AppliedMath

MDPI

Article

Analytical Approximation of the Jackknife Linking Error in Item Response Models Utilizing a Taylor Expansion of the Log-Likelihood Function

Alexander Robitzsch [1,2]

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

Abstract: Linking errors in item response models quantify the dependence on the chosen items in means, standard deviations, or other distribution parameters. The jackknife approach is frequently employed in the computation of the linking error. However, this jackknife linking error could be computationally tedious if many items were involved. In this article, we provide an analytical approximation of the jackknife linking error. The newly proposed approach turns out to be computationally much less demanding. Moreover, the new linking error approach performed satisfactorily for datasets with at least 20 items.

Keywords: item response model; linking error; jackknife

MSC: 62H10; 62H25; 65-04; 65D15

Citation: 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. https://doi.org/10.3390/appliedmath3010004

Academic Editor: Valery Karachik

Received: 21 November 2022
Revised: 12 December 2022
Accepted: 30 December 2022
Published: 5 January 2023

1. Introduction

Item response theory (IRT) models [1–3] are an important class of multivariate statistics methodologies for analyzing dichotomous random variables used to model testing data from educational or psychological applications. This class aims to summarize a high-dimensional contingency table by a few latent factor variables of interest. Of particular relevance is the application of item response models in educational large-scale assessment [4], such as the studies programme for international student assessment (PISA; [5]) or progress in international reading literacy study (PIRLS; [6]).

In this article, only unidimensional IRT models are considered. Let $\boldsymbol{X} = (X_1, \ldots, X_I)$ be the vector of I dichotomous random variables $X_i \in \{0, 1\}$ (also referred to as items). A unidimensional item response model [1,7] is a statistical model for the probability distribution $\mathrm{P}(\boldsymbol{X} = \boldsymbol{x})$ for $\boldsymbol{x} = (x_1, \ldots, x_I) \in \{0, 1\}^I$, where

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

where ϕ 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 $P_i(\theta; \gamma_i) = \mathrm{P}(X_i = 1 | \theta)$.

The one-parameter logistic (1PL) model (also referred to as the Rasch model; [8]) uses 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, the vector of item parameters γ_i only consists of one entry; that is, $\gamma_i = (b_i)$. The two-parameter logistic (2PL) model [9] includes the item discrimination a_i in addition (i.e., $\gamma_i = (a_i, b_i)$), and the item response function is given by $P_i(\theta) = \Psi(a_i(\theta - b_i))$.

Please note that distribution parameters δ and item parameters γ cannot be simultaneously identified. If the parameters $(\mu, \sigma, \{(a_i, b_i)|i \in \{1, \ldots, I\}\})$ parametrize the 2PL model, an equivalent parametrization would be $(\mu = 0, \sigma = 1, \{(a_i\sigma, \sigma^{-1}(b_i - \mu))|i \in \{1, \ldots, I\}\})$. In applications like PISA in which a country mean μ and country a standard deviation σ , item parameters γ_i are often fixed at values γ_i^* that are used for all countries. In this case, μ and σ can be identified. If sample data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ for N persons are available, unknown model parameters in (1) can be estimated by (marginal) maximum likelihood (ML) using an expectation-maximization algorithm [10,11].

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; [12]). DIF effects $\boldsymbol{e}_i$ are defined as deviations $\boldsymbol{e}_i = \gamma_i - \gamma_i^*$. The occurrence of DIF causes additional variability in the estimated (country) mean μ and standard deviation σ [13,14]. Consequently, 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 [15–20]. There exist simple formulas for linking errors based on variance components for the 1PL model [16,18]. For more complex models, resampling techniques [21,22] such as jackknife [16,18] or (balanced) half sampling [19] of items can be employed. In the computation of the jackknife linking error, the model is repeatedly estimated by excluding a single item i at each item resulting in slightly differing estimates $\hat{\mu}_{(-i)}$ and $\hat{\sigma}_{(-i)}$ compared to the estimates $\hat{\mu}$ and $\hat{\sigma}$ in the full sample of items. The jackknife linking error for the estimated mean $\hat{\mu}$ is defined as

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

The disadvantage of the linking error formula (2) is that $I + 1$ model estimations of the IRT model based on the log-likelihood function l are required. In this article, a computational shortcut for determining increments $\hat{\mu}_{(-i)} - \hat{\mu}$ in (2) based on a Taylor expansion of the log-likelihood function is presented. Only second-order derivatives and one additional estimation of the IRT model are required in our proposed approach. Hence, the computational effort is significantly reduced.

The rest of the article is structured as follows. The newly proposed analytical approximation to the jackknife linking error is presented in Section 2. A simulation study compares the performance of our new approach with the jackknife linking error in Section 3. Finally, the article closes with a discussion in Section 4.

2. Analytical Approximation of the Jackknife Linking Error

This section provides details for our analytical approximation to the jackknife linking error. A Taylor expansion of the log-likelihood function l is employed to approximate increments in the jackknife linking error formula.

Let $\delta = (\mu, \sigma)$ be the vector that includes the mean μ and the standard deviation σ. Let $\gamma = (\gamma_1, \ldots, \gamma_I)$ be the vector that includes all item parameters γ_i $(i = 1, \ldots, I)$. Furthermore, let δ_0 and γ_0 be the true distribution parameter and item parameters, respectively. In the computation of $\hat{\gamma}$, the item parameters in the scaling model to $\gamma = \gamma^*$ are fixed. The difference $\boldsymbol{e} = \gamma_0 - \gamma^*$ indicates misspecification. If the scaling model involves data of a country and γ^* are international item parameters, the vector $\boldsymbol{e}$ includes DIF effects.

The approximation of the jackknife linking error relies on a Taylor expansion of the first derivative of the log-likelihood function l with respect to δ (i.e., the score equations) around true data-generating parameters (δ_0, γ_0). In the application of IRT models, the log-likelihood function is typically twice continuously differentiable to guarantee the applicability of the Taylor approximation. Define $l_\delta = (\partial l)/(\partial \delta)$, $l_{\delta\delta} = (\partial^2 l)/(\partial \delta^2)$,

and $l_{\delta\gamma_i} = (\partial^2 l)/(\partial\delta\partial\gamma_i)$. With a sufficiently long test, estimated item parameters $\hat{\gamma}_i$ are independent across items [23]. Hence, l_δ can be approximated around (δ_0, γ_0) as

$$l_\delta(\delta, \gamma) \approx l_\delta(\delta_0, \gamma_0) + l_{\delta\delta}(\delta_0, \gamma_0)(\delta - \delta_0) + \sum_{i=1}^{I} l_{\delta\delta}(\delta_0, \gamma_{i0})(\gamma_i - \gamma_{i0}) \,. \tag{3}$$

The distribution parameter estimates $\hat{\delta} = (\hat{\mu}, \hat{\sigma})$ are obtained by setting (3) to zero and using fixed but misspecified item parameters $\gamma_i = \gamma_i^*$. Hence, we obtain from (3)

$$\mathbf{0} = l_\delta(\delta_0, \gamma_0) + l_{\delta\delta}(\delta_0, \gamma_0)(\hat{\delta} - \delta_0) + \sum_{i=1}^{I} l_{\delta\gamma_i}(\delta_0, \gamma_{i0})(\gamma_i^* - \gamma_{i0}) \,. \tag{4}$$

We now determine the distribution parameter estimate $\hat{\delta}_{(-i)}$ in which item i is omitted from the log-likelihood function. Empirical evidence shows that the distribution parameters can be equivalently estimated if the item parameters of item i were freely estimated. This means that one can set $\gamma_i = \gamma_{i,0}$ for a sufficiently large number of items I. Then, (4) can be rewritten as

$$\mathbf{0} = l_\delta(\delta_0, \gamma_0) + l_{\delta\delta}(\delta_0, \gamma_0)(\hat{\delta}_{(-i)} - \delta_0) + \sum_{\substack{j=1 \\ j\neq i}}^{I} l_{\delta\gamma_j}(\delta_0, \gamma_{j0})(\gamma_j^* - \gamma_{j0}) \,. \tag{5}$$

By subtracting (4) from (5), we obtain

$$\hat{\delta}_{(-i)} - \hat{\delta} = -[l_{\delta\delta}(\delta_0, \gamma_0)]^{-1} l_{\delta\gamma_i}(\delta_0, \gamma_{i0})(\gamma_i^* - \gamma_{i0}) \,. \tag{6}$$

Now, Equation (6) is now specialized for the 2PL model. In this case, $\gamma_i = (a_i, b_i)$ consists of two parameters. We assume that fixed item discriminations were correct and fixed item intercepts b_i^* do not equal true data-generating item intercepts b_{i0}. We obtain from (6)

$$\hat{\delta}_{(-i)} - \hat{\delta} \approx -[l_{\delta\delta}(\delta_0, \gamma_0)]^{-1} l_{\delta b_i}(\delta_0, \gamma_{i0})(b_i^* - b_{i0}) \,. \tag{7}$$

In the following subsections, it is discussed how the finding can be used in the practical implementation (Section 2.1) of the jackknife linking error and how to efficiently compute the necessary derivatives of the log-likelihood function (Section 2.2). The estimation of linking errors is also prone to sampling errors. To circumvent a biased estimation of the linking error, we propose a bias-corrected version of the analytical approximation of the jackknife linking error in Section 2.3. Finally, a variant of the jackknife linking error computation in subsets of items is discussed in Section 2.4.

2.1. Use of the Approximation in Scaling

We now discuss how to apply the analytical approximation formula (6) for the deviations $\hat{\delta}_{(-i)} - \hat{\delta}$ in the jackknife linking error formula. First, we compute the distribution parameters $\hat{\delta}$ by fixing item parameters to γ^*. Second, we estimate item parameters $\hat{\gamma}$ by fixing the distribution parameters to $\hat{\delta}$. The motivation is that differences of $\hat{\delta} - \delta_0$ and $\hat{\gamma} - \gamma_0$ are close to zero for a sufficiently large number of items I. Hence, we replace unknown parameters in (6) with their empirical counterparts, and we arrive at

$$\hat{\delta}_{(-i)} - \hat{\delta} = -\left[l_{\delta\delta}(\hat{\delta}, \hat{\gamma})\right]^{-1} l_{\delta\gamma_i}(\hat{\delta}, \hat{\gamma}_i)(\gamma_i^* - \hat{\gamma}_i) \,. \tag{8}$$

For the special case of the 2PL model with misspecified item intercepts b_i^*, we obtain from (7)

$$\hat{\delta}_{(-i)} - \hat{\delta} = -\boldsymbol{H}_i(b_i^* - \hat{b}_{i0}) \tag{9}$$

for estimated item parameters $\hat{\boldsymbol{b}} = (\hat{b}_1, \ldots, \hat{b}_I)$ and $\boldsymbol{H}_i = \left[l_{\delta\delta}(\hat{\delta}, \hat{\boldsymbol{b}}) \right]^{-1} l_{\delta b_i}(\hat{\delta}, \hat{\boldsymbol{b}})$. The analytical approximation $\mathrm{LE}_{\mathrm{AN}}$ of the jackknife linking error is given as

$$\mathrm{LE}_{\mathrm{AN}}(u) = \sqrt{\frac{I-1}{I} \sum_{i=1}^{I} h_{iu}^2 (b_i^* - \hat{b}_{i0})^2} \;, \; u = \mu \text{ or } u = \sigma \,, \tag{10}$$

where $\boldsymbol{H}_i = (h_{i\mu}, h_{i\sigma})^\top$. For the linking error of $\hat{\mu}$, the first entry $h_{i\mu}$ is chosen. For the linking error of $\hat{\sigma}$, the second entry $h_{i\sigma}$ is chosen.

2.2. Implementation Details for Computing Derivatives of the Log-Likelihood Function

We now discuss how to efficiently compute the necessary derivatives of the log-likelihood function required in the analytical approximation of the jackknife linking error. We evaluate the integral in (1) by a rectangular integration on a finite grid $\theta_1, \ldots, \theta_T$ of θ points. Hence, the continuous normal distribution is approximated by a discretized normal distribution [24]. We set $w_t = w_t(\mu, \sigma) = C\phi(\theta_t; \mu, \sigma)$ using an appropriate scaling constant C that ensures $\sum_{t=1}^{T} w_t = 1$.

Let $l_p = \log L_p$ denote the contribution in the log-likelihood function of person p based on item response data $\boldsymbol{x}_p = (x_{p1}, \ldots, x_{pI})$. It holds that

$$L_p = \sum_{t=1}^{T} w_t \prod_{i=1}^{I} f_{pti} = \sum_{t=1}^{T} w_t f_{pt} \,, \text{where} \tag{11}$$

$$f_{pti} = P_i(\theta_t, \gamma_i)^{x_{pi}} [1 - P_i(\theta_t, \gamma_i)]^{1 - x_{pi}} \tag{12}$$

and $f_{pt} = \prod_{i=1}^{I} f_{pti}$. We now compute the partial derivative of l_p for a scalar parameter u

$$\frac{\partial l_p}{\partial u} = \frac{\dfrac{\partial L_p}{\partial u}}{L_p} \,. \tag{13}$$

The second-order derivative with respect to another parameter v is given as

$$\frac{\partial^2 l_p}{\partial u \partial v} = \frac{\dfrac{\partial^2 L_p}{\partial u \partial v} L_p - \dfrac{\partial L_p}{\partial u} \dfrac{\partial L_p}{\partial v}}{L_p^2} \,. \tag{14}$$

We can compute for $u = \mu$ or $u = \sigma$

$$\frac{\partial L_p}{\partial u} = \sum_{t=1}^{T} \frac{\partial w_t}{\partial u} f_{pt} \,. \tag{15}$$

The second-order derivative for $v = \mu$ or $v = \sigma$ can be obtained as

$$\frac{\partial^2 L_p}{\partial u \partial v} = \sum_{t=1}^{T} \frac{\partial^2 w_t}{\partial u \partial v} f_{pt} \,. \tag{16}$$

In the analytical approximation of the jackknife linking error, we need to compute $(\partial^2 L_p)/(\partial u \partial b_i)$ for an item parameter b_i. We obtain

$$\frac{\partial^2 L_p}{\partial u \partial b_i} = \sum_{t=1}^{T} \frac{\partial w_t}{\partial u} f_{pt} \frac{1}{f_{pit}} \frac{\partial f_{pit}}{\partial b_i} = \sum_{t=1}^{T} \frac{\partial w_t}{\partial u} f_{pt} \frac{\partial \log f_{pit}}{\partial b_i} \,. \tag{17}$$

Equations (15)–(17) indicate that the necessary computations for the first- and second-order derivatives are computationally inexpensive. Hence, the analytical approximation of the linking error is computationally cheap if DIF effects were available.

2.3. Bias Correction due to Sampling Error

The estimation of linking errors is also prone to sampling errors because estimated item parameters are involved in the computation. To avoid a biased estimation of the linking error, we now propose a bias-corrected version of the analytical approximation of the jackknife linking error.

Assume that the variance of $b_i^* - \hat{b}_{i0}$ in (10) is v_i, and the estimates are approximately independent across items [23]. The bias-corrected analytical linking error can be obtained by subtracting variability that is due to sampling error quantified in v_i. We obtain

$$\mathrm{LE}_{\mathrm{AN,bc}}(u) = \mathrm{sqrt}_+\left(\frac{I-1}{I} \sum_{i=1}^{I} h_{iu}^2 \left[(b_i^* - \hat{b}_{i0})^2 - v_i \right] \right) \; , \; u = \mu \text{ or } u = \sigma \, , \tag{18}$$

where $\mathrm{sqrt}_+(x) = \sqrt{\max(x, 0)}$.

A similar bias-correction method was used in the 1PL method in trend estimation [18]. Alternatively, a bias-correction term can also be estimated using resampling techniques regarding persons (e.g., bootstrap or half sampling; [19,22]).

2.4. Jackknife Linking Error Based on Testlets

The jackknife linking error is frequently evaluated at groups of items (so-called testlets; refs. [25,26]) instead of single items. The reason for this is that subsets of items in a test are often presented jointly with a common (text) stimulus. Hence, DIF effects pertain to all items in a testlet and often have the same sign. Therefore, the testlet structure must be taken into account when computing linking errors [17,18,27].

Assume that there are H testlets. That is, the set of item integers $i = 1, \ldots, I$ is partitioned into distinct sets $\mathcal{I}_1, \ldots, \mathcal{I}_H$. The linking error based on testlets for $\hat{\mu}$ is defined as

$$\mathrm{LE}(\hat{\mu}) = \sqrt{ \frac{H-1}{H} \sum_{h=1}^{H} (\hat{\mu}_{(-\mathcal{I}_h)} - \hat{\mu})^2 } \, , \tag{19}$$

where $\hat{\mu}_{(-\mathcal{I}_h)}$ is the estimate in which all items from testlet h were removed. In the analytical approximation, we can approximate the relevant jackknife difference in (19) by

$$\hat{\mu}_{(-\mathcal{I}_h)} - \hat{\mu} = \sum_{i \in \mathcal{I}_h} h_{i\mu} (b_i^* - b_{i0}) \, . \tag{20}$$

A corresponding bias-corrected variant of the linking error (see (18)) can be similarly obtained.

3. Simulation Study

3.1. Method

In this simulation study, we investigate the performance of our analytical approximation of the jackknife linking error. We illustrate the performance based on the 2PL model. Equal discriminations $a_i \equiv 1$ were assumed.

In the simulation study, we varied two factors: the number of items and the DIF standard deviation τ_{DIF}. We chose $I = 10$, 20, 30, and 40 items to cover a range of test lengths that are obtained in empirical practice. The goal is to assess the mean μ and the standard deviation σ in a group (e.g., a country in a large-scale assessment study such as PISA). We defined $\mu = -0.2$ and $\sigma = 0.9$ in the simulation. Assumed item difficulties b_i^* were chosen equispaced in the interval $[-2, 2]$ with increments $4/(I-1)$. For example, for $I = 10$ items, assumed item difficulties were -2.00, -1.56, -1.11, -0.67, -0.22, 0.22, 0.67,

1.11, 1.56, and 2.00. In each replication, data-generating item difficulties b_{i0} were simulated according

$$b_{i0} = b_i^* + e_i \ , \ e_i \sim \mathrm{N}(0, \tau_{\mathrm{DIF}}) \tag{21}$$

Hence, the estimated distribution parameters $\hat{\mu}$ and $\hat{\sigma}$ will vary across replications even with infinite sample sizes of persons because the true data-generating item parameters vary. The standard deviation of DIF effects was either $\tau_{\mathrm{DIF}} = 0.25$ or 0.50.

Item response data were simulated according to a quasi-Monte Carlo simulation method ([28], see also [29] for a similar approach). Our motivation was to assess the performance of the linking error by reducing the uncertainty due to sampling error. Simulated item responses $\boldsymbol{X}$ should be as close as possible to the true distribution $\mathrm{P}(\boldsymbol{X}; \mu, \sigma, \boldsymbol{b}_0)$ (see [29]). To facilitate this, we chose θ values from the same discrete grid $\theta_1 = -3.5, \ldots, \theta_{21} = 3.5$ of equidistant θ points that were also used in fitting the 2PL model. We fixed a pseudo-sample size of 10,000 persons. Then, we computed the number of persons at each θ_t, which is given by $Nw_t(\mu, \sigma)$ (some rounding is necessary). For each $\theta = \theta_t$ and for each item i, we can compute $P_i(\theta_t; a_i, b_{i0})$ according to the item response function. Hence, there are $Nw_t(\mu, \sigma)P_i(\theta_t; a_i, b_{i0})$ persons with $\theta = \theta_t$ with $X_i = 1$ and $Nw_t(\mu, \sigma)(1 - P_i(\theta_t; a_i, b_{i0}))$ with $X_i = 0$. The zeroes and ones for item $i = 1, \ldots, I$ are randomly allocated to the corresponding persons $\theta = \theta_t$. Although the empirical frequencies for multivariate item response patterns $\boldsymbol{X} = \boldsymbol{x}$ do not match the population probabilities, conditional marginal probabilities of the item response functions are (almost) correctly simulated. Hence, one can conclude that this quasi-Monte Carlo approach reduces the impact of sampling errors to a minimum and only reflects variability due to linking errors; that is, using incorrect item parameters b_i^* that differ from the data-generating item parameters b_i.

In each of the $4 \times 2 = 8$ cells of the simulation, 2000 replications were conducted. We fitted the 2PL model with item discriminations fixed to 1 and fixed item difficulties b_i^* and obtained the estimated mean $\hat{\mu}$ and the estimated standard deviation $\hat{\sigma}$. Then, we determined item difficulties $\hat{b}_i$ by fixing the mean and standard deviation to $\hat{\mu}$ and $\hat{\sigma}$, respectively. Using these quantities, we calculated the analytical approximation (AN) of the jackknife linking error given in (10). We compared the analytical linking error with the jackknife linking error (JK). To evaluate the quality of the estimated linking errors, we computed the mean, the standard deviation, the standard error ratio ($\mathrm{SE}_{\mathrm{ratio}}$; defined as the quotient of the average linking error and the empirical standard deviation of the estimate $\hat{\mu}$ or $\hat{\sigma}$), and coverage rates at the 95% confidence level.

The R software [30] was used for simulation and analysis. The R package TAM [31] was used for estimating the 2PL model.

3.2. Results

Table 1 contains the results of the two linking error methods as a function of the standard deviation of DIF effects τ_{DIF} and the number of items I.

It can be seen that the mean can be almost unbiasedly estimated in the presence of DIF effects. As expected, the standard deviation of the estimated mean $\hat{\mu}$ decreases with a larger number of items. Because all item discriminations were equal to one, the linking error of $\hat{\mu}$ can be analytically predicted as $\tau_{\mathrm{DIF}}/\sqrt{I}$ (referred to as EXP in Table 1; [16,18]). It can be seen that empirical standard deviations of $\hat{\mu}$ were close to these expected values. Moreover, the mean of the jackknife linking error JK was very similar to the empirical standard deviation of $\hat{\mu}$, while the linking error AN based on the analytical approximation was slightly too small. This was particularly the case for a low number of items $I = 10$. However, with a larger number of items, the analytical approximation performed well. This behavior is also reflected in the standard error ratio $\mathrm{SE}_{\mathrm{ratio}}$, which attained desired values close to 1 for the jackknife linking error and was smaller than 1 for the linking error based on the analytical approximation. However, for at least 20 items, the analytical approximation might be considered to have satisfactory performance. Overall, it can also be seen that the estimated linking error AN was a bit smaller on average than the jackknife linking error. However, the average absolute deviation (AAD in Table 1) demonstrated that the analytical approximation was very close to the jackknife linking error in each

replication, particularly for a large number of items such as $I = 40$. It can also be seen that coverage rates were satisfactory for the jackknife linking error, while the analytical approximation showed issues in the condition of a few items (i.e., $I = 10$). Finally, we observed that the empirical standard deviation of estimated linking errors was slightly smaller for the analytical approximation compared to the jackknife approach.

Table 1. Simulation study: Results as a function of the standard deviation of DIF effects (τ_{DIF}) and number of items (I).

					Mean		SE_{ratio}			SD		COV95	
τ_{DIF}	I	Bias	SD	EXP	JK	AN	JK	AN	AAD	JK	AN	JK	AN
		Linking error of estimated mean $\hat{\mu}$											
0.25	10	0.000	0.079	0.079	0.079	0.070	0.990	0.875	0.0091	0.0202	0.0177	91.6	88.6
0.25	20	0.003	0.056	0.056	0.056	0.053	1.000	0.945	0.0031	0.0096	0.0091	93.2	91.7
0.25	30	0.000	0.045	0.046	0.046	0.044	1.007	0.970	0.0017	0.0065	0.0062	94.3	93.3
0.25	40	0.003	0.039	0.040	0.040	0.039	1.013	0.986	0.0011	0.0047	0.0045	96.3	95.7
0.5	10	0.003	0.161	0.158	0.153	0.135	0.949	0.837	0.0181	0.0397	0.0354	91.6	87.9
0.5	20	0.002	0.111	0.112	0.111	0.104	1.000	0.942	0.0065	0.0185	0.0177	93.5	91.8
0.5	30	0.004	0.090	0.091	0.091	0.087	1.011	0.972	0.0036	0.0130	0.0127	94.5	93.5
0.5	40	0.008	0.080	0.079	0.078	0.076	0.978	0.948	0.0025	0.0097	0.0095	93.2	92.3
		Linking error of estimated standard deviation $\hat{\sigma}$											
0.25	10	−0.007	0.039	—	0.041	0.033	1.051	0.866	0.0073	0.0111	0.0091	95.2	92.3
0.25	20	−0.008	0.023	—	0.023	0.021	1.009	0.914	0.0022	0.0042	0.0038	94.7	93.0
0.25	30	−0.008	0.019	—	0.017	0.016	0.901	0.848	0.0011	0.0026	0.0024	92.4	91.7
0.25	40	−0.008	0.016	—	0.014	0.014	0.898	0.856	0.0007	0.0019	0.0017	90.9	90.8
0.5	10	−0.030	0.078	—	0.077	0.062	0.985	0.796	0.0155	0.0206	0.0166	94.9	90.8
0.5	20	−0.029	0.053	—	0.043	0.040	0.827	0.752	0.0047	0.0079	0.0068	89.3	87.5
0.5	30	−0.027	0.042	—	0.033	0.031	0.774	0.727	0.0026	0.0048	0.0044	85.9	84.7
0.5	40	−0.027	0.037	—	0.027	0.026	0.723	0.690	0.0019	0.0034	0.0033	85.0	83.0

Note. EXP = expected value of linking error for $\hat{\mu}$; JK = jackknife linking error; AN = linking error estimated by analytical approximation LE_{AN} from Equation (10); SE_{ratio} = standard error ratio; AAD = average absolute difference between JK and AN linking error estimates; SD = standard deviation of estimated linking error; COV95 = coverage rate for confidence level 95%.

The estimated standard deviation $\hat{\sigma}$ showed a small bias in the condition of a larger standard deviation of DIF effects (i.e., $\tau_{DIF} = 0.5$). Hence, one can expect that coverage rates will perform satisfactorily because the expected value deviated from the true value $\sigma = 0.9$. Interestingly, the mean of both linking errors was smaller than the empirical standard deviation of $\hat{\sigma}$. This finding was also reflected in the standard error ratios, which were substantially smaller than 1. Hence, one might conclude that both linking errors (jackknife JK and the analytical approximation AN) were unsatisfactory to reflect the variability in estimated standard deviations in our application. We suspect these results might be explained by the fact that the standard deviation was obtained using fixed but incorrect item parameters. Moreover, such a finding would likely not be observed in a linking approach such as log-mean-mean linking [32].

Across all conditions, the analytical approximation provided linking errors that were slightly too small. In the computation of the jackknife linking error in Equation (10), the multiplication factor $(I-1)/I$ is used. In other applications, the multiplication factor 1 is used [22]. We recomputed linking errors based on the analytical approximation with the multiplication factor of 1, and it turned out that the results grew closer to the jackknife linking error. However, the standard error ratio was still smaller than 1. Nevertheless, there could be a benefit to using the modified formula in empirical applications.

On the Bias in the Estimated Standard Deviation $\hat{\sigma}$

At first sight, the negative bias in the estimated standard deviation $\hat{\sigma}$ is surprising, although such a finding was also found in other studies that use fixed but incorrect item

parameters in estimation [32]. We now present a heuristic derivation of the bias that appears close to the empirically obtained bias. Please note that the estimation of the item response model using a logistic item response function can be approximated by the probit link function [33]. In this case, weighted least squares estimation based on tetrachoric correlations [34] can be used to estimate the standard deviation σ. The approach relies on modeling underlying continuous variables X_i^* for dichotomous items X_i. The variable X_i takes the value if X_i^* exceeds the item parameter b_i. The tetrachoric correlation ρ_{ij} for items i and j is given by

$$\rho_{ij} = \frac{\mathrm{Cov}(X_i^*, X_j^*)}{\sqrt{\mathrm{Var}(X_i^*)\mathrm{Var}(X_j^*)}} \,. \tag{22}$$

Now assume equal item discriminations a_i and a correctly specified model. In this case, (22) simplifies to

$$\rho_{ij} = \frac{\sigma^2}{\sigma^2 + \mathcal{L}} \,, \tag{23}$$

where $\mathcal{L} = \pi^2/3 \approx 3.29$ is the variance of the logistic distribution. If incorrect item parameters b_i^* were used, the covariance $\mathrm{Cov}(X_i^*, X_j^*)$ in (22) is not affected on average. However, the variance $\mathrm{Var}(X_i^*)$ of the underlying latent variable X_i^* increases, and the expected value can be determined by the DIF variance τ^2_{DIF}. The estimated tetrachoric correlation can then be computed as

$$\rho_{ij}^* = \frac{\sigma^2}{\sigma^2 + \tau^2_{\mathrm{DIF}} + \mathcal{L}} \,. \tag{24}$$

In the computation of $\hat{\sigma}$, one, therefore, essentially solves

$$\rho_{ij}^* = \frac{\sigma^2}{\sigma^2 + \tau^2_{\mathrm{DIF}} + \mathcal{L}} = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \mathcal{L}} \,. \tag{25}$$

The estimated standard deviation $\hat{\sigma}$ can be determined as (see [35] for a similar approach)

$$\hat{\sigma} = \sqrt{\mathcal{L}\frac{\rho_{ij}^*}{1-\rho_{ij}^*}} = \sigma\sqrt{\frac{\mathcal{L}}{\mathcal{L} + \tau^2_{\mathrm{DIF}}}} \,. \tag{26}$$

Hence, the estimated standard deviation is negatively biased in the presence of DIF effects. The predicted bias in the estimated standard deviation based on (26) is -0.008 for $\tau_{\mathrm{DIF}} = 0.25$ and -0.032 for $\tau_{\mathrm{DIF}} = 0.50$, which is similar to the empirically obtained bias for $\hat{\sigma}$ in Table 1.

4. Discussion

In this article, an analytical approximation of the jackknife linking error by means of a Taylor expansion of the log-likelihood function has been proposed. It turned out that the analytical approximation performed well for estimated means for at least 20 items. The approximation has the advantage because it only requires one additional estimation of an item response model and second-order derivatives of the log-likelihood function. In contrast, the jackknife linking error requires I additional estimations of the item response model for I items which is computationally much more demanding. One might argue that the analytical approximation provides at least a computationally cheap proxy of the linking error. However, the jackknife linking error would be preferred due to a more reliable statistical inference because our simulation findings indicated that it provided slightly better coverage rates.

In the simulation study, we did not consider sampling errors because a quasi-Monte Carlo simulation method was utilized that minimized the impact of sampling errors. Future studies could simultaneously assess sampling errors and linking errors based on the

analytical approximation. In particular, the performance of our proposed bias-correction method could be evaluated.

In the analytical derivation and the simulation study, we restricted ourselves to the computation of a single group which can be interpreted as country means and standard deviations in a cross-sectional assessment or the computation of trend between two time points. In future research, our proposed simplified computation formula for the linking error might be applied for trend estimates in the country means in educational assessment studies such as PISA [18].

It should be noted that our proposed approach of the analytical approximation of the jackknife linking error bears similarity with the infinitesimal jackknife technique [36–38]. The difference is that the jackknife linking error removes columns (i.e., items) from the dataset, while infinitesimal jackknife addresses contributions of rows (i.e., persons) in a dataset.

In the linking literature [39,40], linking errors are also sometimes referred to as sampling errors (of persons) of obtained linking constants that can be means or standard deviations [41–43]. It is important to note that sampling errors due to the sampling of persons and linking errors due to item choice must be distinguished [19,20]. The computation of linking errors can be justified for random items and fixed items [19]. For random items, the used items are thought to be a (representative) draw from a larger population of items [44,45]. For fixed items, DIF effects can be stochastically modeled by some distribution [19]. The latter case might also be conceived as quantifying model error [46]. We would like to point out that we see great potential in using linking errors by incorporating the extent of model misspecification in the reported parameter uncertainty.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript:

1PL	one-parameter logistic
2PL	two-parameter logistic
DIF	differential item functioning
IRT	item response theory
JK	jackknife
LE	linking error
PIRLS	progress in international reading literacy study
PISA	programme for international student assessment

References

1. Bock, R.D.; Moustaki, I. Item response theory in a general framework. In *Handbook of Statistics, Volume 26: Psychometrics*; Rao, C.R., Sinharay, S., Eds.; Elsevier: Amsterdam, The Netherlands, 2007; pp. 469–513. [CrossRef]
2. van der Linden, W.J.; Hambleton, R.K. (Eds.) *Handbook of Modern Item Response Theory*; Springer: New York, NY, USA, 1997. [CrossRef]
3. van der Linden, W.J. Unidimensional logistic response models. In *Handbook of Item Response Theory, Volume 1: Models*; van der Linden, W.J., Ed.; CRC Press: Boca Raton, FL, USA, 2016; pp. 11–30. [CrossRef]
4. 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]
5. OECD. *PISA 2018. Technical Report*; OECD: Paris, France, 2020.
6. Foy, P.; Yin, L. Scaling the PIRLS 2016 achievement data. In *Methods and Procedures in PIRLS 2016*; Martin, M.O., Mullis, I.V., Hooper, M., Eds.; IEA: Boston College: Chestnut Hill, MA, USA, 2017.

7. 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.
8. Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*; Danish Institute for Educational Research: Copenhagen, Denmark, 1960.
9. 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.
10. Aitkin, M. Expectation maximization algorithm and extensions. In *Handbook of Item Response Theory, Volume 2: Statistical Tools*; van der Linden, W.J., Ed.; CRC Press: Boca Raton, FL, USA, 2016; pp. 217–236. [CrossRef]
11. Bock, R.D.; Aitkin, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **1981**, *46*, 443–459. [CrossRef]
12. Penfield, R.D.; Camilli, G. Differential item functioning and item bias. In *Handbook of Statistics, Volume 26: Psychometrics*; Rao, C.R., Sinharay, S., Eds.; Elsevier: Amsterdam, The Netherlands, 2007; pp. 125–167. [CrossRef]
13. 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]
14. Robitzsch, A.; Lüdtke, O. A review of different scaling approaches under full invariance, partial invariance, and noninvariance for cross-sectional country comparisons in large-scale assessments. *Psychol. Test Assess. Model.* **2020**, *62*, 233–279.
15. Battauz, M. Multiple equating of separate IRT calibrations. *Psychometrika* **2017**, *82*, 610–636. [CrossRef]
16. Monseur, C.; Berezner, A. The computation of equating errors in international surveys in education. *J. Appl. Meas.* **2007**, *8*, 323–335.
17. OECD. *PISA 2012. Technical Report*; OECD: Paris, France, 2014.
18. 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]
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. Wu, M. Measurement, sampling, and equating errors in large-scale assessments. *Educ. Meas.* **2010**, *29*, 15–27. [CrossRef]
21. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; CRC Press: Boca Raton, FL, USA, 1994. [CrossRef]
22. Kolenikov, S. Resampling variance estimation for complex survey data. *Stata J.* **2010**, *10*, 165–199. [CrossRef]
23. 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]
24. Chakraborty, S. Generating discrete analogues of continuous probability distributions—A survey of methods and constructions. *J. Stat. Distrib. Appl.* **2015**, *2*, 6. [CrossRef]
25. Sireci, S.G.; Thissen, D.; Wainer, H. On the reliability of testlet-based tests. *J. Educ. Meas.* **1991**, *28*, 237–247. [CrossRef]
26. Wainer, H.; Bradlow, E.T.; Wang, X. *Testlet Response Theory and Its Applications*; Cambridge University Press: Cambridge, UK, 2007. [CrossRef]
27. Monseur, C.; Sibberns, H.; Hastedt, D. Linking errors in trend estimation for international surveys in education. *IERI Monogr. Ser.* **2008**, *1*, 113–122.
28. Caflisch, R.E. Monte Carlo and quasi-Monte Carlo methods. *Acta Numer.* **1998**, *7*, 1–49. [CrossRef]
29. Robitzsch, A. About the equivalence of the latent D-scoring model and the two-parameter logistic item response model. *Mathematics* **2021**, *9*, 1465. [CrossRef]
30. 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).
31. Robitzsch, A.; Kiefer, T.; Wu, M. TAM: Test Analysis Modules. 2022. R Package Version 4.1-4. Available online: https://CRAN.R-project.org/package=TAM (accessed on 28 August 2022).
32. 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]
33. Robitzsch, A.; Lüdtke, O. Mean comparisons of many groups in the presence of DIF: An evaluation of linking and concurrent scaling approaches. *J. Educ. Behav. Stat.* **2022**, *47*, 36–68. [CrossRef]
34. Muthén, B. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* **1984**, *49*, 115–132. [CrossRef]
35. Ip, E.H. Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *Br. J. Math. Stat. Psychol.* **2010**, *63*, 395–416. [CrossRef]
36. Giordano, R.; Stephenson, W.; Liu, R.; Jordan, M.I.; Broderick, T. A higher-order swiss army infinitesimal jackknife. *arXiv* **2019**, arXiv:1806.00550v5.
37. Jaeckel, L.A. *The Infinitesimal Jackknife*; Bell Telephone Laboratories: Washington, WA, USA, 1972.
38. Jennrich, R.I. Nonparametric estimation of standard errors in covariance analysis using the infinitesimal jackknife. *Psychometrika* **2008**, *73*, 579–594. [CrossRef]
39. Kolen, M.J.; Brennan, R.L. *Test Equating, Scaling, and Linking*; Springer: New York, NY, USA, 2014. [CrossRef]
40. González, J.; Wiberg, M. *Applying Test Equating Methods. Using R*; Springer: New York, NY, USA, 2017. [CrossRef]

41. Andersson, B. Asymptotic variance of linking coefficient estimators for polytomous IRT models. *Appl. Psychol. Meas.* **2018**, *42*, 192–205. [CrossRef] [PubMed]
42. Battauz, M. Factors affecting the variability of IRT equating coefficients. *Stat. Neerl.* **2015**, *69*, 85–101. [CrossRef]
43. Ogasawara, H. Standard errors of item response theory equating/linking by response function methods. *Appl. Psychol. Meas.* **2001**, *25*, 53–67. [CrossRef]
44. Brennan, R.L. *Generalizabilty Theory*; Springer: New York, NY, USA, 2001. [CrossRef]
45. Husek, T.R.; Sirotnik, K. *Item Sampling in Educational Research*; CSEIP Occasional Report No. 2.; University of California: Los Angeles, CA, USA, 1967.
46. Wu, H.; Browne, M.W. Quantifying adventitious error in a covariance structure as a random effect. *Psychometrika* **2015**, *80*, 571–600. [CrossRef] [PubMed]