Linking Error in the 2PL Model

Alexander Robitzsch [1,2]

1 IPN—Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118 Kiel, Germany; robitzsch@leibniz-ipn.de
2 Centre for International Student Assessment (ZIB), 24118 Kiel, Germany

Abstract: The two-parameter logistic (2PL) item response model is likely the most frequently applied item response model for analyzing dichotomous data. Linking errors quantify the variability in means or standard deviations due to the choice of items. Previous research presented analytical work for linking errors in the one-parameter logistic model. In this article, we present linking errors for the 2PL model using the general theory of M-estimation. Linking errors are derived in the case of log-mean-mean linking for linking two groups. The performance of the newly proposed formulas is evaluated in a simulation study. Furthermore, the linking error estimation in the 2PL model is also treated in more complex settings, such as chain linking, trend estimation, fixed item parameter calibration, and concurrent calibration.

Keywords: item response model; 2PL model; linking error; M-estimation

1. Introduction

Item response theory (IRT) models [1,2] are an important class of multivariate statistical models for analyzing dichotomous random variables used to model testing data from educational or psychological applications. Of particular relevance is the application of item response models in educational large-scale assessments [3], such as the programme for international student assessment (PISA; [4]) study.

In this article, we only investigate unidimensional IRT models. Let $X = (X_1, \ldots, X_I)$ be the vector of I dichotomous random variables $X_i \in \{0, 1\}$ (also referred to as items). A unidimensional item response model [5] is a statistical model for the probability distribution $P(X = x)$ for $x = (x_1, \ldots, x_I) \in \{0, 1\}^I$, where

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

where ϕ is the density of the normal distribution with mean μ and standard deviation σ. The vector $\delta = (\mu, \sigma)$ contains the distribution parameters. The vector $\gamma = (\gamma_1, \ldots, \gamma_I)$ contains all estimated item parameters of item response functions $P_i(\theta; \gamma_i) = P(X_i = 1|\theta)$.

The one-parameter logistic (1PL) model (also referred to as the Rasch model; [6]) employs the item response function $P_i(\theta) = \Psi(\theta - b_i)$, where Ψ denotes the logistic distribution function, and b_i is the item difficulty of item i. In this case, $\gamma_i = (b_i)$. The two-parameter logistic (2PL) model [7] additionally includes the item discrimination a_i (i.e., $\gamma_i = (a_i, b_i)$), and the item response function is given by $P_i(\theta) = \Psi(a_i(\theta - b_i))$.

Note that distribution parameters δ and item parameters γ cannot be simultaneously identified. In applications such as PISA in which a country mean μ and a country standard deviation σ, item parameters γ_i are often fixed at values γ_i^* that are used for all countries. In this situation, μ and σ can be identified. If sample data $x_1, \ldots, x_N$ for N persons are available, unknown model parameters in (1) can be estimated by (marginal) maximum likelihood (ML) using an expectation maximization algorithm [8,9].

In practice, data-generating item parameters γ_i differ from assumed fixed item parameters γ_i^*. This property is also referred to as differential item functioning (DIF; [10,11]).

DIF effects e_i are defined as deviations $e_i = \gamma_i - \gamma_i^*$. The occurrence of DIF causes additional variability in the estimated (country) mean μ and standard deviation σ [12,13]. The estimated distribution parameters depend on the choice of selected items, even in infinite sample sizes of persons. This variability is quantified in the linking error [14–19].

There exist simple formulas for linking errors based on variance components for the 1PL model [15,17]. For more complex models, resampling techniques [20,21] such as jackknife [15,17] or the (balanced) half sampling [18] of items can be employed. In this article, we provide closed formulas for the linking error for the 2PL model in various applications based on the M-estimation theory. The proposed formulas have the advantage of avoiding computationally more demanding resampling approaches for computing linking errors.

The rest of this article is structured as follows. The foundation of the M-estimation theory for the application of the computation of the linking error in the 2PL model is presented in Section 2. The specialization of M-estimation to log-mean-mean linking is treated in Section 3. The performance of the newly proposed linking errors is investigated in a simulation study in Section 4. Section 5 provides a further analytical treatment of the linking error in the 2PL model to more complex applications. Finally, this article closes with a discussion in Section 6.

2. Linking Error and M-Estimation

In this section, we discuss the computation of the linking error in the 2PL model for two groups. We do this in a general setting of M-estimation theory [22–24] because our treatment will apply to many of the recently discussed linking methods. However, we focus on log-mean-mean linking in this article as an important example in Sections 3 and 4.

Assume that the 2PL model holds in two groups $g = 1, 2$ or two time points. The goal is to determine the mean μ and the standard deviation σ of the second group, while the first group is assumed to have a mean of 0 and a standard deviation of 1. The DIF effects f_i and e_i for logarithmized item discriminations and item difficulties follow

$$\begin{aligned} \log a_{i2} &= \log a_{i1} + f_i \\ b_{i2} &= b_{i1} + e_i \end{aligned}. \tag{2}$$

It is assumed that f_i and e_i are independently and identically distributed with zero means and variances τ_a^2, τ_b^2, and the covariance is defined as $\mathrm{Cov}(e_i, f_i) = \tau_{ab}$.

In the first step of the linking approach, the 2PL model is separately estimated in each of the two groups. Because the ability θ for the first group has zero mean and a standard deviation of 1, the identified item parameters $\hat{a}_{i1}$ and $\hat{b}_{i1}$ equal the data-generating item parameters a_{i1} and b_{i1}, respectively, $(i = 1, \dots, I)$. In the second group, we fix the mean to 0 and the standard deviation to 1 and obtain identified parameters

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

In the second step of the linking approach, identified item parameters $\{(\hat{a}_{i1}, \hat{b}_{i1})\}$ and $\{(\hat{a}_{i2}, \hat{b}_{i2})\}$ are used in determining the mean μ and the standard deviation σ in the second group. Note that we assume that identified item parameters are known. Hence, we implicitly have infinite sample sizes of persons. In practice, we estimate item parameters from finite sample sizes. Appropriate adjustments are discussed in Section 5.8.

A general estimating equation of the type

$$\sum_{i=1}^{I} \boldsymbol{g}(\delta; \hat{a}_{i1}, \hat{b}_{i1}, \hat{a}_{i2}, \hat{b}_{i2}) = \sum_{i=1}^{I} \boldsymbol{g}(\delta; a_{i1}, b_{i1}, f_i, e_i) = \boldsymbol{0} \tag{4}$$

is employed for determining the parameter $\delta = (\sigma, \mu)$ or $\delta = (s, \mu)$ with $s = \log \sigma$ as the distribution parameters of interest. M-estimation theory provides the asymptotic variance in an estimate $\hat{\delta}$. Because linking errors refer to the uncertainty regarding item choice, this asymptotic variance can be used to compute the linking error.

In the two-group 2PL case, we have two unknown distribution parameters σ (or $s = \log \sigma$) and μ. Hence, $\boldsymbol{g} = (g_1, g_2)$ involves two equations for two unknowns that must be solved. The two M-estimation equations of the linking approaches can be generally written as

$$\begin{aligned} \sum_{i=1}^{I} g_1(\delta; a_{i1}, b_{i1}, f_i, e_i) &= 0 \\ \sum_{i=1}^{I} g_2(\delta; a_{i1}, b_{i1}, f_i, e_i) &= 0 \end{aligned} \tag{5}$$

M-estimation theory provides the asymptotic variance (for $I \to \infty$) for the estimate $\hat{\delta}$ with the sandwich formula [22]

$$\boldsymbol{V}_I(\hat{\delta}) = \boldsymbol{A}_I(\hat{\delta})^{-1} \boldsymbol{B}_I(\hat{\delta}) \boldsymbol{A}_I(\hat{\delta})^{-1} . \tag{6}$$

The matrix $\boldsymbol{A}_I$ is denoted as the bread matrix, while the matrix $\boldsymbol{B}_I$ is referred to as the meat matrix. The latter matrix is given by

$$\boldsymbol{B}_I(\delta) = \sum_{i=1}^{I} \mathrm{Var}\left(\boldsymbol{g}(\delta; a_{i1}, b_{i1}, f_i, e_i) \boldsymbol{g}(\delta; a_{i1}, b_{i1}, f_i, e_i)^\top \right), \tag{7}$$

where Var denotes a covariance matrix. The bread matrix $\boldsymbol{A}_I$ is given as

$$\boldsymbol{A}_I(\delta) = \sum_{i=1}^{I} \mathrm{E}\left(\frac{\partial \boldsymbol{g}}{\partial \delta}(\delta; a_{i1}, b_{i1}, f_i, e_i) \right) . \tag{8}$$

The two matrices in Equations (7) and (8) require the computation of expected values and variances of random variables. If these were unavailable or the quantities could not be algebraically determined, sample-based versions of the bread and the meat matrix are frequently used [23]. The meat matrix $\boldsymbol{B}_I$ can be estimated based on sample data using (7)

$$\hat{\boldsymbol{B}}_I(\hat{\delta}) = \sum_{i=1}^{I} \boldsymbol{g}(\hat{\delta}; \hat{a}_{i1}, \hat{b}_{i1}, \hat{a}_{i2}, \hat{b}_{i2}) \boldsymbol{g}(\hat{\delta}; \hat{a}_{i1}, \hat{b}_{i1}, \hat{a}_{i2}, \hat{b}_{i2})^\top . \tag{9}$$

An empirical version of $\boldsymbol{A}_I$ is given by

$$\hat{\boldsymbol{A}}_I(\hat{\delta}) = \sum_{i=1}^{I} \frac{\partial \boldsymbol{g}}{\partial \delta}(\hat{\delta}; \hat{a}_{i1}, \hat{b}_{i1}, \hat{a}_{i2}, \hat{b}_{i2}) . \tag{10}$$

If $\boldsymbol{A}_I(\hat{\delta})$ and $\boldsymbol{B}_I(\hat{\delta})$ are used for computing the variance matrix for $\hat{\delta}$ (i.e., $\boldsymbol{V}_{I,\mathrm{ESW}}(\hat{\delta})$), the estimate is denoted as the expected sandwich (ESW) estimate. The observed sandwich (OSW) estimate is obtained by using $\hat{\boldsymbol{A}}_I(\hat{\delta})$ and $\hat{\boldsymbol{B}}_I(\hat{\delta})$ in the formula of the variance matrix (i.e., $\boldsymbol{V}_{I,\mathrm{OSW}}(\hat{\delta})$). Finally, a bias-corrected observed sandwich (BOSW) is obtained by using $\boldsymbol{V}_{I,\mathrm{BOSW}}(\hat{\delta}) = I/(I-1)\boldsymbol{V}_{I,\mathrm{OSW}}(\hat{\delta})$ (see [25–27]).

We want to emphasize that M-estimation theory is not restricted to applications of linking approaches for two groups and two distribution parameters. The parameter δ can be of any finite dimensionality and could, for example, involve $2(G-1)$ unknown parameters for linking G groups.

M-estimation theory was applied in the investigation of DIF and linking in [28–30]. The simultaneous treatment of standard errors and linking errors in IRT models relying on M-estimation was presented in [18,31] (see also [32]).

3. Linking Error of Log-Mean-Mean Linking

We now apply the M-estimation of computing linking errors to log-mean-mean linking [33,34] for linking two groups in the 2PL model. The logarithm of the standard deviation $s = \log \sigma$ is estimated by

$$\hat{s} = \frac{1}{I} \sum_{i=1}^{I} (\log \hat{a}_{i2} - \log \hat{a}_{i1}) \, . \tag{11}$$

It can be shown that $\hat{s}$ is an unbiased and consistent estimate for s [34]. Moreover, $\hat{\sigma}$ is obtained by computing $\hat{\sigma} = \exp(\hat{s})$.

An estimate of the group mean μ is obtained using

$$\hat{\mu} = \frac{1}{I} \sum_{i=1}^{I} \left(\exp(\hat{s}) \hat{b}_{i2} - \hat{b}_{i1} \right) \, . \tag{12}$$

We can reformulate (11) and (12) as M-estimators

$$\sum_{i=1}^{I} g_1(\delta; a_{i1}, b_{i1}, f_i, e_i) = \sum_{i=1}^{I} (\log \hat{a}_{i2} - \log \hat{a}_{i1} - s) = 0 \text{ and} \tag{13}$$

$$\sum_{i=1}^{I} g_2(\delta; a_{i1}, b_{i1}, f_i, e_i) = \sum_{i=1}^{I} \left(\exp(s) \hat{b}_{i2} - \hat{b}_{i1} + \mu \right) = 0 \tag{14}$$

using $\boldsymbol{g} = (g_1, g_2)$. Now, we determine the linking errors for $\hat{\sigma}$ and $\hat{\mu}$ in log-mean-mean linking using the sandwich formula (6) of M-estimation. First, we compute the variance matrix of $\boldsymbol{g}$. We obtain using $\hat{a}_{i2} = a_{i1} \exp(s) \exp(f_i)$ and $\hat{b}_{i2} = \sigma^{-1}(b_{i1} - \mu)$

$$\sum_{i=1}^{I} \text{Var}(g_1(\delta; a_{i1}, b_{i1}, f_i, e_i)) = \sum_{i=1}^{I} \text{Var}(\log \hat{a}_{i2} - \log \hat{a}_{i1} - s) = \sum_{i=1}^{I} \text{Var}(f_i) = I\tau_a^2 \, , \tag{15}$$

$$\sum_{i=1}^{I} \text{Cov}(g_1(\delta; a_{i1}, b_{i1}, f_i, e_i), g_2(\delta; a_{i1}, b_{i1}, f_i, e_i)) = \sum_{i=1}^{I} \text{Cov}(f_i, e_i) = I\tau_{ab} \, , \text{ and} \tag{16}$$

$$\sum_{i=1}^{I} \text{Var}(g_2(\delta; a_{i1}, b_{i1}, f_i, e_i)) = \sum_{i=1}^{I} \text{Var}(e_i) = I\tau_b^2 \, . \tag{17}$$

Hence, it follows that

$$\boldsymbol{B}_I(\hat{\delta}) = I \begin{pmatrix} \tau_a^2 & \tau_{ab} \\ \tau_{ab} & \tau_b^2 \end{pmatrix} \tag{18}$$

Now, we compute derivatives of $\boldsymbol{g}$ with respect to s and μ and obtain the bread matrix as

$$\boldsymbol{A}_I(\hat{\delta}) = \begin{pmatrix} \dfrac{\partial g_1}{\partial s} & \dfrac{\partial g_1}{\partial \mu} \\[2mm] \dfrac{\partial g_2}{\partial s} & \dfrac{\partial g_2}{\partial \mu} \end{pmatrix} = I \begin{pmatrix} -1 & 0 \\ -(\mu - \bar{b}_{\bullet 1}) & 1 \end{pmatrix} \tag{19}$$

The inverse of the bread matrix can be determined as

$$\boldsymbol{A}_I(\hat{\delta})^{-1} = -I^{-1} \begin{pmatrix} 1 & 0 \\ \mu - \bar{b}_{\bullet 1} & -1 \end{pmatrix} \tag{20}$$

Formulas (18) and (19) involve unknown parameters τ_a, τ_b, τ_{ab}, and μ that must be estimated. The quantities f_i and e_i can be replaced with their sample estimates to estimate

unknown variances and covariances in the A_I and B_I matrices. We obtain the expected sandwich estimator (ESW) $V_{I,\text{ESW}}$ using Equation (6) by

$$A_{I,\text{ESW}}(\hat{\delta}) = I \begin{pmatrix} -1 & 0 \\ -(\hat{\mu} - \overline{b}_{\bullet 1}) & 1 \end{pmatrix} \quad \text{and} \quad B_{I,\text{ESW}}(\hat{\delta}) = I \begin{pmatrix} \hat{\tau}_a^2 & \hat{\tau}_{ab} \\ \hat{\tau}_{ab} & \hat{\tau}_b^2 \end{pmatrix}. \tag{21}$$

The observed sandwich (OSW) estimate $V_{I,\text{OSW}}$ uses empirical moments in the estimation. First, the bread matrix is estimated by

$$A_{I,\text{OSW}}(\hat{\delta}) = \begin{pmatrix} -I & 0 \\ \exp(s) \sum_{i=1}^{I} \hat{b}_{i2} & I \end{pmatrix}. \tag{22}$$

The meat matrix is estimated by

$$B_{I,\text{OSW}}(\hat{\delta}) = \begin{pmatrix} \sum_{i=1}^{I} (\log \hat{a}_{i2} - \log \hat{a}_{i1} - \hat{s})^2 & \sum_{i=1}^{I} (\log \hat{a}_{i2} - \log \hat{a}_{i1} - \hat{s})(\exp(\hat{s})\hat{b}_{i2} - \hat{b}_{i1} + \hat{\mu}) \\ \sum_{i=1}^{I} (\log \hat{a}_{i2} - \log \hat{a}_{i1} - \hat{s})(\exp(\hat{s})\hat{b}_{i2} - \hat{b}_{i1} + \hat{\mu}) & \sum_{i=1}^{I} (\exp(\hat{s})\hat{b}_{i2} - \hat{b}_{i1} + \hat{\mu})^2 \end{pmatrix} \tag{23}$$

Finally, we use a bias-corrected variant of the observed sandwich estimator [27] as

$$V_{I,\text{BOSW}} = \frac{I}{I-1} V_{I,\text{OSW}}. \tag{24}$$

Linking errors can be obtained as square roots of the diagonal elements of the variance matrices V. The linking error for $\hat{s}$ based on the ESW estimate is given by

$$\text{LE}(\hat{s}) = \frac{\hat{\tau}_a}{\sqrt{I}}. \tag{25}$$

By utilizing the delta method, we can obtain the linking error for $\hat{\sigma} = \exp(\hat{s})$ as

$$\text{LE}(\hat{\sigma}) = \hat{\sigma} \frac{\hat{\tau}_a}{\sqrt{I}}. \tag{26}$$

Finally, the linking error for $\hat{\mu}$ can be estimated by

$$\text{LE}(\hat{\mu}) = \sqrt{\frac{\hat{\tau}_b^2 + (\hat{\mu} - \overline{b}_{\bullet 1})^2 \hat{\tau}_a^2 - 2(\hat{\mu} - \overline{b}_{\bullet 1})\hat{\tau}_{ab}}{I}}. \tag{27}$$

In the absence of a nonuniform DIF, we have $\hat{\tau}_a^2 = \hat{\tau}_{ab} = 0$, and the linking error for the 1PL model is obtained from (27)

$$\text{LE}(\hat{\mu}) = \frac{\hat{\tau}_b}{\sqrt{I}}. \tag{28}$$

Interestingly, the presence of a nonuniform DIF ($\tau_a^2 > 0$) introduces additional uncertainty in computing the group mean. However, there is only an effect of a nonuniform DIF if the average item difficulty does not match the group mean (i.e., $\hat{\mu} - \overline{b}_{\bullet 1} \neq 0$). In typical applications, the third term $2(\hat{\mu} - \overline{b}_{\bullet 1})\hat{\tau}_{ab}$ in (27) will be much more important than the second term $(\hat{\mu} - \overline{b}_{\bullet 1})^2 \hat{\tau}_a^2$. Hence, a nonuniform DIF particularly plays an important role if uniform and nonuniform DIF effects are strongly correlated.

4. Simulation Study

4.1. Method

In this simulation study, we investigate the performance of different linking error estimates for the log-mean-mean linking approach in the 2PL model. In particular, we compare the jackknife linking error (JK) with linking errors obtained by the empirical sandwich (ESW), observed sandwich (OSW), and the bias-corrected observed sandwich (BOSW) estimates. The formulas for the sandwich estimates are presented in Section 3. The jackknife linking error estimate is computed by repeating the linking procedure when omitting the ith item for $i = 1, \ldots, I$. Let $u = \mu$ or $u = \sigma$ be the distribution parameter of interest and $\hat{u}$ be the corresponding estimate. Let $\hat{u}_{(-i)}$ be the estimated parameter if item i was removed from the linking procedure. Then, the jackknife linking error is defined as (see [15])

$$\text{LE}(\hat{u}) = \sqrt{\frac{I-1}{I} \sum_{i=1}^{I} (\hat{u}_{(-i)} - \hat{u})^2} \tag{29}$$

For identification, the first group had a zero mean and a standard deviation of the ability variable θ. For the second group, we defined $\mu = -0.2$ and $\sigma = 0.9$ in the simulation. Item parameters for 10 items are presented in Table 1. In the simulation, we used $I = 10, 20, 40$, or 80 items. For item numbers as multiples of 10, we duplicated the item parameters of the 10 items presented in Table 1 accordingly. The standard deviation of uniform DIF effects e_i was chosen as $\tau_b = 0.25$ or 0.50. The standard deviation of nonuniform DIF effects f_i was chosen as $\tau_a = 0.01$ or 0.25. The first condition mimics the case of the practical absence of nonuniform DIF effects. The correlation of DIF effects between e_i and f_i was set at 0.3 in all simulation conditions (i.e., $\tau_{ab} = 0.3 \cdot \tau_a \tau_b$). Finally, we chose three types of distributions for DIF effects (f_i, e_i). We specified them as a bivariate normal copula model and chose different marginal distributions. First, we chose the normal distribution (i.e., denoted as "Normal") as a marginal distribution appropriately scaled by τ_a and τ_b. Second, we chose a scaled t distribution with four degrees of freedom (i.e., denoted as "t_4") with an appropriate scaling factor to match the desired standard deviation of DIF effects. Third, we use the distribution function F of a normal mixture model (i.e., denoted as "Normal Mixture") of the type

$$F = (1 - \varepsilon) \text{N}(0, \tau^2) + \varepsilon \text{N}(0, k\tau^2) \,, \tag{30}$$

where $k = 3$ and $\varepsilon = 0.05$. This distribution can be interpreted as a contaminated distribution that includes a few outlying DIF effects in $\text{N}(0, k\tau^2)$ with proportion ε. Such a distribution is often employed in robust statistics [35]. For a prespecified DIF effect τ_b, we obtain from (30) the determining equation $\tau = \left((1 - \varepsilon) + \varepsilon k^2\right)^{-1/2} \tau_a$.

Table 1. Simulation Study: Used item parameters of the 2PL model.

Item	a_i	b_i
1	0.73	−1.31
2	1.25	1.44
3	1.20	−1.20
4	1.47	0.10
5	0.97	0.10
6	1.38	−0.74
7	1.05	1.48
8	1.14	−0.61
9	1.15	0.82
10	0.67	−0.07

Note. a_i = item discrimination; b_i = item difficulty.

To disentangle standard errors due to the sampling of persons from linking errors due to item choice, we assumed no sampling error for identified parameters $\{(\hat{a}_{i1}, \hat{b}_{i1})\}$ and $\{(\hat{a}_{i2}, \hat{b}_{i2})\}$. That is, identified item parameters for the second group only vary across replications in the simulation study because different DIF effects f_i and e_i were simulated in each replication. It seems reasonable in the simulation study in the comparison of the different M-estimation approaches with a jackknife to exclude the effects of sampling errors because these are just another source of uncertainty in distribution parameter estimates.

In each of the 4 $(I) \times 2$ $(\tau_b) \times 2$ $(\tau_a) \times 3$ (distribution) = 48 cells of the simulation, 40,000 replications were conducted. We assessed coverage rates at the 95% confidence level based on the normal distribution for distribution parameter estimates. The linking error computation for the estimated standard deviation $\hat{\sigma} = \exp(\hat{s})$ utilized the delta method.

The R software [36] was used for simulation and analysis. We used the `qmixnorm` function from the R package KScorrect [37] for determining quantiles in the data simulation of DIF effects. Because analytical solutions are not available to compute a quantile function for the normal mixture model, the `qmixnorm` function approximates the quantile function using a spline function calculated from cumulative density functions for the specified mixture distribution [37]. Quantiles for probabilities near zero or one are approximated by taking a randomly generated sample.

4.2. Results

In Table 2, the coverage rates for the estimated mean $\hat{\mu}$ are presented as a function of the standard deviation of the DIF effects, the number of items, and the type of distribution for the DIF effects. It turned out that there were no substantial differences in the performance of the different linking error methods with respect to the distribution types of the DIF effects. The jackknife and the ESW estimates were very similar. The OSW estimate did not reach the desired coverage rates in a short test (i.e., $I = 10$) but improved in longer tests. Moreover, the BOSW slightly improved the OSW estimate but still was inferior to the ESW estimate.

Table 2. Simulation Study: Coverage rates for estimated mean $\hat{\mu}$ as a function of the standard deviation of DIF effects for a (τ_a) and b (τ_b), number of items (I), and the type of distribution for DIF effects.

			Normal				t_4				Normal Mixture			
τ_a	τ_b	I	JK	ESW	OSW	BOSW	JK	ESW	OSW	BOSW	JK	ESW	OSW	BOSW
0.01	0.25	10	**92.2**	**92.2**	**90.8**	**92.1**	92.9	92.9	**91.5**	92.9	92.7	92.7	**91.3**	92.7
		20	93.8	93.8	93.2	93.8	94.5	94.5	93.8	94.5	94.4	94.4	93.8	94.4
		40	94.6	94.6	94.3	94.6	94.9	94.9	94.5	94.8	94.9	94.9	94.6	94.9
		80	95.2	95.2	95.0	95.1	95.2	95.2	95.0	95.1	95.2	95.2	95.1	95.2
0.01	0.50	10	92.5	92.5	**91.2**	92.5	92.9	92.9	**91.5**	92.9	93.1	93.1	**91.7**	93.0
		20	94.1	94.1	93.5	94.1	94.6	94.6	94.0	94.6	94.4	94.4	93.8	94.4
		40	94.7	94.7	94.4	94.7	95.0	95.0	94.7	95.0	94.8	94.8	94.5	94.8
		80	95.1	95.1	94.9	95.0	95.3	95.3	95.1	95.3	95.4	95.4	95.2	95.3
0.25	0.25	10	93.9	93.8	**91.1**	92.6	94.6	94.4	**92.0**	93.4	94.4	94.3	**91.8**	93.1
		20	94.8	94.8	93.1	93.8	94.9	94.9	93.2	93.9	95.1	95.1	93.4	94.0
		40	95.1	95.1	93.7	94.0	95.5	95.5	94.0	94.4	95.2	95.2	93.6	94.0
		80	95.2	95.2	93.9	94.0	95.4	95.4	94.2	94.4	95.5	95.5	94.3	94.4
0.25	0.50	10	93.0	92.9	**90.9**	**92.3**	93.7	93.6	**91.6**	93.0	93.5	93.5	**91.3**	92.7
		20	94.3	94.3	93.1	93.8	94.4	94.4	93.2	93.8	94.5	94.5	93.3	94.0
		40	95.0	95.0	94.2	94.6	95.1	95.1	94.3	94.7	95.1	95.1	94.3	94.6
		80	95.2	95.2	94.6	94.8	95.2	95.1	94.6	94.7	95.1	95.1	94.4	94.6

Note. Normal = DIF effects normally distributed; t_4 = DIF effects distributed according to scaled t_4 distribution; Normal Mixture = DIF effects distributed according to contaminated mixture model; JK = linking error (LE) estimated by jackknife; ESW = LE estimated by expected sandwich estimator; OSW = LE estimated by observed sandwich estimator; BOSW = LE estimated by bias-corrected observed sandwich estimator; coverage rates smaller than 92.5% or larger than 97.5% are printed in bold.

In Table 3, the coverage rates for the estimated standard deviation $\hat{\sigma}$ are presented as a function of the standard deviation of the DIF effects, the number of items, and the type of distribution for the DIF effects. The OSW and BOSW had particular issues in the coverage rates with very small nonuniform DIF effects. These issues also remained in the longer tests. However, the estimated standard deviations of the nonuniform DIF effects $\hat{\tau}_a$ turned out to be unbiased and can be detectable in such situations to indicate that linking errors for $\hat{\sigma}$ would be tiny in this situation. Hence, the practical absence of nonuniform DIF effects using the simulation condition $\tau_a = 0.01$ might not be very realistic, and future studies could investigate the performance using $\tau_a = 0.10$.

Table 3. Simulation Study: Coverage rates for estimated standard deviation $\hat{\sigma}$ as a function of the standard deviation of DIF effects for a (τ_a) and b (τ_b), number of items (I), and the type of distribution for DIF effects.

			Normal				t_4				Normal Mixture			
τ_a	τ_b	I	JK	ESW	OSW	BOSW	JK	ESW	OSW	BOSW	JK	ESW	OSW	BOSW
0.01	0.25	10	92.6	92.6	**86.1**	**87.3**	92.9	92.9	**84.9**	**86.2**	92.9	92.9	**85.3**	**86.7**
		20	93.8	93.8	**83.0**	**83.8**	94.5	94.5	**82.1**	**82.8**	94.4	94.4	**82.5**	**83.3**
		40	94.6	94.6	**80.2**	**80.6**	95.0	95.0	**79.1**	**79.5**	94.9	94.9	**79.2**	**79.5**
		80	95.2	95.2	**77.9**	**78.1**	95.1	95.1	**76.2**	**76.4**	95.0	95.0	**76.3**	**76.6**
0.01	0.50	10	**92.1**	**92.2**	**92.3**	93.0	93.2	93.2	**91.8**	92.6	93.0	93.0	**92.2**	92.9
		20	94.1	94.1	**91.3**	**91.7**	94.4	94.4	**90.7**	**91.2**	94.2	94.2	**90.8**	**91.3**
		40	94.5	94.5	**91.3**	**91.6**	94.8	94.8	**90.6**	**90.8**	94.9	94.9	**90.7**	**90.9**
		80	95.1	95.1	93.7	93.7	95.0	95.0	92.5	92.6	95.3	95.3	92.9	93.0
0.25	0.25	10	**92.3**	**92.3**	**89.8**	**91.3**	93.1	93.1	**90.5**	**92.0**	92.6	92.6	**90.0**	**91.4**
		20	93.8	93.8	**92.4**	93.0	94.5	94.5	93.0	93.6	94.2	94.2	92.7	93.3
		40	94.7	94.7	93.6	93.9	95.1	95.1	93.9	94.3	94.7	94.7	93.7	94.0
		80	94.9	94.9	94.0	94.2	95.2	95.2	94.3	94.5	95.2	95.2	94.3	94.5
0.25	0.50	10	**92.3**	**92.3**	**87.7**	**89.3**	93.0	92.9	**88.2**	**89.7**	92.7	92.6	**88.0**	**89.5**
		20	94.1	94.1	**91.2**	**91.8**	94.3	94.3	**91.2**	**91.9**	94.1	94.1	**90.9**	**91.6**
		40	94.8	94.8	92.7	93.0	94.9	94.9	92.7	93.0	94.9	94.9	92.7	93.1
		80	95.1	95.1	93.3	93.4	95.4	95.4	93.5	93.7	95.1	95.1	93.3	93.4

Note. Normal = DIF effects normally distributed; t_4 = DIF effects distributed according to scaled t_4 distribution; Normal Mixture = DIF effects distributed according to contaminated mixture model; JK = linking error (LE) estimated by jackknife; ESW = LE estimated by expected sandwich estimator; OSW = LE estimated by observed sandwich estimator; BOSW = LE estimated by bias-corrected observed sandwich estimator; coverage rates smaller than 92.5% or larger than 97.5% are printed in bold.

Overall, the findings of this simulation study indicate that the sandwich estimates (in particular the ESW) are as effective as the jackknife estimates for linking errors.

5. Further Applications of the Linking Error in the 2PL Model

In this section, several applications of the linking error computations in the 2PL model are presented. In Section 5.1, the M-estimation theory is applied to linking approaches other than the log-mean-mean linking. Section 5.2 discusses the computation of the linking errors if the items are nested within testlets. The linking errors for chain linking and trend estimation are discussed in Sections 5.3 and 5.4, respectively. In Section 5.5, the linking error under a fixed item parameter calibration is derived. Section 5.6 presents the linking error in the 2PL model for a concurrent calibration. Section 5.7 investigates the computation of the linking errors of derived parameters. Finally, Section 5.8 focuses on the computation of the total error and sampling error corrections in the linking error estimation.

5.1. Different Linking Methods

We now illustrate how the sandwich estimates in the M-estimation theory from Section 2 can be used for other linking approaches.

5.1.1. Robust Log-Mean-Mean Linking

Log-mean-mean linking involves two steps that compute the mean for determining the logarithm of the standard deviation $s = \log \sigma$ and the mean μ. A few outlying items might introduce bias in the estimated distribution parameters [18]; robust estimators for the location measures can be preferred. In this case, the estimating functions for $\delta = (s, \mu)$ are given by

$$\sum_{i=1}^{I} g_1(\delta; a_{i1}, b_{i1}, f_i, e_i) = \sum_{i=1}^{I} \rho(\log \hat{a}_{i2} - \log \hat{a}_{i1} - s) = 0 \text{ and} \tag{31}$$

$$\sum_{i=1}^{I} g_2(\delta; a_{i1}, b_{i1}, f_i, e_i) = \sum_{i=1}^{I} \rho\left(\exp(s)\hat{b}_{i2} - \hat{b}_{i1} + \mu\right) = 0 \tag{32}$$

using a robust function that fulfills the property $\rho(x)/|x| \to \infty$ for $|x| \to \infty$ (see [29,30]). For example, if the median was used, ρ would be the sign function $\rho(x) = \mathbf{1}_{\{x>0\}} - \mathbf{1}_{\{x<0\}}$. A differentiable approximation of this function is given by $\tilde{\rho}(x) = \frac{\mathrm{d}}{\mathrm{d}x}(x^2 + \varepsilon)^{p/2}$ for $p = 1$ and $\varepsilon = 0.01$. The observed sandwich formula can be easily applied to obtain linking errors for a wide class of robust linking approaches.

5.1.2. Haebara Linking

Haebara (HAE) linking [38] aligns item response functions instead of directly aligning item parameters. The linking function in HAE linking is given as

$$H(\mu, \sigma) = \sum_{i=1}^{I} \int \left[\Psi\left(\hat{a}_{i1}(\theta - \hat{b}_{i1})\right) - \Psi\left(\sigma^{-1}\hat{a}_{i2}(\theta - \sigma\hat{b}_{i2} - \mu_2)\right) \right]^2 \omega(\theta)\,\mathrm{d}\theta\,, \tag{33}$$

where ω is a weighting function. By defining the difference

$$h(\theta, \mu, \sigma; \hat{a}_{i1}, \hat{b}_{i1}, \hat{a}_{i2}, \hat{b}_{i2}) = \Psi\left(\hat{a}_{i1}(\theta - \hat{b}_{i1})\right) - \Psi\left(\sigma^{-1}\hat{a}_{i2}(\theta - \sigma\hat{b}_{i2} - \mu_2)\right)\,, \tag{34}$$

we can rewrite (33) as

$$H(\mu, \sigma) = \sum_{i=1}^{I} \int \left[h(\theta, \mu, \sigma; \hat{a}_{i1}, \hat{b}_{i1}, \hat{a}_{i2}, \hat{b}_{i2}) \right]^2 \omega(\theta)\,\mathrm{d}\theta\,. \tag{35}$$

The estimating equations for σ and μ can be determined by

$$\sum_{i=1}^{I} g_1(\delta; a_{i1}, b_{i1}, f_i, e_i) = \sum_{i=1}^{I} \int h(\theta, \mu, \sigma; \hat{a}_{i1}, \hat{b}_{i1}, \hat{a}_{i2}, \hat{b}_{i2}) \left[\frac{\partial}{\partial\sigma} h(\theta, \mu, \sigma; \hat{a}_{i1}, \hat{b}_{i1}, \hat{a}_{i2}, \hat{b}_{i2}) \right] \omega(\theta)\,\mathrm{d}\theta = 0 \text{ and} \tag{36}$$

$$\sum_{i=1}^{I} g_2(\delta; a_{i1}, b_{i1}, f_i, e_i) = \sum_{i=1}^{I} \int h(\theta, \mu, \sigma; \hat{a}_{i1}, \hat{b}_{i1}, \hat{a}_{i2}, \hat{b}_{i2}) \left[\frac{\partial}{\partial\mu} h(\theta, \mu, \sigma; \hat{a}_{i1}, \hat{b}_{i1}, \hat{a}_{i2}, \hat{b}_{i2}) \right] \omega(\theta)\,\mathrm{d}\theta = 0\,. \tag{37}$$

Again, linking errors can be easily obtained using the observed sandwich (OSW) formula.

5.2. Linking Error with Testlets

In educational large-scale assessment studies such as PISA, several items frequently share a common item stimulus. In this situation, items are nested within testlets [39–41]. It was demonstrated that DIF effects are also pronounced at the testlet level and not only at the item level [17,42]. This additional source of uncertainty should be included in the computation of linking errors to avoid negatively biased linking error estimates.

We illustrate how to apply the theory of M-estimation in the case of testlets. The linking error for $\hat{\sigma}$ and $\hat{\mu}$ in log-mean-mean linking is derived. Let there be H testlets, and there are $I_h \geq 1$ items within each testlet h. It holds that $\sum_{h=1}^{H} I_h = I$. The data-generating

model for DIF effects in the 2PL model in Equation (2) is adapted to include DIF effects at the item and the testlet level. DIF effects for logarithmized item discriminations now include two terms referring to testlets (i.e., h) and items within testlets (i.e., item i nested within testlet h). We assume

$$\begin{aligned}
\log a_{ih2} &= \log a_{ih1} + f_h + f_{ih} \\
b_{ih2} &= b_{ih1} + e_h + e_{ih}
\end{aligned} \tag{38}$$

Note that item parameters now possess an item index i and a testlet index h in (38). Again, it is assumed that the DIF effects are independently distributed across items, and we define $\tau_{a,\text{testlet}}^2 = \text{Var}(f_h)$, $\tau_{a,\text{item}}^2 = \text{Var}(f_{ih})$, $\tau_{b,\text{testlet}}^2 = \text{Var}(e_h)$, $\tau_{b,\text{item}}^2 = \text{Var}(e_{ih})$, $\tau_{ab,\text{testlet}} = \text{Cov}(e_h, f_h)$, and $\tau_{ab,\text{item}} = \text{Cov}(e_{ih}, f_{ih})$.

The estimating equations in log-mean-mean linking are the same. However, they must include the testlet structure of items. The estimating Equations (13) and (14) are rearranged as

$$\sum_{h=1}^{H} \sum_{i=1}^{I_h} g_1(\delta; a_{ih1}, b_{ih1}, f_h, f_{ih}, e_h, e_{ih}) = \sum_{h=1}^{H} \sum_{i=1}^{I_h} (\log \hat{a}_{ih2} - \log \hat{a}_{ih1} - s) = 0 \text{ and} \tag{39}$$

$$\sum_{h=1}^{H} \sum_{i=1}^{I_h} g_2(\delta; a_{ih1}, b_{ih1}, f_h, f_{ih}, e_h, e_{ih}) = \sum_{h=1}^{H} \sum_{i=1}^{I_h} \left(\exp(s) \hat{b}_{ih2} - \hat{b}_{ih1} + \mu \right) = 0 . \tag{40}$$

The essential change in the computation of the sandwich variance in the testlet case is that the variance matrix $\boldsymbol{B}_I$ (i.e., the meat matrix) requires the computation of the variance that is carried out at each testlet h instead of each individual item i (see [27]). To indicate the dependency from the testlet structure, it is more appropriate to label the meat matrix $\boldsymbol{B}_H$ because testlets are independent units, not items. The entry involving the variance in DIF effects in item discriminations in the meat matrix can be computed as

$$\sum_{h=1}^{H} \text{Var} \left(\sum_{i=1}^{I_h} g_1(\delta; a_{i1}, b_{i1}, f_h, f_{ih}, e_h, e_{ih}) \right) = \sum_{h=1}^{H} \text{Var} \left(\sum_{i=1}^{I_h} (\log \hat{a}_{i2} - \log \hat{a}_{i1} - s) \right) = \left(\sum_{h=1}^{H} I_h^2 \right) \tau_{a,\text{testlet}}^2 + I \tau_{a,\text{item}}^2 . \tag{41}$$

The other variance and the covariance can be derived similarly. Consequently, the meat matrix is given by

$$\boldsymbol{B}_H(\hat{\delta}) = \left(\sum_{h=1}^{H} I_h^2 \right) \begin{pmatrix} \tau_{a,\text{testlet}}^2 & \tau_{ab,\text{testlet}} \\ \tau_{ab,\text{testlet}} & \tau_{b,\text{testlet}}^2 \end{pmatrix} + I \begin{pmatrix} \tau_{a,\text{item}}^2 & \tau_{ab,\text{item}} \\ \tau_{ab,\text{item}} & \tau_{b,\text{item}}^2 \end{pmatrix} . \tag{42}$$

The bread matrix $\boldsymbol{A}_H$ only involves the expected value of the derivatives of the estimating equations. Hence, this matrix remains unchanged in a testlet structure [27].

The unknown variance and covariance components in (42) can be replaced by sample estimates. Then, an expected sandwich variance estimate can be obtained. A sample estimate of the meat matrix can be obtained by replacing the population variance in (41) with an empirical variance. Like in the case of independent items, the observed sandwich variance estimate $\hat{\boldsymbol{V}}_H$ can be modified to obtain a bias-corrected variant. In the testlet case, one should use correction factor $H/(H-1)$ instead of $I/(I-1)$.

5.3. Linking Error in Chain Linking

In this subsection, we discuss the computation of the linking error in chain linking [33,43,44] in log-mean-mean linking. Figure 1 illustrates the test design in the chain linking. The items are administered at three time points, T1, T2, and T3 (or in three groups). The goal is to determine the distribution parameters at T3. The distribution parameters of the ability variable θ at T3 can be compared with those at T1 by carrying out the linking step T1↔T2 and T2↔T3. The set of all items is denoted as $\mathcal{J} = \{1, \ldots, I\}$ and is partitioned into three distinct sets, $\mathcal{J}_0$, $\mathcal{J}_1$, and $\mathcal{J}_2$. The set $\mathcal{J}_0$ contains items that were

administered at all three time points. Items in $\mathcal{J}_1$ and $\mathcal{J}_2$ are administered at T1 and T2 and T2 and T3, respectively. We fix the distribution parameters at the first time point T1 to a mean of 0 and a standard deviation of 1. The mean of θ at T2 is denoted by μ_1 and the standard deviation by σ_1. The mean of θ at T3 is denoted by μ_2 and the standard deviation by σ_2.

	T1	T2	T3
Items $\mathcal{J}_0$	x	x	x
Items $\mathcal{J}_1$	x	x	
Items $\mathcal{J}_2$		x	x

Figure 1. Illustration of chain linking at three time points: T1, T2, and T3.

We assume that the 2PL model holds at the three time points. For longitudinal data, DIF is referred to as item parameter drift (IPD; [45,46]). The data-generating model involves IPD effects f_{it} for item discriminations and e_{it} for item intercepts ($t = 1, 2$).

$$\begin{aligned}
\log a_{i2} &= \log a_{i1} + f_{i1} \\
b_{i2} &= b_{i1} + e_{i1} \\
\log a_{i3} &= \log a_{i2} + f_{i2} \\
b_{i3} &= b_{i2} + e_{i2}
\end{aligned} \tag{43}$$

All IPD effects are allowed to be correlated within each item i but are uncorrelated across items.

In chain linking, the 2PL model is separately estimated for the three time points. The identified item parameters are given as $\hat{a}_{i1} = a_{i1}$, $\hat{b}_{i1} = b_{i1}$ and

$$\begin{aligned}
\hat{a}_{i2} &= a_{i2}\sigma_1 \\
\hat{b}_{i2} &= \sigma_1^{-1}(b_{i2} - \mu_1) \\
\hat{a}_{i3} &= a_{i3}\sigma_2 \\
\hat{b}_{i3} &= \sigma_2^{-1}(b_{i3} - \mu_2)
\end{aligned} \tag{44}$$

In the first linking step T1↔T2, the mean μ_1 and the standard deviation $\sigma_1 = \exp(s_1)$ are determined in log-mean-mean linking. In the second linking step T1↔T2, linking constants s_2 and m_2 are derived that refer to the linear transformation $\theta \mapsto \exp(s_2)\theta + m_2$.

In the chain linking approach, the unknown parameters are collected in the vector $\delta = (\mu_1, s_1, m_2, s_2)$. The parameters of interest (μ_2, σ_2) are computed as a derived parameter given by

$$\mu_2 = \exp(s_2)\mu_1 + m_2 \text{ and } \sigma_2 = \exp(s_2)\sigma_1 . \tag{45}$$

Let $I_k = |\mathcal{J}_k|$ be the number of items in the sets $\mathcal{J}_k$ for $k = 0, 1, 2$. We define $\kappa_0 = I_0/I$ and $\kappa_k = I_k/I$ for $k = 1, 2$. The two successive linking steps are formulated as a joint linking problem involving four estimating equations

$$
\begin{pmatrix}
\sum_{i=1}^{I} g_1(\delta; X_i) \\
\sum_{i=1}^{I} g_2(\delta; X_i) \\
\sum_{i=1}^{I} g_3(\delta; X_i) \\
\sum_{i=1}^{I} g_4(\delta; X_i)
\end{pmatrix}
=
\begin{pmatrix}
\sum_{i=1}^{I} d_{i1}(\log \hat{a}_{i2} - \log \hat{a}_{i1} - s_1) \\
\sum_{i=1}^{I} d_{i1}(\exp(s_1)\hat{b}_{i2} - \hat{b}_{i1} + \mu_1) \\
\sum_{i=1}^{I} d_{i2}(\log \hat{a}_{i3} - \log \hat{a}_{i2} - s_2) \\
\sum_{i=1}^{I} d_{i2}(\exp(s_2)\hat{b}_{i3} - \hat{b}_{i2} + m_2)
\end{pmatrix}
=
\begin{pmatrix}
0 \\ 0 \\ 0 \\ 0
\end{pmatrix},
\tag{46}
$$

where X_i includes all identified item parameters and design variables for item i. The first two estimating equations in (46) refer to log-mean-mean linking of the step T1↔T2, while the last two refer to step T2↔T3. In Equation (46), dummy indicators d_{ik} are used that take the value of one if item i is contained in $\mathcal{J}_0$ or $\mathcal{J}_k$ for $k = 1, 2$. Note that $\sum_{i=1}^{I} d_{ik} = I_k$ for $k = 1, 2$ and $\sum_{i=1}^{I} d_{i1}d_{i2} = I_0$.

We simplify the terms in (46) to

$$
g_1(\delta; X_i) = d_{i1}\{\log \hat{a}_{i2} - \log \hat{a}_{i1} - s_1\} = d_{i1}f_{i1}
\tag{47}
$$

$$
g_2(\delta; X_i) = d_{i1}\left\{\exp(s_1)\hat{b}_{i2} - \hat{b}_{i1} + \mu_1\right\} = d_{i1}e_{i1}
\tag{48}
$$

$$
g_3(\delta; X_i) = d_{i2}\{\log \hat{a}_{i3} - \log \hat{a}_{i2} - s_2\} = d_{i2}f_{i2}
\tag{49}
$$

$$
g_4(\delta; X_i) = d_{i2}\left\{\exp(s_2)\hat{b}_{i3} - \hat{b}_{i2} + m_2\right\} = d_{i2}\sigma_1^{-1}e_{i2}
\tag{50}
$$

We can now compute the meat matrix B_I

$$
B_I(\delta) = I
\begin{pmatrix}
\kappa_1 \tau_{f_1}^2 & \kappa_1 \tau_{e_1 f_1} & \kappa_0 \tau_{f_1 f_2} & \kappa_0 \sigma_1^{-1} \tau_{f_1 e_2} \\
\kappa_1 \tau_{e_1 f_1} & \kappa_1 \tau_{e_1}^2 & \kappa_0 \tau_{e_1 f_2} & \kappa_0 \sigma_1^{-1} \tau_{e_1 e_2} \\
\kappa_0 \tau_{f_1 f_2} & \kappa_0 \tau_{e_1 f_2} & \kappa_2 \tau_{f_2}^2 & \kappa_2 \sigma_1^{-1} \tau_{e_2 f_2} \\
\kappa_0 \sigma_1^{-1} \tau_{f_1 e_2} & \kappa_0 \sigma_1^{-1} \tau_{e_1 e_2} & \kappa_2 \sigma_1^{-1} \tau_{e_2 f_2} & \kappa_2 \sigma_1^{-2} \tau_{e_2}^2
\end{pmatrix}.
\tag{51}
$$

Moreover, we can determine the expected value of the bread matrix $\mathrm{E}(A_I(\delta))$ as

$$
A_I(\delta) =
\begin{pmatrix}
-I\kappa_1 & 0 & 0 & 0 \\
I\kappa_1 D_1 & I\kappa_1 & 0 & 0 \\
0 & 0 & -I\kappa_2 & 0 \\
0 & 0 & I\kappa_2 D_2 & I\kappa_2
\end{pmatrix}, \text{ where}
\tag{52}
$$

$$
\begin{aligned}
D_1 &= I^{-1}\kappa_1^{-1} \sum_{i=1}^{I} d_{i1}(b_{i1} - \mu_1) \\
D_2 &= I^{-1}\kappa_2^{-1} \sum_{i=1}^{I} d_{i2}(b_{i1} - \mu_2)
\end{aligned}.
\tag{53}
$$

The inverse of A_I can be computed as

$$[A_I(\delta)]^{-1} = \begin{pmatrix} -I^{-1}\kappa_1^{-1} & 0 & 0 & 0 \\ I^{-1}\kappa_1^{-1}D_1 & I^{-1}\kappa_1^{-1} & 0 & 0 \\ 0 & 0 & -I^{-1}\kappa_2^{-1} & 0 \\ 0 & 0 & I^{-1}\kappa_2^{-1}D_2 & I^{-1}\kappa_2^{-1} \end{pmatrix} . \tag{54}$$

Using the matrices $B_I(\delta)$ and $[A_I(\delta)]^{-1}$, the variance matrix $V_I(\delta)$ of $\hat{\delta}$ can be computed using the sandwich formula (6).

We now explicitly derive linking errors for $\hat{\sigma}_2$ and $\hat{\mu}_2$. First, the standard deviation σ_2 at T3 is given as $\sigma_2 = h(\delta) = \exp(s_1)\exp(s_2)$. We derive the variance by applying the delta method to the nonlinear transformation h and using the variance matrix V_I. The first-order partial derivatives of h are given by

$$u^\top = \left(\frac{\partial\sigma_2}{\partial\delta}\right) = \begin{pmatrix} \exp(s_1)\exp(s_2) & 0 & \exp(s_1)\exp(s_2) & 0 \end{pmatrix} . \tag{55}$$

Hence, the linking error of the estimated standard deviation $\hat{\sigma}_2$ can be determined as

$$\mathrm{LE}(\hat{\sigma}_2) = \sqrt{u^\top V_I(\hat{\delta})u} . \tag{56}$$

The algebraic derivation for the linking error formula was somewhat more intricate, which is why the R package rSymPy package [47] as a wrapper to the SymPy computer algebra system [48] was used. The square of the linking error for $\hat{\sigma}_2$ (i.e., the quantity $\mathrm{LE}(\hat{\sigma}_2)^2$) is computed using (56)

$$\begin{aligned} \mathrm{LE}(\hat{\sigma}_2)^2 &= \frac{\tau_{f_1}^2}{I}\left[\frac{\sigma_1^2\sigma_2^2}{\kappa_1}\right] + \frac{\tau_{f_2}^2}{I}\left[\frac{\sigma_1^2\sigma_2^2}{\kappa_2}\right] - \frac{\tau_{e_1f_2}}{I}\left[\frac{D_1\kappa_0\sigma_1^2\sigma_2^2}{\kappa_1\kappa_2}\right] \\ &\quad - \frac{\tau_{e_2f_2}}{I}\left[\frac{D_2\sigma_1\sigma_2^2}{\kappa_2}\right] - \frac{\tau_{f_1e_2}}{I}\left[\frac{D_2\kappa_0\sigma_1\sigma_2^2}{\kappa_1\kappa_2}\right] + \frac{\tau_{f_1f_2}}{I}\left[\frac{2\kappa_0\sigma_1^2\sigma_2^2}{\kappa_1\kappa_2}\right] . \end{aligned} \tag{57}$$

The linking error for the estimated mean $\hat{\mu}_2$ can be determined as a derived parameter using the transformation $\mu_2 = h(\delta) = \exp(s_2)\mu_1 + m_2$. The first-order partial derivatives of h are given by

$$u^\top = \left(\frac{\partial\mu_2}{\partial\delta}\right) = \begin{pmatrix} 0 & \exp(s_2) & \exp(s_2)\mu_1 & 1 \end{pmatrix} . \tag{58}$$

The linking error for $\hat{\mu}_2$ can be computed by using (56)

$$\begin{aligned} \mathrm{LE}(\hat{\mu}_2)^2 &= \frac{\tau_{f_2}^2}{I}\left[\frac{\mu_1^2\sigma_2^2 - D_2\mu_1\sigma_2}{\kappa_2}\right] + \frac{\tau_{e_1}^2}{I}\left[\frac{\sigma_2^2}{\kappa_1}\right] + \frac{\tau_{e_2}^2}{I}\left[\frac{1 + D_2\mu_1\sigma_2}{\kappa_2\sigma_1^2}\right] \\ &\quad + \frac{\tau_{e_1f_2}}{I}\left[\frac{D_2\kappa_0\sigma_2 - 2\kappa_0\mu_1\sigma_2^2}{\kappa_1\kappa_2}\right] + \frac{\tau_{e_2f_2}}{I}\left[\frac{D_2 - 2\mu_1\sigma_2 + \mu_1\sigma_2D_2^2 - D_2\mu_1^2\sigma_2^2}{\kappa_2\sigma_1}\right) \\ &\quad + \frac{\tau_{f_1e_2}}{I}\left[\frac{D_1\kappa_0\sigma_2 + D_1D_2\kappa_0\mu_1\sigma_2^2}{\kappa_1\kappa_2\sigma_1}\right] - \frac{\tau_{f_1f_2}}{I}\left[D_1\frac{\kappa_0\mu_1\sigma_2^2}{\kappa_1\kappa_2}\right] . \end{aligned} \tag{59}$$

In chain linking in the 1PL model, the linking error for $\hat{\mu}_2$ substantially simplifies by setting $\sigma_1 = \sigma_2 = 1$, $D_2 = 0$ (see also [49])

$$\text{LE}(\hat{\mu}_2)^2 = \frac{\tau_{e_1}^2}{I}\left[\frac{1}{\kappa_1}\right] + \frac{\tau_{e_2}^2}{I}\left[\frac{1}{\kappa_2}\right] . \tag{60}$$

Note that all DIF effects referring to item discriminations vanish in this case.

5.4. Linking Error for Trend Estimates in Educational Large-Scale Assessment Studies

We now turn to the important application of linking errors for trend estimates [13,15,50,51] in educational LSA studies such as PISA [16,52]. The main goal is to compute a linking error for a trend estimate in country means or country standard deviations between two successive assessments. Again, we rely on the 2PL model and use log-mean-mean linking for the derivation of the linking errors. Previous research derived closed formulas for linking errors in the 1PL model [17]. It was stated in official PISA publications that there does not exist a simple generalization to the 2PL model [52]. However, this section provides a closed formula for the 2PL model.

Figure 2 illustrates the problem of trend estimation in LSA studies. The label "NAT" refers to a nation (i.e., a country) c. The label "INT" refers to the international metric that is defined as a pooled sample comprising all students of participating countries in the LSA study. A trend estimate for a country between two LSA assessments can involve means and standard deviations at T1 and T2 (i.e., it compares NAT2 and NAT1). The first linking step NAT1↔INT1 maps country-specific results onto an international metric at T1. This step allows a cross-sectional comparison of countries on an international reference metric. The second linking step INT1↔INT2 links results of two LSA studies at T1 and T2 at the international metric. The third linking step NAT2↔INT2 maps country-specific results to the international metric at T2.
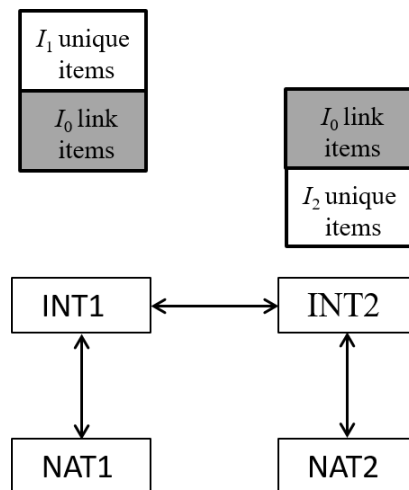

Figure 2. Trend estimation for country means and standard deviations at two time points in an international large-scale assessment study.

The set of administered items typically differs across assessments [53]. There are I_0 link items that are administered at both time points. A set of I_1 unique items is only administered at T1, while a set of I_2 unique items is only administered at T2. For identification reasons, we assume that the θ ability variable has zero mean and a standard deviation of one at T1. Then, we can identify the mean μ_{c1} and the standard deviation σ_{c1} of country c at T1. Furthermore, we assume that the mean at the international metric at T2 is μ_0 and the standard deviation is σ_0. We can also identify the mean μ_{c2} and the standard deviation σ_{c2} of country c at T2.

The first linking step NAT1↔INT1 in log-mean-mean linking estimates μ_{c1} and $s_{c1} = \log \sigma_{c2}$. The second linking step INT1↔INT2 estimates μ_0 and $s_0 = \log \sigma_0$. The third linking step NAT2↔INT2 estimates linking constants m_{c2} and s_{c2} to put the results of country c at the international metric at T2.

The country mean and standard deviation of country c at T2 are derived functions of the estimated linking parameters

$$\mu_{c2} = \exp(s_0)\mu_0 + m_{c2} \text{ and } \sigma_{c2} = \exp(s_{c2})\exp(s_0) \tag{61}$$

The linking constants for the link INT2↔NAT2 are recomputed using (61) as

$$m_{c2} = \mu_{c2} - \exp(s_0)\mu_0 \text{ and } s_{c2} = \log \sigma_{c2} - s_0 \tag{62}$$

The main idea is to apply the M-estimation theory and the sandwich formula for deriving linking errors for trend estimates. The three-step linking procedure can also be written as a simultaneous estimation problem involving six estimating equations for the vector of unknown linking parameters $\delta = (s_{c1}, \mu_{c1}, s_0, m_0, s_{c2}, m_{c2})$. The trend estimate in means is given as

$$\Delta\mu_c = \mu_{c2} - \mu_{c1} = \exp(s_0)\mu_0 + m_{c2} - \mu_{c1} \ . \tag{63}$$

The trend estimate in standard deviations is given as

$$\Delta\sigma_c = \sigma_2 - \sigma_1 = \exp(s_0 + s_{c2}) - \exp(s_{c1}) \ . \tag{64}$$

The source of linking errors in trend estimates is the presence of DIF effects. We now present a data-generating model for DIF effects in item parameters in the 2PL model. Let a_{ict} and b_{ict} for $t = 1, 2$ be item discriminations and item difficulties for country c. These item parameters are referred to as national item parameters [17]. International item parameters that result from item response models at the international metric that involves students from all countries are denoted by α_{it} and β_{it}. We use the same random effects model as in [17] for item difficulties

$$\begin{aligned} b_{ict} &= b_i + e_{it} + e_{ic} + e_{ict} \\ \beta_{ict} &= b_i + e_{it} \end{aligned} \tag{65}$$

The variance component $\text{Var}(e_{it}) = \tau_{b,\text{IPD}}^2$ refers to item parameter drift (IPD; [54]). The variance $\text{Var}(e_{ic}) = \tau_{b,\text{DIF}}^2$ is referred to as cross-sectional country DIF, and the variance $\text{Var}(e_{ict}) = \tau_{b,\text{DIF}\times\text{IPD}}^2$ refers to time-point-specific country DIF. All DIF effects are uncorrelated with each other.

We now extend the random effects model in [17] to item discriminations

$$\begin{aligned} \log a_{ict} &= \log a_i + f_{it} + f_{ic} + f_{ict} \\ \log \alpha_{ict} &= \log a_i + f_{it} \end{aligned} \tag{66}$$

All DIF effects for logarithmized item discriminations are uncorrelated, but DIF effects e and f can be correlated for the same type of heterogeneity (i.e., $\text{Cov}(e_{it}, f_{it}) = \tau_{ab,\text{IPD}}$, $\text{Cov}(e_{ic}, f_{ic}) = \tau_{ab,\text{DIF}}$, and $\text{Cov}(e_{ict}, f_{ict}) = \tau_{ab,\text{DIF}\times\text{IPD}}$).

Identified national item parameters are given as $\hat{a}_{ict} = a_{ict}\sigma_{ct}$ and $\hat{b}_{ict} = \sigma_{ct}^{-1}(b_{ict} - \mu_{ct})$ for $t = 1, 2$. Moreover, for international item parameters, it holds that $\hat{\alpha}_{i1} = \alpha_{i1}$ and $\hat{\beta}_{i1} = \beta_{i1}$ at T1, and $\hat{\alpha}_{i2} = \alpha_{i2}\sigma_1$ and $\hat{\beta}_{i2} = \sigma_0^{-1}(\beta_{i2} - \mu_0)$.

We now apply the M-estimation theory to the estimation of linking errors. The three log-mean-mean linking steps can be formalized as the following six estimating equations:

$$
\begin{pmatrix}
\sum_{i=1}^{I} g_1(\delta; X_i) \\
\sum_{i=1}^{I} g_2(\delta; X_i) \\
\sum_{i=1}^{I} g_3(\delta; X_i) \\
\sum_{i=1}^{I} g_4(\delta; X_i) \\
\sum_{i=1}^{I} g_5(\delta; X_i) \\
\sum_{i=1}^{I} g_6(\delta; X_i)
\end{pmatrix}
=
\begin{pmatrix}
\sum_{i=1}^{I} d_{i1}(\log \hat{a}_{ic1} - \log \hat{\alpha}_{i1} - s_1) \\
\sum_{i=1}^{I} d_{i1}(\exp(s_1)\hat{b}_{ic1} - \hat{\beta}_{i1} + \mu_1) \\
\sum_{i=1}^{I} d_{i0}(\log \hat{\alpha}_{i2} - \log \hat{\alpha}_{i1} - s_0) \\
\sum_{i=1}^{I} d_{i0}(\exp(s_0)\hat{\beta}_{i2} - \hat{\beta} b_{i1} + \mu_0) \\
\sum_{i=1}^{I} d_{i2}(\log \hat{a}_{ic2} - \log \hat{\alpha}_{i2} - s_2) \\
\sum_{i=1}^{I} d_{i2}(\exp(s_2)\hat{b}_{ic2} - \hat{\beta}_{i2} + m_2)
\end{pmatrix}
=
\begin{pmatrix}
0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0
\end{pmatrix}
\tag{67}
$$

The estimating equations in (67) define an estimate of the parameter of interest $\delta = (s_{c1}, \mu_{c1}, s_0, m_0, s_{c2}, m_{c2})$. We now use the sandwich formula to derive the asymptotic variance of $\hat{\delta}$ by means of the sandwich Formula (6).

The entries of the meat matrix $B_I(\delta)$ in the sandwich formula are denoted by

$$
B_I(\delta) =
\begin{pmatrix}
B_{11} & B_{21} & 0 & 0 & B_{51} & B_{61} \\
B_{21} & B_{22} & 0 & 0 & B_{52} & B_{62} \\
0 & 0 & B_{33} & B_{43} & 0 & 0 \\
0 & 0 & B_{43} & B_{44} & 0 & 0 \\
B_{51} & B_{52} & 0 & 0 & B_{55} & B_{65} \\
B_{61} & B_{62} & 0 & 0 & B_{65} & B_{66}
\end{pmatrix}.
\tag{68}
$$

The non-zero entries in $B_I(\delta)$ are given as

$$
\begin{aligned}
B_{11} &= I\kappa_1(\tau_{a,\mathrm{DIF}}^2 + \tau_{a,\mathrm{DIF}\times\mathrm{IPD}}^2) \\
B_{21} &= I\kappa_1(\tau_{ab,\mathrm{DIF}} + \tau_{ab,\mathrm{DIF}\times\mathrm{IPD}}) \\
B_{22} &= I\kappa_1(\tau_{b,\mathrm{DIF}}^2 + \tau_{b,\mathrm{DIF}\times\mathrm{IPD}}^2) \\
B_{33} &= 2I\kappa_0\tau_{a,\mathrm{IPD}}^2 \\
B_{43} &= 2I\kappa_0\tau_{ab,\mathrm{IPD}}^2 \\
B_{44} &= 2I\kappa_0\tau_{b,\mathrm{IPD}}^2 \\
B_{51} &= I\kappa_0\tau_{a,\mathrm{DIF}}^2 \\
B_{52} &= I\kappa_0\tau_{ab,\mathrm{DIF}} \\
B_{61} &= I\kappa_0\exp(s_0)^{-1}\tau_{ab,\mathrm{DIF}} \\
B_{62} &= I\kappa_0\exp(s_0)^{-1}\tau_{b,\mathrm{DIF}}^2 \\
B_{55} &= I\kappa_2(\tau_{a,\mathrm{DIF}}^2 + \tau_{a,\mathrm{DIF}\times\mathrm{IPD}}^2) \\
B_{65} &= I\kappa_2\exp(s_0)^{-1}(\tau_{ab,\mathrm{DIF}} + \tau_{ab,\mathrm{DIF}\times\mathrm{IPD}}) \\
B_{66} &= I\kappa_2\exp(s_0)^{-2}(\tau_{b,\mathrm{DIF}}^2 + \tau_{b,\mathrm{DIF}\times\mathrm{IPD}}^2)
\end{aligned}
\tag{69}
$$

Moreover, we can determine the bread matrix $A_I(\delta)$ as

$$
A_I(\delta) = \begin{pmatrix}
-I\kappa_1 & 0 & 0 & 0 & 0 & 0 \\
I\kappa_1 D_1 & I\kappa_1 & 0 & 0 & 0 & 0 \\
0 & 0 & -I\kappa_0 & 0 & 0 & 0 \\
0 & 0 & I\kappa_0 D_0 & I\kappa_0 & 0 & 0 \\
0 & 0 & 0 & 0 & -I\kappa_2 & 0 \\
0 & 0 & 0 & 0 & I\kappa_2 D_2 & I\kappa_2
\end{pmatrix} \text{, where}
\tag{70}
$$

$$
\begin{aligned}
D_1 &= I^{-1}\kappa_1^{-1} \sum_{i=1}^{I} d_{i1}(b_i - \mu_1) \\
D_0 &= I^{-1}\kappa_0^{-1} \sum_{i=1}^{I} d_{i0}(b_i - \mu_0) \\
D_2 &= I^{-1}\kappa_2 \sum_{i=1}^{I} d_{i2}(b_i - m_2)
\end{aligned}
\tag{71}
$$

The inverse of $A_I(\delta)$ can be computed as

$$
(A_I(\delta))^{-1} = \begin{pmatrix}
-I^{-1}\kappa_1^{-1} & 0 & 0 & 0 & 0 & 0 \\
I^{-1}\kappa_1^{-1}D_1 & I^{-1}\kappa_1^{-1} & 0 & 0 & 0 & 0 \\
0 & 0 & -I^{-1}\kappa_0^{-1} & 0 & 0 & 0 \\
0 & 0 & I^{-1}\kappa_0^{-1}D_0 & I^{-1}\kappa_0^{-1} & 0 & 0 \\
0 & 0 & 0 & 0 & -I^{-1}\kappa_2^{-1} & 0 \\
0 & 0 & 0 & 0 & I^{-1}\kappa_2^{-1}D_2 & I^{-1}\kappa_2^{-1}
\end{pmatrix}
\tag{72}
$$

Using (69) and (72), the variance matrix V_I can be computed.

We now derive the linking error of the trend estimate in standard deviations that is a nonlinear function of $\delta = (s_{c1}, \mu_{c1}, s_0, m_0, s_{c2}, m_{c2})$

$$
\Delta\sigma_c = h(\delta) = \sigma_2 - \sigma_1 = \exp(s_0 + s_{c2}) - \exp(s_{c1}) .
\tag{73}
$$

The first-order partial derivatives of h are given by

$$
\left(\frac{\partial h}{\partial \delta} \right) = \left(-\exp(s_{c1}) \quad 0 \quad \exp(s_0)\exp(s_{c2}) \quad 0 \quad \exp(s_0)\exp(s_{c2}) \quad 0 \right) .
\tag{74}
$$

Using (56), we can determine the square of the linking error $\mathrm{LE}(\Delta\sigma_c)^2$ by using computer algebra software [47,48] as

$$
\begin{aligned}
\mathrm{LE}(\Delta\sigma_c)^2 &= \frac{\tau_{a,\mathrm{DIF}}^2}{I}\left[\frac{\sigma_1^2}{\kappa_1} + \frac{\sigma_2^2}{\kappa_2} - \frac{2\sigma_1\sigma_2\kappa_0}{\kappa_1\kappa_2}\right] + \frac{\tau_{a,\mathrm{IPD}}^2}{I}\left[\frac{2\sigma_2^2}{\kappa_0}\right] + \frac{\tau_{a,\mathrm{DIF}\times\mathrm{IPD}}^2}{I}\left[\frac{\sigma_1^2}{\kappa_1} + \frac{\sigma_2^2}{\kappa_2}\right] \\
&\quad + \frac{\tau_{ab,\mathrm{DIF}}}{I}\left[\frac{D_2\sigma_1\sigma_2\kappa_0}{\kappa_1\kappa_2\sigma_0} + \frac{D_1\sigma_1\sigma_2\kappa_0}{\kappa_1\kappa_2} - \frac{D_1\sigma_1^2}{\kappa_1} - \frac{D_2\sigma_2^2}{\kappa_2\sigma_0}\right] - \frac{\tau_{ab,\mathrm{IPD}}}{I}\left[\frac{2D_0\sigma_2^2}{\kappa_0}\right] \\
&\quad - \frac{\tau_{ab,\mathrm{DIF}\times\mathrm{IPD}}}{I}\left[\frac{D_1\sigma_1^2}{\kappa_1} + \frac{D_2\sigma_2^2}{\kappa_2\sigma_0}\right] .
\end{aligned}
\tag{75}
$$

We now compute the linking error for the trend estimate in the mean in the 2PL model. The country mean difference can be computed as

$$\Delta\mu_c = h(\delta) = \mu_{c2} - \mu_{c1} = \exp(s_0)\mu_0 + m_{c2} - \mu_{c1} \,. \tag{76}$$

The first-order partial derivatives are given as

$$\left(\frac{\partial h}{\partial \delta}\right) = \left(\begin{array}{cccccc} 0 & -1 & \exp(s_0)\mu_0 & \exp(s_0) & 0 & 1 \end{array}\right) . \tag{77}$$

The linking error for the trend estimate in means is given by

$$
\begin{aligned}
\mathrm{LE}(\Delta\mu_c)^2 = {} & \frac{\tau_{b,\mathrm{DIF}}^2}{I}\left[\frac{1}{\kappa_1} + \frac{1}{\kappa_2\sigma_0^2} - \frac{2\kappa_0}{\sigma_0\kappa_1\kappa_2}\right] + \frac{\tau_{b,\mathrm{IPD}}^2}{I}\left[\frac{2\sigma_0^2(1 + D_0\mu_0)}{\kappa_0}\right] + \frac{\tau_{b,\mathrm{DIF}\times\mathrm{IPD}}^2}{I}\left[\frac{1}{\kappa_1} + \frac{1}{\sigma_0^2\kappa_2}\right] \\
& + \frac{\tau_{ab,\mathrm{DIF}}}{I}\left[\frac{D_1}{\kappa_1} + \frac{D_2}{\sigma_0\kappa_2} - \frac{D_2\kappa_0}{\kappa_1\kappa_2} - \frac{D_1\kappa_0}{\sigma_0\kappa_1\kappa_2}\right] + \frac{\tau_{a,\mathrm{IPD}}^2}{I}\left[\frac{2\sigma_0^2\mu_0(\mu_0 - D_0)}{\kappa_0}\right] \\
& + \frac{\tau_{ab,\mathrm{IPD}}}{I}\left[\frac{2D_0\sigma_0^2 + 2\mu_0 D_0^2\sigma_0^2 - 2D_0\sigma_0^2\mu_0^2 - 4\mu_0\sigma_0^2}{\kappa_0}\right] + \frac{\tau_{ab,\mathrm{DIF}\times\mathrm{IPD}}}{I}\left[\frac{D_1}{\kappa_1} + \frac{D_2}{\sigma_0\kappa_2}\right] .
\end{aligned} \tag{78}
$$

The linking error formula for the 1PL model derived in [17] can be obtained from the terms in (78) that involve the variance components $\tau_{b,\mathrm{DIF}}^2$, $\tau_{b,\mathrm{IPD}}^2$, and $\tau_{b,\mathrm{DIF}\times\mathrm{IPD}}^2$ by setting $\sigma_0^2 = 1$ and $D_0 = 0$:

$$\mathrm{LE}(\Delta\mu_c)^2 = \frac{\tau_{b,\mathrm{DIF}}^2}{I}\left[\frac{1}{\kappa_1} + \frac{1}{\kappa_2} - \frac{2\kappa_0}{\kappa_1\kappa_2}\right] + \frac{\tau_{b,\mathrm{IPD}}^2}{I}\left[\frac{2}{\kappa_0}\right] + \frac{\tau_{b,\mathrm{DIF}\times\mathrm{IPD}}^2}{I}\left[\frac{1}{\kappa_1} + \frac{1}{\kappa_2}\right] \tag{79}$$

All other components in (78) vanish in the case of the 1PL model.

5.5. Linking Error in Fixed Item Parameter Calibration

In this subsection, linking errors for the estimated mean and the estimated standard deviations under fixed item parameter calibration (FIPC) are derived. It is assumed that one uses fixed item parameters a_{i1} and b_{i1}, but the true item parameters a_{i2} and b_{i2} have DIF effects and follow the data-generating model (2).

The FIPC is typically applied using marginal maximum likelihood estimation [55]. However, we derive the linking error for a diagonally weighted least squares (DWLS) estimation that approximates the former estimation method [56]. The DWLS minimizes the weighted sums of the differences between the estimated and model-implied item thresholds as well as the tetrachoric correlations. For the simplicity of exposition in this subsection, we assume that there are only DIF effects in item difficulties (i.e., uniform DIF). Hence, we can assume that the standard deviation σ can be estimated without bias.

The item-specific weights in the DWLS estimation are precision weights ω that are defined as the inverse of the variance in thresholds τ. When omitting a factor that involves the sample size, precision weights ω are determined by

$$\omega = \omega(\tau) = \frac{\phi(\tau)^2}{\Phi(\tau)(1 - \Phi(\tau))} \,. \tag{80}$$

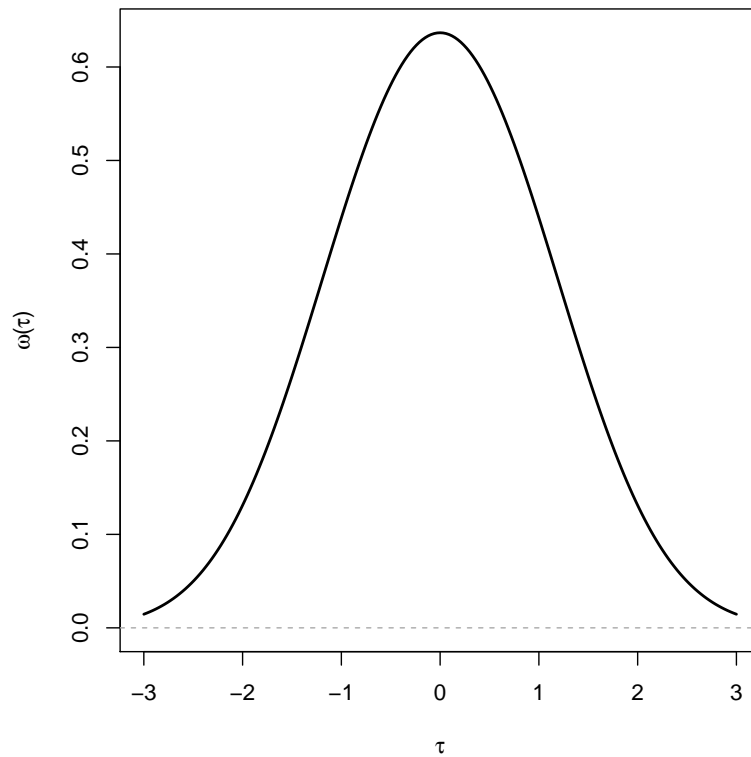Equation (80) is displayed in Figure 3. It can be seen that extreme thresholds are downweighted.

Figure 3. Precision weights ω as a function of the threshold τ (see Equation (80)).

The optimization function for μ in DWLS is given by

$$H(\mu) = \sum_{i=1}^{I} \omega_i (\hat{\tau}_{i2} - \tau_{i2}(\mu))^2 \, , \tag{81}$$

where $\hat{\tau}_{i2}$ is the identified threshold of item i and τ_{i2} is the model-implied threshold. Now, define $\alpha_i = a_i \sigma (a_i^2 \sigma^2 + \mathcal{L})^{-1/2}$, where $\mathcal{L}$ is the logistic variance that is the byproduct of using the logistic instead of the probit link function in the 2PL model. Then, (81) can be rewritten as

$$H(\mu) = \sum_{i=1}^{I} \omega_i \alpha_i (\tilde{b}_{i2} - (\mu - b_{i1}))^2 = \sum_{i=1}^{I} w_i (\hat{b}_{i2} - (\mu - b_{i1}))^2 \, , \tag{82}$$

where $w_i = C \omega_i \alpha_i$ (C is an appropriate scaling constant) and $\hat{b}_{i2} = \alpha_i^{-1} \hat{\tau}_{i2}$. Note that $\tilde{b}_{i2} = \mu_0 - b_{i2} = \mu_0 - b_{i1} - e_i$ for a data-generating true mean μ_0.

The minimization of (82) leads to a weighted mean

$$\hat{\mu} = \frac{\sum_{i=1}^{I} w_i (\tilde{b}_{i2} - b_{i1})}{W} = \mu_0 - \frac{\sum_{i=1}^{I} w_i e_i}{W} \, , \tag{83}$$

where $W = \sum_{i=1}^{I} w_i$, and μ_0 in (83) denotes the true mean.

Hence, one can determine the linking error as the variance in the second term on the right-hand side in (83). We then obtain the linking error in fixed item parameter calibration

$$\mathrm{LE}(\hat{\mu}) = \frac{\sqrt{\sum_{i=1}^{I} w_i^2}}{W} \tau_b \, . \tag{84}$$

By the Cauchy–Schwarz inequality, it holds that

$$W = \sum_{i=1}^{I} w_i \leq \sqrt{I} \cdot \sqrt{\sum_{i=1}^{I} w_i^2} \; . \tag{85}$$

Hence, we get from (84) by using (85)

$$\mathrm{LE}(\hat{\mu}) \geq \frac{1}{\sqrt{I}} \tau_b \; . \tag{86}$$

Hence, the linking error has a lower bound in which all item-specific weights were set equal to one. This case corresponds to the linking error in the Rasch model obtained for mean-mean linking [15,17]

$$\mathrm{LE}(\hat{\mu}) = \frac{\tau_b}{\sqrt{I}} \; . \tag{87}$$

This linking error is also obtained in FIPC of the 1PL model using unweighted least squares (ULS) estimation.

Interestingly, the finding in (86) illustrates that using incorrect item parameters in the FIPC in the presence of DIF effects results in a precision loss. Hence, the maximum likelihood estimation can only achieve the most efficient estimates under correctly specified models. Because it is almost always expected that there are some intentionally unmodeled DIF effects in real datasets, the dominance of the maximum likelihood estimation in LSA practice can be questioned.

5.6. Linking Error in Concurrent Calibration

In this subsection, we derive the linking error for the estimation of the 2PL model in two groups. The mean and the standard deviation of the ability variable θ are fixed to 0 and 1, respectively. The mean μ and the standard deviation σ of the second group are estimated.

Like in Section 5.5, we assume that there are only DIF effects e_i in item difficulties and no DIF effects in item discriminations. We assume that σ can be estimated without bias and derive the linking error under concurrent calibration using DWLS. DIF effects e_i with zero means follow $b_{i2} = b_{i1} + e_i$. The identified parameters are given by $\hat{b}_{i1} = b_{i1}$ and $\hat{b}_{i2} = b_{i2} - \mu$. The estimation of the vector of common item difficulties $\boldsymbol{b} = (b_1, \dots, b_I)$ and the mean μ is conducted using a weighted square loss of differences in estimated and model-implied item thresholds. By simplifying the setting while assuming known item discriminations and standard deviation, the optimization function is given by

$$H(\mu, \boldsymbol{b}) = \sum_{i=1}^{I} w_{i1}(\hat{b}_{i1} - b_i)^2 + \sum_{i=1}^{I} w_{i2}(\hat{b}_{i2} + \mu - b_i)^2 \; , \tag{88}$$

where weights w_{ig} are allowed to be group specific ($g = 1, 2$). The estimating equations for μ and $\boldsymbol{b}$ are given as

$$w_{i1}(\hat{b}_{i1} - \hat{b}_i) + w_{i2}(\hat{b}_{i2} + \hat{\mu} - \hat{b}_i) = 0 \text{ and} \tag{89}$$

$$\sum_{i=1}^{I} w_{i2}(\hat{b}_{i2} + \hat{\mu} - \hat{b}_i) = 0 \; . \tag{90}$$

By inserting (89) into (90), we obtain

$$\sum_{i=1}^{I} w_{i2}\hat{b}_{i2} + \sum_{i=1}^{I} w_{i2}\hat{\mu} - \sum_{i=1}^{I} w_{i2} \frac{w_{i1}\hat{b}_{i1} + w_{i2}\hat{b}_{i2} + w_{i2}\hat{\mu}}{w_{i1} + w_{i2}} = 0 \; . \tag{91}$$

Equation (91) can be further simplified to

$$\sum_{i=1}^{I} \frac{w_{i1}w_{i2}}{w_{i1}+w_{i2}} \hat{b}_{i2} + \sum_{i=1}^{I} \frac{w_{i1}w_{i2}}{w_{i1}+w_{i2}} \hat{\mu} - \sum_{i=1}^{I} \frac{w_{i1}w_{i2}}{w_{i1}+w_{i2}} \hat{b}_{i1} = 0 \,. \tag{92}$$

By defining $w_i^* = (w_{i1}+w_{i2})^{-1}w_{i1}w_{i2}$, we obtain from (92)

$$\hat{\mu} = -\frac{\sum_{i=1}^{I} w_i^*(\hat{b}_{i2}-\hat{b}_{i1})}{\sum_{i=1}^{I} w_i^*} = \mu - \frac{\sum_{i=1}^{I} w_i^* e_i}{\sum_{i=1}^{I} w_i^*} \,. \tag{93}$$

Hence, we obtain the linking error for $\hat{\mu}$

$$\mathrm{LE}(\hat{\mu}) = \frac{\sqrt{\sum_{i=1}^{I}(w_i^*)^2}}{W^*} \tau_b \,, \tag{94}$$

where $W^* = \sum_{i=1}^{I} w_i^*$. Again, one can conclude $\mathrm{LE}(\hat{\mu}) \geq \tau_b/\sqrt{I}$. Hence, using the linking error $\tau_b/\sqrt{I}$ that does not take item-specific weights into account provides a lower bound of the true linking error.

If the weights w_{i1} and w_{i2} would be equal across both groups, we obtain the same linking error like under fixed item parameter calibration.

5.7. Linking Error for Derived Parameters

In previous sections of this paper, we derived the linking error for μ and σ for different applications of the 2PL model. Sometimes, other distribution parameters might be estimated. In this subsection, we assume that the ability variable θ is approximately normally distributed and derive the linking error for nonlinear functions h of μ and σ. Let $v = h(\delta) = h(\mu, \sigma)$ be a nonlinear function of the mean μ and the standard deviation σ. Let $\boldsymbol{V}$ denote the variance matrix of δ regarding item choice (i.e., quantifying linking errors). The linking error estimate $\hat{v} = h(\hat{\mu}, \hat{\sigma})$ is given as

$$\mathrm{LE}(\hat{v}) = \sqrt{\boldsymbol{u}^\top \boldsymbol{V} \boldsymbol{u}} \,, \text{ where } \boldsymbol{u}^\top = \left(\begin{array}{cc} \dfrac{\partial h}{\partial \mu} & \dfrac{\partial h}{\partial \sigma} \end{array} \right) \text{ and } \boldsymbol{V} = \left(\begin{array}{cc} v_\mu^2 & v_{\mu\sigma} \\ v_{\mu\sigma} & v_\sigma^2 \end{array} \right) . \tag{95}$$

We now illustrate the computation in two examples.

5.7.1. Proportions

First, we are interested in computing the probability p

$$p = h(\mu, \sigma) = \mathrm{P}(\theta \in [c_1, c_2]) = \Phi\left(\sigma^{-1}(c_2-\mu)\right) - \Phi\left(\sigma^{-1}(c_1-\mu)\right) . \tag{96}$$

The partial derivatives of h are given as

$$\frac{\partial h}{\partial \mu} = -\sigma^{-1}\left[\phi\left(\sigma^{-1}(c_2-\mu)\right) - \phi\left(\sigma^{-1}(c_1-\mu)\right)\right] = -\sigma^{-1}d(\mu,\sigma,c_1,c_2) \text{ and} \tag{97}$$

$$\frac{\partial h}{\partial \sigma} = -\sigma^{-2}\left[\phi\left(\sigma^{-1}(c_2-\mu)\right) - \phi\left(\sigma^{-1}(c_1-\mu)\right)\right] = -\sigma^{-2}d(\mu,\sigma,c_1,c_2) \,, \tag{98}$$

where $d(\mu, \sigma, c_1, c_2) = \phi(\sigma^{-1}(c_2 - \mu)) - \phi(\sigma^{-1}(c_1 - \mu))$ Hence, we can determine the linking error of $\hat{p} = g(\hat{\mu}, \hat{\sigma})$ using (95) as

$$\text{LE}(\hat{p}) = \hat{\sigma}^{-1} d(\hat{\mu}, \hat{\sigma}, c_1, c_2) \sqrt{\hat{v}_\mu^2 + \hat{\sigma}^{-2} \hat{v}_\sigma^2 + 2\hat{\sigma}^{-1} \hat{v}_{\mu\sigma}}, \tag{99}$$

where $\hat{v}_\mu^2$, $\hat{v}_\sigma^2$, and $\hat{v}_{\mu\sigma}$ are estimated linking variances and covariances.

5.7.2. Percentiles

Second, the linking error of the pth percentile Q_p should be computed. The pth percentile is defined by

$$\text{P}(\theta \leq Q_p) = \Phi(\sigma^{-1}(Q_p - \mu)) = p. \tag{100}$$

Hence, we can solve (100) for Q_p and obtain

$$Q_p = h(\mu, \sigma) = \mu + \sigma z_p, \tag{101}$$

where $z_p = \Phi^{-1}(p)$ is the pth percentile for the standard normal distribution.

Now, we can compute the linking error of $\hat{Q}_p$. We obtain

$$\frac{\partial h}{\partial \mu} = 1 \text{ and } \frac{\partial h}{\partial \sigma} = z_p. \tag{102}$$

The linking error of the percentile Q_p using (95) is given by

$$\text{LE}(\hat{Q}_p) = \sqrt{\hat{v}_\mu^2 + z_p^2 \hat{v}_\sigma^2 + 2z_p \hat{v}_{\mu\sigma}}. \tag{103}$$

It can be seen that for more extreme percentiles, the absolute value of z_p gets larger and the linking error for σ (i.e., $\hat{v}_\sigma$ in (103)) becomes more relevant.

5.8. Computation of Total Error and Sampling Error Correction for Linking Error Estimates

In practical applications, the sampling error SE due to the sampling of persons must also be taken into account to quantify the uncertainty in the σ and μ estimates. The total error (TE) is given by (see [18,19])

$$\text{TE} = \sqrt{\text{SE}^2 + \text{LE}^2}. \tag{104}$$

A critical issue might be that estimated linking errors also include a portion of variability that can be attributed to the sampling error of persons. We illustrate an analytical bias correction method for the case of log-mean-mean linking. The issue in the sandwich estimate is the variance matrix (i.e., the meat matrix) $\boldsymbol{B}_I$ (see (31) for the computation). The identified item parameters also include a sampling error variance that can be estimated by fitting an item response model. Let v_{ia}^* be the variance of $\log \hat{a}_{i2} - \log \hat{a}_{i1}$ due to person sampling (i.e., $v_{ia}^* = \text{Var}(\log \hat{a}_{i2} - \log \hat{a}_{i1})$). Then, we can determine a corresponding entry in the meat matrix $\boldsymbol{B}_I$ by modifying (31) into

$$\sum_{i=1}^{I} \text{Var}(g_1(\delta; a_{i1}, b_{i1}, f_i, e_i)) = \sum_{i=1}^{I} \text{Var}(\log \hat{a}_{i2} - \log \hat{a}_{i1} - s) = \sum_{i=1}^{I} (\text{Var}(f_i) + v_{ia}^*) = I\tau_a^2 + \sum_{i=1}^{I} v_{ia}^*. \tag{105}$$

In the derivation of (105), we relied on the property that item parameters of different items are approximately uncorrelated in sufficiently long tests [57]. The other entries in $\boldsymbol{B}_I$ can be similarly determined. By defining $v_{ib}^* = \text{Var}(\hat{b}_{i2} - \hat{b}_{i1})$ and $v_{iab}^* = \text{Cov}(\log \hat{a}_{i2} - \log \hat{a}_{i1}, \hat{b}_{i2} - \hat{b}_{i1})$, Equations (16) and (17) can be modified to

$$\sum_{i=1}^{I} \text{Cov}(g_1(\delta; a_{i1}, b_{i1}, f_i, e_i), g_2(\delta; a_{i1}, b_{i1}, f_i, e_i)) = I\tau_{ab} + \sum_{i=1}^{I} v_{iab}^*, \text{ and} \tag{106}$$

$$\sum_{i=1}^{I} \text{Var}(g_2(\delta; a_{i1}, b_{i1}, f_i, e_i)) = I\tau_b^2 + \sum_{i=1}^{I} v_{ib}^* \, . \tag{107}$$

Hence, the originally obtained meat matrix $\boldsymbol{B}_I$ can be removed from sampling error contributions by defining a bias-corrected estimate

$$\boldsymbol{B}_I^*(\hat{\delta}) = \boldsymbol{B}_I(\hat{\delta}) - \boldsymbol{V}^* \, , \text{ where } \boldsymbol{V}^* = \begin{pmatrix} \displaystyle\sum_{i=1}^{I} v_{ia}^* & \displaystyle\sum_{i=1}^{I} v_{iab}^* \\ \displaystyle\sum_{i=1}^{I} v_{iab}^* & \displaystyle\sum_{i=1}^{I} v_{ib}^* \end{pmatrix} \, . \tag{108}$$

This approach was implemented in the simplified setting in the 1PL model [17].

We now show how to generalize the bias-corrected estimate of the meat matrix $\boldsymbol{B}_I$. We can rewrite the estimate from Equation (9) as

$$\hat{\boldsymbol{B}}_I(\hat{\delta}) = \sum_{i=1}^{I} \boldsymbol{g}(\hat{\delta}; \hat{\gamma}_i) \boldsymbol{g}(\hat{\delta}; \hat{\gamma}_i)^\top \, , \tag{109}$$

where $\hat{\gamma}_i = (\hat{a}_{i1}, \hat{b}_{i1}, \hat{a}_{i2}, \hat{b}_{i2})$. The estimate $\hat{\gamma}_i$ includes the sampling error, and the sampling variance is denoted as

$$\boldsymbol{V}_i = \text{Var}(\hat{\gamma}_i) = \text{E}[(\hat{\gamma}_i - \gamma_i)(\hat{\gamma}_i - \gamma_i)^\top] \, . \tag{110}$$

The estimating function $\boldsymbol{g}$ in (109) can be viewed as a function $\boldsymbol{g} = \boldsymbol{g}(\delta, \gamma)$. Denote by $\boldsymbol{g}_\gamma = (\partial \boldsymbol{g})/(\partial \gamma)$ the matrix of partial derivatives with respect to γ. We can now apply a Taylor expansion and obtain

$$\boldsymbol{g}(\hat{\delta}, \hat{\gamma}_i) = \boldsymbol{g}(\hat{\delta}, \gamma_i) + \boldsymbol{g}_\gamma(\hat{\delta}, \gamma_i)(\hat{\gamma}_i - \gamma_i) \, . \tag{111}$$

Using $\text{E}(\hat{\gamma}_i) = \gamma$, we can obtain a bias-corrected estimate of the bread matrix as

$$\boldsymbol{B}_I^*(\hat{\delta}) = \hat{\boldsymbol{B}}_I(\hat{\delta}) - \sum_{i=1}^{I} \boldsymbol{g}_\gamma(\hat{\delta}, \gamma_i) \boldsymbol{V}_i \boldsymbol{g}_\gamma(\hat{\delta}, \gamma_i)^\top \, . \tag{112}$$

6. Discussion

In this article, we have shown that the sandwich formula from the M-estimation theory can be successfully employed for computing the linking error in the 2PL model in a variety of situations. It was shown in a simulation study for the log-mean-mean linking of two groups that the expected sandwich estimate of the linking error produced satisfactory coverage rates. Interestingly, it had a comparable performance to the jackknife linking error in the 2PL model.

As with any simulation study, some limitations of our study can be stated. More comprehensive studies could involve a different range of standard deviations of the DIF effects, different test lengths, or other linking approaches. It might be interesting to compare the performance of the M-estimation approach with the jackknife linking error for the more complex linking problems of chain linking and trend estimation.

In the simulation study, we only considered uncertainty in distribution parameters due to DIF effects (i.e., linking errors). In practical applications, there will also be a sampling of persons, and the simultaneous assessment of sampling errors and linking errors would be an exciting extension of this study.

In most of the applications involving instruments with cognitive and noncognitive items, linking errors are not reported even if linking approaches were utilized [43,58,59]. There might be two reasons why this is the case. First, simulation studies typically presuppose that the IRT model perfectly fits the data. That is, the DIF is absent, and the IRT model is correctly specified. Second, items could be treated as fixed, and the model

misspecification is taken for granted but is not included in the uncertainty quantification of the linking approach. In our view, linking errors provide additional information about the impact of heterogeneous item functioning on parameters of interest and should, therefore, (almost always) be additionally reported. In general, we do not think that the presence of DIF or IPD threatens the validity of group differences or trend estimates.

It should be emphasized that our proposed linking error for the trend estimates in the 2PL model differs from a newly proposed linking error estimate since PISA 2015 [4]. The latter relies on a recalibration approach of the item response data from the first time point [34,60]. The new PISA linking error rather assesses the extent of the assumed noninvariant item parameters across the two PISA studies instead of quantifying the variability due to heterogeneous item functioning.

The derived linking errors rely on the assumption that item parameters are identified. In LSA studies such as PISA, balanced incomplete block designs are employed in which only a subset of items is administered to each student [61]. If item parameters can be identified in such test designs, linking error formulas would not change because country means and standard deviations are based on all the items administered in the test, irrespective of the proportion of items administered to each student.

As argued by an anonymous reviewer, the computation of linking errors relies on the assumption of a representative item sample from an item population. Notably, the identification of item parameters in the joint maximum likelihood estimation also requires asymptotic regimes regarding the sample size and the number of items [62–64] but does not require that items are a random sample from an item population. However, the main difference in the computation of linking errors is that linking errors reflect the variability in the distribution parameter estimates due to the DIF. Infinite item samples are not required for model identification; they are only used as a statistical tool for justifying the statistical inference with respect to modeled or unmodeled DIF.

Finally, we assumed that DIF effects had zero means throughout this paper. However, this assumption is not essential in deriving linking errors because the M-estimation theory does not require that estimated parameters converge to true parameters. The M-estimation theory will nevertheless provide the asymptotic variance in a potentially biased estimator. However, DIF effects could also result in a mean that is different from zero. For example, Ref. [28] assumes that the median of the DIF effects equals zero. In this case, log-mean-mean linking in the 2PL model could be replaced by an alternative robust linking method in which the mean is substituted with the median. Notably, the choice of an adequate estimating function to provide unbiased distribution parameters depending on the distribution of DIF effects is a different topic though [18].

Abbreviations

The following abbreviations are used in this manuscript:

1PL	one-parameter logistic
2PL	two-parameter logistic
DIF	differential item functioning
DWLS	diagonally weighted least squares
FIPC	fixed item parameter calibration
IPD	item parameter drift
IRT	item response theory
JK	jackknife
LE	linking error
LSA	large-scale assessment studies
PIRLS	progress in international reading literacy study
PISA	programme for international student assessment
ULS	unweighted least squares

References

1. Chen, Y.; Li, X.; Liu, J.; Ying, Z. Item response theory—A statistical framework for educational and psychological measurement. *arXiv* **2021**, arXiv:2108.08604.
2. 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]
3. 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: London, UK; CRC Press: Boca Raton, FL, USA, 2013. [CrossRef]
4. OECD. *PISA 2018. Technical Report*; OECD: Paris, France, 2020.
5. 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.
6. Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*; Danish Institute for Educational Research: Copenhagen, Denmark, 1960.
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, Vol. 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. Holland, P.W.; Wainer, H. (Eds.) *Differential Item Functioning: Theory and Practice*; Lawrence Erlbaum: Hillsdale, NJ, USA, 1993. [CrossRef]
11. 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]
12. 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]
13. 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]
14. Battauz, M. Multiple equating of separate IRT calibrations. *Psychometrika* **2017**, *82*, 610–636. [CrossRef]
15. Monseur, C.; Berezner, A. The computation of equating errors in international surveys in education. *J. Appl. Meas.* **2007**, *8*, 323–335.
16. OECD. *PISA 2012. Technical Report*; OECD: Paris, France, 2014. Available online: https://bit.ly/2YLG24g (accessed on 3 December 2022).
17. 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]
18. 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]
19. Wu, M. Measurement, sampling, and equating errors in large-scale assessments. *Educ. Meas.* **2010**, *29*, 15–27. [CrossRef]
20. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; CRC Press: Boca Raton, FL, USA, 1994. [CrossRef]
21. Kolenikov, S. Resampling variance estimation for complex survey data. *Stata J.* **2010**, *10*, 165–199. [CrossRef]
22. Boos, D.D.; Stefanski, L.A. *Essential Statistical Inference*; Springer: New York, NY, USA, 2013. [CrossRef]
23. Stefanski, L.A.; Boos, D.D. The calculus of M-estimation. *Am. Stat.* **2002**, *56*, 29–38. [CrossRef]
24. Zeileis, A. Object-oriented computation of sandwich estimators. *J. Stat. Softw.* **2006**, *16*, 1–16. [CrossRef]

25. Fay, M.P.; Graubard, B.I. Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics* **2001**, *57*, 1198–1206. [CrossRef]
26. Li, P.; Redden, D.T. Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. *Stat. Med.* **2015**, *34*, 281–296. [CrossRef]
27. Zeileis, A.; Köll, S.; Graham, N. Various versatile variances: An object-oriented implementation of clustered covariances in R. *J. Stat. Softw.* **2020**, *95*, 1–36. [CrossRef]
28. Chen, Y.; Li, C.; Xu, G. DIF statistical inference and detection without knowing anchoring items. *arXiv* **2021**, arXiv:2110.11112.
29. Halpin, P.F. Differential item functioning via robust scaling. *arXiv* **2022**, arXiv:2207.04598.
30. Wang, W.; Liu, Y.; Liu, H. Testing differential item functioning without predefined anchor items using robust regression. *J. Educ. Behav. Stat.* **2022**, *47*, 666–692. [CrossRef]
31. Robitzsch, A. L_p loss functions in invariance alignment and Haberman linking with few or many groups. *Stats* **2020**, *3*, 246–283. [CrossRef]
32. Hunter, J.E. Probabilistic foundations for coefficients of generalizability. *Psychometrika* **1968**, *33*, 1–18. [CrossRef]
33. Kolen, M.J.; Brennan, R.L. *Test Equating, Scaling, and Linking*; Springer: New York, NY, USA, 2014. [CrossRef]
34. Robitzsch, A. A comparison of linking methods for two groups for the two-parameter logistic item response model in the presence and absence of random differential item functioning. *Foundations* **2021**, *1*, 116–144. [CrossRef]
35. Maronna, R.A.; Martin, R.D.; Yohai, V.J. *Robust Statistics: Theory and Methods*; Wiley: New York, NY, USA, 2006. [CrossRef]
36. R Core Team. *R: A Language and Environment for Statistical Computing*; R Core Team: Vienna, Austria, 2022. Available online: https://www.R-project.org/ (accessed on 11 January 2022).
37. Novack-Gottshall, P.; Wang, S.C. *KScorrect: Lilliefors-Corrected Kolmogorov-Smirnov Goodness-of-Fit Tests*; R Package Version 1.4-0. 2019. Available online: https://CRAN.R-project.org/package=KScorrect (accessed on 3 July 2019).
38. Haebara, T. Equating logistic ability scales by a weighted least squares method. *Jpn. Psychol. Res.* **1980**, *22*, 144–149. [CrossRef]
39. Bradlow, E.T.; Wainer, H.; Wang, X. A Bayesian random effects model for testlets. *Psychometrika* **1999**, *64*, 153–168. [CrossRef]
40. Sireci, S.G.; Thissen, D.; Wainer, H. On the reliability of testlet-based tests. *J. Educ. Meas.* **1991**, *28*, 237–247. [CrossRef]
41. Wainer, H.; Bradlow, E.T.; Wang, X. *Testlet Response Theory and Its Applications*; Cambridge University Press: Cambridge, UK, 2007. [CrossRef]
42. Monseur, C.; Sibberns, H.; Hastedt, D. Linking errors in trend estimation for international surveys in education. *IERI Monogr. Ser.* **2008**, *1*, 113–122. https://bit.ly/38aTVeZ.
43. Battauz, M. IRT test equating in complex linkage plans. *Psychometrika* **2013**, *78*, 464–480. [CrossRef]
44. Battauz, M. Factors affecting the variability of IRT equating coefficients. *Stat. Neerl.* **2015**, *69*, 85–101. [CrossRef]
45. Arce-Ferrer, A.J.; Bulut, O. Investigating separate and concurrent approaches for item parameter drift in 3PL item response theory equating. *Int. J. Test.* **2017**, *17*, 1–22. [CrossRef]
46. Taherbhai, H.; Seo, D. The philosophical aspects of IRT equating: Modeling drift to evaluate cohort growth in large-scale assessments. *Educ. Meas.* **2013**, *32*, 2–14. [CrossRef]
47. Grothendieck, G. *rSymPy: R Interface to SymPy Computer Algebra System*. R Package Version 0.2-1.2. 2010. Available online: https://CRAN.R-project.org/package=rSymPy (accessed on 31 July 2010).
48. Meurer, A.; Smith, C.P.; Paprocki, M.; Čertík, O.; Kirpichev, S.B.; Rocklin, M.; Kumar, A.; Ivanov, S.; Moore, J.K.; Singh, S.; et al. SymPy: Symbolic computing in Python. *PeerJ Comput. Sci.* **2017**, *3*, e103. [CrossRef]
49. Fischer, L.; Gnambs, T.; Rohm, T.; Carstensen, C.H. Longitudinal linking of Rasch-model-scaled competence tests in large-scale assessments: A comparison and evaluation of different linking methods and anchoring designs based on two tests on mathematical competence administered in grades 5 and 7. *Psych. Test Assess. Model.* **2019**, *61*, 37–64. https://bit.ly/3F1Hd0C.
50. 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]
51. Sachse, K.A.; Mahler, N.; Pohl, S. When nonresponse mechanisms change: Effects on trends and group comparisons in international large-scale assessments. *Educ. Psychol. Meas.* **2019**, *79*, 699–726. [CrossRef]
52. OECD. *PISA 2015. Technical Report*; OECD: Paris, France, 2017. Available online: https://bit.ly/32buWnZ (accessed on 3 December 2022).
53. Weeks, J.; von Davier, M.; Yamamoto, K. Design considerations for the program for international student assessment. 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: London, UK; CRC Press: Boca Raton, FL, USA, 2013; pp. 259–276. [CrossRef]
54. Kang, H.A.; Lu, Y.; Chang, H.H. IRT item parameter scaling for developing new item pools. *Appl. Meas. Educ.* **2017**, *30*, 1–15. [CrossRef]
55. 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]
56. Cai, L.; Moustaki, I. Estimation methods in latent variable models for categorical outcome variables. In *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test*; Irwing, P., Booth, T., Hughes, D.J., Eds.; Wiley: New York, NY, USA, 2018; pp. 253–277. [CrossRef]
57. 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]

58. González, J.; Wiberg, M. *Applying Test Equating Methods. Using R*; Springer: New York, NY, USA, 2017. [CrossRef]
59. Jewsbury, P.A. *Error Variance in Common Population Linking Bridge Studies*; (Research Report No. RR-19-42); Educational Testing Service: Princeton, NJ, USA, 2019. [CrossRef]
60. Martin, M.O.; Mullis, I.V.S.; Foy, P.; Brossman, B.; Stanco, G.M. Estimating linking error in PIRLS. *IERI Monogr. Ser.* **2012**, *5*, 35–47. Available online: https://bit.ly/2Vx3el8 (accessed on 3 December 2022).
61. 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]
62. Chen, Y.; Li, X.; Zhang, S. Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. *Psychometrika* **2019**, *84*, 124–146. [CrossRef]
63. Chen, Y.; Li, X.; Zhang, S. Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications. *J. Am. Stat. Assoc.* **2020**, *115*, 1756–1770. [CrossRef]
64. Haberman, S.J. Maximum likelihood estimates in exponential response models. *Ann. Stat.* **1977**, *5*, 815–841. [CrossRef]