foundations

MDPI

Article

Linking Error Estimation in Fixed Item Parameter Calibration: Theory and Application in Large-Scale Assessment Studies

Alexander Robitzsch [1,2]

[1] IPN—Leibniz Institute for Science and Mathematics Education (Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik), Olshausenstraße 62, 24118 Kiel, Germany; robitzsch@leibniz-ipn.de
[2] Centre for International Student Assessment (ZIB-Zentrum für Internationale Bildungsvergleichsstudien), Olshausenstraße 62, 24118 Kiel, Germany

Abstract: In fixed item parameter calibration (FIPC), an item response theory (IRT) model is estimated with item parameters fixed at reference values to estimate the distribution parameters within a specific group. The presence of random differential item functioning (DIF) within this group introduces additional variability in the distribution parameter estimates, which is captured by the linking error (LE). Conventional LE estimates, based on item jackknife methods, are subject to positive bias due to sampling errors. To address this, this article introduces a bias-corrected LE estimate. Moreover, the use of statistical inference is examined using the newly proposed bias-corrected total error, which includes both the sampling error and LE. The proposed error estimates were evaluated through a simulation study, and their application is illustrated using PISA 2006 data for the reading domain.

Keywords: fixed item parameter calibration; linking error; differential item functioning; item response model; resampling

MSC: 62H25; 62P25

check for updates

Academic Editor: Luigi Rodino

Received: 25 November 2024
Revised: 20 January 2025
Accepted: 6 February 2025
Published: 11 February 2025

Citation: Robitzsch, A. Linking Error Estimation in Fixed Item Parameter Calibration: Theory and Application in Large-Scale Assessment Studies. *Foundations* **2025**, *5*, 4. https://doi.org/10.3390/foundations5010004

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] are statistical models for analyzing multivariate binary (i.e., dichotomous), polytomous, or continuous data. These models are widely applied in the social sciences, primarily to reduce the dataset dimensionality to one or a few interpretable factors that provide a meaningful summary of the data.

In this article, we only focus on IRT models for dichotomous items. Consider a vector, $\boldsymbol{X} = (X_1, \ldots, X_I)$, of $I > 1$ dichotomous (i.e., binary) random variables, $X_i \in \{0, 1\}$, commonly referred to as (scored) items or (scored) item responses. A unidimensional IRT model [5] specifies the probability distribution $\mathsf{P}(\boldsymbol{X} = \boldsymbol{x})$ for $\boldsymbol{x} = (x_1, \ldots, x_I) \in \{0, 1\}^I$ as the following parametrized statistical model:

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

where $f(\theta; \mu, \sigma)$ represents the probability density function of a normal distribution with the mean μ and standard deviation (SD) σ. Both parameters are collected in the vector of the distribution parameters $\delta = (\mu, \sigma)$ for the latent variable θ, which is also called the latent factor, trait, or ability. The vector $\gamma = (\gamma_1, \ldots, \gamma_I)$ includes all estimated item

parameters for the item response functions (IRFs), $P_i(\theta;\gamma_i) = \mathsf{P}(X_i = 1|\theta)$ $(i = 1,\ldots,I)$. In the two-parameter logistic (2PL) model [6], AN IRF takes the form

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

where a_i represents the item discrimination and b_i the item difficulty for item i (i.e., $\gamma_i = (a_i, b_i)$). The logistic distribution function Ψ is defined as $\Psi(x) = (1 + \exp(-x))^{-1}$. Using this link function comes with the advantage that the weighted sum score $\sum_{i=1}^{I} a_i X_i$ is a sufficient statistic for the factor variable θ.

The model parameters in the IRT model (1) are typically estimated using marginal maximum likelihood estimation (MML [7]). Since no closed-form solutions exist for δ and γ, MML estimation relies on iterative procedures. To distinguish ability characteristics from item properties, additional identification constraints are typically applied to the model (1), ensuring proper separation between the distribution parameters, δ, and the item parameters, γ [8]. For example, to separate distribution parameters from item parameters, the mean μ could be fixed at 0, and the SD σ could be set to 1 to identify the item parameters in the 2PL model.

IRT models are often applied to compare the distribution of a specific group, such as a country, in a test (i.e., a set of items) against a reference distribution, focusing on the latent variable θ within the IRT model (1). Fixed item parameter calibration (FIPC [9–12]) is a widely employed approach, which assumes item parameters are known (or estimated from the total sample), leaving only the distribution parameters to be estimated in the IRT model. As emphasized by an anonymous reviewer, FIPC is also commonly used to calibrate new items on an existing scale.

In educational large-scale assessment (LSA) studies [13], such as the programme for international student assessment (PISA; [14]), FIPC is used to estimate country means and country SDs. Fixed item parameters are obtained from a pooled sample that includes students from all participating countries in the study. These international item parameters are then treated as fixed item parameters in FIPC applied to compute scores of the ability variable θ within a specific country.

When applied with the 2PL model, the FIPC method is expected to yield consistent estimates for the distribution parameters, provided the item parameters also hold for the group under investigation. However, item parameters may differ across groups, a phenomenon known as differential item functioning (DIF [15,16]) in the literature.

The presence of random DIF [17–19] introduces additional variability into the estimates of the (country) mean μ and SD σ [20,21]. As a consequence, the estimated distribution parameters depend on the choice of selected items, even in infinite sample sizes of persons. This variability has been termed the linking error (LE [19,21–25]).

Note that linking errors were initially proposed for trend estimates involving linking across two LSA studies [21,22]. However, the concept of a linking error similarly applies to cross-sectional comparisons [21,25,26].

When applying FIPC to a specific group, such as a country, a linking error arises because the group mean or group SD depends on the item selection in the presence of random DIF. If no random DIF were present, the linking error would be zero. Thus, linking errors differ conceptually from sampling errors (SEs), which result from the sampling or selection of persons. Consequently, FIPC for a group involves both sources of uncertainty: LEs and SEs. This article introduces a newly proposed unbiased estimate of linking errors in FIPC.

In the Rasch model, simple formulas based on variance component analysis are available for calculating LEs [21,22]. For more complex models, resampling methods [27,28]

such as jackknife [21,22,29], (balanced) half-sampling [19], or bootstrap [30] methods can be used.

The jackknife approach involves reestimating the IRT model by systematically leaving out one item, i, at a time. This yields slightly altered estimates, $\widehat{\mu}_{(-i)}$ and $\widehat{\sigma}_{(-i)}$, compared to the estimates $\widehat{\mu}$ and $\widehat{\sigma}$ based on all items. The jackknife linking error for the estimated mean $\widehat{\mu}$ is defined as [22]:

$$\mathrm{LE}(\widehat{\mu}) = \sqrt{\frac{I-1}{I}\sum_{i=1}^{I}(\widehat{\mu}_{(-i)} - \widehat{\mu})^2}. \tag{3}$$

However, the jackknife LE estimate (3) is also susceptible to sampling errors due to the sampling of subjects, as reflected in SEs. This issue introduces a positive bias in the conventional LE estimate (3). The current article addresses this limitation by proposing a bias-corrected LE estimate, which eliminates the portion of the variance in the jackknife LE estimate that is attributable to sampling errors.

The overall uncertainty in the estimated distribution parameters $\widehat{\mu}$ and $\widehat{\sigma}$ is captured by the total errors (TEs [22,26]), which include both SEs and LEs. The statistical inference for $\widehat{\mu}$ and $\widehat{\sigma}$ in this article is based on these TEs.

Additionally, this article emphasizes the application of LE estimation methods in educational large-scale assessment studies [13] that involve the stratified clustered sampling of students. In such studies, the assumption of subject independence cannot be made when deriving estimators. To address this, the proposed LE estimation techniques were adapted for complex subject sampling by employing replication methods.

The rest of this article is organized as follows. Section 2 outlines LE estimation for the independent sampling of subjects. The section formally introduces the LE for FIPC, followed by LE estimation based on jackknifing items. As this approach is susceptible to sampling errors, a newly proposed bias-corrected LE estimate is presented. Finally, the two sources of error—items and subjects—are combined into the total error, capturing the overall uncertainty in the estimated distribution parameters in FIPC. Section 3 shows how we adapted these methods to handle stratified clustered sampling using replication techniques. The uncertainty regarding the sampling of persons was estimated using a replication method that involved repeated applications of FIPC. Section 4 presents the findings from a simulation study that examined the accuracy of LE and TE estimates and evaluated the coverage rates for estimated distribution parameters. Section 5 demonstrates the LE estimation approaches through an empirical example involving PISA 2006 data. Finally, the article closes with a discussion in Section 6 and a conclusion in Section 7.

2. Linking Error Estimation for Independent Sampling

Let $\delta = (\mu, \sigma)$ denote the vector that contains the mean μ and the SD σ. Let $\gamma = (\gamma_1, \ldots, \gamma_I)$ represent the vector of the item parameters, γ_i ($i = 1, \ldots, I$). The true distribution parameter and item parameters are denoted by δ_0 and γ, respectively. In the computation of $\widehat{\gamma}$ in the scaling model, the item parameters, γ, are fixed at $\gamma^* = (\gamma_1^*, \ldots, \gamma_I^*)$. The difference $\boldsymbol{e} = \gamma - \gamma^* = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_I)$ captures all DIF effects and represents the misspecification of the IRT model. When the scaling model incorporates the item responses of students from a certain country, the vector γ^* reflects the international item parameters, and the DIF effects, $\boldsymbol{e}$, quantify country-specific DIF [31]. This article assumes random DIF effects with zero means, expressed as

$$\mathrm{E}(\boldsymbol{e}_i) = \mathrm{E}(\gamma_i - \gamma_i^*) = 0. \tag{4}$$

Let $l(\delta, \gamma)$ be the log likelihood function. Define the partial derivatives $l_\delta = (\partial l)/(\partial \delta)$, $l_{\delta\delta} = (\partial^2 l)/(\partial \delta \partial \delta^\top)$, $l_{\delta\gamma_i} = (\partial^2 l)/(\partial \delta \partial \gamma_i)$, and $l_{\delta\gamma_i\gamma_i} = (\partial^3 l)/(\partial \delta \partial \gamma_i \partial \gamma_i^\top)$ for $i = 1, \ldots, I$. In FIPC, the parameter estimate $\widehat{\delta} = (\widehat{\mu}, \widehat{\sigma})$ of the distribution parameter is determined by

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

where γ^* is the vector of the fixed item parameters. Equivalently, the estimate $\widehat{\delta}$ satisfies the estimating equation

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

2.1. Standard Error

The SE of $\widehat{\delta}$ can be obtained from the inverse of the observed information matrix. In detail, the variance matrix $\boldsymbol{V}_{\mathsf{SE}}$ due to sampling errors is given by

$$\boldsymbol{V}_{\mathsf{SE}} = \left(l_{\delta\delta}(\widehat{\delta}, \gamma^*) \right)^{-1}. \tag{7}$$

2.2. Bias

The bias of $\widehat{\delta}$ in FIPC is now derived (see [12]). The derivation requires that a second-order Taylor expansion of l_δ around (δ_0, γ) is carried out. Assuming the independence of items, we obtain

$$l_\delta(\widehat{\delta}, \gamma^*) = l_\delta(\delta_0, \gamma) + l_{\delta\delta}(\delta_0, \gamma)(\widehat{\delta} - \delta) + \sum_{i=1}^{I} l_{\delta\gamma_i}(\delta_0, \gamma)(\gamma_i^* - \gamma_i) + \frac{1}{2} \sum_{i=1}^{I} (\gamma_i^* - \gamma_i)^\top l_{\delta\gamma_i\gamma_i}(\delta_0, \gamma)(\gamma_i^* - \gamma_i). \tag{8}$$

The identity $l_\delta(\delta_0, \gamma) = \mathbf{0}$ holds because δ_0 and γ are the true parameters. Utilizing this identity along with (4) and (6), the expected bias of $\widehat{\delta}$ can be expressed as

$$\mathsf{Bias}(\widehat{\delta}) = \mathsf{E}(\widehat{\delta} - \delta) = -\frac{1}{2} l_{\delta\delta}(\delta_0, \gamma)^{-1} \sum_{i=1}^{I} \mathsf{E}\left[(\gamma_i^* - \gamma_i)^\top l_{\delta\gamma_i\gamma_i}(\delta_0, \gamma)(\gamma_i^* - \gamma_i) \right]. \tag{9}$$

Equation (9) illustrates that the variance due to DIF $\gamma_i^* - \gamma_i$ can induce bias in FIPC estimation. It should be noted that the bias of $\widehat{\delta}$ is independent of the number of items I and, therefore, cannot be reduced by increasing the number of items. Specifically, it has been shown in [12] that the SD estimate $\widehat{\sigma}$ is typically more biased than the mean estimate $\widehat{\mu}$. Notably, similar biases are also observed in Haebara linking [32], which relies on the alignment of IRFs in the linking process [33].

2.3. Linking Error

The linking error of $\widehat{\delta}$, representing the variance in the distribution parameter due to random DIF, is now derived (see [34]). By neglecting the terms involving $l_{\delta\gamma_i\gamma_i}$ in (8), the following expression is obtained

$$\widehat{\delta} - \delta = -\sum_{i=1}^{I} l_{\delta\delta}(\delta_0, \gamma)^{-1} l_{\delta\gamma_i}(\delta_0, \gamma)(\gamma_i^* - \gamma_i). \tag{10}$$

The variance matrix $\boldsymbol{V}_{\mathrm{LE}}$ due to linking errors can be determined as (see [12] for a similar derivation)

$$\boldsymbol{V}_{\mathrm{LE}} = \mathsf{Var}(\widehat{\delta} - \delta) = \sum_{i=1}^{I} \boldsymbol{A}_i \mathsf{Var}(\gamma_i - \gamma_i^*) \boldsymbol{A}_i^\top, \text{ where } \boldsymbol{A}_i = l_{\delta\delta}(\delta_0, \gamma)^{-1} l_{\delta\gamma_i}(\delta_0, \gamma), \tag{11}$$

and $\mathrm{Var}(\gamma_i - \gamma_i^*)$ is the variance matrix of the random DIF effects. This variance matrix can also be approximated empirically by relying on the estimated differences $\widehat{\gamma}_i - \gamma_i^*$, provided that the estimates, $\widehat{\gamma}_i$, of the true item parameters are available. However, these estimates are not readily available. This article focuses on statistical inference for the linking error based on the jackknife method.

2.4. Jackknife Linking Error

The jackknife LE involves the repeated maximization of the log likelihood function, excluding one item at a time. Let $\widehat{\delta}_{(-i)}$ denote the distribution parameter estimate if item i is removed from the dataset. The jackknife estimate of the variance matrix due to linking errors is given by

$$\boldsymbol{V}_{\mathrm{LE}} = \frac{I-1}{I} \sum_{i=1}^{I} (\widehat{\delta}_{(-i)} - \widehat{\delta})(\widehat{\delta}_{(-i)} - \widehat{\delta})^{\top}. \tag{12}$$

Neglecting the terms involving $l_{\delta\gamma_i\gamma_i}$ in (8), the estimate $\widehat{\delta}_{(-i)}$ approximately satisfies

$$\widehat{\delta}_{(-i)} - \delta = -\sum_{j \neq i} \boldsymbol{A}_j (\gamma_j^* - \gamma_j). \tag{13}$$

Moreover, (10) can be rewritten as

$$\widehat{\delta} - \delta = -\boldsymbol{A}_i (\gamma_i^* - \gamma_i) - \sum_{j \neq i} \boldsymbol{A}_j (\gamma_j^* - \gamma_j). \tag{14}$$

Subtracting (14) from (13), the deviations $\widehat{\delta}_{(-i)} - \widehat{\delta}$ appearing in the jackknife LE (12) can be approximated as

$$\widehat{\delta}_{(-i)} - \widehat{\delta} = \boldsymbol{A}_i (\gamma_i^* - \gamma_i). \tag{15}$$

Substituting this approximation into (12), the jackknife LE can be reformulated as

$$\boldsymbol{V}_{\mathrm{LE}} = \frac{I-1}{I} \sum_{i=1}^{I} \boldsymbol{A}_i (\gamma_i^* - \gamma_i)(\gamma_i^* - \gamma_i)^{\top} \boldsymbol{A}_i^{\top}. \tag{16}$$

2.5. Bias-Corrected Linking Error

Previous research has demonstrated that conventional linking error (LE) estimates are susceptible to sampling errors, as discussed in [26,35]. Bias-corrected LE estimates offer a potential solution to mitigate this issue. However, the deviation, $\widehat{\delta}_{(-i)} - \widehat{\delta}$, in the jackknife LE (see (12)) is also affected by sampling errors. These errors introduce additional variation, resulting in a positive bias in the LE estimate.

The derivation in (16) illustrates how sampling errors contribute to the deviation $\widehat{\delta}_{(-i)} - \widehat{\delta}$. Formally, this can be expressed as

$$\widehat{\gamma}_i - \gamma_i^* = (\widehat{\gamma}_i - \gamma_i) + (\gamma_i - \gamma_i^*). \tag{17}$$

In (17), the second term $(\gamma_i - \gamma_i^*)$ on the right-hand side represents the true variation associated with the LE, while the first term $(\widehat{\gamma}_i - \gamma_i)$ corresponds to sampling errors. To address this, a computational shortcut was introduced to estimate the variance contribution from the LE variance estimate.

The distribution parameter estimate $\widehat{\delta}_{(-i)}$, obtained by omitting item i from the log likelihood function, can be approximated using an alternative model. In this model, the distribution parameters, δ, are estimated alongside the item parameters for item i, while the parameters for all other items are fixed at their values in γ^*. The resulting item

parameter vector is denoted as $\widehat{\gamma}^*_{(+i)}$, which includes the entries from γ^* for all items except item i and the estimated item parameters for item i.

The estimation of this model yields the quantities $l_{\delta\delta}(\widehat{\delta}_{(-i)}, \widehat{\gamma}^*_{(+i)})$, $l_{\delta\gamma_i}(\widehat{\delta}_{(-i)}, \widehat{\gamma}^*_{(+i)})$, and $l_{\gamma_i\gamma_i}(\widehat{\delta}_{(-i)}, \widehat{\gamma}^*_{(+i)})$. The term represents the part of the observed information matrix associated with the item parameter estimate $\widehat{\gamma}_i$ for item i. The quantity $l_{\gamma_i\gamma_i}(\widehat{\delta}_{(-i)}, \widehat{\gamma}^*_{(+i)})^{-1}$ is an estimate of the variance matrix of the item parameters, $\widehat{\gamma}^*_{(+i)}$, associated with item i. The three quantities mentioned above can be utilized to estimate the bias contribution in $\boldsymbol{V}_{\text{LE}}$ in (16).

$$\mathcal{V} = \frac{I-1}{I} \sum_{i=1}^{I} \widehat{\boldsymbol{A}}_i \left(l_{\gamma_i\gamma_i}(\widehat{\delta}_{(-i)}, \widehat{\gamma}^*_{(+i)}) \right)^{-1} \widehat{\boldsymbol{A}}_i^\top, \text{ where} \tag{18}$$

$$\widehat{\boldsymbol{A}}_i = l_{\delta\delta}(\widehat{\delta}_{(-i)}, \widehat{\gamma}^*_{(+i)})^{-1} l_{\delta\gamma_i}(\widehat{\delta}_{(-i)}, \widehat{\gamma}^*_{(+i)}). \tag{19}$$

Based on this, a bias-corrected estimate of the LE variance matrix is defined as

$$\boldsymbol{V}_{\text{LE}_{\text{bc}}} = \boldsymbol{V}_{\text{LE}} - \mathcal{V}. \tag{20}$$

The LE estimates for the individual parameters $\widehat{\mu}$ and $\widehat{\sigma}$ from $\widehat{\delta}$ are obtained by taking the square root of the diagonal elements in $\boldsymbol{V}_{\text{LE}_{\text{bc}}}$. If any negative diagonal entries appear in $\boldsymbol{V}_{\text{LE}_{\text{bc}}}$, the corresponding LE estimate is set to 0. In the absence of random DIF effects, the linking error variance $\boldsymbol{V}_{\text{LE}}$ should be zero at the population level. Since the bias-corrected variance matrix $\boldsymbol{V}_{\text{LE}_{\text{bc}}}$ is designed to estimate the true linking error variance matrix, negative diagonal entries in $\boldsymbol{V}_{\text{LE}_{\text{bc}}}$ are expected in no DIF situations.

It is important to note that the derivation of the bias-corrected LE estimate assumes that DIF effects on the items are independent. This assumption is necessary for both the jackknife LE (Section 2.4) and the determination of the variance inflation in the jackknife LE estimate due to sampling errors (see Section 2.5). If multiple items are associated with a single-item stimulus, such as a reading text, the independence assumption of DIF effects may be violated.

2.6. Total Error

The total error (TE) encompasses both sources of uncertainty: the SE due to randomness due to persons and the LE due to randomness (or random DIF) in items [19,26,35,36]. The conventional estimate of the variance matrix for the TE, based on estimates of $\boldsymbol{V}_{\text{SE}}$ and $\boldsymbol{V}_{\text{LE}}$, is given by

$$\boldsymbol{V}_{\text{TE}} = \boldsymbol{V}_{\text{SE}} + \boldsymbol{V}_{\text{LE}}. \tag{21}$$

A bias-corrected variance estimate of the TE, based on estimates of $\boldsymbol{V}_{\text{SE}}$ and $\boldsymbol{V}_{\text{LE}_{\text{bc}}}$, is given by

$$\boldsymbol{V}_{\text{TE}_{\text{bc}}} = \boldsymbol{V}_{\text{SE}} + \boldsymbol{V}_{\text{LE}_{\text{bc}}}. \tag{22}$$

3. Linking Error Estimation Based on Resampling Methods

In educational LSA studies [13] like the PISA [14] or the trends in international mathematics and science study (TIMSS [37]), statistical inference is typically conducted using a repeated replication methodology for accounting for stratified clustered sampling within countries [28]. The rth replication sample uses a modified set of person sampling weights, $w_p^{(r)}$. For example, the PISA employs $R = 80$ replication samples to perform statistical inference for a parameter of interest (i.e., $\widehat{\boldsymbol{\beta}}$) based on student weights, w_p. In each replica-

tion sample, the analysis is repeated using sampling weights, $w_p^{(r)}$, resulting in parameter estimates, $\widehat{\boldsymbol{\beta}}^{(r)}$. The variance matrix $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}$ for $\widehat{\boldsymbol{\beta}}$ is then calculated as [28]

$$\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = A \sum_{r=1}^{R} (\widehat{\boldsymbol{\beta}}^{(r)} - \widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}^{(r)} - \widehat{\boldsymbol{\beta}})^{\top}, \tag{23}$$

where the scaling factor A depends on the replication method used. In the PISA, which uses balanced repeated replication, the factor A, also referred to as the Fay factor, equals 0.05 [14]. An SE for an individual parameter is obtained by taking the square root from the corresponding diagonal entries in $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}$.

In the following, the error estimation methods presented in Section 2 are adapted to the situation using a replication method.

First, variance estimates, $\boldsymbol{V}_{\mathsf{SE}}$, for δ, as presented in Section 2.1, due to sampling errors (i.e., SEs) based on the observed information matrix must be replaced with the repeated variance estimation using replication samples. In this case, FIPC is reestimated using different sampling weights in each of the R replication samples, yielding $\widehat{\delta}^{(r)}$ distribution parameter estimates for $r = 1, \ldots, R$. These estimates are inserted into (23) to compute $\boldsymbol{V}_{\mathsf{SE}}$ in the case of resampling methods.

Second, the computation of the jackknife LE variance matrix $\boldsymbol{V}_{\mathrm{LE}}$, as presented in Section 2.4, does not require modification for replication methods, as the items are still regarded as independent.

Third, the bias correction term $\mathcal{V}$ in the LE variance matrix, as presented in Section 2.5, involves the observed information matrix $l_{\gamma_i\gamma_i}(\widehat{\delta}_{(-i)}, \widehat{\gamma}^*_{(+i)})$ in (18) for item i. The model freely estimates the distribution parameter $\widehat{\delta}_{(-i)}$ and the item parameter vector $\widehat{\gamma}^*_{(+i)}$, where only parameters from item i are freely estimated while the item parameters of the other items are fixed. This model is reestimated for all the replication samples, providing a variance matrix for the item parameters involving item i, which substitutes the matrix $(l_{\gamma_i\gamma_i}(\widehat{\delta}_{(-i)}, \widehat{\gamma}^*_{(+i)}))^{-1}$ in (18).

Fourth, no changes are required in the formulas for computing the TE and the bias-corrected TE, as outlined in Section 2.6. The modifications related to the sampling method apply solely to the defining matrices $\boldsymbol{V}_{\mathsf{SE}}$ and $\boldsymbol{V}_{\mathrm{LE}_{\mathrm{bc}}}$.

4. Simulation Study

4.1. Method

In this simulation study, item responses were generated for a single group of subjects. The data-generating model followed the 2PL IRFs. The ability variable θ was assumed to follow a normal distribution with a mean of $\mu = 0.3$ and SD of $\sigma = 1.2$. In the context of an LSA study, such as the PISA, these values refer to a well-performing country that is above average.

The simulation study was conducted for $I = 20$ and $I = 40$ items. The item parameters for each replication were based on fixed base parameters used in the FIPC scaling method, along with newly simulated random DIF effects in terms of item difficulties and discriminations. The same set of 10 base item parameters from [23,26] was used. For the $I = 20$ and $I = 40$ item conditions, the item parameters were duplicated two and four times, respectively. The base item discrimination values, a_i^*, were set to 0.73, 1.25, 1.20, 1.47, 0.97, 1.38, 1.05, 1.14, 1.15, and 0.67, resulting in a mean of $M = 1.101$ and an SD of 0.257 of the item discriminations. The base item difficulties, b_i^*, were set to -1.31, 1.44, -1.20, 0.10, 0.10, -0.74, 1.48, -0.61, 0.82, and -0.07, with a mean item of $M = 0.001$ and an SD of 1.002. The item parameters can also be retrieved from https://osf.io/hb3ck (accessed

on 25 November 2024). The distribution of item parameters reflected typical values found in educational LSA studies.

The item parameters used in the data-generating 2PL model were defined as

$$a_i = a_i^* \exp(f_i) \text{ and } b_i = b_i^* + e_i, \tag{24}$$

where e_i and f_i are normally distributed DIF effects with zero means and SDs of τ and $0.35 \times \tau$, respectively. The uniform DIF effect e_i and the nonuniform DIF effect f_i were assumed to be uncorrelated. Two DIF conditions were examined: $\tau = 0$, indicating no DIF, and $\tau = 0.3$, representing moderate DIF. A DIF SD of $\tau = 0.3$ could be considered an average country DIF observed in the PISA study [21,24].

Item responses were generated for sample sizes, N, of 500, 1000, and 2000, reflecting typical applications of the 2PL model in large-scale assessment studies [13].

FIPC based on the 2PL model was applied to the simulated dataset, with item parameters fixed at the values of the base item parameters. Subsequently, the different SE, LE and TE estimators described in Section 2 were calculated.

In each of the 3 (sample size of N) $\times$ 2 (DIF SD of τ) $\times$ 2 (number of items of I) = 12 cells of the simulation, 5000 replications were conducted. The empirical bias and the empirical SD of the estimates $\widehat{\mu}$ and $\widehat{\sigma}$ were calculated. We evaluated the coverage rate for $\widehat{\mu}$ and $\widehat{\sigma}$ at a 95% confidence level for the TE and the bias-corrected TE, using the normal distribution to calculate the percentage of instances in which the confidence interval included the true values $\mu = 0.3$ or $\sigma = 1.2$, respectively. Additionally, we computed the median of the LE and the bias-corrected LE, as well as the median of the TE and the bias-corrected TE.

All analyses for this simulation study were conducted using the open-source statistical software R (Version 4.4.1 [38]). The 2PL model and its SE estimates were obtained using the `sirt::xxirt()` function from the R package sirt (Version 4.2-89 [39]). A custom R function was developed to compute the different LE and TE estimates. This function, along with replication material for this simulation study, is available at https://osf.io/hb3ck (accessed on 25 November 2024).

4.2. Results

Table 1 presents the bias, SD, median TE and LE error estimates, and coverage rates for the estimated mean $\widehat{\mu}$ and the estimated SD $\widehat{\sigma}$ as a function of the DIF SD τ, the number of items I, and the sample size N. The distribution parameters were unbiased in the no-DIF condition with $\tau = 0$. However, $\widehat{\mu}$ exhibited a small bias in the DIF condition with $\tau = 0.3$, while the bias for the estimated SD $\widehat{\sigma}$ was more pronounced. Note that the bias in $\widehat{\sigma}$ was independent of the sample size and the number of items in the condition with a DIF SD of $\tau = 0.3$. As expected, the SD of the estimated $\widehat{\mu}$ and $\widehat{\sigma}$ parameters decreased with an increasing sample size.

The median bias-corrected TE (i.e., TE_{bc}) more closely aligned with the empirical SDs of the estimates compared to the conventional TE. This was also reflected in the error ratio (ER), computed as the ratio of the corresponding TE and SD. The mean ER for the TE was 1.073 (ranging between 0.963 and 1.208), while it was 1.004 for the TE_{bc} (ranging between 0.945 and 1.097), highlighting that the TE_{bc} better reflected the variability of the distribution parameter estimated.

The conventional LE showed a positive bias in the no-DIF condition with $\tau = 0$, particularly with smaller sample sizes. In contrast, the bias-corrected LE (i.e., LE_{bc}) aligned closely with the expected value of 0 for $\tau = 0$, except in some cells for the $\widehat{\sigma}$ estimate. Coverage rates based on the TE_{bc} outperformed those for the TE in the no-DIF condition, where the TE led to overcoverage. However, slight undercoverage was observed for the

TE_{bc} in the DIF condition with $\tau = 0.3$. However, undercoverage for $\widehat{\sigma}$ was also expected as $\widehat{\sigma}$ exhibited slight bias in the DIF condition.

Table 1. Simulation study: the bias, standard deviation (SD), median error estimates and coverage rates for the estimated mean $\widehat{\mu}$ and the estimated SD $\widehat{\sigma}$ as a function of the DIF SD τ, the number of items I, and the sample size N.

Par	τ	I	N	Bias	SD	Median Error Estimate				Coverage Rate	
						TE	TE_{bc}	LE	LE_{bc}	TE	TE_{bc}
$\widehat{\mu}$	0	20	500	0.000	0.058	0.062	0.059	0.022	0.000	96.1	95.0
			1000	−0.001	0.041	0.044	0.041	0.014	0.000	96.3	95.1
			2000	0.000	0.029	0.031	0.029	0.011	0.000	96.5	95.3
		40	500	0.001	0.056	0.058	0.056	0.016	0.000	96.3	95.6
			1000	0.000	0.040	0.041	0.040	0.009	0.000	95.6	94.7
			2000	0.000	0.028	0.029	0.028	0.008	0.000	96.3	95.6
	0.3	20	500	−0.004	0.091	0.092	0.088	0.071	0.067	94.8	93.9
			1000	−0.004	0.083	0.081	0.079	0.070	0.068	93.4	92.8
			2000	−0.007	0.076	0.076	0.075	0.070	0.069	93.9	93.4
		40	500	−0.004	0.074	0.075	0.073	0.051	0.048	95.5	95.0
			1000	−0.006	0.064	0.062	0.061	0.048	0.046	94.0	93.6
			2000	−0.006	0.057	0.057	0.057	0.050	0.049	95.0	94.8
$\widehat{\sigma}$	0	20	500	−0.001	0.049	0.058	0.050	0.031	0.000	98.1	95.6
			1000	0.000	0.034	0.041	0.035	0.022	0.000	98.1	96.0
			2000	0.000	0.025	0.030	0.025	0.017	0.007	97.8	95.5
		40	500	−0.002	0.044	0.048	0.044	0.020	0.000	96.5	94.8
			1000	0.000	0.031	0.034	0.031	0.015	0.001	97.4	95.6
			2000	0.000	0.022	0.024	0.022	0.009	0.000	96.5	95.4
	0.3	20	500	−0.024	0.062	0.069	0.062	0.050	0.039	95.2	92.2
			1000	−0.025	0.052	0.056	0.051	0.044	0.038	93.5	90.3
			2000	−0.024	0.046	0.052	0.049	0.046	0.043	94.1	92.8
		40	500	−0.024	0.050	0.054	0.049	0.032	0.025	92.9	90.3
			1000	−0.025	0.040	0.042	0.040	0.030	0.026	91.9	89.6
			2000	−0.024	0.034	0.038	0.037	0.032	0.030	93.9	92.9

Note: Par = parameter; TE = total error; TE_{bc} = bias-corrected total error; LE = linking error; LE_{bc} = bias-corrected linking error.

5. Empirical Example: PISA 2006 Reading

5.1. Method

To illustrate the computation of the different error estimates in FIPC, data from the cognitive domain reading in the 2006 programme for international student assessment (PISA) were analyzed (PISA 2006 [40]). The dataset included participants from 26 countries (see Table 2) who participated in 2006. The PISA 2006 dataset is available at https://www.oecd.org/en/data/datasets/pisa-2006-database.html (accessed on 25 November 2024).

Items for the reading domain were administered only to a subset of the participating students. The analysis included only those students who had been administered at least one reading item. This resulted in a total sample size of 110,236 students (ranging from 2010 to 12,142 students per country with a mean of $M = 4239.9$). In total, 28 reading items were included in the PISA 2006 reading test. Six of the twenty-eight items were polytomous and were dichotomously recoded, with only the highest category being recoded as correct.

In all analyses, student weights were incorporated. Within each country, the student weights were normalized to a sum of 5000, ensuring an equal contribution from all countries. In the first step, international item parameters were obtained by applying the 2PL model

to the weighted pooled dataset of all the students. In the second step, FIPC using the 2PL model was applied in each country, with the item parameters fixed at the values of the international item parameters. Next, SE, LE, and TE error estimates, as described in Section 3, were computed using the balanced repeated replication zones provided by the PISA 2006 study. Point estimates (i.e., country means, $\widehat{\mu}$, and country SDs, $\widehat{\sigma}$), as well as error estimates, were then linearly transformed into the metric used in the official PISA reports, such that the weighted pooled sample of all the students had a mean of 500 and an SD of 100. Since the countries contributed equally to this analysis, the average of the country means was also 500.

The R code from the simulation study was adapted for this example. The steps in which the variances due to sampling errors were computed were replaced with variance estimation using replication methods (see (23)).

Table 2. PISA 2006 reading: point estimates and error estimates for country means and country standard deviations (SDs).

CNT	*N*	*I*	Country Mean, $\widehat{\mu}$						Country SD, $\widehat{\sigma}$					
			Est	SE	LE	LE_{bc}	TE	TE_{bc}	Est	SE	LE	LE_{bc}	TE	TE_{bc}
AUS	7562	28	517.0	2.25	5.36	5.32	5.81	5.77	95.8	1.48	2.45	2.26	2.86	2.70
AUT	2646	27	496.3	3.75	4.39	4.25	5.77	5.67	103.1	2.69	3.33	2.94	4.28	3.99
BEL	4840	28	505.9	3.08	4.32	4.24	5.31	5.24	107.0	2.69	3.65	3.42	4.53	4.35
CAN	12,142	28	527.6	2.11	5.64	5.58	6.02	5.96	93.3	1.60	3.82	3.69	4.14	4.02
CHE	6578	28	502.3	3.13	4.55	4.47	5.52	5.46	95.7	2.33	3.03	2.81	3.83	3.65
CZE	3246	28	483.2	4.44	5.84	5.73	7.34	7.25	112.8	3.11	4.02	3.72	5.08	4.85
DEU	2701	28	496.1	4.96	4.60	4.48	6.76	6.68	113.9	2.80	5.09	4.86	5.81	5.61
DNK	2431	27	500.1	3.14	7.19	7.11	7.84	7.77	89.0	1.97	4.72	4.49	5.11	4.90
ESP	10,506	28	465.0	2.12	5.70	5.60	6.08	5.99	81.4	1.24	6.40	6.26	6.52	6.38
EST	2630	28	499.4	2.94	6.36	6.26	7.01	6.92	83.6	1.88	3.81	3.52	4.25	3.99
FIN	2536	28	551.6	2.37	6.26	6.15	6.69	6.59	85.3	1.92	4.52	4.26	4.91	4.67
FRA	2524	28	499.1	3.80	5.76	5.66	6.90	6.82	98.2	2.89	5.10	4.89	5.86	5.67
GBR	7061	28	498.4	2.22	6.00	5.93	6.40	6.33	98.3	1.77	4.68	4.50	5.01	4.84
GRC	2606	28	456.9	3.59	6.57	6.49	7.49	7.42	95.0	2.54	4.18	3.92	4.89	4.67
HUN	2399	28	485.2	3.32	4.77	4.57	5.81	5.65	91.7	2.40	4.80	4.49	5.37	5.09
IRL	2468	28	518.4	3.49	4.59	4.45	5.77	5.66	94.5	2.19	3.04	2.67	3.75	3.45
ISL	2010	28	493.2	1.96	5.36	5.25	5.71	5.60	91.3	2.09	3.38	3.03	3.97	3.68
ITA	11,629	28	471.6	2.15	5.35	5.30	5.76	5.72	98.1	1.91	3.29	3.13	3.81	3.67
JPN	3203	28	502.8	3.61	9.33	9.28	10.00	9.96	103.2	2.15	3.93	3.73	4.48	4.30
KOR	2790	27	556.0	3.75	9.12	9.05	9.86	9.79	95.8	3.19	5.43	5.23	6.30	6.13
LUX	2443	27	482.1	2.12	4.26	4.14	4.76	4.65	101.0	1.94	2.68	2.23	3.31	2.96
NLD	2666	28	509.2	3.16	7.08	7.00	7.75	7.68	101.6	3.00	4.29	4.00	5.24	5.00
NOR	2504	28	489.3	2.79	6.52	6.37	7.09	6.95	101.6	1.93	4.44	4.04	4.84	4.48
POL	2968	28	506.8	2.79	6.06	5.98	6.68	6.60	99.8	2.24	3.45	3.18	4.11	3.89
PRT	2773	28	475.9	3.41	5.94	5.84	6.85	6.77	95.3	2.55	3.90	3.59	4.66	4.40
SWE	2374	28	510.7	2.99	4.51	4.36	5.41	5.29	100.2	2.55	3.26	2.91	4.14	3.87

Note: CNT = country label (see Appendix A); *N* = sample size per country; *I* = number of items per country; SE = standard error; LE = linking error; LE_{bc} = bias-corrected linking error; TE = total error; TE_{bc} = bias-corrected total error.

5.2. Results

Table 2 presents the estimates for the country means, $\widehat{\mu}$, and country SDs, $\widehat{\sigma}$, in the PISA 2006 reading study. It is evident that the LE estimates substantially exceeded the SE estimates for both the mean and the SD. The ratio of the LE_{bc} and SE was computed for each country for both $\widehat{\mu}$ and $\widehat{\sigma}$, yielding an average of 1.97 for $\widehat{\mu}$ and 1.76 for $\widehat{\sigma}$ across countries. This indicates that variability in the country distribution parameter estimates was more strongly impacted by the item choice than by the sampling error. Nevertheless,

the bias-corrected LE (i.e., LE_{bc}) was frequently substantially smaller than the conventional LE estimate, particularly for country SDs.

It should be noted that official PISA reports consider only SE estimates as error estimates for statistical inference in country comparisons [14,40]. If the total errors, as computed in our analysis, were used, the error estimates would be much larger, resulting in fewer significant country differences. For example, Austria (AUT) significantly differed in its mean compared to 8 out of 26 countries when the SE was used for statistical inference. However, only five significant differences were observed when the TE_{bc} was used. Specifically, the differences between Austria and the Netherlands (NLD), Poland (POL), and Sweden (SWE) were statistically significant based on the SE but not based on the TE_{bc}.

6. Discussion

This article investigated the computation of LEs in FIPC focusing on educational large-scale assessment studies that rely on resampling methods for statistical inference. Computing LEs in addition to SEs in LSA studies is crucial for accurate reporting, as the heterogeneous functioning of items (i.e., random DIF) introduces an additional source of uncertainty in the country means and country SDs in FIPC. A bias-corrected LE estimate was proposed as an alternative to the commonly used jackknife LE. The bias-corrected LE was shown to eliminate the variance contribution from sampling errors present in the conventional LE estimate. This result was illustrated through a simulation study.

An empirical example using PISA 2006 reading data demonstrated that bias-corrected LE estimates were notably smaller than the commonly used jackknife LE estimates. The findings emphasize that LEs play a much larger role than SEs in estimating country distribution parameters [21]. This raises the argument that official PISA reports should incorporate LE estimates alongside SEs, combined as the total error, for statistical inference. If LEs were ignored, the statistical uncertainty regarding the country means and country SDs would be significantly underestimated. As a result, differences between countries could be misinterpreted as significant when they fall within the margins of error. It should be noted, however, that LEs in more recent PISA studies are likely smaller than those in the PISA 2006 due to the increased number of administered items in later cycles.

As noted by an anonymous reviewer, the fixed item parameters used in FIPC for educational studies like the PISA are not free of uncertainty. These parameters are typically international item parameters derived from a calibration process that includes data from all participating countries in a PISA study. While the standard errors of these parameters are small, they are not zero. Additionally, the item responses from a particular country are used twice in FIPC: once for scaling that country and once for deriving the international item parameters from the pooled sample of all the countries. The extra uncertainty in these fixed item parameters was not accounted for in this paper but should be addressed in future research.

The treatment of LEs in this study assumed independence among items. However, items are often grouped into item clusters that share a common stimulus (i.e., testlets [41]). Future research could extend LE estimation to account for this additional dependence structure among items (see [22,23]). In such cases, the item jackknife in the linking error formula should be replaced with a jackknife of testlets.

In this article, stratified clustered sampling was incorporated into SE, LE, and TE estimation using replication methods for persons. Alternatively, the variance estimates of item parameters could be obtained using cluster-robust standard errors, which do not require the repeated application of FIPC.

As an alternative to jackknife-based LEs, balanced half-sampling [19] or double jackknife [42] methods could be used. However, the latter methods are computationally more

intensive than the approach proposed in this article, as FIPC must be applied more times. Future research could compare our proposed bias-corrected LE and TE estimates with those obtained from half-sampling, double jackknife, or alternative jackknife [35] approaches.

LEs could also be computed for non-educational applications, such as group comparisons based on questionnaire items in psychological or sociological studies. The primary motivation for using LEs remains the same: uncertainties in distribution parameters due to item selection and heterogeneous item functioning should be accounted for with an additional error component complementary to the SE from the sampling of individuals. In this sense, the inference extends to a larger or potentially infinite set of items beyond the specific set used in this study.

When computing LEs, it is implicitly assumed that the variation in item parameters due to DIF is construct-relevant [43]. Recent PISA developments have shown a tendency to exclude items with large DIF effects in certain countries from comparisons [44], reflecting the view that DIF effects are construct-irrelevant. However, an alternative approach retains all items in the comparisons and quantifies heterogeneity in item selection through linking errors. This approach avoids the potential validity concerns associated with item removal [45,46].

7. Conclusions

This article investigated the estimation of the linking error and total error in fixed item parameter calibration. Bias-corrected estimates for these quantities were proposed based on analytical derivations. The usefulness of the error estimates was demonstrated through a simulation study and an empirical application using PISA data.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Replication material for creating the simulated datasets in the simulation study (Section 4) is available at https://osf.io/hb3ck (accessed on 25 November 2024). The PISA 2006 dataset used in Section 5 can be assessed at https://www.oecd.org/en/data/datasets/pisa-2006-database.html (accessed on 25 November 2024).

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

Abbreviations

The following abbreviations are used in this manuscript:

2PL	two-parameter logistic
DIF	differential item functioning
ER	error ratio
FIPC	fixed item parameter calibration
IRF	item response function
IRT	item response theory
LE	linking error
LSA	large-scale assessment
LE_{bc}	bias-corrected linking error
MML	marginal maximum likelihood
PISA	programme for international student assessment
SD	standard deviation
SE	standard error
TE	total error
TE_{bc}	bias-corrected total error

TIMSS trends in international mathematics and science study

Appendix A. Country Labels for PISA 2006 Reading Study

The country labels used in Table 2 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. *Handb. Stat.* **2007**, *26*, 469–513. [CrossRef]
2. Formann, A.K. Linear logistic latent class analysis for polytomous data. *J. Am. Stat. Assoc.* **1992**, *87*, 476–486. [CrossRef]
3. Lord, F.M.; Novick, R. *Statistical Theories of Mental Test Scores*; Addison-Wesley: Reading, MA, USA, 1968.
4. Mellenbergh, G.J. Generalized linear item response theory. *Psychol. Bull.* **1994**, *115*, 300–307. [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. 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.
7. Bock, R.D.; Aitkin, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **1981**, *46*, 443–459. [CrossRef]
8. San Martin, E. Identification of item response theory models. In *Handbook of Item Response Theory, Volume 2: Statistical Tools*; van der Linden, W.J., Ed.; CRC Press: Boca Raton, FL, USA, 2016; pp. 127–150. [CrossRef]
9. Kim, S. A comparative study of IRT fixed parameter calibration methods. *J. Educ. Meas.* **2006**, *43*, 355–381. [CrossRef]
10. Kim, S.; Kolen, M.J. Application of IRT fixed parameter calibration to multiple-group test data. *Appl. Meas. Educ.* **2019**, *32*, 310–324. [CrossRef]
11. 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]
12. Robitzsch, A. Bias and linking error in fixed item parameter calibration. *AppliedMath* **2024**, *4*, 1181–1191. [CrossRef]
13. 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]
14. OECD. *PISA 2018. Technical Report*; OECD: Paris, France, 2020. Available online: https://bit.ly/3zWbidA (accessed on 25 November 2024).
15. Mellenbergh, G.J. Item bias and item response theory. *Int. J. Educ. Res.* **1989**, *13*, 127–143. [CrossRef]
16. Penfield, R.D.; Camilli, G. 5 Differential item functioning and item bias. *Handb. Stat.* **2007**, *26*, 125–167. [CrossRef]
17. De Boeck, P. Random item IRT models. *Psychometrika* **2008**, *73*, 533–559. [CrossRef]
18. 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]
19. Robitzsch, A. Robust and nonrobust linking of two groups for the Rasch model with balanced and unbalanced random DIF: A comparative simulation study and the simultaneous assessment of standard errors and linking errors with resampling techniques. *Symmetry* **2021**, *13*, 2198. [CrossRef]
20. Joo, S.; Ali, U.; Robin, F.; Shin, H.J. Impact of differential item functioning on group score reporting in the context of large-scale assessments. *Large-Scale Assess. Educ.* **2022**, *10*, 18. [CrossRef]
21. Robitzsch, A.; Lüdtke, O. Linking errors in international large-scale assessments: Calculation of standard errors for trend estimation. *Assess. Educ.* **2019**, *26*, 444–465. [CrossRef]
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 25 November 2024).
23. Robitzsch, A. Linking error in the 2PL model. *J* **2023**, *6*, 58–84. [CrossRef]
24. 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]
25. Wu, M. Measurement, sampling, and equating errors in large-scale assessments. *Educ. Meas.* **2010**, *29*, 15–27. [CrossRef]
26. 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]
27. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; CRC Press: Boca Raton, FL, USA, 1994. [CrossRef]

28. Kolenikov, S. Resampling variance estimation for complex survey data. *Stata J.* **2010**, *10*, 165–199. [CrossRef]
29. Monseur, C.; Sibberns, H.; Hastedt, D. Linking errors in trend estimation for international surveys in education. *IERI Monogr. Ser.* **2008**, *1*, 113–122. Available online: https://bit.ly/38aTVeZ (accessed on 25 November 2024).
30. 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]
31. 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]
32. Haebara, T. Equating logistic ability scales by a weighted least squares method. *Jpn. Psychol. Res.* **1980**, *22*, 144–149. [CrossRef]
33. Robitzsch, A. Bias-reduced Haebara and Stocking-Lord linking. *J* **2024**, *7*, 373–384. [CrossRef]
34. 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]
35. Battauz, M. Multiple equating of separate IRT calibrations. *Psychometrika* **2017**, *82*, 610–636. [CrossRef]
36. Haberman, S.J.; Lee, Y.H.; Qian, J. *Jackknifing Techniques for Evaluation of Equating Accuracy*; (Research Report No. RR-09-02); Educational Testing Service: Princeton, NJ, USA, 2009. [CrossRef]
37. Foy, P.; Fishbein, B.; von Davier, M.; Yin, L. Implementing the TIMSS 2019 scaling methodology. In *Methods and Procedures: TIMSS 2019 Technical Report*; Martin, M.O., von Davier, M., Mullis, I.V., Eds.; IEA, Boston College: Chestnut Hill, MA, USA, 2020.
38. R Core Team. *R: A Language and Environment for Statistical Computing*; The R Foundation: Vienna, Austria, 2024. Available online: https://www.R-project.org (accessed on 15 June 2024).
39. Robitzsch, A. sirt: Supplementary Item Response Theory Models. R Package Version 4.2-89. 2024. Available online: https://github.com/alexanderrobitzsch/sirt (accessed on 13 November 2024).
40. OECD. *PISA 2006. Technical Report*; OECD: Paris, France, 2009. Available online: https://bit.ly/38jhdzp (accessed on 25 November 2024).
41. Wainer, H.; Bradlow, E.T.; Wang, X. *Testlet Response Theory and Its Applications*; Cambridge University Press: Cambridge, UK, 2007. [CrossRef]
42. Xu, X.; von Davier, M. *Linking Errors in Trend Estimation in Large-Scale Surveys: A Case Study*; (Research Report No. RR-10-10); Educational Testing Service: Princeton, NJ, USA, 2010. [CrossRef]
43. Camilli, G. The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In *Differential Item Functioning: Theory and Practice*; Holland, P.W., Wainer, H., Eds.; Erlbaum: Hillsdale, NJ, USA, 1993; pp. 397–417.
44. 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]
45. Adams, R.J. Response to 'Cautions on OECD's recent educational survey (PISA)'. *Oxf. Rev. Educ.* **2003**, *29*, 379–389. [CrossRef]
46. 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]

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.