt()` function from the R package sirt (Version 4.2-106; [43]). The BCFIPC method was implemented in a custom function developed specifically for this study. This function, along with replication materials, is available at https://osf.io/wcsdm (accessed on 19 March 2025).

3.2. Results

Figure 1 displays the bias of the estimated DIF SD $\widehat{\tau}$ as a function of the true DIF SD τ, the number of items I and sample size N for $\mu = 0.6$ and $\sigma = 1.2$. In the absence of random DIF (i.e., $\tau = 0$), there was a slight positive bias in $\widehat{\tau}$, while the bias was negative in scenarios in which DIF was present (i.e., $\tau = 0.2$ and $\tau = 0.4$). Moreover, it was observed that the sample size and number of items converged to zero with increasing sample size.

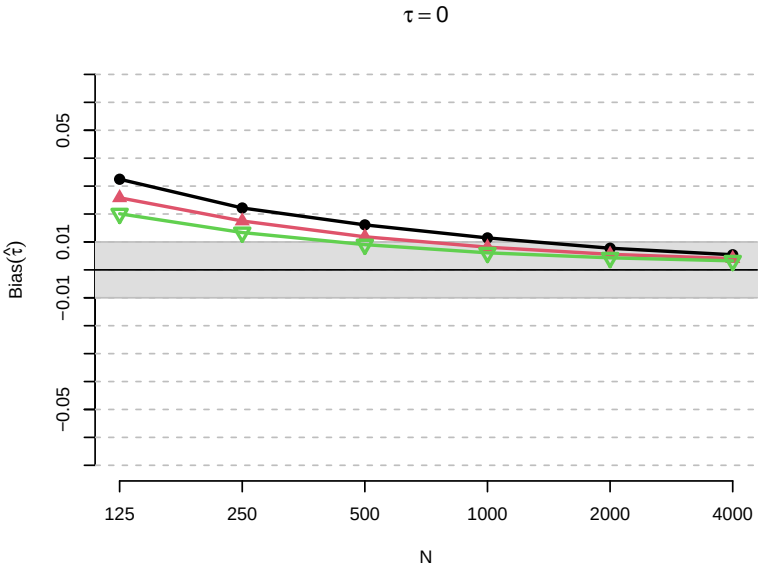


Figure 1. *Cont.*

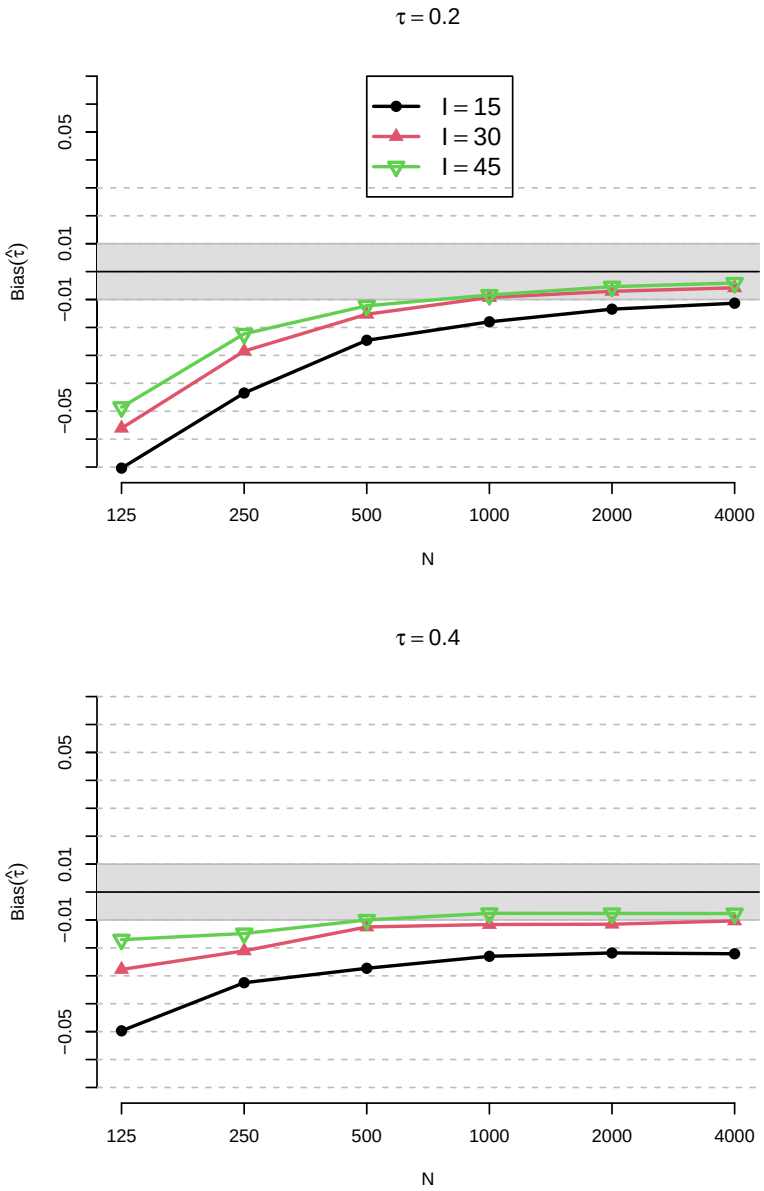


Figure 1. Simulation Study: Bias of the estimated DIF SD $\widehat{\tau}$ as a function of the DIF SD τ, the number of items I and sample size N for $\mu = 0.6$ and $\sigma = 1.2$. The range of bias values between -0.01 and 0.01 are shown in a gray background color.

Table 1 presents the estimated mean $\widehat{\mu}$ for FIPC and BCFIPC as a function of the DIF SD τ, the true mean μ, the number of items I and sample size N for $\sigma = 1.2$. In the absence of DIF ($\tau = 0$), both methods were unbiased across all simulation conditions. When DIF was present ($\tau = 0.2$ or $\tau = 0.4$), slight biases for $\widehat{\mu}$ were observed for the FIPC method. These biases were relatively independent of the sample size N, meaning the bias of FIPC did not decrease as the sample size increased. Furthermore, the bias of FIPC did not converge to zero with an increasing number of items I. Notably, a positive bias for $\widehat{\mu}$ was observed in the difficult test condition ($\mu = -0.3$), while a negative bias was found in the easy test condition ($\mu = 0.6$). The bias for $\widehat{\mu}$ was smaller in the $\mu = 0$ and $\mu = 0.3$ conditions. In contrast, BCFIPC consistently reduced the bias of FIPC, with absolute bias values not exceeding 0.007. Therefore, BCFIPC is generally preferred over FIPC in terms of bias correction when random DIF effects are present in item difficulties.

Table 1. Simulation Study: Bias of the estimated mean $\widehat{\mu}$ as a function of the DIF SD τ, the true mean μ, the number of items I and sample size N for $\sigma = 1.2$.

			FIPC for $N =$						**BCFIPC for $N =$**					
τ	μ	I	**125**	**250**	**500**	**1000**	**2000**	**4000**	**125**	**250**	**500**	**1000**	**2000**	**4000**
0	−0.3	15	0.001	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000
		30	−0.002	0.001	−0.002	0.000	−0.001	0.000	−0.003	0.001	−0.002	0.000	−0.001	0.000
		45	−0.003	−0.001	0.000	0.001	0.000	−0.001	−0.003	−0.001	0.000	0.001	0.000	−0.001
	0	15	−0.001	0.002	−0.001	−0.001	0.000	0.000	−0.002	0.002	−0.001	−0.001	0.000	0.000
		30	−0.001	0.000	−0.001	−0.001	0.000	−0.001	−0.001	0.000	−0.001	−0.001	0.000	−0.001
		45	−0.002	0.000	−0.003	0.000	0.000	0.000	−0.002	0.000	−0.003	0.000	0.000	0.000
	0.3	15	−0.002	0.002	−0.003	−0.001	0.000	0.000	−0.002	0.002	−0.003	−0.001	0.000	0.000
		30	−0.001	0.001	0.002	0.001	0.001	0.000	−0.001	0.001	0.002	0.001	0.001	0.000
		45	0.001	0.002	0.002	0.000	0.001	0.000	0.001	0.002	0.002	0.000	0.001	0.000
	0.6	15	−0.001	0.001	0.000	−0.001	0.000	0.000	0.000	0.001	0.000	−0.001	0.000	0.000
		30	0.004	−0.001	0.001	0.001	−0.001	0.000	0.004	−0.001	0.001	0.001	−0.001	0.000
		45	0.001	0.002	0.000	−0.001	0.001	0.000	0.001	0.002	0.000	−0.001	0.001	0.000
0.2	−0.3	15	0.001	0.000	0.001	0.005	0.005	0.003	−0.001	−0.002	−0.002	0.002	0.002	0.000
		30	0.008	0.006	0.003	0.002	0.003	0.003	0.006	0.003	0.000	−0.001	−0.001	0.000
		45	0.001	0.004	0.003	0.004	0.004	0.003	−0.001	0.001	0.000	0.001	0.001	0.000
	0	15	0.002	0.003	0.001	0.004	0.000	0.001	0.002	0.001	0.000	0.003	−0.001	0.000
		30	0.004	−0.001	0.001	0.003	0.003	0.001	0.003	−0.002	0.000	0.002	0.002	0.000
		45	0.001	0.002	0.001	−0.001	0.002	0.002	0.000	0.001	0.000	−0.002	0.001	0.000
	0.3	15	0.000	−0.003	0.003	0.001	0.000	−0.002	0.001	−0.003	0.004	0.002	0.001	−0.001
		30	−0.003	0.001	−0.001	0.000	−0.002	−0.002	−0.002	0.002	0.000	0.001	−0.001	−0.001
		45	−0.002	−0.001	−0.001	−0.002	0.000	−0.001	−0.001	0.000	0.000	−0.001	0.001	0.000
	0.6	15	−0.005	−0.004	−0.002	−0.004	−0.002	−0.003	−0.003	−0.001	0.000	−0.001	0.001	0.000
		30	−0.001	−0.003	−0.003	−0.004	−0.003	−0.001	0.001	0.000	0.000	−0.001	0.000	0.001
		45	−0.001	−0.003	−0.005	−0.004	−0.002	−0.003	0.002	−0.001	−0.002	−0.001	0.001	0.000
0.4	−0.3	15	0.011	0.009	**0.019**	**0.011**	**0.012**	**0.014**	0.000	−0.003	0.007	−0.001	0.000	0.002
		30	0.013	**0.017**	**0.010**	**0.013**	**0.014**	**0.013**	0.001	0.005	−0.002	0.001	0.001	0.001
		45	0.014	**0.010**	**0.014**	**0.012**	**0.013**	**0.014**	0.002	−0.002	0.002	0.000	0.001	0.002
	0	15	0.006	0.006	0.004	0.002	0.007	0.006	0.001	0.002	−0.001	−0.004	0.002	0.001
		30	0.006	0.002	0.005	0.005	0.003	0.005	0.002	−0.003	0.000	0.000	−0.002	0.000
		45	0.004	0.005	0.004	0.004	0.006	0.003	0.000	0.000	−0.001	−0.001	0.001	−0.002
	0.3	15	−0.002	−0.001	−0.001	−0.002	−0.006	−0.004	0.000	0.002	0.001	0.001	−0.004	−0.002
		30	−0.003	−0.001	0.001	−0.003	−0.001	−0.003	0.000	0.002	0.004	0.000	0.002	0.000
		45	−0.005	−0.005	−0.004	−0.006	−0.003	−0.001	−0.002	−0.002	−0.001	−0.003	0.000	0.002
	0.6	15	−0.007	**−0.010**	**−0.010**	**−0.014**	**−0.013**	**−0.010**	0.004	0.000	0.001	−0.003	−0.002	0.001
		30	−0.014	−0.008	**−0.010**	**−0.013**	**−0.014**	−0.009	−0.004	0.003	0.001	−0.002	−0.003	0.002
		45	−0.007	**−0.012**	**−0.012**	**−0.011**	**−0.012**	−0.008	0.004	−0.001	−0.001	0.001	0.000	0.003

Note. SD = standard deviation; FIPC = fixed item parameter calibration; BCFIPC = bias-corrected fixed item parameter calibration; Values of absolute bias larger than 0.010 are printed in bold font.

Table A1 in Appendix A reports the estimated mean $\widehat{\mu}$ for FIPC and BCFIPC as a function of the DIF SD τ, the true mean μ, the number of items I and sample size N for $\sigma = 0.8$. Overall, the pattern of bias closely resembled that observed in the case of $\sigma = 1.2$. Absolute differences in bias values were computed using the corresponding entries in Tables 1 and A1. The mean absolute difference was $M = 0.0016$ with an $SD = 0.0015$. The maximum observed difference in bias was 0.007; this occurred four times, with three instances corresponding to the smallest sample size $N = 125$. These results suggest that the bias of FIPC and BCFIPC for $\widehat{\mu}$ is largely unaffected by the population value of σ.

Table 2 displays the estimated SD $\widehat{\sigma}$ for FIPC and BCFIPC as a function of the DIF SD τ, the true mean μ, the number of items I and sample size N for $\sigma = 1.2$. It is evident that the bias for the FIPC method was more pronounced for $\widehat{\sigma}$ than for $\widehat{\mu}$. Importantly, a consistent negative bias for FIPC was observed for $\widehat{\sigma}$ independent of the population value of μ. In contrast, the bias of FIPC was essentially eliminated with the BCFIPC method, in which all absolute biases were smaller than 0.010 with only one exception. Hence, bias correction of FIPC is much more relevant for $\widehat{\sigma}$.

Table 2. Simulation Study: Bias of the estimated SD $\widehat{\sigma}$ as a function of the DIF SD τ, the true mean μ, the number of items I and sample size N for $\sigma = 1.2$.

			FIPC for N =						**BCFIPC for N =**					
τ	μ	I	**125**	**250**	**500**	**1000**	**2000**	**4000**	**125**	**250**	**500**	**1000**	**2000**	**4000**
0	−0.3	15	−0.006	−0.004	−0.003	−0.001	−0.001	0.000	−0.004	−0.003	−0.002	−0.001	−0.001	0.000
		30	−0.004	−0.005	−0.003	−0.001	0.000	0.000	−0.004	−0.005	−0.003	−0.001	0.000	0.000
		45	−0.007	−0.004	−0.002	−0.001	0.000	0.000	−0.007	−0.004	−0.002	−0.001	0.000	0.000
	0	15	−0.004	−0.003	0.001	0.000	−0.001	0.000	−0.003	−0.003	0.001	0.000	−0.001	0.000
		30	−0.004	−0.002	−0.002	−0.002	0.000	0.000	−0.003	−0.002	−0.002	−0.002	0.000	0.000
		45	−0.008	−0.003	−0.002	−0.001	0.000	−0.001	−0.008	−0.002	−0.001	−0.001	0.000	−0.001
	0.3	15	**−0.011**	−0.004	−0.002	−0.001	0.000	0.000	−0.009	−0.004	−0.002	−0.001	0.000	0.000
		30	−0.008	−0.001	−0.002	−0.001	−0.001	0.000	−0.008	−0.001	−0.002	−0.001	−0.001	0.000
		45	−0.004	−0.003	−0.001	0.000	0.000	0.000	−0.004	−0.003	−0.001	0.000	0.000	0.000
	0.6	15	−0.006	−0.002	−0.001	−0.002	0.000	−0.001	−0.004	−0.001	0.000	−0.001	0.000	−0.001
		30	−0.005	−0.003	−0.003	0.000	−0.001	0.000	−0.004	−0.003	−0.003	0.000	−0.001	0.000
		45	−0.005	−0.004	−0.001	0.000	0.000	−0.001	−0.005	−0.003	−0.001	0.000	0.000	−0.001
0.2	−0.3	15	**−0.016**	−0.009	**−0.010**	**−0.010**	−0.009	−0.008	−0.009	−0.001	−0.001	0.000	0.001	0.002
		30	**−0.011**	**−0.013**	**−0.010**	−0.008	−0.009	−0.008	−0.005	−0.005	−0.002	0.000	0.000	0.001
		45	**−0.013**	**−0.013**	**−0.011**	−0.009	−0.008	−0.008	−0.007	−0.006	−0.003	−0.001	0.000	0.000
	0	15	**−0.019**	−0.009	**−0.012**	**−0.012**	−0.009	**−0.010**	**−0.011**	−0.001	−0.003	−0.002	0.001	0.000
		30	**−0.015**	**−0.010**	**−0.011**	−0.009	−0.009	−0.009	−0.008	−0.002	−0.002	0.000	0.000	0.000
		45	**−0.016**	**−0.010**	−0.009	−0.008	−0.009	−0.009	−0.009	−0.002	−0.001	0.000	0.000	0.000
	0.3	15	**−0.014**	**−0.013**	**−0.010**	**−0.011**	**−0.010**	**−0.010**	−0.006	−0.004	0.000	−0.001	0.001	0.001
		30	**−0.015**	**−0.012**	**−0.011**	**−0.011**	**−0.010**	−0.009	−0.008	−0.004	−0.003	−0.002	0.000	0.000
		45	**−0.020**	**−0.013**	**−0.010**	**−0.010**	−0.008	−0.008	**−0.013**	−0.005	−0.001	−0.001	0.001	0.001
	0.6	15	**−0.012**	**−0.010**	**−0.011**	**−0.010**	**−0.011**	−0.009	−0.004	−0.001	−0.001	0.000	0.000	0.001
		30	**−0.016**	**−0.013**	**−0.010**	−0.009	**−0.010**	−0.009	−0.008	−0.005	−0.001	0.000	0.000	0.001
		45	**−0.016**	**−0.012**	**−0.010**	−0.009	**−0.010**	**−0.010**	−0.009	−0.004	−0.002	−0.001	−0.001	0.000
0.4	−0.3	15	**−0.038**	**−0.035**	**−0.035**	**−0.035**	**−0.033**	**−0.031**	−0.002	0.003	0.004	0.004	0.006	0.008
		30	**−0.040**	**−0.037**	**−0.035**	**−0.031**	**−0.032**	**−0.031**	−0.008	−0.004	−0.001	0.003	0.003	0.004
		45	**−0.036**	**−0.033**	**−0.032**	**−0.030**	**−0.032**	**−0.031**	−0.005	−0.001	0.000	0.003	0.001	0.002
	0	15	**−0.042**	**−0.041**	**−0.037**	**−0.037**	**−0.036**	**−0.034**	−0.004	−0.002	0.003	0.004	0.005	0.007
		30	**−0.037**	**−0.038**	**−0.034**	**−0.035**	**−0.033**	**−0.033**	−0.002	−0.003	0.002	0.000	0.003	0.003
		45	**−0.040**	**−0.037**	**−0.036**	**−0.033**	**−0.034**	**−0.033**	−0.007	−0.004	−0.002	0.001	0.001	0.001
	0.3	15	**−0.042**	**−0.040**	**−0.038**	**−0.036**	**−0.039**	**−0.035**	−0.003	0.001	0.004	0.006	0.003	0.006
		30	**−0.040**	**−0.038**	**−0.036**	**−0.034**	**−0.036**	**−0.033**	−0.005	−0.001	0.001	0.003	0.002	0.004
		45	**−0.039**	**−0.036**	**−0.035**	**−0.034**	**−0.034**	**−0.034**	−0.005	0.000	0.001	0.002	0.001	0.002
	0.6	15	**−0.039**	**−0.042**	**−0.040**	**−0.037**	**−0.039**	**−0.037**	0.001	−0.001	0.002	0.006	0.005	0.006
		30	**−0.041**	**−0.038**	**−0.037**	**−0.035**	**−0.035**	**−0.036**	−0.005	−0.001	0.001	0.003	0.003	0.002
		45	**−0.044**	**−0.038**	**−0.036**	**−0.035**	**−0.035**	**−0.034**	−0.009	−0.002	0.000	0.002	0.002	0.003

Note. SD = standard deviation; FIPC = fixed item parameter calibration; BCFIPC = bias-corrected fixed item parameter calibration; Values of absolute bias larger than 0.010 are printed in bold font.

Table A2 in Appendix A shows the estimated SD $\widehat{\sigma}$ for FIPC and BCFIPC as a function of the DIF SD τ, the true mean μ, the number of items I and sample size N for $\sigma = 0.8$. While the overall pattern of results resembled that in Table 2, the biases of $\widehat{\sigma}$ under the FIPC method were smaller in magnitude for $\sigma = 0.8$ compared to $\sigma = 1.2$. This difference was particularly pronounced for $\tau = 0.4$.

Figure 2 illustrates the bias in the estimated mean $\widehat{\mu}$ and standard deviation $\widehat{\sigma}$ of the FIPC method as a function of the DIF SD τ, the number of items I, and the sample size N, for $\mu = 0.6$ and $\sigma = 1.2$. Notably, the bias in $\widehat{\mu}$ and $\widehat{\sigma}$ remains relatively stable across different values of I and N. Thus, applying the bias correction method BCFIPC remains essential even in settings with large samples and many items.

Figure 3 displays the RMSE in the estimated mean $\widehat{\mu}$ and SD $\widehat{\sigma}$ of the FIPC method as a function of the DIF SD τ, the number of items I, and the sample size N, for $\mu = 0.6$ and $\sigma = 1.2$. The RMSE decreased with increasing N or I and was lower in scenarios with smaller DIF SD τ.

Table 3 reports the relative RMSE for BCFIPC as a function of the DIF SD τ, the true mean μ, the number of items I and sample size N for $\sigma = 1.2$. Notably, no efficiency loss was observed when using BCFIPC instead of FIPC in the no DIF condition ($\tau = 0$). However, in the presence of DIF, slight efficiency losses for BCFIPC emerged for $\widehat{\mu}$. This occurred because the small reduction in bias was outweighed by a larger variance in the BCFIPC method. For $\tau = 0.4$, the average relative RMSE was 102.2, with a maximum of 102.8, indicating the small efficiency loss. In contrast, the efficiency gains in terms of RMSE for BCFIPC were substantial for the estimated SD $\widehat{\sigma}$, especially for larger sample sizes. These gains were more pronounced in longer tests; that is, in scenarios with more items.

Table A3 in Appendix A shows the relative RMSE for BCFIPC as a function of the DIF SD τ, the true mean μ, the number of items I and sample size N for $\sigma = 0.8$. The pattern in Table A3 for $\sigma = 0.8$ closely mirrored that in Table 3 for $\sigma = 1.2$.

Figure 4 illustrates the relative RMSE in the estimated mean $\widehat{\mu}$ and SD $\widehat{\sigma}$ of the BCFIPC method as a function of the DIF SD τ, the number of items I, and the sample size N, for $\mu = 0.6$ and $\sigma = 1.2$. Efficiency gains of BCFIPC are especially pronounced with large DIF ($\tau = 0.4$), large sample sizes N, and a greater number of items I. In larger samples, these gains can be attributed to the bias reduction achieved by BCFIPC, which offsets its higher variance relative to FIPC.

Overall, BCFIPC is generally recommended over FIPC in empirical applications when random DIF effects are present. The BCFIPC method effectively reduces the bias observed in FIPC. While some simulation conditions showed negligible efficiency losses in $\widehat{\mu}$ for BCFIPC compared to FIPC, these were outweighed by the efficiency gains for $\widehat{\sigma}$. It should be emphasized that the BCFIPC method also proved successful even in the small sample size scenario $N = 125$.

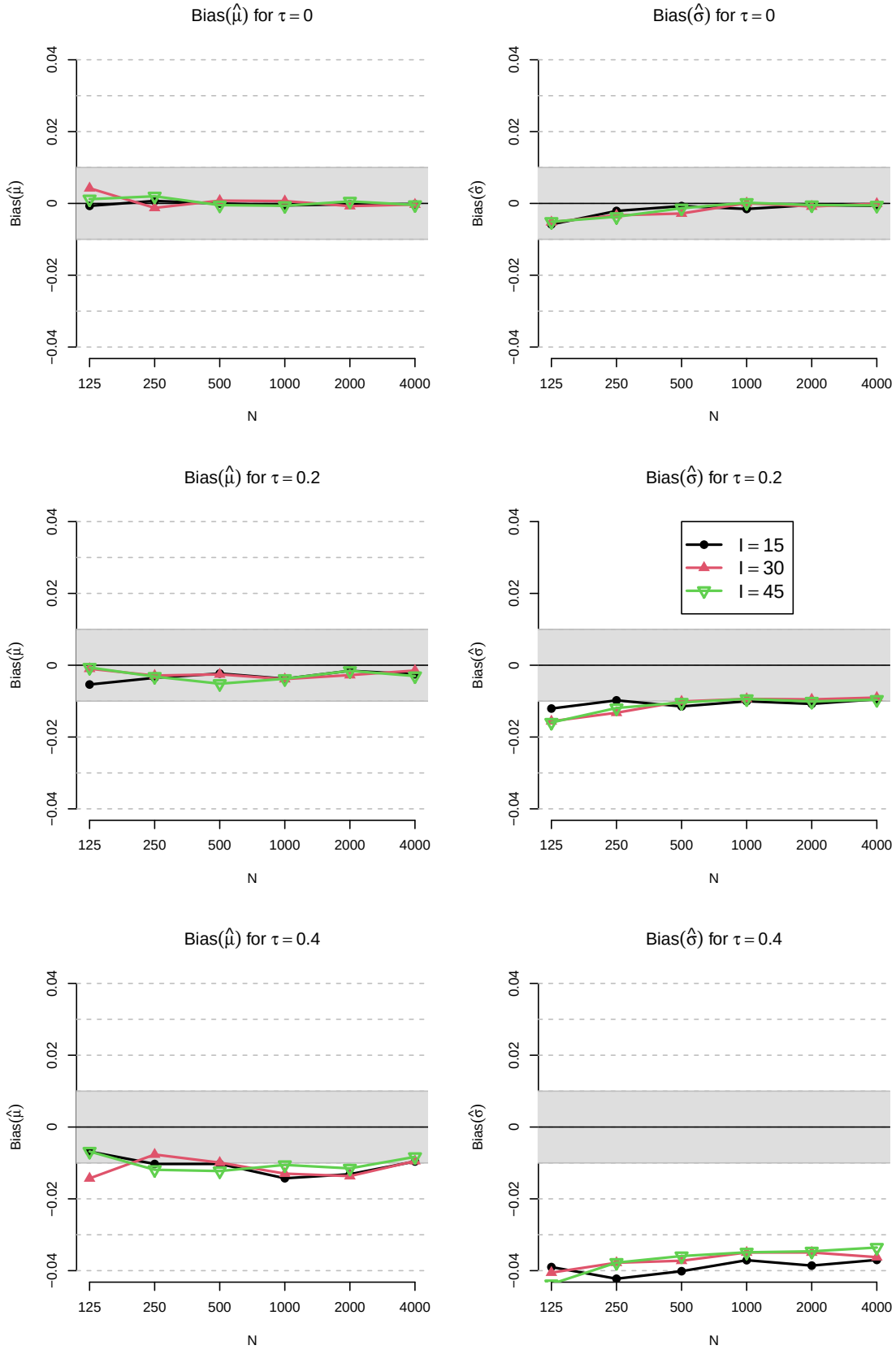


Figure 2. Simulation Study: Bias of the estimated mean $\widehat{\mu}$ and the estimated SD $\widehat{\sigma}$ of the FIPC method as a function of the DIF SD τ, the number of items I and sample size N for $\mu = 0.6$ and $\sigma = 1.2$. The range of bias values between -0.01 and 0.01 are shown in a gray background color.

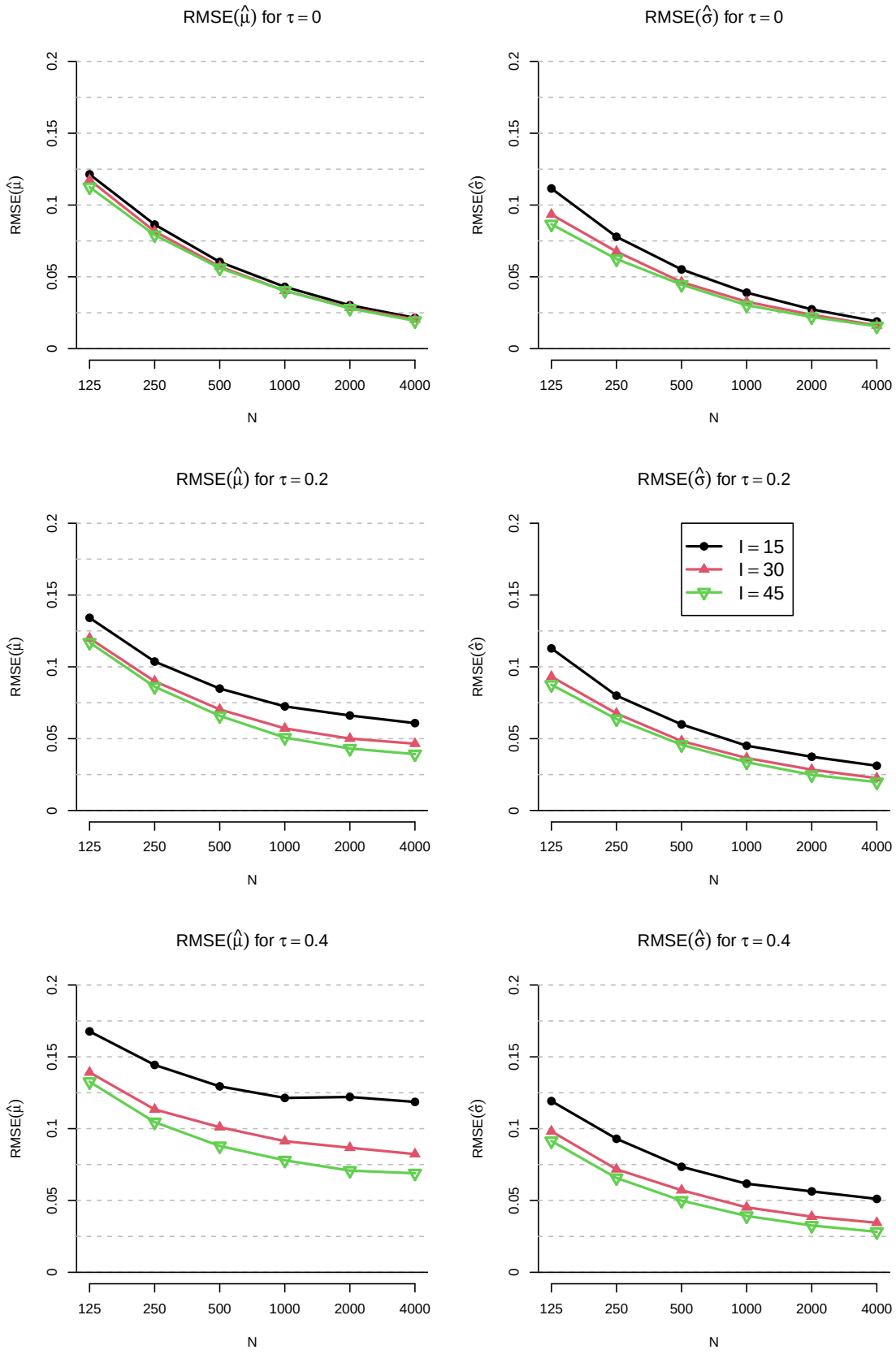


Figure 3. Simulation Study: Root mean square error (RMSE) of the estimated mean $\widehat{\mu}$ and the estimated SD $\widehat{\sigma}$ of the FIPC method as a function of the DIF SD τ, the number of items I and sample size N for $\mu = 0.6$ and $\sigma = 1.2$.

Table 3. Simulation Study: Relative root mean square error (RMSE) for bias-corrected fixed item parameter calibration (BCFIPC) as a function of the DIF SD τ, the true mean μ, the number of items I and sample size N for $\sigma = 1.2$.

			$\widehat{\mu}$ for $N =$						$\widehat{\sigma}$ for $N =$					
τ	μ	I	**125**	**250**	**500**	**1000**	**2000**	**4000**	**125**	**250**	**500**	**1000**	**2000**	**4000**
0	−0.3	15	100.1	100.0	100.0	100.0	100.0	100.0	99.9	100.0	99.9	100.0	100.0	100.0
		30	100.1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		45	100.0	100.0	100.0	100.0	100.0	100.0	99.9	99.9	100.0	100.0	100.0	100.0
	0	15	100.1	100.0	100.0	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0
		30	100.0	100.0	100.0	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0
		45	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	0.3	15	100.1	100.1	100.0	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0
		30	100.1	100.0	100.0	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0
		45	100.0	100.0	100.0	100.0	100.0	100.0	99.9	99.9	100.0	100.0	100.0	100.0
	0.6	15	100.1	100.1	100.0	100.0	100.0	100.0	99.9	99.9	100.0	99.9	100.0	100.0
		30	100.1	100.0	100.0	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0
		45	100.1	100.0	100.0	100.0	100.0	100.0	99.9	99.9	100.0	100.0	100.0	100.0
0.2	−0.3	15	100.5	100.5	100.6	100.3	100.3	100.5	99.0	99.7	99.1	97.9	97.9	97.5
		30	100.4	100.4	100.5	100.5	100.4	100.4	99.3	98.5	98.3	98.1	96.1	95.5
		45	100.5	100.5	100.5	100.3	100.2	100.3	99.5	98.5	97.9	96.9	95.5	92.6
	0	15	100.5	100.5	100.6	100.5	100.7	100.7	99.0	99.8	98.3	96.5	97.5	95.3
		30	100.5	100.6	100.6	100.5	100.5	100.6	99.2	99.1	97.9	97.5	95.6	92.6
		45	100.5	100.6	100.6	100.7	100.6	100.6	99.2	98.9	98.3	97.2	94.4	91.1
	0.3	15	100.5	100.5	100.7	100.7	100.7	100.7	99.6	99.0	98.8	98.0	96.7	95.6
		30	100.5	100.6	100.6	100.7	100.7	100.6	99.0	98.7	97.5	96.1	94.1	92.5
		45	100.5	100.6	100.6	100.6	100.7	100.7	98.7	98.4	98.2	96.4	95.0	92.4
	0.6	15	100.5	100.6	100.6	100.6	100.7	100.7	99.6	99.4	98.6	97.8	96.3	95.7
		30	100.5	100.6	100.6	100.5	100.6	100.7	99.1	98.7	98.2	97.1	94.9	92.8
		45	100.5	100.5	100.4	100.4	100.6	100.4	99.0	98.8	97.7	96.7	93.0	89.9
0.4	−0.3	15	102.0	**102.1**	101.4	101.9	101.7	101.6	96.0	94.9	91.6	88.6	86.7	87.0
		30	101.9	101.3	101.8	101.1	101.1	101.1	93.3	89.9	86.1	83.5	77.6	76.7
		45	101.8	101.8	100.9	101.0	100.6	100.1	93.9	90.9	84.8	80.5	72.2	68.7
	0	15	**102.4**	**102.4**	**102.5**	**102.8**	**102.4**	**102.5**	95.4	91.9	90.3	86.9	85.3	85.7
		30	**102.4**	**102.6**	**102.5**	**102.4**	**102.7**	**102.5**	94.7	89.7	86.2	78.5	77.2	72.6
		45	**102.5**	**102.5**	**102.5**	**102.6**	**102.4**	**102.7**	92.9	88.0	81.9	76.6	70.2	65.5
	0.3	15	**102.5**	**102.7**	**102.8**	**102.8**	**102.7**	**102.8**	95.6	92.3	89.5	86.9	80.6	83.4
		30	**102.6**	**102.7**	**102.8**	**102.8**	**102.8**	**102.7**	93.0	89.6	85.0	79.9	73.4	72.2
		45	**102.6**	**102.6**	**102.6**	**102.6**	**102.7**	**102.8**	92.8	89.3	82.2	76.0	68.5	64.1
	0.6	15	**102.8**	**102.5**	**102.4**	**102.1**	**102.2**	**102.5**	96.3	91.4	88.4	86.3	83.0	81.6
		30	**102.2**	**102.6**	**102.4**	101.7	101.6	**102.1**	93.3	88.7	84.3	79.3	74.4	68.6
		45	**102.7**	**102.1**	101.8	101.9	101.4	102.0	91.4	87.3	81.4	74.7	68.7	64.1

Note. SD = standard deviation; The fixed item parameter calibration (FIPC) method was used as the reference method to compute the relative RMSE.; RMSE value larger than 102.0 are printed in bold font. RMSE values smaller than 98.0 are highlighted in yellow.

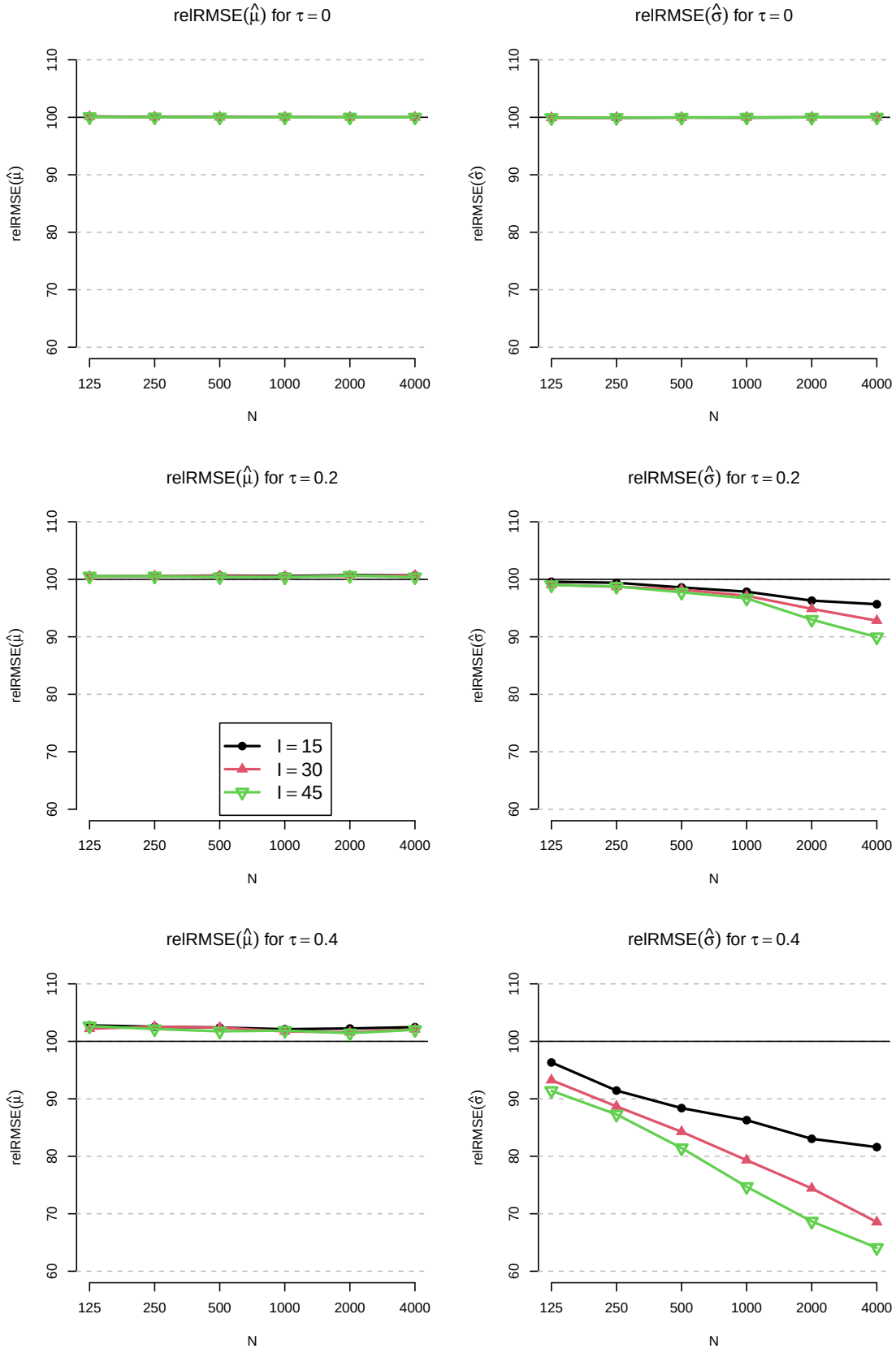


Figure 4. Simulation Study: Relative root mean square error (RMSE) of the estimated mean $\widehat{\mu}$ and the estimated SD $\widehat{\sigma}$ of the BCFIPC method as a function of the DIF SD τ, the number of items I and sample size N for $\mu = 0.6$ and $\sigma = 1.2$.

4. Empirical Example: PISA 2006 Data

The present study sought to illustrate the consequences of the FIPC and BCFIPC scaling methods. To this end, data from the cognitive domains of mathematics, reading and science in the programme for international student assessment (PISA) 2006 study were analyzed (PISA 2006; [13]).

4.1. Method

4.1.1. Sample and Instruments

The PISA dataset used in this analysis included participants from 26 selected countries (see Table 4) that took part in the PISA 2006 study. The complete PISA 2006 dataset is publicly available at https://www.oecd.org/en/data/datasets/pisa-2006-database.html (accessed on 19 March 2025).

Table 4. Empirical Example, PISA 2006 mathematics: Estimated country mean $\widehat{\mu}$ and estimated country SD $\widehat{\sigma}$ for FIPC and BCFIPC methods, along with their differences and estimated standard errors.

				$\widehat{\mu}$			$\widehat{\sigma}$		
CNT	***N***	***I***	$\widehat{\tau}$	**FIPC**	**BCFIPC**	**Diff**	**FIPC**	**BCFIPC**	**Diff**
AUS	10,838	48	0.24 (0.01)	515.0 (2.4)	514.7 (2.4)	−0.25 (0.03)	95.8 (1.3)	97.4 (1.3)	1.56 (0.10)
AUT	3784	48	0.25 (0.01)	501.5 (4.3)	501.1 (4.3)	−0.40 (0.06)	105.9 (2.6)	107.5 (2.6)	1.61 (0.17)
BEL	6851	48	0.16 (0.02)	518.9 (2.7)	518.8 (2.8)	−0.09 (0.03)	106.9 (2.3)	107.7 (2.4)	0.76 (0.18)
CAN	17,349	48	0.18 (0.01)	524.0 (1.9)	523.9 (1.9)	−0.10 (0.02)	88.6 (1.2)	89.5 (1.2)	0.87 (0.09)
CHE	9384	48	0.22 (0.01)	527.5 (3.2)	527.5 (3.3)	−0.07 (0.03)	102.3 (1.7)	103.6 (1.8)	1.31 (0.15)
CZE	4600	48	0.26 (0.02)	504.6 (3.7)	504.2 (3.8)	−0.41 (0.07)	106.6 (2.4)	108.5 (2.5)	1.88 (0.26)
DEU	3795	48	0.20 (0.01)	499.0 (4.2)	498.7 (4.3)	−0.27 (0.04)	104.9 (2.7)	105.9 (2.6)	1.03 (0.13)
DNK	3441	48	0.25 (0.01)	509.9 (2.5)	509.5 (2.5)	−0.38 (0.05)	87.8 (2.0)	89.5 (2.0)	1.65 (0.16)
ESP	15,043	48	0.22 (0.01)	474.8 (2.4)	474.3 (2.4)	−0.59 (0.05)	93.0 (1.3)	94.2 (1.3)	1.15 (0.09)
EST	3751	48	0.29 (0.01)	509.8 (2.9)	509.3 (3.0)	−0.51 (0.07)	86.3 (2.1)	88.4 (2.1)	2.18 (0.19)
FIN	3644	48	0.26 (0.01)	546.8 (2.1)	546.9 (2.1)	0.13 (0.04)	84.1 (1.6)	86.0 (1.6)	1.88 (0.21)
FRA	3629	48	0.27 (0.01)	487.8 (3.5)	487.0 (3.6)	−0.74 (0.07)	102.3 (2.6)	104.3 (2.6)	1.96 (0.19)
GBR	10,074	48	0.30 (0.01)	487.8 (2.3)	487.0 (2.3)	−0.88 (0.09)	96.6 (1.5)	98.9 (1.5)	2.26 (0.19)
GRC	3732	48	0.26 (0.01)	450.3 (3.0)	449.3 (3.1)	−1.06 (0.10)	99.8 (2.1)	101.3 (2.1)	1.52 (0.14)
HUN	3445	48	0.23 (0.02)	483.2 (3.1)	482.6 (3.1)	−0.56 (0.08)	97.8 (2.1)	99.1 (2.1)	1.34 (0.19)
IRL	3540	48	0.28 (0.01)	496.2 (3.0)	495.5 (3.1)	−0.70 (0.08)	88.8 (2.0)	90.8 (2.0)	2.00 (0.17)
ISL	2888	48	0.27 (0.02)	501.4 (2.2)	500.8 (2.2)	−0.54 (0.06)	96.0 (1.7)	98.0 (1.7)	1.94 (0.22)
ITA	16,740	48	0.28 (0.01)	456.7 (2.4)	455.6 (2.4)	−1.17 (0.07)	101.6 (1.7)	103.5 (1.7)	1.87 (0.11)
JPN	4565	48	0.55 (0.01)	523.1 (3.7)	522.3 (3.9)	−0.80 (0.22)	94.6 (2.7)	102.9 (2.8)	8.23 (0.37)
KOR	4004	47	0.54 (0.02)	542.6 (3.8)	543.1 (4.1)	0.44 (0.27)	96.1 (3.5)	104.6 (3.5)	8.49 (0.65)
LUX	3503	48	0.17 (0.01)	484.1 (1.5)	483.8 (1.6)	−0.31 (0.04)	100.2 (1.4)	101.0 (1.4)	0.76 (0.09)
NLD	3768	48	0.27 (0.01)	523.8 (2.9)	523.6 (2.9)	−0.21 (0.04)	94.8 (2.5)	96.8 (2.5)	2.00 (0.14)
NOR	3575	48	0.28 (0.01)	486.7 (2.7)	485.9 (2.8)	−0.79 (0.10)	97.5 (2.0)	99.4 (1.9)	1.93 (0.20)
POL	4258	48	0.29 (0.01)	487.0 (2.6)	486.1 (2.6)	−0.82 (0.09)	96.0 (1.8)	98.1 (1.8)	2.10 (0.19)
PRT	3938	48	0.31 (0.01)	459.4 (3.1)	458.0 (3.2)	−1.41 (0.15)	98.6 (2.1)	101.0 (2.1)	2.31 (0.22)
SWE	3419	48	0.24 (0.01)	497.4 (2.8)	496.9 (2.8)	−0.46 (0.07)	96.5 (2.0)	98.0 (2.0)	1.52 (0.18)

Note. SD = standard deviation; CNT = country label (see Appendix B); N = sample size; I = number of items; $\widehat{\tau}$ = estimated DIF SD; FIPC = fixed item parameter calibration; BCFIPC = bias-corrected fixed item parameter calibration; Diff = difference between estimates of FIPC and BCFIPC.

Items in the mathematics, reading, and science domains were administered only to a subset of students. The analysis included students in each domain who had been administered at least one item in the respective cognitive domain.

This approach resulted in a total sample size of 157,558 students for mathematics, with country sample sizes ranging from 2888 to 17,349 (with a mean of $M = 6059.9$ and an SD of $SD = 4368.1$). The reading domain included 110,236 students, with sample sizes per country ranging from 2010 to 12,142 ($M = 4239.8$, $SD = 3046.7$). For science, 204,786

students were analyzed, with country sample sizes between 3778 and 22,602 ($M = 7876.4$, $SD = 5686.3$).

All items from the three cognitive domains were included in the analysis. A small portion of the items, originally polytomous, were recoded dichotomously, with only the highest category considered correct. The remaining items were already dichotomous in the original PISA scoring. In total, 48 mathematics items, 28 reading items, and 103 science items were used.

4.1.2. Sampling Weights and Standard Errors

Student (sampling) weights were incorporated in all analyses. Within each country, student weights were normalized to sum to 5000, ensuring equal contribution across countries. The specific value of 5000 is arbitrary; any constant would suffice, but equal weighting across countries is essential. In the first step, international item parameters were estimated using the 2PL model applied to the weighted pooled dataset for each domain. In the second step, FIPC and BCFIPC scaling based on the 2PL model were conducted separately for each country, with item parameters fixed to the international estimates. In the scaling model, item responses omitted due to the planned missingness design in PISA were excluded from the likelihood estimation. This approach aligns with the operational procedures used for reporting in PISA.

Point estimates, including country means $\widehat{\mu}$ and country SDs $\widehat{\sigma}$, as well as error estimates, were linearly transformed onto the PISA reporting metric. The weighted pooled sample of all students was scaled to have a mean of 500 and an SD of 100 in each cognitive domain. Because all countries contributed equally, the average of the country means was also 500.

Standard errors referring to the sampling of students were computed using the balanced repeated replication zones provided in the PISA 2006 study [13], accounting for the stratified clustered sampling within countries [44,45]. Each rth replication sample used modified student sampling weights $w_n^{(r)}$. The PISA study employs $R = 80$ replication samples to perform statistical inference for a parameter of interest (i.e., $\widehat{\boldsymbol{\beta}}$) based on student weights w_n. In each replication, the analysis was repeated using sampling weights $w_n^{(r)}$, producing estimates $\widehat{\boldsymbol{\beta}}^{(r)}$. The variance matrix $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}$ for $\widehat{\boldsymbol{\beta}}$ was computed as [44]

$$\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = A \sum_{r=1}^{R} (\widehat{\boldsymbol{\beta}}^{(r)} - \widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}^{(r)} - \widehat{\boldsymbol{\beta}})^{\top}, \tag{18}$$

where the scaling factor A depends on the replication method. PISA employs balanced repeated replication, with a Fay factor of 0.05 [46]. The standard error for an individual parameter is obtained from the square root of the corresponding diagonal entries in $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}$. This resampling approach also allows for the computation of standard errors for differences between estimates from the same dataset. Moreover, the standard error for the difference in country means between FIPC and BCFIPC scaling can be directly obtained. To illustrate the approach in more detail, let $\widehat{\mu}_{\text{FIPC}}$ and $\widehat{\mu}_{\text{BCFIPC}}$ denote the estimates of the mean for the FIPC and BCFIPC methods, respectively. Furthermore, let $\widehat{\mu}_{\text{FIPC}}^{(r)}$ and $\widehat{\mu}_{\text{BCFIPC}}^{(r)}$ denote the corresponding estimates in the rth replication sample. Using these estimates, the difference $\Delta = \widehat{\mu}_{\text{BCFIPC}} - \widehat{\mu}_{\text{FIPC}}$ between the two scaling methods can be computed, along with the differences $\widehat{\Delta}^{(r)} = \widehat{\mu}_{\text{BCFIPC}}^{(r)} - \widehat{\mu}_{\text{FIPC}}^{(r)}$ in each replication sample. The variance of $\widehat{\Delta}$, which corresponds to the squared standard error, can then be expressed according to (18) as

$$\text{Var}(\widehat{\Delta}) = A \sum_{r=1}^{R} (\widehat{\Delta}^{(r)} - \widehat{\Delta})^2. \tag{19}$$

4.1.3. Analysis

The R code from the Simulation Study (see Section 3.1) was adapted for this example. Standard errors for country means and country SDs, along with the differences between the FIPC and BCFIPC approaches, were estimated using the balanced repeated replication method (see (18)), requiring only a loop over scaling models with different inputs of sampling weights. To compute the bias correction term in BCFIPC, the standard errors of country-specific item parameters (and consequently the DIF effects) were computed under the assumption of independent sampling of students. This might be a reasonable approximation for statistical inference, but it is likely not for distribution parameters.

4.2. Results

4.2.1. Mathematics

Table 4 reports the results for PISA 2006 mathematics. The estimated DIF SD $\widehat{\tau}$ in item difficulties had a mean of $M = 0.27$ with an SD of $SD = 0.09$, ranging from 0.16 (BEL) to 0.55 (JPN). Notably, substantial variation in DIF SD was observed across countries, with some showing low DIF (BEL, CAN, LUX) and others exhibiting high DIF (JPN, KOR). The average item discrimination in the international item parameters (available at https://osf.io/wcsdm, accessed on 19 March 2025) was 1.321, while the average item difficulty was -0.036, corresponding to a mean p value of 0.493. Table 4 also shows that, with few exceptions, estimated country means under BCFIPC were lower than those under FIPC. The absolute difference in country means $\widehat{\mu}$ between the two scaling methods had a mean of $M = 0.54$ with an SD of $SD = 0.35$, ranging from 0.07 (CHE) to 1.41 (PRT). Country SDs $\widehat{\sigma}$ were consistently larger under BCFIPC than under FIPC, with an average absolute difference of $M = 2.16$ (SD $SD = 1.88$), ranging from 0.76 (BEL, LUX) to 8.49 (KOR). Nearly all differences between the two scaling methods were statistically significant.

4.2.2. Reading

Table 5 presents the results for PISA 2006 reading. The estimated DIF SD $\widehat{\tau}$ in item difficulties had a mean of $M = 0.37$ with an SD of $SD = 0.09$, ranging between 0.25 (AUS) and 0.59 (KOR). The average item discrimination in the international item parameters (available at https://osf.io/wcsdm, accessed on 19 March 2025) was 1.402, while the average item difficulty was -0.163, yielding a mean p value of 0.569. All estimated country means were higher under BCFIPC than under FIPC. The mean absolute difference in country means was $M = 1.80$ with $SD = 1.77$, ranging from 0.18 (GRC) to 9.74 (KOR). As in mathematics, all country SDs were larger under BCFIPC than under FIPC, with a mean of $M = 4.94$ and an SD of $SD = 2.60$, ranging from 2.03 (AUS) to 10.48 (PRT).

4.2.3. Science

Table 6 shows the results for PISA 2006 science. The estimated DIF SD $\widehat{\tau}$ had a mean $M = 0.36$ with $SD = 0.09$, ranging from 0.23 (BEL) to 0.63 (KOR). The average item discrimination in the international item parameters (available at https://osf.io/wcsdm, accessed on 19 March 2025) was 1.122, while the average item difficulty was -0.239, resulting in an average p value of 0.565. With two exceptions, the estimated country means under BCFIPC were larger than under FIPC. The mean absolute average difference in country means was $M = 0.75$ with $SD = 0.82$ and ranged between 0.00 (LUX) and 3.22 (KOR). All country SDs were all larger under BCFIPC than under FIPC, with a mean absolute difference of $M = 3.46$ with $SD = 1.89$, ranging from 1.42 (CHE) to 10.15 (KOR).

Table 5. Empirical Example, PISA 2006 reading: Estimated country mean $\widehat{\mu}$ and estimated country SD $\widehat{\sigma}$ for FIPC and BCFIPC methods, along with their differences and estimated standard errors.

				$\widehat{\mu}$			$\widehat{\sigma}$		
CNT	***N***	***I***	$\widehat{\tau}$	**FIPC**	**BCFIPC**	**Diff**	**FIPC**	**BCFIPC**	**Diff**
AUS	7562	28	0.25 (0.01)	517.0 (2.3)	518.1 (2.3)	1.09 (0.08)	96.0 (1.5)	98.0 (1.5)	2.03 (0.12)
AUT	2646	27	0.27 (0.01)	496.3 (3.8)	497.1 (3.8)	0.81 (0.12)	103.3 (2.7)	106.0 (2.8)	2.70 (0.30)
BEL	4840	28	0.27 (0.01)	505.9 (3.1)	506.9 (3.1)	0.98 (0.12)	107.1 (2.7)	109.7 (2.7)	2.57 (0.27)
CAN	12,142	28	0.28 (0.01)	527.6 (2.1)	529.2 (2.2)	1.62 (0.11)	93.4 (1.6)	95.9 (1.6)	2.50 (0.13)
CHE	6578	28	0.33 (0.01)	502.3 (3.1)	503.8 (3.2)	1.49 (0.15)	95.8 (2.3)	99.4 (2.4)	3.60 (0.28)
CZE	3246	28	0.33 (0.02)	483.2 (4.4)	484.0 (4.6)	0.85 (0.18)	113.0 (3.1)	117.4 (3.2)	4.36 (0.43)
DEU	2701	28	0.52 (0.03)	496.1 (5.0)	498.9 (5.4)	2.84 (0.49)	114.0 (2.8)	124.5 (3.2)	10.48 (1.05)
DNK	2431	27	0.40 (0.01)	500.1 (3.1)	502.3 (3.3)	2.18 (0.19)	89.1 (2.0)	94.3 (2.0)	5.17 (0.32)
ESP	10,506	28	0.41 (0.01)	464.9 (2.1)	465.6 (2.3)	0.70 (0.14)	81.6 (1.2)	87.5 (1.4)	5.97 (0.37)
EST	2630	28	0.34 (0.01)	499.4 (3.0)	501.0 (3.0)	1.65 (0.14)	83.8 (1.9)	87.6 (1.9)	3.80 (0.32)
FIN	2536	28	0.33 (0.01)	551.6 (2.4)	554.5 (2.5)	2.92 (0.28)	85.4 (1.9)	88.5 (2.0)	3.13 (0.28)
FRA	2524	28	0.33 (0.02)	499.0 (3.8)	500.4 (3.9)	1.39 (0.19)	98.4 (2.9)	102.3 (3.1)	3.88 (0.48)
GBR	7061	28	0.34 (0.01)	498.4 (2.2)	499.8 (2.3)	1.45 (0.12)	98.5 (1.8)	102.5 (1.8)	4.03 (0.25)
GRC	2606	28	0.49 (0.01)	456.8 (3.6)	457.0 (3.9)	0.18 (0.28)	95.2 (2.6)	104.2 (2.6)	9.02 (0.53)
HUN	2399	28	0.32 (0.02)	485.2 (3.3)	486.2 (3.4)	0.98 (0.18)	91.8 (2.4)	95.5 (2.5)	3.62 (0.49)
IRL	2468	28	0.27 (0.01)	518.4 (3.5)	519.8 (3.6)	1.37 (0.15)	94.6 (2.2)	97.1 (2.2)	2.45 (0.23)
ISL	2010	28	0.32 (0.02)	493.1 (2.0)	494.4 (2.0)	1.21 (0.14)	91.5 (2.1)	95.0 (2.2)	3.47 (0.40)
ITA	11,629	28	0.35 (0.01)	471.5 (2.2)	472.2 (2.2)	0.63 (0.09)	98.4 (1.9)	102.9 (2.0)	4.55 (0.38)
JPN	3203	28	0.44 (0.01)	502.8 (3.6)	505.4 (3.8)	2.57 (0.26)	103.4 (2.2)	110.1 (2.2)	6.73 (0.38)
KOR	2790	27	0.59 (0.02)	556.1 (3.7)	565.8 (4.3)	9.74 (0.81)	95.9 (3.2)	106.0 (3.2)	10.02 (0.60)
LUX	2443	27	0.33 (0.02)	482.0 (2.1)	482.8 (2.2)	0.76 (0.13)	101.2 (1.9)	105.2 (2.1)	4.01 (0.56)
NLD	2666	28	0.43 (0.01)	509.2 (3.2)	512.0 (3.4)	2.80 (0.27)	101.7 (3.0)	108.2 (3.1)	6.45 (0.41)
NOR	2504	28	0.45 (0.02)	489.3 (2.8)	491.3 (3.0)	2.03 (0.29)	101.8 (1.9)	109.1 (2.1)	7.37 (0.65)
POL	2968	28	0.31 (0.01)	506.8 (2.8)	508.2 (2.9)	1.40 (0.15)	99.9 (2.2)	103.2 (2.3)	3.25 (0.33)
PRT	2773	28	0.53 (0.02)	475.8 (3.4)	477.7 (3.7)	1.86 (0.34)	95.5 (2.6)	106.0 (2.7)	10.48 (0.75)
SWE	2374	28	0.29 (0.01)	510.7 (3.0)	512.0 (3.1)	1.31 (0.14)	100.4 (2.6)	103.2 (2.6)	2.83 (0.28)

Note. SD = standard deviation; CNT = country label (see Appendix B); N = sample size; I = number of items; $\widehat{\tau}$ = estimated DIF SD; FIPC = fixed item parameter calibration; BCFIPC = bias-corrected fixed item parameter calibration; Diff = difference between estimates of FIPC and BCFIPC.

Table 6. Empirical Example, PISA 2006 science: Estimated country mean $\widehat{\mu}$ and estimated country SD $\widehat{\sigma}$ for FIPC and BCFIPC methods, along with their differences and estimated standard errors.

				$\widehat{\mu}$			$\widehat{\sigma}$		
CNT	***N***	***I***	$\widehat{\tau}$	**FIPC**	**BCFIPC**	**Diff**	**FIPC**	**BCFIPC**	**Diff**
AUS	14,142	103	0.33 (0.01)	517.9 (2.2)	518.8 (2.2)	0.88 (0.06)	100.9 (1.0)	103.7 (1.0)	2.81 (0.10)
AUT	4927	103	0.30 (0.01)	502.2 (3.9)	502.7 (4.0)	0.43 (0.09)	101.3 (2.6)	103.7 (2.6)	2.31 (0.15)
BEL	8850	103	0.23 (0.01)	503.1 (2.4)	503.4 (2.5)	0.28 (0.04)	103.5 (1.9)	104.9 (1.9)	1.44 (0.09)
CAN	22,602	103	0.27 (0.01)	527.1 (2.0)	527.9 (2.0)	0.79 (0.05)	95.4 (1.2)	97.4 (1.2)	1.93 (0.09)
CHE	12,188	103	0.23 (0.01)	505.1 (3.1)	505.4 (3.2)	0.30 (0.04)	101.4 (1.7)	102.8 (1.7)	1.42 (0.09)
CZE	5931	103	0.30 (0.01)	505.3 (3.5)	505.8 (3.5)	0.52 (0.08)	101.5 (2.1)	103.9 (2.1)	2.43 (0.18)
DEU	4881	103	0.29 (0.01)	508.8 (3.8)	509.3 (3.9)	0.52 (0.08)	102.7 (2.2)	104.9 (2.2)	2.18 (0.13)
DNK	4529	103	0.32 (0.01)	486.9 (3.1)	487.1 (3.1)	0.17 (0.07)	95.2 (1.5)	97.8 (1.5)	2.60 (0.15)
ESP	19,569	103	0.27 (0.01)	482.1 (2.3)	482.2 (2.4)	0.04 (0.04)	90.7 (0.7)	92.6 (0.7)	1.83 (0.10)
EST	4865	103	0.42 (0.01)	523.5 (2.5)	525.2 (2.7)	1.71 (0.14)	87.4 (1.4)	91.8 (1.5)	4.36 (0.21)
FIN	4712	103	0.40 (0.01)	554.4 (1.9)	557.1 (2.0)	2.64 (0.18)	88.4 (1.1)	92.3 (1.1)	3.86 (0.20)
FRA	4702	103	0.43 (0.01)	489.8 (3.4)	490.2 (3.5)	0.41 (0.15)	105.4 (2.3)	110.4 (2.3)	5.01 (0.29)
GBR	13,099	103	0.42 (0.01)	506.5 (2.0)	507.5 (2.1)	0.99 (0.09)	107.3 (1.3)	112.1 (1.3)	4.76 (0.19)
GRC	4866	103	0.41 (0.01)	469.7 (3.1)	469.3 (3.2)	−0.40 (0.11)	95.7 (1.8)	100.1 (1.9)	4.38 (0.24)

Table 6. *Cont.*

				$\widehat{\mu}$			$\widehat{\sigma}$		
CNT	***N***	***I***	$\widehat{\tau}$	**FIPC**	**BCFIPC**	**Diff**	**FIPC**	**BCFIPC**	**Diff**
HUN	4489	102	0.47 (0.01)	496.1 (2.7)	496.9 (2.9)	0.76 (0.14)	91.5 (1.6)	97.0 (1.7)	5.49 (0.28)
IRL	4582	103	0.41 (0.01)	498.4 (3.1)	499.0 (3.3)	0.66 (0.13)	95.4 (1.7)	99.6 (1.7)	4.25 (0.19)
ISL	3778	103	0.40 (0.01)	484.3 (1.7)	484.5 (1.7)	0.16 (0.06)	96.6 (1.3)	100.7 (1.4)	4.06 (0.22)
ITA	21,752	103	0.29 (0.01)	469.3 (2.0)	469.1 (2.0)	−0.19 (0.04)	98.4 (1.2)	100.6 (1.2)	2.19 (0.09)
JPN	5940	102	0.48 (0.01)	526.2 (3.4)	528.5 (3.7)	2.28 (0.23)	103.1 (2.1)	109.2 (2.0)	6.07 (0.22)
KOR	5174	103	0.63 (0.01)	515.5 (3.3)	518.7 (3.7)	3.22 (0.35)	93.6 (2.5)	103.8 (2.6)	10.15 (0.36)
LUX	4566	103	0.25 (0.01)	479.4 (1.2)	479.4 (1.2)	0.00 (0.02)	101.7 (1.2)	103.4 (1.2)	1.64 (0.12)
NLD	4867	103	0.37 (0.01)	515.4 (2.7)	516.5 (2.8)	1.08 (0.12)	99.5 (1.9)	103.1 (1.9)	3.60 (0.18)
NOR	4684	101	0.30 (0.01)	478.8 (2.9)	478.9 (2.9)	0.06 (0.05)	99.4 (2.2)	101.7 (2.2)	2.30 (0.16)
POL	5547	102	0.33 (0.01)	490.8 (2.4)	491.1 (2.5)	0.26 (0.06)	93.6 (1.3)	96.3 (1.3)	2.71 (0.16)
PRT	5107	103	0.38 (0.01)	465.9 (3.0)	465.4 (3.1)	−0.49 (0.10)	90.8 (1.7)	94.5 (1.8)	3.67 (0.20)
SWE	4437	102	0.31 (0.01)	496.2 (2.3)	496.6 (2.4)	0.34 (0.06)	96.0 (1.7)	98.4 (1.8)	2.44 (0.17)

Note. SD = standard deviation; CNT = country label (see Appendix B); N = sample size; I = number of items; $\widehat{\tau}$ = estimated DIF SD; FIPC = fixed item parameter calibration; BCFIPC = bias-corrected fixed item parameter calibration; Diff = difference between estimates of FIPC and BCFIPC.

4.2.4. Summary

This reanalysis of the PISA 2006 dataset showed moderate differences in country means $\widehat{\mu}$ when BCFIPC was used instead of FIPC as a scaling model. In contrast, the differences in country SDs $\widehat{\sigma}$ between BCFIPC and FIPC were more pronounced.

Figure 5 shows the differences between BCFIPC and FIPC estimates (Diff) for country means and country SDs as a function of the squared estimated DIF SD, $\widehat{\tau}^2$. Consistent with the analytical and simulation findings, a strong relationship emerged between Diff and $\widehat{\tau}^2$ for the bias correction in the SD. As observed in the simulation study, both the sign and magnitude of the bias correction for country means depend on the country mean itself, which accounts for the more scattered pattern of results for country means in Figure 5.

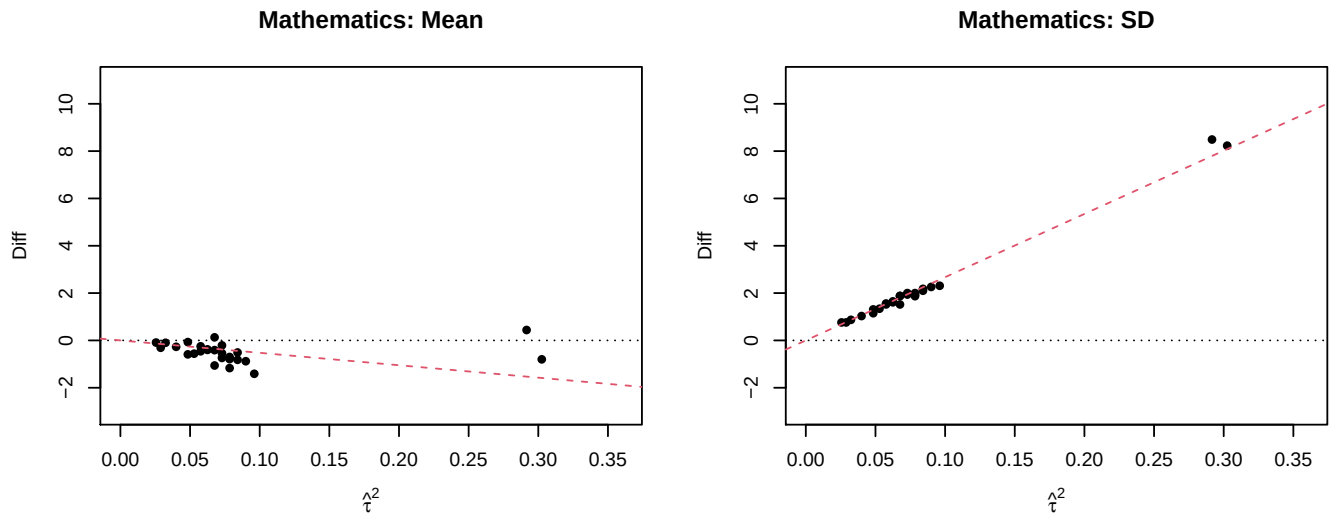


Figure 5. *Cont.*

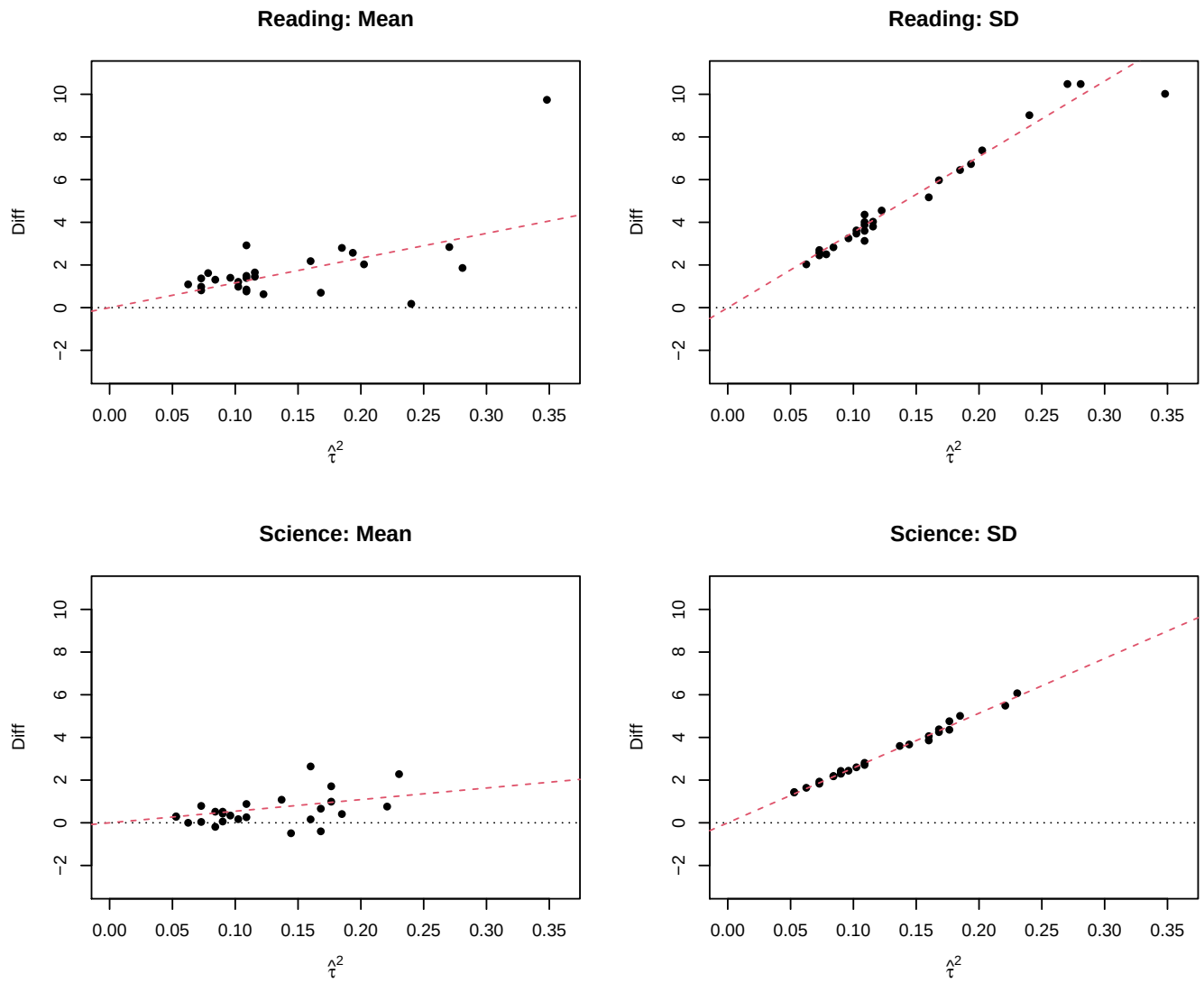


Figure 5. Empirical Example, PISA 2006: Estimated differences (Diff) in country means and country standard deviations (SD) between the BCFIPC and FIPC methods. The fit of a robust regression of Diff on $\widehat{\tau}^2$ is displayed with dotted red lines.

5. Discussion

This article examined a bias correction approach for FIPC, namely the BCFIPC scaling method, in a simulation study. The results showed that the bias in the estimated mean and SD under FIPC was substantially reduced, if not eliminated, with BCFIPC. While applying BCFIPC led to a slight efficiency loss for the mean, it was significantly more efficient than FIPC for the estimated SD.

The BCFIPC method was also compared to FIPC using the PISA 2006 dataset across three cognitive domains. The differences in country means and SD estimates between BCFIPC and FIPC were not negligible, suggesting that BCFIPC could be a viable scaling method for operational use.

It is worth noting that the BCFIPC approach is non-iterative and requires only one additional calibration step using estimated item difficulties, their standard errors, and the necessary partial derivatives of the log-likelihood function. As a result, the increase in computational complexity relative to FIPC remains modest.

The BCFIPC method in this study addressed DIF in item difficulties. DIF in item discriminations tends to occur less frequently and with a smaller magnitude compared to

DIF in item difficulties [47–50]. Nevertheless, the proposed BCFIPC method can be easily extended to incorporate bias correction for FIPC due to DIF in both item discriminations and item difficulties. This extension requires the estimation of the variance matrix $\boldsymbol{V}_{\text{DIF}}$, which captures the joint DIF effects in item discriminations and difficulties. The bias correction terms for $\widehat{\mu}$ and $\widehat{\sigma}$ depend on $\boldsymbol{V}_{\text{DIF}}$.

The simulation study did not evaluate confidence interval coverage. Standard errors for FIPC can be derived using M-estimation theory [38,51]. The two-step BCFIPC method is also amenable to M-estimation by defining a parameter vector that includes both FIPC and BCFIPC estimates [38]. Alternatively, standard errors may be obtained through resampling techniques such as the jackknife, balanced repeated replication, or bootstrap [44,52,53], as demonstrated in the empirical example in Section 4.

The bias of the FIPC method has been demonstrated in this article only for the 2PL model. However, such bias also arises under the simpler Rasch model [54–57]. Future research could examine the performance of BCFIPC relative to FIPC in the context of the Rasch model. Conversely, FIPC under more complex models such as the three-parameter logistic model [58,59] or alternative IRT models [60–64] could also serve as a basis for bias correction within a BCFIPC method.

The analytical bias correction implemented in BCFIPC could be compared to simulation extrapolation (SIMEX; [65–69]) as an alternative method for bias-corrected FIPC. A key advantage of SIMEX is its fully numerical approach, which replaces the Taylor expansion of the log-likelihood function used in deriving the bias correction. This method resimulates item responses and repeatedly applies the FIPC procedure to the resimulated data.

An anonymous reviewer suggested the use of additional model diagnostic tools [70], such as model fit indices [71] or residual analysis [72,73], for practical applications. However, in operational contexts such as PISA, FIPC is applied under an intentionally misspecified IRT model, and model fit assessment is generally not required [74]. Modifying the IRT model to improve fit would alter the interpretation of the latent variable θ, potentially resulting in a model misaligned with the underlying measurement objective [75–77].

Interestingly, the bias in FIPC due to random DIF is also observed in the Haebara linking [78] and Stocking-Lord linking [79] methods (see [80]). A similar bias-correction approach for these linking methods, based on a Taylor expansion of the linking function, has been proposed in [80]. Importantly, mean-geometric mean (MGM) linking [81] is unbiased in the presence of random DIF because the corresponding second-order derivatives in the linking function with respect to DIF in item difficulties vanish for MGM, but do not vanish for Haebara or Stocking-Lord (see [80]). Similarly, the second-order derivatives of the log-likelihood function with respect to DIF in item difficulties do not vanish in FIPC.

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 in Section 3 simulation study can be found at https://osf.io/wcsdm (accessed on 19 March 2025). The PISA 2006 dataset used in Section 4 can be downloaded from https://www.oecd.org/en/data/datasets/pisa-2006-database.html (accessed on 19 March 2025).

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

Abbreviations

The following abbreviations are used in this manuscript:

2PL	two-parameter logistic
BCFIPC	bias-corrected fixed item parameter calibration
DIF	differential item functioning
FIPC	fixed item parameter calibration
IRF	item response function
IRT	item response theory
LSA	large-scale assessment
MGM	mean-geometric mean
MML	marginal maximum likelihood
PISA	programme for international student assessment
RMSE	root mean square error
SD	standard deviation
SIMEX	simulation extrapolation

Appendix A. Additional Results for the Simulation Study

In this section, additional results for the Simulation Study presented in Section 3 are reported.

Table A1 presents the estimated mean $\widehat{\mu}$ for FIPC and BCFIPC as a function of the DIF SD τ, the true mean μ, the number of items I and sample size N for $\sigma = 0.8$. Table A2 contains the estimated SD $\widehat{\sigma}$ for FIPC and BCFIPC as a function of the DIF SD τ, the true mean μ, the number of items I and sample size N for $\sigma = 0.8$. Table A3 shows the relative RMSE for BCFIPC as a function of the DIF SD τ, the true mean μ, the number of items I and sample size N for $\sigma = 0.8$.

Table A1. Simulation Study: Bias of the estimated mean $\widehat{\mu}$ as a function of the DIF SD τ, the true mean μ, the number of items I and sample size N for $\sigma = 0.8$.

			FIPC for $N =$						**BCFIPC for $N =$**					
τ	μ	I	**125**	**250**	**500**	**1000**	**2000**	**4000**	**125**	**250**	**500**	**1000**	**2000**	**4000**
0	−0.3	15	0.001	0.000	−0.001	0.000	0.000	0.000	0.000	0.000	−0.001	0.000	0.000	0.000
		30	−0.001	0.001	0.001	0.000	0.000	0.000	−0.001	0.001	0.000	0.000	0.000	0.000
		45	0.001	0.002	0.001	0.000	0.001	0.000	0.001	0.002	0.001	0.000	0.001	0.000
	0	15	0.000	−0.002	−0.001	0.000	−0.001	0.000	0.000	−0.002	−0.001	0.000	−0.001	0.000
		30	0.001	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000
		45	−0.001	0.000	−0.001	0.001	0.000	0.000	−0.002	0.000	−0.001	0.001	0.000	0.000
	0.3	15	0.002	0.000	0.000	−0.001	0.000	0.000	0.002	0.000	0.000	−0.001	0.000	0.000
		30	0.000	−0.002	−0.001	0.000	0.000	0.000	0.000	−0.002	−0.001	0.000	0.000	0.000
		45	0.000	0.001	−0.002	0.000	0.000	0.000	0.000	0.001	−0.002	0.000	0.000	0.000
	0.6	15	0.002	−0.001	0.000	0.000	0.001	0.000	0.002	−0.001	0.001	0.000	0.001	0.000
		30	0.001	−0.001	0.001	0.000	0.000	0.000	0.001	−0.001	0.001	0.000	0.000	0.000
		45	0.001	−0.002	−0.001	0.001	0.000	0.000	0.001	−0.002	0.000	0.001	0.000	0.000

Table A1. *Cont.*

			FIPC for $N =$						BCFIPC for $N =$					
τ	μ	I	**125**	**250**	**500**	**1000**	**2000**	**4000**	**125**	**250**	**500**	**1000**	**2000**	**4000**
0.2	−0.3	15	0.005	0.005	0.004	0.003	0.003	0.003	0.002	0.001	0.000	−0.001	−0.001	−0.001
		30	0.005	0.004	0.004	0.005	0.004	0.005	0.001	0.001	0.000	0.001	0.000	0.001
		45	0.001	0.004	0.006	0.004	0.005	0.004	−0.002	0.000	0.002	0.000	0.001	0.000
	0	15	0.004	0.004	0.001	0.002	0.001	0.003	0.002	0.003	−0.001	0.001	−0.001	0.001
		30	−0.002	0.000	0.001	0.003	0.001	0.001	−0.004	−0.001	−0.001	0.001	0.000	0.000
		45	0.000	0.002	0.004	0.002	0.001	0.002	−0.001	0.000	0.002	0.001	−0.001	0.000
	0.3	15	0.003	−0.003	−0.001	0.001	0.000	−0.001	0.004	−0.002	0.000	0.002	0.000	0.000
		30	−0.001	0.001	0.000	0.001	−0.002	0.000	0.000	0.001	0.000	0.002	−0.001	0.001
		45	0.001	−0.002	−0.001	0.001	−0.001	−0.001	0.001	−0.002	−0.001	0.001	0.000	0.000
	0.6	15	−0.002	−0.003	−0.003	−0.004	−0.002	−0.003	0.001	0.000	0.000	−0.001	0.001	0.000
		30	−0.005	−0.003	−0.005	−0.005	−0.003	−0.003	−0.002	−0.001	−0.002	−0.002	0.000	0.000
		45	−0.004	−0.002	−0.004	−0.002	−0.003	−0.003	−0.002	0.001	−0.001	0.001	0.000	−0.001
0.4	−0.3	15	0.016	**0.014**	**0.016**	**0.017**	**0.016**	**0.017**	0.002	−0.001	0.000	0.002	0.001	0.002
		30	0.015	**0.015**	**0.016**	**0.015**	**0.015**	**0.014**	0.000	0.000	0.001	0.000	−0.001	−0.002
		45	0.015	**0.014**	**0.013**	**0.012**	**0.014**	**0.014**	0.000	−0.001	−0.002	−0.004	−0.001	−0.001
	0	15	0.005	0.009	0.008	0.005	0.005	0.004	−0.002	0.002	0.001	−0.002	−0.002	−0.003
		30	0.009	0.008	0.007	**0.010**	0.007	0.009	0.002	0.002	0.000	0.003	0.000	0.002
		45	0.006	0.007	0.009	0.008	0.006	0.007	−0.001	0.001	0.002	0.001	−0.001	0.000
	0.3	15	0.000	−0.004	0.000	−0.002	−0.001	−0.003	0.002	−0.002	0.002	0.000	0.001	−0.001
		30	−0.001	−0.005	−0.002	−0.005	0.000	−0.001	0.001	−0.003	0.000	−0.003	0.002	0.001
		45	−0.003	−0.003	−0.003	−0.002	0.000	0.001	−0.001	−0.001	−0.001	0.000	0.002	0.003
	0.6	15	−0.014	**−0.010**	**−0.014**	−0.009	**−0.010**	**−0.014**	−0.003	0.002	−0.003	0.003	0.001	−0.003
		30	−0.012	**−0.010**	**−0.011**	−0.009	**−0.013**	**−0.010**	−0.001	0.001	0.000	0.002	−0.002	0.001
		45	−0.009	**−0.012**	**−0.011**	**−0.011**	**−0.011**	−0.009	0.002	−0.001	0.000	0.000	0.000	0.002

Note. SD = standard deviation; FIPC = fixed item parameter calibration; BCFIPC = bias-corrected fixed item parameter calibration; Values of absolute bias larger than 0.010 are printed in bold font.

Table A2. Simulation Study: Bias of the estimated SD $\widehat{\sigma}$ as a function of the DIF SD τ, the true mean μ, the number of items I and sample size N for $\sigma = 0.8$.

			FIPC for $N =$						BCFIPC for $N =$					
τ	μ	I	**125**	**250**	**500**	**1000**	**2000**	**4000**	**125**	**250**	**500**	**1000**	**2000**	**4000**
0	−0.3	15	−0.006	−0.004	−0.002	0.000	−0.001	0.000	−0.004	−0.004	−0.001	0.000	−0.001	0.000
		30	−0.003	−0.002	−0.001	0.000	−0.001	0.000	−0.002	−0.002	−0.001	0.000	−0.001	0.000
		45	−0.006	−0.004	−0.002	−0.001	0.000	0.000	−0.005	−0.003	−0.001	−0.001	0.000	0.000
	0	15	−0.005	−0.004	−0.003	0.000	−0.001	0.000	−0.004	−0.003	−0.002	0.000	−0.001	0.000
		30	−0.004	−0.003	0.000	−0.001	0.000	0.000	−0.003	−0.002	0.000	−0.001	0.000	0.000
		45	−0.006	−0.002	0.000	−0.001	−0.001	0.000	−0.005	−0.001	0.000	−0.001	0.000	0.000
	0.3	15	−0.006	−0.004	0.000	0.000	0.000	0.000	−0.004	−0.003	0.000	0.000	0.000	0.000
		30	−0.005	−0.003	0.000	0.000	0.000	0.000	−0.004	−0.002	0.000	0.000	0.000	0.000
		45	−0.005	−0.002	0.000	−0.001	−0.001	0.000	−0.004	−0.002	0.000	−0.001	0.000	0.000
	0.6	15	−0.004	−0.003	−0.002	0.000	−0.001	0.000	−0.002	−0.002	−0.001	0.000	0.000	0.000
		30	−0.006	−0.002	−0.002	−0.001	0.000	0.000	−0.005	−0.002	−0.001	−0.001	0.000	0.000
		45	−0.006	−0.004	−0.001	0.000	0.000	0.000	−0.005	−0.004	−0.001	0.000	0.000	0.000

Table A2. *Cont.*

			FIPC for $N =$						BCFIPC for $N =$					
τ	μ	I	**125**	**250**	**500**	**1000**	**2000**	**4000**	**125**	**250**	**500**	**1000**	**2000**	**4000**
0.2	−0.3	15	**−0.011**	**−0.011**	−0.007	−0.007	−0.008	−0.007	−0.004	−0.003	0.001	0.002	0.001	0.002
		30	**−0.011**	**−0.010**	−0.008	−0.007	−0.007	−0.006	−0.004	−0.004	−0.001	0.000	0.001	0.001
		45	**−0.010**	−0.009	−0.007	−0.006	−0.006	−0.006	−0.004	−0.003	0.000	0.001	0.000	0.000
	0	15	**−0.016**	**−0.011**	−0.009	−0.009	−0.008	−0.008	−0.008	−0.003	0.000	0.000	0.002	0.002
		30	**−0.011**	−0.008	−0.009	−0.008	−0.007	−0.007	−0.005	−0.001	−0.002	0.000	0.001	0.001
		45	**−0.013**	−0.008	−0.009	−0.007	−0.007	−0.006	−0.007	−0.001	−0.002	0.000	0.000	0.001
	0.3	15	**−0.016**	**−0.012**	−0.009	−0.008	−0.008	−0.008	−0.007	−0.003	0.000	0.001	0.001	0.002
		30	**−0.013**	−0.009	−0.009	−0.008	−0.007	−0.007	−0.006	−0.002	−0.001	0.000	0.001	0.001
		45	**−0.013**	−0.009	−0.008	−0.007	−0.007	−0.007	−0.007	−0.002	−0.001	0.001	0.000	0.001
	0.6	15	**−0.016**	**−0.011**	−0.009	−0.009	−0.007	−0.008	−0.007	−0.001	0.001	0.001	0.003	0.003
		30	**−0.013**	**−0.010**	−0.009	−0.008	−0.009	−0.007	−0.005	−0.002	−0.001	0.000	0.000	0.001
		45	**−0.013**	−0.008	−0.008	−0.008	−0.007	−0.007	−0.006	−0.001	−0.001	0.000	0.000	0.001
0.4	−0.3	15	**−0.031**	**−0.030**	**−0.029**	**−0.027**	**−0.027**	**−0.028**	0.003	0.006	0.006	0.008	0.009	0.008
		30	**−0.029**	**−0.027**	**−0.025**	**−0.025**	**−0.025**	**−0.024**	−0.001	0.002	0.004	0.004	0.004	0.004
		45	**−0.030**	**−0.027**	**−0.024**	**−0.023**	**−0.023**	**−0.023**	−0.004	−0.001	0.002	0.003	0.003	0.003
	0	15	**−0.037**	**−0.033**	**−0.030**	**−0.030**	**−0.029**	**−0.030**	0.000	0.005	0.008	0.009	0.009	0.009
		30	**−0.030**	**−0.027**	**−0.027**	**−0.027**	**−0.027**	**−0.026**	0.000	0.003	0.004	0.004	0.004	0.005
		45	**−0.030**	**−0.027**	**−0.026**	**−0.027**	**−0.025**	**−0.025**	−0.002	0.001	0.002	0.002	0.004	0.004
	0.3	15	**−0.038**	**−0.033**	**−0.031**	**−0.030**	**−0.031**	**−0.031**	0.001	0.005	0.008	0.009	0.009	0.009
		30	**−0.034**	**−0.028**	**−0.028**	**−0.028**	**−0.027**	**−0.027**	−0.003	0.004	0.004	0.004	0.005	0.005
		45	**−0.032**	**−0.029**	**−0.027**	**−0.027**	**−0.026**	**−0.027**	−0.003	0.000	0.002	0.002	0.004	0.003
	0.6	15	**−0.037**	**−0.034**	**−0.034**	**−0.029**	**−0.030**	**−0.031**	0.003	0.006	0.007	**0.011**	**0.010**	0.009
		30	**−0.035**	**−0.030**	**−0.029**	**−0.028**	**−0.028**	**−0.028**	−0.003	0.002	0.003	0.004	0.005	0.005
		45	**−0.029**	**−0.030**	**−0.027**	**−0.027**	**−0.027**	**−0.026**	0.000	0.000	0.003	0.003	0.003	0.004

Note. SD = standard deviation; FIPC = fixed item parameter calibration; BCFIPC = bias-corrected fixed item parameter calibration; Values of absolute bias larger than 0.010 are printed in bold font.

Table A3. Simulation Study: Relative root mean square error (RMSE) for bias-corrected fixed item parameter calibration (BCFIPC) as a function of the DIF SD τ, the true mean μ, the number of items I and sample size N for $\sigma = 0.8$.

			$\widehat{\mu}$ for $N =$						$\widehat{\sigma}$ for $N =$					
τ	μ	I	**125**	**250**	**500**	**1000**	**2000**	**4000**	**125**	**250**	**500**	**1000**	**2000**	**4000**
0	−0.3	15	100.1	100.1	100.1	100.0	100.0	100.0	100.1	99.8	100.0	100.0	100.0	100.0
		30	100.1	100.1	100.0	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0
		45	100.0	100.0	100.0	100.0	100.0	100.0	99.9	99.9	100.0	100.0	100.0	100.0
	0	15	100.1	100.1	100.0	100.0	100.0	100.0	99.9	99.9	100.0	100.1	100.0	100.0
		30	100.1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		45	100.1	100.0	100.0	100.0	100.0	100.0	99.9	99.9	100.0	100.0	100.0	100.0
	0.3	15	100.1	100.1	100.0	100.0	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0
		30	100.1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		45	100.1	100.0	100.0	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0
	0.6	15	100.2	100.1	100.0	100.0	100.0	100.0	99.9	99.9	99.9	100.0	100.0	100.0
		30	100.2	100.1	100.0	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0
		45	100.1	100.0	100.0	100.0	100.0	100.0	99.9	99.9	99.9	100.0	100.0	100.0

Table A3. *Cont.*

			$\widehat{\mu}$ for $N =$						$\widehat{\sigma}$ for $N =$					
τ	μ	I	**125**	**250**	**500**	**1000**	**2000**	**4000**	**125**	**250**	**500**	**1000**	**2000**	**4000**
0.2	−0.3	15	100.6	100.5	100.4	100.5	100.5	100.6	99.6	98.9	99.2	98.7	97.0	96.6
		30	100.5	100.4	100.5	100.2	100.3	100.0	99.3	98.4	97.9	97.1	95.5	94.9
		45	100.7	100.4	100.0	100.2	99.9	100.2	99.1	98.7	98.2	97.5	95.2	92.5
	0	15	100.6	100.6	100.8	100.7	100.7	100.6	98.9	98.6	98.5	97.4	97.2	95.5
		30	100.7	100.7	100.7	100.6	100.7	100.7	99.0	98.9	97.5	96.2	94.6	93.2
		45	100.7	100.6	100.5	100.6	100.7	100.6	98.9	99.0	96.8	96.0	93.8	92.0
	0.3	15	100.8	100.7	100.8	100.8	100.8	100.8	99.1	98.5	98.7	97.5	95.9	96.6
		30	100.7	100.7	100.7	100.8	100.7	100.8	98.8	98.8	97.6	95.6	94.6	91.8
		45	100.7	100.7	100.7	100.8	100.8	100.8	98.6	98.5	97.3	96.6	92.7	90.1
	0.6	15	100.7	100.7	100.7	100.7	100.8	100.6	98.9	98.8	98.0	97.4	98.3	96.1
		30	100.7	100.7	100.5	100.4	100.7	100.5	99.4	98.8	97.3	96.1	92.9	92.0
		45	100.6	100.7	100.5	100.7	100.5	100.4	98.7	98.9	97.5	95.4	92.9	90.8
0.4	−0.3	15	101.8	102.0	101.9	101.6	101.6	101.6	95.6	93.6	90.5	89.6	88.7	86.8
		30	101.8	101.2	101.0	100.9	100.9	101.2	94.1	91.2	86.9	82.2	78.2	77.0
		45	101.7	101.3	101.2	101.3	100.4	100.5	92.2	88.1	84.3	78.9	74.7	69.6
	0	15	**102.8**	**102.8**	**102.9**	**103.0**	**102.9**	**103.0**	93.2	92.0	90.7	88.1	86.2	83.9
		30	**102.7**	**102.6**	**102.6**	**102.4**	**102.7**	**102.5**	93.1	91.0	85.0	79.2	74.3	73.3
		45	**102.7**	**102.6**	**102.3**	**102.3**	**102.6**	**102.5**	91.5	87.3	82.3	73.5	71.8	68.1
	0.3	15	**103.2**	**103.1**	**103.2**	**103.2**	**103.2**	**103.3**	93.3	91.3	88.7	87.8	84.6	82.0
		30	**103.2**	**103.1**	**103.1**	**103.1**	**103.2**	**103.1**	91.3	90.1	83.2	77.8	74.1	72.5
		45	**103.1**	**103.1**	**103.1**	**103.1**	**103.2**	**103.3**	90.7	86.1	79.8	72.5	68.4	63.1
	0.6	15	**102.7**	**103.2**	**102.8**	**103.3**	**102.9**	**102.3**	93.3	91.6	87.5	88.1	84.8	82.2
		30	**102.6**	**102.8**	**102.4**	**102.6**	101.7	**102.4**	90.6	88.2	82.4	77.0	74.2	70.0
		45	**102.8**	**102.1**	102.0	101.9	101.8	**102.3**	92.5	85.3	80.4	73.4	67.9	64.8

Note. SD = standard deviation; The fixed item parameter calibration (FIPC) method was used as the reference method to compute the relative RMSE.; RMSE value larger than 102.0 are printed in bold font. RMSE values smaller than 98.0 are highlighted in yellow.

Appendix B. Country Labels for PISA 2006 Mathematics Study

The country labels used in Tables 4–6 are as follows: AUS = Australia; AUT = Austria; BEL = Belgium; CAN = Canada; CHE = Switzerland; CZE = Czech Republic; DEU = Germany; DNK = Denmark; ESP = Spain; EST = Estonia; FIN = Finland; FRA = France; GBR = United Kingdom; GRC = Greece; HUN = Hungary; IRL = Ireland; ISL = Iceland; ITA = Italy; JPN = Japan; KOR = Korea; LUX = Luxembourg; NLD = Netherlands; NOR = Norway; POL = Poland; PRT = Portugal; SWE = Sweden.

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. Bock, R.D.; Gibbons, R.D. *Item Response Theory*; Wiley: New York, NY, USA, 2021. [CrossRef]
3. 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 19 March 2025).
4. Sijtsma, K.; van der Ark, L.A. *Measurement Models for Psychological Attributes*; CRC Press: Boca Raton, FL, USA, 2020. [CrossRef]
5. 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]

6. Yen, W.M.; Fitzpatrick, A.R. Item response theory. In *Educational Measurement*; Brennan, R.L., Ed.; Praeger Publishers: Westport, CT, USA, 2006; pp. 111–154.
7. 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.
8. Aitkin, M. Expectation maximization algorithm and extensions. In *Handbook of Item Response Theory, Volume 2: Statistical Tools*; van der Linden, W.J., Ed.; CRC Press: Boca Raton, FL, USA, 2016; pp. 217–236. [CrossRef]
9. Bock, R.D.; Aitkin, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **1981**, *46*, 443–459. [CrossRef]
10. Glas, C.A.W. Maximum-likelihood estimation. In *Handbook of Item Response Theory, Volume 2: Statistical Tools*; van der Linden, W.J., Ed.; CRC Press: Boca Raton, FL, USA, 2016; pp. 197–216. [CrossRef]
11. Lietz, P.; Cresswell, J.C.; Rust, K.F.; Adams, R.J. (Eds.) *Implementation of Large-Scale Education Assessments*; Wiley: New York, NY, USA, 2017. [CrossRef]
12. Rutkowski, L.; von Davier, M.; Rutkowski, D. (Eds.) *A Handbook of International Large-scale Assessment: Background, Technical Issues, and Methods of Data Analysis*; Chapman Hall/CRC Press: London, UK, 2013. [CrossRef]
13. OECD. *PISA 2006. Technical Report*; OECD: Paris, France, 2009. Available online: https://bit.ly/3xfxdwD (accessed on 19 March 2025).
14. Kim, S. A comparative study of IRT fixed parameter calibration methods. *J. Educ. Meas.* **2006**, *43*, 355–381. [CrossRef]
15. König, C.; Khorramdel, L.; Yamamoto, K.; Frey, A. The benefits of fixed item parameter calibration for parameter accuracy in small sample situations in large-scale assessments. *Educ. Meas.* **2021**, *40*, 17–27. [CrossRef]
16. Mellenbergh, G.J. Item bias and item response theory. *Int. J. Educ. Res.* **1989**, *13*, 127–143. [CrossRef]
17. Meredith, W. Measurement invariance, factor analysis and factorial invariance. *Psychometrika* **1993**, *58*, 525–543. [CrossRef]
18. Millsap, R.E. *Statistical Approaches to Measurement Invariance*; Routledge: New York, NY, USA, 2011. [CrossRef]
19. 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]
20. 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] [PubMed]
21. 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]
22. 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 19 March 2025).
23. Monseur, C.; Sibberns, H.; Hastedt, D. Linking errors in trend estimation for international surveys in education. *IERI Monogr. Ser.* **2008**, *1*, 113–122.
24. Robitzsch, A. Linking error in the 2PL model. *J* **2023**, *6*, 58–84. [CrossRef]
25. 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. [CrossRef]
26. 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]
27. 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]
28. Wu, M. Measurement, sampling, and equating errors in large-scale assessments. *Educ. Meas.* **2010**, *29*, 15–27. [CrossRef]
29. De Boeck, P. Random item IRT models. *Psychometrika* **2008**, *73*, 533–559. [CrossRef]
30. Fox, J.P.; Verhagen, A.J. Random item effects modeling for cross-national survey data. In *Cross-Cultural Analysis: Methods and Applications*; Davidov, E., Schmidt, P., Billiet, J., Eds.; Routledge: London, UK, 2010; pp. 461–482. [CrossRef]
31. de Jong, M.G.; Steenkamp, J.B.E.M.; Fox, J.P. Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *J. Consum. Res.* **2007**, *34*, 260–278. [CrossRef]
32. 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]
33. Robitzsch, A. Bias and linking error in fixed item parameter calibration. *AppliedMath* **2024**, *4*, 1181–1191. [CrossRef]
34. Robitzsch, A. Linking error estimation in fixed item parameter calibration: Theory and application in large-scale assessment studies. *Foundations* **2025**, *5*, 4. [CrossRef]
35. Glas, C.A.W.; Jehangir, M. Modeling country-specific differential functioning. In *A Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*; Rutkowski, L., von Davier, M., Rutkowski, D., Eds.; Chapman Hall/CRC Press: London, UK, 2013; pp. 97–115. [CrossRef]
36. 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]

37. Yuan, K.H.; Cheng, Y.; Patton, J. Information matrices and standard errors for MLEs of item parameters in IRT. *Psychometrika* **2014**, *79*, 232–254. [CrossRef]
38. Boos, D.D.; Stefanski, L.A. *Essential Statistical Inference*; Springer: New York, NY, USA, 2013. [CrossRef]
39. White, H. Maximum likelihood estimation of misspecified models. *Econometrica* **1982**, *50*, 1–25. [CrossRef]
40. Frey, A.; Hartig, J.; Rupp, A.A. An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educ. Meas.* **2009**, *28*, 39–53. [CrossRef]
41. Pokropek, A. Missing by design: Planned missing-data designs in social science. *ASK Res. Meth.* **2011**, *20*, 81–105. Available online: https://tinyurl.com/3px352sy (accessed on 19 March 2025).
42. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2024. Available online: https://www.R-project.org (accessed on 15 June 2024).
43. Robitzsch, A. sirt: Supplementary Item Response Theory Models. R Package Version 4.2-106. 2024. Available online: https://github.com/alexanderrobitzsch/sirt (accessed on 31 December 2024).
44. Kolenikov, S. Resampling variance estimation for complex survey data. *Stata J.* **2010**, *10*, 165–199. [CrossRef]
45. Särndal, C.E.; Swensson, B.; Wretman, J. *Model Assisted Survey Sampling*; Springer: New York, NY, USA, 1992. [CrossRef]
46. OECD. *PISA 2018. Technical Report*; OECD: Paris, France, 2020. Available online: https://bit.ly/3zWbidA (accessed on 19 March 2025).
47. Boer, D.; Hanke, K.; He, J. On detecting systematic measurement error in cross-cultural research: A review and critical reflection on equivalence and invariance tests. *J. Cross-Cult. Psychol.* **2018**, *49*, 713–734. [CrossRef]
48. He, J.; Barrera-Pedemonte, F.; Buchholz, J. Cross-cultural comparability of noncognitive constructs in TIMSS and PISA. *Assess. Educ.* **2019**, *26*, 369–385. [CrossRef]
49. Kankaraš, M.; Moors, G. Analysis of cross-cultural comparability of PISA 2009 scores. *J. Cross-Cult. Psychol.* **2014**, *45*, 381–399. [CrossRef]
50. Rutkowski, L.; Svetina, D. Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educ. Psychol. Meas.* **2014**, *74*, 31–57. [CrossRef]
51. Stefanski, L.A.; Boos, D.D. The calculus of M-estimation. *Am. Stat.* **2002**, *56*, 29–38. [CrossRef]
52. Lohr, S.L. *Sampling: Design and Analysis*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2021. [CrossRef]
53. Lumley, T. *Complex Surveys: A Guide to Analysis Using R*; John Wiley & Sons: New York, NY, USA, 2011. [CrossRef]
54. Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*; Danish Institute for Educational Research: Copenhagen, Denmark, 1960.
55. Bond, T.; Yan, Z.; Heene, M. *Applying the Rasch Model*; Routledge: New York, NY, USA, 2020. [CrossRef]
56. Debelak, R.; Strobl, C.; Zeigenfuse, M.D. *An Introduction to the Rasch Model with Examples in R*; CRC Press: Boca Raton, FL, USA, 2022. [CrossRef]
57. Engelhard, G. *Invariant Measurement*; Routledge: New York, NY, USA, 2012. [CrossRef]
58. De Ayala, R.J. *The Theory and Practice of Item Response Theory*; Guilford Publications: New York, NY, USA, 2022.
59. Lord, F.M. *Applications of Item Response Theory to Practical Testing Problems*; Erlbaum: Hillsdale, NJ, USA, 1980. [CrossRef]
60. Culpepper, S.A. The prevalence and implications of slipping on low-stakes, large-scale assessments. *J. Educ. Behav. Stat.* **2017**, *42*, 706–725. [CrossRef]
61. Falk, C.F.; Cai, L. Semiparametric item response functions in the context of guessing. *J. Educ. Meas.* **2016**, *53*, 229–247. [CrossRef]
62. Feuerstahler, L. Flexible item response modeling in R with the flexmet package. *Psych* **2021**, *3*, 447–478. [CrossRef]
63. Lee, S.; Bolt, D.M. An alternative to the 3PL: Using asymmetric item characteristic curves to address guessing effects. *J. Educ. Meas.* **2018**, *55*, 90–111. [CrossRef]
64. Liao, X.; Bolt, D.M. Item characteristic curve asymmetry: A better way to accommodate slips and guesses than a four-parameter model? *J. Educ. Behav. Stat.* **2021**, *46*, 753–775. [CrossRef]
65. Carroll, R.J.; Küchenhoff, H.; Lombard, F.; Stefanski, L.A. Asymptotics for the SIMEX estimator in nonlinear measurement error models. *J. Am. Stat. Assoc.* **1996**, *91*, 242–250. [CrossRef]
66. Carroll, R.J.; Ruppert, D.; Stefanski, L.A.; Crainiceanu, C.M. *Measurement Error in Nonlinear Models*; CRC Press: Boca Raton, FL, USA, 2006. [CrossRef]
67. Cook, J.R.; Stefanski, L.A. Simulation-extrapolation estimation in parametric measurement error models. *J. Am. Stat. Assoc.* **1994**, *89*, 1314–1328. [CrossRef]
68. Robitzsch, A. SIMEX-based and analytical bias corrections in Stocking-Lord linking. *Analytics* **2024**, *3*, 368–388. [CrossRef]
69. Robitzsch, A. Implementation aspects in simulation extrapolation-based Stocking–Lord linking. *Appl. Sci.* **2025**, *15*, 901. [CrossRef]
70. Levy, R. The rise of Markov chain Monte Carlo estimation for psychometric modeling. *J. Probab. Stat.* **2009**, *2009*, 537139. [CrossRef]

71. Levy, R.; Mislevy, R.J.; Sinharay, S. Posterior predictive model checking for multidimensionality in item response theory. *Appl. Psychol. Meas.* **2009**, *33*, 519–537. [CrossRef]
72. Haberman, S.J.; Sinharay, S.; Chon, K.H. Assessing item fit for unidimensional item response theory models using residuals from estimated item response functions. *Psychometrika* **2013**, *78*, 417–440. [CrossRef]
73. van Rijn, P.W.; Sinharay, S.; Haberman, S.J.; Johnson, M.S. Assessment of fit of item response theory models used in large-scale educational survey assessments. *Large-Scale Assess. Educ.* **2016**, *4*, 10. [CrossRef]
74. 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]
75. Brennan, R.L. Misconceptions at the intersection of measurement theory and practice. *Educ. Meas.* **1998**, *17*, 5–9. [CrossRef]
76. Camilli, G. IRT scoring and test blueprint fidelity. *Appl. Psychol. Meas.* **2018**, *42*, 393–400. [CrossRef]
77. Chiu, T.W.; Camilli, G. Comment on 3PL IRT adjustment for guessing. *Appl. Psychol. Meas.* **2013**, *37*, 76–86. [CrossRef]
78. Haebara, T. Equating logistic ability scales by a weighted least squares method. *Jpn. Psychol. Res.* **1980**, *22*, 144–149. [CrossRef]
79. Stocking, M.L.; Lord, F.M. Developing a common metric in item response theory. *Appl. Psychol. Meas.* **1983**, *7*, 201–210. [CrossRef]
80. Robitzsch, A. Bias-reduced Haebara and Stocking-Lord linking. *J* **2024**, *7*, 373–384. [CrossRef]
81. Kolen, M.J.; Brennan, R.L. *Test Equating, Scaling, and Linking*; Springer: New York, NY, USA, 2014. [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.