foundations

MDPI

Article

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

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: This article investigates the comparison of two groups based on the two-parameter logistic item response model. It is assumed that there is random differential item functioning in item difficulties and item discriminations. The group difference is estimated using separate calibration with subsequent linking, as well as concurrent calibration. The following linking methods are compared: mean-mean linking, log-mean-mean linking, invariance alignment, Haberman linking, asymmetric and symmetric Haebara linking, different recalibration linking methods, anchored item parameters, and concurrent calibration. It is analytically shown that log-mean-mean linking and mean-mean linking provide consistent estimates if random DIF effects have zero means. The performance of the linking methods was evaluated through a simulation study. It turned out that (log-)mean-mean and Haberman linking performed best, followed by symmetric Haebara linking and a newly proposed recalibration linking method. Interestingly, linking methods frequently found in applications (i.e., asymmetric Haebara linking, recalibration linking used in a variant in current large-scale assessment studies, anchored item parameters, concurrent calibration) perform worse in the presence of random differential item functioning. In line with the previous literature, differences between linking methods turned out be negligible in the absence of random differential item functioning. The different linking methods were also applied in an empirical example that performed a linking of PISA 2006 to PISA 2009 for Austrian students. This application showed that estimated trends in the means and standard deviations depended on the chosen linking method and the employed item response model.

Keywords: linking; 2PL model; item response model; concurrent calibration; differential item functioning; large-scale assessments

check for updates

Citation: 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. https://doi.org/10.3390/foundations1010009

Academic Editor: Ravi P. Agarwal

Received: 12 July 2021
Accepted: 10 September 2021
Published: 15 September 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

The analysis of educational and psychological tests is an important field in the social sciences. The test items (i.e., tasks presented in these tests) are often modeled using item response theory (IRT; [1–3]; for applications, see, e.g., [4–14]) models. In this article, the two-parameter logistic (2PL; [15]) IRT model is investigated to compare two groups on test items. For example, groups could be demographic groups, countries, studies, or time points. The group comparisons were carried out using linking methods [16]. A significant obstacle in applying linking methods is that the test items could behave differently in the two groups (i.e., differential item functioning), that is it cannot be expected that the two groups share a common set of statistical parameters for the test items. Such a situation is particularly important in educational large-scale assessment (LSA; [17–19]) studies in which several countries are compared. It can be expected that test items function differently because there are curricular differences in those countries.

The paper is structured as follows. In Section 2, the 2PL model with differential item functioning is introduced. In Section 3, several linking methods are discussed. In Section 4, we present the results of a simulation study in which these linking methods are compared. In Section 5, an empirical example using the PISA datasets from 2006 and 2009 for Austria is presented. Finally, the paper closes with a discussion in Section 6.

2. Linking Two Groups with the 2PL Model

2.1. 2PL Model

For dichotomous items $i = 1, \dots, I$, the item response function (IRF) of the 2PL model is given by [3,20]:

$$P(X_i = x|\theta) = P_i(x; a_i, b_i) = \Psi((2x-1)a_i(\theta - b_i)) \qquad x = 0, 1 \quad , \tag{1}$$

where $\Psi(y) = \exp(y)/(1+\exp(y))$ is the logistic link function, a_i is the item discrimination, and b_i is the item difficulty. The one-parameter logistic (1PL) IRT model (also referred to as the Rasch model; [21]) fixes all item discriminations a_i to one.

The 1PL or the 2PL model is usually estimated under a local independence assumption, that is:

$$P(\boldsymbol{X} = \boldsymbol{x}) = \int \prod_{i=1}^{I} P_i(x_i, a_i, b_i) f(\theta) \, \mathrm{d}\theta \quad , \tag{2}$$

where f is the density of θ and $\boldsymbol{x} = (x_1, \dots, x_I)$. In many applications, θ is assumed to be normally distributed (i.e., $\theta \sim N(\mu, \sigma^2)$). In (2), the multivariate contingency table of $\boldsymbol{X}$ with corresponding probabilities $P(\boldsymbol{X} = \boldsymbol{x})$ involving 2^I item response patterns was summarized by a unidimensional latent variable θ.

2.2. Linking Design

To assess the distributional differences between two groups, an appropriate linking design must be established to identify group differences. Such a linking design is displayed in Figure 1. For two groups $g = 1, 2$, there is a set I_0 of common items (also referred to as anchor items or link items) that are administered in both groups. There are also group-specific items I_g that are uniquely administered in each of the two groups.

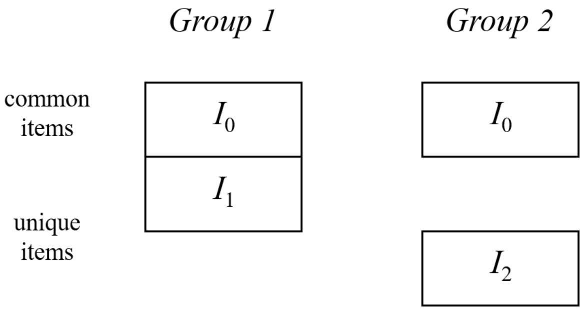

Figure 1. Linking design for two groups with common items I_0 and group-specific unique items I_1 and I_2.

This linking design is also referred to as a common items nonequivalent group design [22]. In equating, no common items exist in many applications, but two test forms are administered to equivalent groups [23]. Equating is often performed with the goal of producing an equivalency table for the sum score in a test, while linking determines linking constants in order to identify group differences with respect to the latent variable θ. In this article, we are interested in determining distributional differences between the two groups for the latent variable θ.

It has to be emphasized that differences between the two groups are mainly determined by the set of common items if the linking relies on an IRT model. The employed IRT models can predict the expected performance of students on items that were not administered. The crucial assumption is that item responses on nonadministered unique items can be inferred (or imputed) from common items. This corresponds to an ignorability assumption in the missing data literature and translates into the conditional independence of item responses on unique items conditional on item responses on common items (see [24,25]). If a unidimensional IRT model with a local independence assumption (2) holds, this condition is automatically fulfilled.

2.3. Random Differential Item Functioning

Common items in the set I_0 are administered in two groups $g = 1, 2$ (see Figure 1). It is assumed that the 2PL model (see Section 2.1) holds in both groups. The situation in which item parameters do not differ between groups is called measurement invariance [26–28]. However, it is likely in many applications that items function differently in the two groups. The existence of group-specific item parameters is labeled differential item functioning (DIF; [29–33]; for applications, see, e.g., [34–38]). In this case, items possess group-specific item discriminations a_{ig} and item difficulties b_{ig}. DIF in item discriminations is referred to as nonuniform DIF (NUDIF), while DIF in item difficulties is referred to as uniform DIF (UDIF) if there is no DIF in item discriminations (see [32,39,40]). Note that the local independence assumption (2) holds within each group even though the item parameters can differ across groups.

For the rest of this article, we assumed random DIF [41–50]. The main idea is that the difference of item parameters between groups is modeled by a distribution. Hence, a population perspective to a universe of items was adopted. Random DIF for item discriminations a_{ig} $(i = 1, \ldots, I; g = 1, 2)$ is defined as:

$$a_{i1} = a_i - f_i \quad \text{and} \quad a_{i2} = a_i + f_i \quad , \tag{3}$$

where DIF effects f_i are considered as a random variable that follows a distribution F_f. Common item discriminations a_i can be considered either fixed or random. In the following, we considered them as fixed. Item difficulties b_{ig} are also considered as random DIF, and DIF effects e_i follow a random distribution F_e:

$$b_{i1} = b_i - e_i \quad \text{and} \quad b_{i2} = b_i + e_i \quad , \tag{4}$$

where b_i are common item difficulties.

It is important to emphasize that there is an inherent indefinability of a simultaneous determination of average DIF effects and average group differences [51–54]. Hence, some structural assumptions must be posed on the distributions of DIF effects F_e and F_f. One possible choice would be to assume $\mathrm{E}(e_i) = \mathrm{E}(f_i) = 0$; that is, random DIF is is described by unsystematic differences in the item parameters (see Appendix A for further details). A frequent assumption is that DIF effects are normally distributed with zero means:

$$e_i \sim \mathrm{N}(0, \tau_b^2) \quad \text{and} \quad f_i \sim \mathrm{N}(0, \tau_a^2) \tag{5}$$

It is impossible to assume the means of DIF effects different from zero because average DIF effects are confounded with average group differences. Note that it holds that $a_{i2} - a_{i1} = 2f_i \sim \mathrm{N}(0, 2\tau_a^2)$ and $b_{i2} - b_{i1} = 2e_i \sim \mathrm{N}(0, 2\tau_b^2)$.

As an alternative, a sparsity assumption might be posed on DIF effects. In this case, the majority of items have DIF effects of zero, while only a few items have DIF effects different from zero [44,55]. If DIF effects are considered as fixed, this situation is known as partial invariance [56–58]. The random DIF distribution is a mixture distribution with two classes: one class with zero effects and the other class containing DIF effects different from zero. The mixture distribution can be simultaneously estimated in an IRT model with two groups ([55]; see also [59]). Alternatively, DIF detection methods can be used to identify items whose DIF effects differ from zero [32,60,61]. These items can be removed from linking in subsequent analysis (see, e.g., [62]). However, some research has shown that simultaneous treatment of the linking and modeling of DIF effects can result in superior statistical performance [54,63–66].

We would also like to point out that DIF effects in item discriminations in Equation (3) follow an additive model. Alternatively, a multiplicative model for DIF effects can be assumed:

$$a_{i1} = a_i / f_i \quad \text{and} \quad a_{i2} = a_i f_i \quad , \tag{6}$$

which corresponds to an additive model in logarithmized item discriminations:

$$\log a_{i1} = \log a_i - \log f_i \quad \text{and} \quad \log a_{i2} = \log a_i + \log f_i \quad . \tag{7}$$

It is hard to decide in empirical applications whether DIF effects for item discriminations should be modeled in the untransformed metric (see Equation (3)) or a logarithmized metric (see Equation (7)).

The simulation study only considered normally distributed DIF effects with zero means and an additive model for DIF effects in item discriminations.

2.3.1. Identified Item Parameters in Separate Calibrations in the Two Groups

In Section 3, we discuss some linking methods that rely on item parameters that were obtained from separate calibrations. That is, the 2PL model was separately fitted for the two groups. For reasons of identifiability, it has to be assumed that in the first group, it holds that $\mu_1 = 0$ and $\sigma_1 = 1$. In a separate estimation for the first group with an infinite sample size, the estimated item discriminations $\hat{a}_{i1}$ are equal to the data-generating parameters a_{i1}. The same holds for estimated item difficulties, that is $\hat{b}_{i1} = b_{i1}$.

In the second group, there are group-specific item parameters a_{i2} and b_{i2} and distribution parameters $\mu_2 \neq 0$ and $\sigma_2 \neq 1$. In a separate estimation for the second group, it was assumed that the mean was set to zero, and the standard deviation (SD) was set to one. Hence, estimated item parameters also include the distribution parameters. We obtain:

$$a_{i2}(\theta - b_{i2}) = a_{i2}(\sigma_2\theta^* + \mu_2 - b_{i2}) = (a_{i2}\sigma_2)(\theta^* - \sigma_2^{-1}(b_{i2} - \mu_2)) \quad , \tag{8}$$

where the standardized ability θ^* is standard normally distributed (i.e., $\mathrm{N}(0,1)$). From Equation (8), it follows that:

$$\hat{a}_{i2} = a_{i2}\sigma_2 \qquad \text{and} \qquad \hat{b}_{i2} = \sigma_2^{-1}(b_{i2} - \mu_2) \quad . \tag{9}$$

2.3.2. The Role of Normally Distributed Random DIF in Educational Assessment

It should be noted that the presence of random DIF implies that there is no single item for which the two groups share the same item parameters. Hence, all items are allowed to have different item parameters. By posing a normal distribution on DIF effects, it was assumed that DIF across groups vanishes on average for an infinite number of items. The presence of normally distributed random DIF is strongly different from the situation of partial invariance in which only a few items (or a few item parameters) possess DIF, while the rest of the items (or item parameters) do not have DIF. The two kinds of DIF effects can also simultaneously occur [65]. In our experience, we mainly observed random fluctuations of group-specific item parameters, which would correspond to normally distributed random DIF instead of to the case of partial invariance. However, it seems that ensuring partial invariance is seen as a measurement ideal in educational LSA studies [67–70]. We have argued against such a perspective elsewhere [54,65,71,72] and think that the random DIF perspective with normally distributed DIF effects is more relevant in real-world applications. Moreover, we think removing items due to DIF from group comparisons is unwise because DIF can be interpreted as construct-relevant [32,52,73–75]. In this situation, a group difference that relies on a purified set of items can bias estimated group differences with respect to the means and standard deviations.

3. Linking Methods

In the following, we discuss linking methods that allow estimating the distribution parameters $\mu_2 = \mathrm{E}(\theta)$ and $\sigma_2 = \mathrm{SD}(\theta)$ of the second group. Let γ_0 denote item parameters a_i and b_i of common items I_0 and γ_g item parameters from the group-specific sets of unique items I_g for $g = 1, 2$. Furthermore, denote by $\boldsymbol{X}_g$ the matrix of observed item responses from group g for items $i \in I_0 \cup I_g$. For group $g = 1, 2$, the log-likelihood function is defined by:

$$l(\mu, \sigma, \gamma_0, \gamma_g; \boldsymbol{X}_g) = \sum_{p=1}^{N_g} \log\left(\int \prod_{i \in I_0 \cup I_g} P_i(x_{pgi}; a_i, b_i) f(\theta; \mu, \sigma)\, \mathrm{d}\theta \right) \quad , \tag{10}$$

where x_{pgi} is the item response of person p in group g at item i. The IRT model in (10) can be estimated by marginal maximum likelihood (MML) estimation and an expectation–maximization algorithm [76–79].

For reasons of identification, we define $\mu_1 = 0$ and $\sigma_1 = 1$ and identify the distribution parameters μ_2 and σ_2 of the second group. Separate calibrations for the two groups can be carried out and result in group-specific item parameter estimates $\hat{a}_{ig}$ and $\hat{b}_{ig}$ (see Section 2.3.1). These item parameters were subsequently transformed utilizing linking methods in order to obtain estimates $\hat{\mu}_2$ and $\hat{\sigma}_2$. See [16,22,80–82] for an overview of linking methods. In the next subsection, we discuss several linking methods.

3.1. Log-Mean-Mean Linking

In log-mean-mean linking (logMM; [16]), the means of the logarithmized item discriminations and item difficulties in the two groups are set equal for identifying group means and group SDs. The SD σ_2 of the second group is estimated as:

$$\hat{\sigma}_2 = \exp\left(\frac{1}{|I_0|} \sum_{i \in I_0} \log\, \hat{a}_{i2} - \frac{1}{|I_0|} \sum_{i \in I_0} \log\, \hat{a}_{i1} \right) \quad , \tag{11}$$

where $|I_0|$ is the number of items in the set I_0. The estimation in (11) corresponds to the assumption of additive DIF effects in logarithmized item discriminations (see (7)). The mean μ_2 of the second group is estimated by:

$$\hat{\mu}_2 = -\hat{\sigma}_2 \frac{1}{|I_0|} \sum_{i \in I_0} \hat{b}_{i2} + \frac{1}{|I_0|} \sum_{i \in I_0} \hat{b}_{i1} \quad . \tag{12}$$

It can be shown that logMM estimates are consistent under weak conditions. Here, consistency means that the number of common items (i.e., $|I_0|$) is tending toward infinity. Moreover, it was assumed that group-specific item parameters would have been estimated with an infinite sample size (i.e., $N \to \infty$). Consistency proofs can be easily modified to the case of finite sample sizes (use $\xrightarrow{p}$ as the mathematical symbol for consistency in the folllowing). Then, consistency is meant under a double sampling scheme in which the number of persons and number of items tends to infinity (i.e., $N \to \infty$ and $|I_0| \to \infty$). We now present the consistency result.

Proposition 1. *Assume that DIF effects e_i fulfill $\mathrm{E}(e_i) = 0$. For DIF effects f_i, one of the following conditions holds:*

(i) For additive DIF effects f_i (Equation (3)), it holds that $\mathrm{E}(f_i) = 0$ and f_i has a symmetric distribution;

(ii) For multiplicative DIF effects f_i (Equation (7)), it holds that $\mathrm{E}(\log f_i) = 0$.

Then, logMM estimators $\hat{\mu}_2$ and $\hat{\sigma}_2$ are consistent for μ_2 and σ_2, respectively:

$$\hat{\mu}_2 \xrightarrow{p} \mu_2 \quad \text{and} \quad \hat{\sigma}_2 \xrightarrow{p} \sigma_2 \quad \text{for } |I_0| \to \infty \quad . \tag{13}$$

Proof. See Appendix B. □

3.2. Mean-Mean Linking

In mean-mean linking (MM; [16]), the means of untransformed item discriminations are matched. Hence, the estimation in (11) is substituted by:

$$\hat{\sigma}_2 = \frac{\frac{1}{|I_0|}\sum_{i\in I_0}\hat{a}_{i2}}{\frac{1}{|I_0|}\sum_{i\in I_0}\hat{a}_{i1}} \quad . \tag{14}$$

This estimation corresponds to additive DIF effects in untransformed item discriminations (see (3)). The estimation formula for $\hat{\mu}_2$ in (12) remains unaltered.

It can also be shown that MM estimates are consistent under weak conditions.

Proposition 2. *Assume that DIF effects e_i fulfill $\mathrm{E}(e_i) = 0$. For DIF effects f_i, one of the following conditions holds:*

(i) For additive DIF effects f_i (Equation (3)), it holds that $\mathrm{E}(f_i) = 0$;

(ii) For multiplicative DIF effects f_i (Equation (7)), it holds that $\mathrm{E}(\log f_i) = 0$ and $\log f_i$ has a symmetric distribution.

Then, MM estimators $\hat{\mu}_2$ and $\hat{\sigma}_2$ are consistent for μ_2 and σ_2, respectively:

$$\hat{\mu}_2 \xrightarrow{p} \mu_2 \quad \text{and} \quad \hat{\sigma}_2 \xrightarrow{p} \sigma_2 \quad \text{for } |I_0| \to \infty \quad . \tag{15}$$

Proof. See Appendix C. □

Finally, it is instructive to study the effects on the estimates in MM linking if the true effects do not have zero means. For additive DIF effects f_i, we assumed $\mathrm{E}(f_i) = \delta_f$ and $\mathrm{E}(e_i) = \delta_e$. Using similar derivations as in Appendix C, we obtain:

$$\hat{\sigma}_2 \xrightarrow{p} \sigma_2 \frac{A+\delta_f}{A-\delta_f} = \sigma_2 \frac{1+\delta_f/A}{1-\delta_f/A} \quad , \tag{16}$$

where $A = \lim_{|I_0|\to\infty} \frac{1}{|I_0|}\sum_{i=1}^{|I_0|} a_i$. Setting $B = \lim_{|I_0|\to\infty} \frac{1}{|I_0|}\sum_{i=1}^{|I_0|} b_i$, we obtain:

$$\hat{\mu}_2 \xrightarrow{p} \mu_2 \frac{1+\delta_f/A}{1-\delta_f/A} - 2B\frac{\delta_f/A}{1-\delta_f/A} - 2\delta_e \frac{1}{1-\delta_f/A} \quad . \tag{17}$$

If DIF effects do not have zero means, Equations (16) and (17) show that biased estimates for the group mean and the group standard deviation can be expected.

3.3. Haberman Linking (HAB and HAB-Nolog)

Haberman linking (HAB; [71,83,84]) provides a generalization of logMM and MM to multiple groups while simultaneously estimating common item parameters $\boldsymbol{a}$ and $\boldsymbol{b}$. HAB consists of two steps. In the first step, group-specific SDs are estimated. In the second step, group-specific means are estimated.

The originally proposed HAB linking [83] operates on logarithmized item discriminations (HAB). To estimate σ_2, the linking function in HAB for the particular case of two groups is:

$$\begin{aligned} H_{1,\log}(\sigma_2, \boldsymbol{a}) &= \sum_{i\in I_0\cup I_1} (\log\hat{a}_{i1} - \log a_i)^2 \\ &\quad + \sum_{i\in I_0\cup I_2} (\log\hat{a}_{i2} - \log a_i - \log\sigma_2)^2 \quad . \end{aligned} \tag{18}$$

Parameters in HAB linking are estimated as the minimizer of (18):

$$(\hat{\sigma}_2, \hat{\boldsymbol{a}}) = \underset{\sigma_2, \boldsymbol{a}}{\arg\min}\, H_{1,\log}(\sigma_2, \boldsymbol{a}) \quad . \tag{19}$$

For $i \in I_g$ for $g = 1, 2$, we obtain $\hat{a}_i = \hat{a}_{ig}$. Furthermore, for $i \in I_0$, the minimization of (18) corresponds to a two-way analysis of variance. We obtain (see Equation (A31) in Appendix D):

$$\hat{\sigma}_2 = \exp\left(\frac{1}{|I_0|} \sum_{i \in I_0} \log \hat{a}_{i2} - \frac{1}{|I_0|} \sum_{i \in I_0} \log \hat{a}_{i1} \right) \quad , \tag{20}$$

which shows the equivalence to logMM linking with respect to the estimation of σ_2.

In the second step, group means are estimated. The linking function is given as:

$$H_2(\mu_2, \boldsymbol{b}) = \sum_{i \in I_0 \cup I_1} \left(\hat{b}_{i1} - b_i \right)^2 + \sum_{i \in I_0 \cup I_2} \left(\hat{\sigma}_2 \hat{b}_{i2} - b_i + \mu_2 \right)^2 \quad . \tag{21}$$

The parameters are estimated as:

$$(\hat{\mu}_2, \hat{\boldsymbol{b}}) = \underset{\mu_2, \boldsymbol{b}}{\arg\min}\, H_2(\mu_2, \boldsymbol{b}) \quad . \tag{22}$$

Using the same derivation as for $H_{1,\log}$, we obtain:

$$\hat{\mu}_2 = -\hat{\sigma}_2 \frac{1}{|I_0|} \sum_{i \in I_0} \hat{b}_{i2} + \frac{1}{|I_0|} \sum_{i \in I_0} \hat{b}_{i1} \quad . \tag{23}$$

Notably, Equation (23) coincides with the estimation in logMM linking (see Equation (12)).

As an alternative, Haberman linking can also be conducted based on untransformed item discriminations [71]. This method is labeled as HAB-nolog. It turned out that HAB-nolog outperformed HAB in some situations for multiple groups [71,84]. The linking function of HAB-nolog is given by:

$$H_{1,\text{nolog}}(\sigma_2, \boldsymbol{a}) = \sum_{i \in I_0 \cup I_1} (\hat{a}_{i1} - a_i - 1)^2 + \sum_{i \in I_0 \cup I_2} (\hat{a}_{i2} - a_i - \sigma_2)^2 \quad . \tag{24}$$

Parameter estimates in HAB-nolog linking are determined by:

$$(\hat{\sigma}_2, \hat{\boldsymbol{a}}) = \underset{\sigma_2, \boldsymbol{a}}{\arg\min}\, H_{1,\text{nolog}}(\sigma_2, \boldsymbol{a}) \quad . \tag{25}$$

The SD of the second group is given by (see Equation (A36) in Appendix D):

$$\hat{\sigma}_2 = 1 + \frac{1}{|I_0|} \sum_{i \in I_0} \hat{a}_{i2} - \frac{1}{|I_0|} \sum_{i \in I_0} \hat{a}_{i1} \quad . \tag{26}$$

The linking function for μ_2 in HAB-nolog is the same as in HAB (see Equation (21)).

3.4. *Invariance Alignment with* $p = 2$

Asparouhov and Muthén [85,86] proposed the method of invariance alignment (IA) to define a linking that maximizes the extent of invariant item parameters. IA can also be regarded as a linking method that handles noninvariant item parameters [71].

The IA method is based on estimated group-specific item intercepts $\hat{\nu}_{ig} = \hat{a}_{ig}\hat{b}_{ig}$ and item discriminations $\hat{a}_{ig}$ obtained from separate calibrations. IA was originally formulated to detect only a few noninvariant items. It has been pointed out that IA in its original proposal is not an acceptable linking method in the presence of normally distributed DIF effects [87–89]. In [71], IA was studied using a general class of so-called L_p-type robust linking functions. The original IA formulation used $p = 0.5$ [85]. For normally distributed DIF effects, $p = 2$ is an adequate choice [71,89]. Hence, we investigate IA with the loss function using $p = 2$ (IA2).

It has been shown that IA estimation originally formulated as a joint estimation problem can be reformulated as a two-step estimation method [71]. In the first step, group SDs are computed. In the second step, group means are computed. For two groups, σ_2 is estimated in the first step and μ_2 in the second step. Note that the first group serves as the reference group (i.e., it holds that $\mu_1 = 0$ and $\sigma_1 = 1$). The formulas in [71] can be simplified to the case of two groups for providing estimates $\hat{\mu}_2$ and $\hat{\sigma}_2$:

$$\hat{\sigma}_2 = \arg\min_{\sigma_2} \sum_{i \in I_0} \left(\hat{a}_{i1} - \frac{\hat{a}_{i2}}{\sigma_2} \right)^2 \quad \text{and} \tag{27}$$

$$\hat{\mu}_2 = \arg\min_{\mu_2} \sum_{i \in I_0} \left(\hat{v}_{i1} - \hat{v}_{i2} + \mu_2 \frac{\hat{a}_{i2}}{\hat{\sigma}_2} \right)^2 \quad . \tag{28}$$

The estimates in (27) and (28) have close expressions (see Equations (A41) and (A45) in Appendix E).

3.5. Haebara Linking Methods (HAE-Asymm, HAE-Symm, HAE-Joint)

In contrast to the MM, logMM, HAB, HAB-nolog, and IA linking methods, Haebara (HAE) linking [90] aligns IRFs instead of directly aligning item parameters. The linking function in asymmetric HAE (HAE-asymm; [90]) linking is given as:

$$H_{\text{asymm}}(\mu_2, \sigma_2) = \sum_{i \in I_0} \int \left[\Psi\left(\hat{a}_{i1}(\theta - \hat{b}_{i1}) \right) - \Psi\left(\sigma_2^{-1} \hat{a}_{i2}(\theta - \sigma_2 \hat{b}_{i2} - \mu_2) \right) \right]^2 \omega(\theta) \, \mathrm{d}\theta \quad . \tag{29}$$

Estimated distribution parameters are obtained by minimizing (29):

$$(\hat{\mu}_2, \hat{\sigma}_2) = \arg\min_{\mu_2, \sigma_2} H_{\text{asymm}}(\mu_2, \sigma_2) \quad . \tag{30}$$

The linking function (29) aligns IRFs of the second group to those in the first group. Hence, it is asymmetric because parameters in the second group are expected to behave similarly to the first group. However, the first group could alternatively be aligned to the second group. To robustify the HAE linking, both directions of alignment are considered in symmetric Haebara linking (HAE-symm; [91,92]), which employs the linking function:

$$\begin{aligned} H_{\text{symm}}(\mu_2, \sigma_2) &= \sum_{i \in I_0} \int \left[\Psi\left(\hat{a}_{i1}(\theta - \hat{b}_{i1}) \right) - \Psi\left(\sigma_2^{-1} \hat{a}_{i2}(\theta - \sigma_2 \hat{b}_{i2} - \mu_2) \right) \right]^2 \omega(\theta) \, \mathrm{d}\theta \\ &\quad + \sum_{i \in I_0} \int \left[\Psi\left(\hat{a}_{i1} \sigma_2 (\theta - \sigma_2^{-1}(\hat{b}_{i1} - \mu_2)) \right) - \Psi\left(\hat{a}_{i2}(\theta - \hat{b}_{i2}) \right) \right]^2 \omega(\theta) \, \mathrm{d}\theta \quad . \end{aligned} \tag{31}$$

In a similar vein, the estimated mean and the SD for the second group is given by and defined as:

$$(\hat{\mu}_2, \hat{\sigma}_2) = \arg\min_{\mu_2, \sigma_2} H_{\text{symm}}(\mu_2, \sigma_2) \quad . \tag{32}$$

A generalization of HAE linking to the general case of multiple groups was proposed in [93]. In this joint Haebara (HAE-joint) linking approach, distribution parameters are simultaneously estimated with common item parameters. The linking function in HAE-joint is defined as [65,93,94]:

$$\begin{aligned} H_{\text{joint}}(\mu_2, \sigma_2, \boldsymbol{a}, \boldsymbol{b}) &= \sum_{i \in I_0} \int \left[\Psi\left(\hat{a}_{i1}(\theta - \hat{b}_{i1}) \right) - \Psi\left(a_i(\theta - b_i) \right) \right]^2 \omega(\theta) \, \mathrm{d}\theta \\ &\quad + \sum_{i \in I_0} \int \left[\Psi\left(\hat{a}_{i2}(\theta - \hat{b}_{i2}) \right) - \Psi\left(a_i(\sigma_2 \theta - b_i + \mu_2) \right) \right]^2 \omega(\theta) \, \mathrm{d}\theta \quad , \end{aligned} \tag{33}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ denote common item parameters. Parameter estimates are given by minimizing (33):

$$(\hat{\mu}_2, \hat{\sigma}_2, \hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}) = \underset{\mu_2, \sigma_2, \boldsymbol{a}, \boldsymbol{b}}{\arg\min}\ H_{\text{joint}}(\mu_2, \sigma_2, \boldsymbol{a}, \boldsymbol{b}) \quad . \tag{34}$$

A variant of joint Haebara linking was proposed in [84]. Note that in joint Haebara linking, common IRFs are aligned to group-specific IRFs.

3.6. Recalibration Linking (RC1, RC2, and RC3)

A further linking technique is recalibration (RC) linking. This is based on item parameters obtained from separate calibrations. The core idea is that group differences in distributions can be inferred by recalibrating a study with one group using item parameters from the other group. RC methods are mainly used in LSA studies such as the Programme for International Student Assessment (PISA; [95]), Progress in International Reading Literacy Study (PIRLS; [96]), and Trends in International Mathematics and Science Study (TIMSS; [97,98]).

In PISA, RC linking was employed until PISA 2012 for the 1PL model to handle model misspecifications (i.e., the data-generating IRT model deviates from the 1PL model) in linking that could artificially impact the estimated SDs [99]. RC linking in PIRLS and TIMSS operates on the 3PL model [98,100,101]. Notably, recalibration methods have also been proposed for determining linking errors in PIRLS/TIMSS [101], as well as in PISA ([69], p. 176ff.), but the two approaches turned out to be different.

RC linking is based on estimated item parameters from the first and the second groups. Item parameters are obtained assuming $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 1$ in separate calibrations. The group-specific parameter estimates are defined as:

$$(\hat{\gamma}_0^{(1)}, \hat{\gamma}_1^{(1)}) = \underset{\gamma_0, \gamma_1}{\arg\max}\ l(0, 1, (\gamma_0, \gamma_1); \boldsymbol{X}_1) \quad \text{and} \tag{35}$$

$$(\hat{\gamma}_0^{(2)}, \hat{\gamma}_2^{(2)}) = \underset{\gamma_0, \gamma_2}{\arg\max}\ l(0, 1, (\gamma_0, \gamma_2); \boldsymbol{X}_2) \quad . \tag{36}$$

To obtain the mean and the SD of the second group, the data of the first group (i.e., $\boldsymbol{X}_1$) are recalibrated using item parameters from the second group (i.e., $\hat{\gamma}_0^{(2)}$):

$$(m_1, s_1, \hat{g}_1) = \underset{\mu, \sigma, \gamma_1}{\arg\max}\ l(\mu, \sigma, (\hat{\gamma}_0^{(2)}, \gamma_1); \boldsymbol{X}_1) \quad . \tag{37}$$

A recalibrated mean m_1 and a recalibrated SD s_2 are obtained. These two parameters indicate differences between the two groups. Similarly, data of the second group (i.e., $\boldsymbol{X}_2$) can be recalibrated using item parameters from the first group (i.e., $\hat{\gamma}_0^{(1)}$):

$$(m_2, s_2, \hat{g}_2) = \underset{\mu, \sigma, \gamma_2}{\arg\max}\ l(\mu, \sigma, (\hat{\gamma}_0^{(1)}, \gamma_2); \boldsymbol{X}_2) \quad . \tag{38}$$

Based on these estimates, the distribution parameters for the second group are defined as:

$$\hat{\mu}_2 = -s m_1 \quad , \quad \hat{\sigma}_2 = s \tag{39}$$

where the scaling factor s can take different forms. Note that $\hat{\mu}_2$ relies on the recalibrated mean m_1 for the first group and the scaling factor s. The three RC linking methods differ concerning the scaling constant used:

$$\text{Method RC1}: s = \frac{1}{s_1} \quad , \quad \text{Method RC2}: s = s_2 \quad , \quad \text{Method RC3}: s = \sqrt{\frac{s_2}{s_1}} \tag{40}$$

The linking methods RC1 and RC2 are asymmetric, while method RC3 relies on both recalibrated SDs. The scaling factor in RC3 linking is defined as the geometric mean of the scaling factors in RC1 and RC2. The definition is motivated by the fact that the impact on recalibrated SDs is symmetrically treated. The linking method RC1 is currently used in PIRLS and TIMSS [96,98] and was used in PISA until 2012 [99]. To our knowledge, linking methods RC2 and RC3 have not yet been investigated in the literature.

3.7. Anchored Item Parameters

The linking method based on anchored item parameters (ANCH; [16,22,102]) estimates the distribution parameters in the second group by fixing the item parameters of the common items to those of the first group. Assume that item parameter estimates in the first group are computed as:

$$(\hat{\gamma}_0, \hat{\gamma}_1) = \underset{\gamma_0, \gamma_1}{\arg\max}\ l(0, 1, (\gamma_0, \gamma_1); \boldsymbol{X}_1) \quad . \tag{41}$$

In ANCH linking, μ_2 and σ_2 are estimated by maximizing the log-likelihood function while fixing γ_0 (i.e., $\gamma_0 = \hat{\gamma}_0$):

$$(\hat{\mu}_2, \hat{\sigma}_2, \hat{\gamma}_2) = \underset{\mu_2, \sigma_2, \gamma_2}{\arg\max}\ l(\mu_2, \sigma_2, (\hat{\gamma}_0, \gamma_2); \boldsymbol{X}_2) \quad . \tag{42}$$

It should be noted that RC linking can be considered a variant of ANCH linking because, in RC linking, the distribution parameters of the first group are re-estimated using anchored item parameters from the second group. However, the distribution parameters of the second group are indirectly obtained by transforming the re-estimated distribution parameters of the first group (see Section 3.6). In contrast, ANCH provides an estimate of μ_2 and σ_2 directly.

3.8. Concurrent Calibration

Concurrent calibration (CC; [65,70,103]) is based on a multiple-group IRT model and presupposes invariant item groups across groups. The distribution parameters of the second group are determined by maximizing the joint likelihood:

$$(\hat{\mu}_2, \hat{\sigma}_2, \hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2) = \underset{\mu_2, \sigma_2, \gamma_0, \gamma_1, \gamma_2}{\arg\max} \left\{ l(0, 1, (\gamma_0, \gamma_1); \boldsymbol{X}_1) + l(\mu_2, \sigma_2, (\gamma_0, \gamma_2); \boldsymbol{X}_2) \right\} \tag{43}$$

In the presence of random DIF, the log-likelihood function in (43) will typically be misspecified. The estimated mean and SD can typically be biased due to this misspecification. It has frequently been pointed out that separate calibration with subsequent linking can be more robust to the presence of DIF than CC [16,54,102]. CC can only be expected to be more efficient than linking based on separate calibrations in small to moderate sample sizes and in the absence of DIF [65]. Notably, CC is more computationally demanding than linking based on separate calibration [65,104].

4. Simulation Study

4.1. Purpose

The purpose of this simulation study was to investigate the performance of linking methods in the two-group case for the 2PL model under different sample sizes, different numbers of items, and different amounts of uniform and nonuniform DIF. Most simulation studies either assume invariant item parameters (i.e., no DIF) or presuppose partial invariance in which only a few item parameters differ between groups (e.g., [63,105–112]). There is a lack of research in the presence of random DIF, although there is some initial work for continuous items [88,89]. Moreover, although recalibration linking is in operational use in LSA studies, they have not yet been systematically compared to alternative linking methods.

We expected from the previous research and our analytical findings in Propositions 1 and 2 that moment-based linking methods could be competitive with the CC and HAE linking methods [54,65]. We did not have specific hypotheses regarding the performance of recalibration linking.

4.2. Design

We simulated a design with two groups and only common items (i.e., no unique items). Data were simulated according to the 2PL model with different amounts of random DIF. In the simulation, it was assumed that there were only common items and no group-specific unique items. The 2PL model with random DIF was simulated according to Equations (3) and (4). Table A1 in Appendix F shows item parameters a_i and b_i for 20 items used in the simulation. The first group served as a reference group assuming $\mu_1 = 0$ and $\sigma_1 = 1$, while the distribution parameters for the second group were $\mu_2 = 0.3$ and $\sigma_2 = 1.2$.

In the simulation, four factors were simulated. First, we chose sample sizes $N = 500$, 1000, and 5000 (3 factor levels). Second, the item number was either $I = 20$ or $I = 40$ (2 factor levels). For 40 items, the item parameters from Table A1 were duplicated. The random DIF SD τ_b for item difficulties b_i was 0, 0.1, 0.3, or 0.5 (4 factor levels). The random DIF SD τ_a for item discriminations a_i was 0, 0.15, 0.25 (3 factor levels). In total, there were $3 \times 2 \times 4 \times 3 = 96$ conditions employed in the simulation.

In total, 1000 datasets were simulated and analyzed in each condition.

4.3. Analysis Methods

The 2PL model was separately estimated in the two groups. Afterward, 13 linking methods (logMM, HAB-log, MM, HAB-nolog, IA2, HAE-asymm, HAE-symm, HAE-joint, RC1, RC2, RC3, ANCH, CC; see Section 3) were applied.

The parameters of interest were the estimated mean $\hat{\mu}_2$ and SD $\hat{\sigma}_2$ for the second group. For the two parameters, the bias and root-mean-squared error (RMSE) were computed. To decrease the dependence of the RMSE on the sample size and the number of items, we computed a relative RMSE for which the RMSE of a linking method was divided by the RMSE of the linking method with the best performance. Hence, this relative RMSE had 100 for the best linking method as its lowest value.

To summarize the contribution of each of the manipulated factors in the simulation, we conducted an analysis of variance (ANOVA) and used a variance decomposition to assess the importance.

Moreover, we classified linking methods as whether or not they showed satisfactory performance in a particular condition. We defined satisfactory performance for the bias if the absolute bias of a parameter (i.e., the estimated mean $\hat{\mu}_2$ or estimated SD $\hat{\sigma}_2$) was smaller than 0.01. In LSA studies such as the Programme for International Student Assessment (PISA), standard errors were about 0.02 or 0.03 for standardized ability variables θ. It should be required that the bias only be a fraction of the variability introduced by the sampling error, which motivated the value of 0.01 as a cutoff. An estimator had satisfactory performance concerning the RMSE if the relative RMSE was smaller than 120. Here, an alternative estimator to the best-performing estimation should not lose too much precision. With an RMSE of 120, the relative-mean-squared error (MSE) was 144 (note that $1.2^2 = 1.44$), which corresponded to a loss in precision of 44%. Estimators with worse performance might not be considered as satisfactory in such a situation.

In all the analyses, the statistical software R [113] was used. The R package TAM [114] was used to estimate the 2PL model with marginal maximum likelihood as the estimation method. The linking methods were estimated using dedicated R functions or the existing functionality in the R packages sirt [115] and TAM [114]. The ANOVA model was estimated with the R package lme4 [116].

4.4. Results

In Table 1, the variance decomposition of the ANOVA is presented. All terms up to three-way interactions are included. From the size of the residual variance, it can be concluded that the first three orders are sufficient to capture the most important factors in the simulation.

It turned out that sample size (N) and the number of items (I) were only of minor importance in the first three-order terms in ANOVA. However, the linking method, as well as the size of random DIF were important factors. Importantly, the random DIF SD (τ_b) in item difficulties b_i had a large influence on the estimated means, while random DIF SD (τ_a) in item discriminations a_i strongly impacted the estimated SDs.

Due to these observations, we decided to provide aggregated results for three important groups of cells in the simulation. First, we aggregated the results for six conditions with no DIF (NODIF; $\tau_b = 0$ and $\tau_a = 0$). Because there were two estimated parameters (mean and SD), this resulted in aggregation across twelve measures for the bias and RMSE, respectively. Second, we considered all 18 conditions with uniform DIF (UDIF; $\tau_b > 0$ and $\tau_a = 0$). Third, we provide summaries across all 48 conditions with nonuniform DIF (NUDIF; $\tau_a > 0$).

Table 1. Variance proportions of different factors in the simulation study for the bias and RMSE for the estimated mean $\hat{\mu}_2$ and estimated SD $\hat{\sigma}_2$ for the second group.

Source	$\hat{\mu}_2$		$\hat{\sigma}_2$	
	Bias	**RMSE**	**Bias**	**RMSE**
N	0.3	**1.1**	0.6	**3.9**
I	0.0	0.3	0.0	0.0
Meth	**10.2**	**14.9**	**19.1**	0.0
τ_b	**13.0**	0.0	0.8	**1.8**
τ_a	**4.3**	**9.0**	**12.3**	0.0
$N \times I$	0.0	0.0	0.0	0.0
$N \times$ Meth	0.0	**3.7**	0.8	0.0
$N \times \tau_b$	0.0	**2.4**	0.0	0.0
$N \times \tau_a$	0.0	0.6	0.0	**5.0**
$I \times$ Meth	0.4	0.1	0.1	0.0
$I \times \tau_b$	0.0	0.1	0.0	0.0
$I \times \tau_a$	0.0	0.0	0.0	0.6
Meth $\times \tau_b$	**58.1**	**13.1**	**17.5**	**14.2**
Meth $\times \tau_a$	**8.2**	**12.1**	**47.7**	**13.2**
$\tau_a \times \tau_b$	0.0	**4.1**	0.0	**17.7**
$N \times I \times$ Meth	0.0	0.0	0.1	0.0
$N \times I \times \tau_b$	0.0	0.2	0.0	0.0
$N \times I \times \tau_a$	0.0	0.1	0.0	0.4
$N \times$ Meth $\times \tau_b$	0.2	**7.5**	0.0	**4.2**
$N \times$ Meth $\times \tau_a$	0.0	**4.0**	0.0	**9.1**
$N \times \tau_a \times \tau_b$	0.1	**10.0**	0.0	**13.8**
$I \times$ Meth $\times \tau_b$	0.5	0.0	0.0	0.4
$I \times$ Meth $\times \tau_a$	0.1	0.0	0.2	**1.1**
$I \times \tau_a \times \tau_b$	0.1	0.3	0.0	0.7
Meth $\times \tau_a \times \tau_b$	**1.0**	**10.1**	0.1	**8.2**
Residual	**3.7**	**6.4**	0.6	**5.7**

Note. N = sample size; I = number of items; Meth = linking method; τ_a = standard deviation of DIF effects in item discriminations a_i; τ_b = standard deviation of DIF effects in item difficulties b_i. Percentage values larger than 1.0 are printed in bold.

In Table 2, the performance and the linking methods are summarized across these three groups of conditions by classifying all linking methods as either satisfactory or nonsatisfactory. Table 2 shows the proportion of conditions in which a linking method provided satisfactory results. In the absence of DIF (columns "NODIF"), all methods performed well in most of the conditions. Notably, IA2, the asymmetric recalibration methods RC1 and RC2, as well as ANCH provided biases in some instances. In the case of uniform DIF (columns "UDIF"), HAE-asymm, HAE-joint and CC cannot be recommended in terms of the bias. Moreover, only moment-based methods (logMM, HAB, MM, HAB-nolog) can be recommended in terms of the RMSE. Finally, in the presence of nonuniform DIF (columns "NUDIF"), moment-based methods, as well as HAE-symm and RC3 were satisfactory in terms of the bias. If the RMSE is considered as an additional criterion, utilizing linking methods MM, HAB-nolog, HAE-symm, and RC3 can be suggested.

Table 2. Summary of the satisfactory performance of linking methods for the absolute bias and RMSE across parameters (mean $\hat{\mu}_2$ and standard deviation $\hat{\sigma}_2$) and conditions.

	Bias			**RMSE**		
	NODIF	**UDIF**	**NUDIF**	**NODIF**	**UDIF**	**NUDIF**
logMM	100	97	94	100	100	**45**
HAB	100	97	94	100	100	**44**
MM	100	94	95	92	100	72
HAB-nolog	100	94	96	100	100	78
IA2	75	78	**8**	100	100	**4**
HAE-asymm	100	**42**	**42**	100	**61**	78
HAE-symm	100	97	94	100	**61**	81
HAE-joint	100	**42**	**60**	100	**42**	**61**
RC1	83	78	**16**	100	**61**	**29**
RC2	83	78	**8**	100	**61**	**48**
RC3	100	94	96	100	**61**	79
ANCH	83	78	**13**	100	**61**	**48**
CC	100	**50**	**45**	100	**33**	**46**

Note. DIF = differential item functioning; NODIF = no DIF; UDIF = uniform DIF; NUDIF = nonuniform DIF; logMM = log-mean-mean linking; HAB = Haberman linking with logarithmized item discriminations; MM = mean-mean linking; HAB-nolog = Haberman linking with untransformed item discriminations; IA2 = invariance alignment with power $p = 2$; HAE-asymm = asymmetric Haebara linking; HAE-symm = symmetric Haebara linking; HAE-joint = Haebara linking with joint item parameters; RC = recalibration linking (see Equation (40)); ANCH = anchored item parameters; CC = concurrent calibration. Values smaller than 70 are printed in bold.

In Table 3, the bias and RMSE for the mean $\hat{\mu}_2$ and the SD $\hat{\sigma}_2$ of the second group for $N = 1000$ students and $I = 40$ items are shown. It can be seen that all linking methods performed well in the absence of DIF (see columns "NODIF"). In the case of uniform DIF (columns "UDIF"), HAE-asymm, HAE-joint, and CC produced nonsatisfactory results in terms of the bias for the mean or the SD. Interestingly, the RMSE for the SD was much larger for HAE, RC, ANCH, and CC than the moment-based methods logMM, HAB, MM, and logMM. In the case of nonuniform DIF (column "NUDIF"), only moment-based methods (except IA2) and HAE-symm and the newly proposed recalibration linking method RC3 produced satisfactory results. Furthermore, note that the RMSE for the SD was larger for logMM and HAB compared to MM and HAB-nolog. This finding can be explained by the fact that the data-generating model for DIF was an additive model that operated on nontransformed item discriminations and favored MM and HAB-nolog. To conclude, linking methods that are not moment-based can only be recommended as a default linking method when the mean is of interest and there is an absence of random DIF. It depends on the extent of UDIF (i.e., size of τ_b) whether the additional variance introduced by moment-based methods is compensated by smaller biases.

Table 3. Bias and RMSE for mean $\hat{\mu}_2$ and standard deviation $\hat{\sigma}_2$ for the second group for a sample size $N = 1000$ and $I = 40$ items as a function of the type of differential item functioning and linking method.

	Bias			**RMSE**		
	NODIF	**UDIF**	**NUDIF**	**NODIF**	**UDIF**	**NUDIF**
	$\tau_b = 0$	$\tau_b = 0.5$	$\tau_b = 0.5$	$\tau_b = 0$	$\tau_b = 0.5$	$\tau_b = 0.5$
	$\tau_a = 0$	$\tau_a = 0$	$\tau_a = 0.25$	$\tau_a = 0$	$\tau_a = 0$	$\tau_a = 0.25$
	Mean $\hat{\mu}_2$					
logMM	0.000	0.007	0.008	108.2	104.4	106.1
HAB	0.000	0.007	0.008	108.2	104.4	106.1
MM	0.000	0.007	0.007	108.1	103.7	104.7
HAB-nolog	0.001	0.007	0.007	108.5	103.5	104.5
IA2	−0.001	0.001	**0.045**	103.2	107.5	**133.3**
HAE-asymm	−0.002	**−0.030**	**−0.032**	102.3	100.0	100.0
HAE-symm	−0.001	0.002	0.005	102.7	105.0	105.2
HAE-joint	−0.002	**0.067**	**0.064**	100.9	**136.1**	**132.4**
RC1	−0.001	0.001	**0.028**	100.2	104.8	**120.5**
RC2	−0.006	−0.004	**−0.022**	100.0	104.0	100.1
RC3	−0.003	−0.001	0.002	100.1	103.9	109.4
ANCH	−0.003	−0.004	**−0.021**	101.4	104.2	103.9
CC	−0.002	**0.095**	**0.109**	101.3	**149.2**	**157.7**
	Standard Deviation $\hat{\sigma}_2$					
logMM	0.000	0.003	0.008	110.2	112.6	**128.9**
HAB	0.000	0.003	0.008	110.2	112.6	**129.4**
MM	−0.001	0.001	0.005	108.5	109.4	107.7
HAB-nolog	0.001	0.002	0.007	100.0	100.0	100.0
IA2	0.009	0.009	**0.147**	113.2	111.6	**197.9**
HAE-asymm	−0.002	**−0.120**	**−0.134**	107.2	**378.8**	**185.6**
HAE-symm	0.001	−0.003	0.003	108.3	**233.7**	119.9
HAE-joint	−0.001	**0.020**	**0.029**	107.5	**317.0**	**146.6**
RC1	0.006	0.008	**0.105**	109.8	**243.8**	**174.5**
RC2	−0.009	−0.008	**−0.097**	108.5	**217.2**	**148.3**
RC3	−0.002	0.000	0.002	106.6	**228.3**	110.2
ANCH	−0.009	−0.008	**−0.097**	108.5	**217.2**	**148.3**
CC	−0.001	0.015	**0.029**	107.4	**220.4**	**129.0**

Note: DIF = differential item functioning; NODIF = no DIF; UDIF = uniform DIF; NUDIF = nonuniform DIF; τ_a = standard deviation of DIF effects in item discriminations a_i; τ_b = standard deviation of DIF effects in item difficulties b_i; logMM = log-mean-mean linking; HAB = Haberman linking with logarithmized item discriminations; MM = mean-mean linking; HAB-nolog = Haberman linking with untransformed item discriminations; IA2 = invariance alignment with power $p = 2$; HAE-asymm = asymmetric Haebara linking; HAE-symm = symmetric Haebara linking; HAE-joint = Haebara linking with joint item parameters; RC = recalibration linking (see Equation (40)); ANCH = anchored item parameters; CC = concurrent calibration. Absolute biases larger than 0.02 are printed in bold. RMSE values larger than 120 are printed in bold.

5. Empirical Example: Linking PISA 2006 and PISA 2009 for Austria

5.1. Method

In order to illustrate the consequences of different scaling models (i.e., 1PL and 2PL models), as well as different linking methods (see Section 3), we analyzed the data for Austrian students from the PISA study conducted in 2006 (PISA 2006; [95]) and 2009 (PISA 2009; [117]). In Table 4, the sample sizes (i.e., N) and item numbers (i.e., I) used are presented. There were common items in the two PISA studies (Mathematics: $I_0 = 35$; Reading: $I_0 = 27$; Science: $I_0 = 53$). Moreover, the officially reported means and SDs for Austrian students in PISA 2006 and PISA 2009 are displayed. Note that PISA 2006 and PISA 2009 employed the 1PL model and utilized the MM linking method with a subsequent recalibration linking (method RC1; [117]).

Table 4. Sample information and descriptive results for PISA 2006 and PISA 2009 Austria.

Domain	*N*		*I*		M		SD	
	P06	**P09**	**P06**	**P09**	**P06**	**P09**	**P06**	**P09**
Mathematics	3784	4575	48	35	506.8	495.9	96.8	96.1
Reading	2646	6585	27	99	491.2	470.3	107.7	100.1
Science	4927	4577	103	53	511.7	494.3	97.3	101.8

Note: M = mean; SD = standard deviation; P06 = PISA 2006; P09 = PISA 2009; *N* = number of students; *I* = number of items; M = mean; SD = standard deviation.

In both PISA studies, we included only those students who received a test booklet with at least one item in a respective domain. For simplicity, all polytomous items (i.e., items with maximum scores larger than one) were dichotomously recoded, with only the highest category being recoded as correct. The 1PL and the 2PL model were used as scaling models, and student weights were taken into account. In total, 13 linking methods (logMM, HAB-log, MM, HAB-nolog, IA2, HAE-asymm, HAE-symm, HAE-joint, RC1, RC2, RC3, ANCH, CC; see Section 3) were applied. In the linking procedure, the distribution parameters of the first study (PISA 2006) were fixed (i.e., $\mu_1 = 0$, $\sigma_1 = 0$), and the distribution parameters of the second study (PISA 2009) were estimated. The two ability distributions for the three domains (and the distribution parameters) were linearly transformed such that the mean equaled the officially reported mean and the SD equaled the officially reported SD in PISA 2006 in the respective domain. For example, for Mathematics, it holds that $\hat{\mu}_1 = \text{M} = 506.8$ and $\hat{\sigma}_1 = \text{SD} = 96.8$ in PISA 2006 for all linking methods for the 1PL and the 2PL model.

5.2. Results

In Table 5, the trend estimates for Austria from PISA 2006 and PISA 2009 are shown. For the linearly transformed scores, the trend estimate is given as $\hat{\Delta} = \hat{\mu}_2 - \hat{\mu}_1$. There was a significant and large negative trend for Mathematics and Science, while the trend in Reading turned out to be smaller. Across both models (1PL and 2PL) and linking methods, the average trend in Mathematics was $\text{M} = -14.4$ ($\text{SD} = 1.1$, $\text{Range} = 3.6$). There were slight differences between the 1PL and the 2PL models for the moment-based linking methods (i.e., logMM, HAB, MM, HAB-nolog). Notably, the variation of estimates across linking methods was larger for the 2PL model ($\text{SD} = 1.3$) than for the 1PL model ($\text{SD} = 0.7$). The trend in Reading was smaller ($\text{M} = -5.4$) and less variable ($\text{SD} = 0.8$, $\text{Range} = 2.4$) than in Mathematics. Finally, the trend in Science was also strongly negative ($\text{M} = -14.5$). The variability ($\text{SD} = 1.3$, $\text{Range} = 5.2$) was mainly caused by the variability in linking methods of the 2PL model. In contrast, linking methods for the 1PL model were very similar ($\text{SD} = 0.3$, $\text{Range} = 0.8$).

In Table 6, the estimated SD $\hat{\sigma}_2$ in PISA 2009 is displayed for the 1PL and the 2PL models as a function of the linking method. Interestingly, the variability across linking methods turned out to be larger than for the mean estimate. Relatively large differences of the 1PL model and the 2PL model were observed for Reading (1PL: $\text{M} = 100.4$, 2PL: $\text{M} = 105.2$) and Science (1PL: $\text{M} = 104.1$, 2PL: $\text{M} = 107.4$). Note that the differences between the 1PL and the 2PL model were particularly pronounced for the IA2 method. The difference between the two scaling models might be explained by different average item discriminations for common items and unique items. Reading was a minor domain in PISA 2006 (no unique items) and a major domain in PISA 2009 in which a large set of unique items were introduced. Science was a major domain in PISA 2006 (many unique items) and a minor domain in PISA 2009 (no unique items). In contrast, Mathematics was a minor domain in PISA 2006 and PISA 2009, and there was only a small number of unique items in PISA 2006 and no unique items in PISA 2009. These findings indicate that an adjustment for the average differences in item discriminations has to be conducted to avoid biased estimates of the means and SDs when a misspecified 1PL model is employed (see [99]).

Table 5. Trend estimate for Austrian students in average achievement from PISA 2006 to PISA 2009.

Method	Mathematics		Reading		Science	
	1PL	2PL	1PL	2PL	1PL	2PL
logMM	−15.5	−12.4	−5.8	−6.3	−14.7	−16.8
HAB	−15.5	−12.4	−5.8	−6.3	−14.7	−16.8
MM	−15.5	−12.4	−5.8	−6.3	−14.7	−16.7
HAB-nolog	−15.5	−12.3	−6.0	−6.3	−14.5	−16.6
IA2	−15.5	−15.9	−5.8	−6.1	−14.7	−11.6
HAE-asymm	−14.4	−14.6	−4.9	−6.4	−14.2	−15.9
HAE-symm	−14.6	−15.0	−5.0	−6.6	−14.2	−15.7
HAE-joint	−13.5	−14.1	−4.1	−5.0	−13.9	−14.0
RC1	−14.3	−14.5	−4.4	−5.1	−14.0	−13.2
RC2	−14.3	−14.3	−4.3	−5.0	−14.2	−12.9
RC3	−14.3	−14.4	−4.4	−5.0	−14.1	−13.1
ANCH	−14.4	−15.7	−4.5	−5.4	−14.5	−14.1
CC	−14.3	−14.9	−4.3	−5.3	−14.2	−13.6
M	−14.8	−14.1	−5.0	−5.8	−14.3	−14.7
SD	0.7	1.3	0.7	0.6	0.3	1.8
Min	−15.5	−15.9	−6.0	−6.6	−14.7	−16.8
Max	−13.5	−12.3	−4.1	−5.0	−13.9	−11.6

Note. logMM = log-mean-mean linking; HAB = Haberman linking with logarithmized item discriminations; MM = mean-mean linking; HAB-nolog = Haberman linking with untransformed item discriminations; IA2 = invariance alignment with power $p = 2$; HAE-asymm = asymmetric Haebara linking; HAE-symm = symmetric Haebara linking; HAE-joint = Haebara linking with joint item parameters; RC = recalibration linking (see Equation (40)); ANCH = anchored item parameters; CC = concurrent calibration.

Table 6. Standard deviation for Austrian students in PISA 2009. for domains Mathematics, Reading and Science for the 1PL and the 2PL model as a function of the linking method.

Method	Mathematics		Reading		Science	
	1PL	2PL	1PL	2PL	1PL	2PL
logMM	97.7	98.3	98.6	103.2	103.2	106.8
HAB	97.7	98.3	98.6	103.2	103.2	106.8
MM	97.7	98.7	98.6	103.8	103.2	106.9
HAB-nolog	97.9	99.3	94.6	102.0	103.9	108.1
IA2	97.7	99.5	98.6	104.6	103.2	109.2
HAE-asymm	94.1	95.0	102.6	105.4	105.0	107.5
HAE-symm	95.0	96.2	103.1	105.9	105.3	107.8
HAE-joint	95.0	95.7	105.1	107.5	104.7	107.4
RC1	96.0	96.9	103.1	107.2	103.9	108.6
RC2	96.0	95.6	99.9	106.2	104.7	105.9
RC3	96.0	96.3	101.5	106.7	104.3	107.2
ANCH	96.0	95.6	99.9	106.2	104.7	105.9
CC	95.9	96.7	101.3	106.4	104.1	107.5
M	96.3	97.1	100.4	105.2	104.1	107.4
SD	1.2	1.5	2.7	1.7	0.7	0.9
Min	94.1	95.0	94.6	102.0	103.2	105.9
Max	97.9	99.5	105.1	107.5	105.3	109.2

Note. logMM = log-mean-mean linking; HAB = Haberman linking with logarithmized item discriminations; MM = mean-mean linking; HAB-nolog = Haberman linking with untransformed item discriminations; IA2 = invariance alignment with power $p = 2$; HAE-asymm = asymmetric Haebara linking; HAE-symm = symmetric Haebara linking; HAE-joint = Haebara linking with joint item parameters; RC = recalibration linking (see Equation (40)); ANCH = anchored item parameters; CC = concurrent calibration.

6. Discussion

In this article, several linking methods for two groups were evaluated for the 2PL model in the case of normally distributed random DIF effects with zero means for item difficulties and item discriminations. Somehow surprisingly and contrary to the recommendations in the literature, moment-based linking methods (mean-mean and log-mean-mean, as well as Haberman linking) performed best in terms of the bias and RMSE. The unbiasedness of moment-based methods in the case of many items was expected due to our consistency results for the estimators presented in Section 3. When the primary criterion is the bias, symmetric Haebara (HAE-symm) linking and the newly proposed recalibration method RC3 can only be recommended among the nonmoment-based methods. In contrast, the commonly used asymmetric Haebara (HAE-asymm) linking and the recalibration linking RC1 used in PIRLS and TIMSS had substantially worse performance. Furthermore, note that concurrent calibration—which incorrectly assumes invariant item parameters across groups—and the anchored item parameters method provided biased estimates and cannot be recommended in operational use. Concurrent calibration can only achieve the promised highest efficiency [67,70] in small-to-moderate-sized samples, the absence of random DIF, and a correctly specified IRT model. If data in educational large-scale assessment studies were indeed to follow a random DIF distribution, the currently used linking methods (concurrent calibration, recalibration linking RC1) could be replaced by better alternatives as proposed in this study (moment-based linking, recalibration linking RC3). Of course, the extent of the bias and loss in the precision of the current methods depends on the variance of the DIF effects and can vary from study to study.

Our study assumed that the utilized scaling model (i.e., the 2PL model) was correctly specified. This assumption might be unrealistic in practice, and data could have been generated with much more complex item response functions [118–122] or multidimensional IRT models [123,124]. The performance of the linking methods for misspecified IRT models [125,126] in the presence of random DIF might be an exciting topic for future research [127,128].

In this article, we only considered random DIF effects with a normal distribution and zero means. For relevance in practical applications, it might be interesting for future research to study the DIF effects under different data-generating models. First, random DIF could be simulation in a partial invariance situation in which only a few items have DIF effects, while the majority of DIF effects do not have DIF effects. Second, the case of normally distributed random DIF and partial invariance could occur in tandem. DIF effects could be simulated from a mixture distribution in which the first class includes normally distributed DIF effects with zero means, while the second class of smaller proportion includes outlying and large DIF effects. Due to the presence of outliers, the average of DIF effects will typically differ from zero. For this kind of DIF effects, robust linking methods must be employed for removing these outlying items from group comparisons [65,71,89,94,107].

In this article, we only studied linking in the two-group case. In the case of multiple groups, different linking methods can be employed [65,84,93,129,130]. However, the findings from two groups are likely to generalize to the multiple group case if the linking were to be performed based on sequences of pairwise linking approaches [131–133]. Comparing the performance of pairwise linking approaches and simultaneous linking of multiple groups would be an exciting topic for future research.

Our findings are likely to have even more impact on vertical scaling in which groups constitute time points or grades in the school career [134]. In this situation, differences in the means and SDs between groups are expected to be larger, and the consequences of choosing an inappropriate linking method can be much more pronounced [128,135–138]. As pointed out by an anonymous reviewer, vertical linking poses more assumptions than cross-sectional linking because younger test-takers could incorporate guessing strategies if they had a chance to respond to harder items designed for older test-takers (see also [25]).

It should be noted that we did not investigate the computation of standard errors for the linking methods. There is a rich literature that derives standard error formulas for linking due to sampling of persons (e.g., [85,104,132,139–141]). In addition, variability in estimated group means and SDs due to selecting items has been studied as linking errors in the literature [66,72,99,142–145]. It might be interesting in future research to investigate standard errors that reflect these sources of uncertainty [72,84,132,146]. Procedures that rely on resampling of persons and items will likely correctly reflect uncertainty due to persons and items in the parameters of interest.

Funding: This research received no external funding.

Data Availability Statement: The PISA 2006 and 2009 datasets are available from https://www.oecd.org/pisa/data/ (accessed on 12 July 2021).

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

Abbreviations

The following abbreviations are used in this manuscript:

1PL	one-parameter logistic model
2PL	two-parameter logistic model
ANCH	anchored item parameters
CC	concurrent calibration
DIF	differential item functioning
HAB	Haberman linking with logarithmized item discriminations
HAB-nolog	Haberman linking with untransformed item discriminations
HAE	Haebara linking
HAE-asymm	asymmetric Haebara linking
HAE-joint	Haebara linking with joint item parameters
HAE-symm	symmetric Haebara linking
IA2	invariance alignment with power $p = 2$
IRF	item response function
IRT	item response theory
logMM	log-mean-mean linking
LSA	large-scale assessment
MM	mean-mean linking
MML	marginal maximum likelihood
MSE	mean-squared error
NUDIF	nonuniform differential item functioning
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
RC	recalibration linking
RMSE	root-mean-squared error
SD	standard deviation
TIMSS	Trends in International Mathematics and Science Study
UDIF	uniform differential item functioning

Appendix A. Nonidentifiability of DIF Effects Distributions

In this Appendix, we show that some constraints on the distributions of DIF effects e_i and f_i have to be posed in order to disentangle group differences in the ability distribution from average DIF effects. In Appendixes A.1 and A.2, we discuss nonidentifiability for item difficulties and item discriminations, respectively.

Appendix A.1. DIF Effects for Item Difficulties

First, we show that the mean $\mathrm{E}(e_i)$ must be set to zero for reasons of identification. Assume that there are additive DIF effects for item difficulties (see Equation (4)). Let us assume that DIF effects e_i have a mean different from zero (i.e., $\mathrm{E}(e_i) = \delta_e$. Then, we can write $e_i = \delta_e + e_i^*$ with $E(e_i^*) = 0$. The IRF for item i for persons in the first group (i.e., $g = 1$) can be written as:

$$\mathrm{P}(X_i = 1|\theta) = \Psi(a_i(\theta - b_i^* + e_i^*)) \quad , \quad \theta \sim \mathrm{N}(0,1) \quad , \tag{A1}$$

where $b_i^* = b_i - \delta_e$. For persons in the second group, the ability distribution is given by $\theta \sim \mathrm{N}(\mu_2, \sigma_2^2)$. However, the IRF can be equivalently formulated as:

$$\mathrm{P}(X_i = 1|\theta) = \Psi(a_i(\theta - b_i^* - e_i^*)) \quad , \quad \theta \sim \mathrm{N}(\mu_2 - 2\delta_e, \sigma_2^2) \quad . \tag{A2}$$

Hence, the parameterization involving common item difficulties b_i, DIF effects e_i with $\mathrm{E}(e_i) = \delta_e \neq 0$, and $\theta \sim \mathrm{N}(\mu_2, \sigma_2^2)$ for persons in the second group can be equivalently parameterized with common item difficulties b_i^*, DIF effects e_i^* with $\mathrm{E}(e_i^*) = 0$, and $\theta \sim \mathrm{N}(\mu_2, \sigma_2^2)$ for persons in the second group. This demonstrates that group mean differences for abilities cannot be identified if the average DIF effect $\mathrm{E}(e_i)$ is assumed to differ from zero.

Appendix A.2. DIF Effects for Item Discriminations

Now, we show nonidentifiability for the DIF effects distribution of item discriminations. We assumed multiplicative DIF effects (see Equation (6)). It is shown that $\mathrm{E}(\log f_i)$ cannot be identified from data if the standard deviation σ_2 is simultaneously estimated. Assume $\mathrm{E}(\log f_i) = \delta_f \neq 0$. We can reparametrize the DIF effects as $f_i = \exp(\delta_f) f_i^*$ with $\mathrm{E}(\log f_i^*) = 0$. Then, we obtain the IRF for item i in the first group:

$$P(X_i = 1|\theta) = \Psi\left(\frac{a_i}{f_i}(\theta - b_i + e_i)\right) = \Psi\left(\frac{a_i^*}{f_i^*}(\theta - b_i + e_i)\right) \quad , \quad \theta \sim \mathrm{N}(0,1) \quad , \tag{A3}$$

where $a_i^* = a_i \exp(-\delta_f)$. The IRF for persons in the second group is given as:

$$P(X_i = 1|\theta) = \Psi(a_i^* f_i^* (\theta - b_i + e_i)) \quad , \quad \theta \sim \mathrm{N}(\mu_2, \sigma_2 \exp(2\delta_f)) \quad , \tag{A4}$$

Hence, the mean of DIF effects f_i (or $\log f_i$, respectively) must be fixed for identification reasons.

Appendix B. Proof of Proposition 1

Appendix B.1. Consistency of Additive DIF Effects f_i with Condition (I)

Define:

$$\hat{T}_g = \frac{1}{|I_0|} \sum_{i \in I_0} \log \hat{a}_{ig} \quad \text{for } g = 1,2 \quad . \tag{A5}$$

From Section 2.3.1 (Equation (9)), we know that $\hat{a}_{i1} = a_{i1} = a_i - f_i$ and $\hat{a}_{i2} = a_{i2}\sigma_2 = \sigma_2(a_i + f_i)$. Hence, we obtain:

$$\mathrm{E}(\hat{T}_2) = \frac{1}{|I_0|} \sum_{i \in I_0} \mathrm{E}(\log(\sigma_2(a_i + f_i))) = \log \sigma_2 + \frac{1}{|I_0|} \sum_{i \in I_0} \mathrm{E}(\log(a_i + f_i)) \quad \text{and} \tag{A6}$$

$$\mathrm{E}(\hat{T}_1) = \frac{1}{|I_0|} \sum_{i \in I_0} \mathrm{E}(\log(a_i - f_i)) \quad . \tag{A7}$$

Because the distribution of f_i is symmetric with $\mathrm{E}(f_i) = 0$, we obtain $\mathrm{E}(\log(a_i + f_i)) = \mathrm{E}(\log(a_i - f_i))$. Therefore, we obtain:

$$\mathrm{E}(\hat{T}_2 - \hat{T}_1) = \log \sigma_2 \quad . \tag{A8}$$

Trivially, it follows that $\hat{T}_2 - \hat{T}_1 \xrightarrow{p} \log \sigma_2$. Because $\hat{\sigma}_2 = \exp(\hat{T}_2 - \hat{T}_1)$, we obtain by the continuous mapping theorem ([147], p. 7):

$$\hat{\sigma}_2 = \exp(\hat{T}_2 - \hat{T}_1) \xrightarrow{p} \exp(\log \sigma_2) = \sigma_2 \quad . \tag{A9}$$

To derive the consistency of $\hat{\mu}_2$, define:

$$\hat{B}_g = \frac{1}{|I_0|} \sum_{i \in I_0} \hat{b}_{i2} \quad \text{for } g = 1, 2 \quad . \tag{A10}$$

From Equation (9) in Section 2.3.1, it follows that $\hat{b}_{i1} = b_i - e_i$ and $\hat{b}_{i2} = \sigma_2^{-1}(b_i + e_i - \mu_2)$. Thus,

$$\mathrm{E}(\hat{B}_2) = \frac{\mu_2}{\sigma_2} + \frac{1}{\sigma_2} \frac{1}{|I_0|} \sum_{i \in I_0} b_i \quad \text{and} \tag{A11}$$

$$\mathrm{E}(\hat{B}_1) = \frac{1}{|I_0|} \sum_{i \in I_0} b_i \quad . \tag{A12}$$

We can rewrite:

$$\hat{\mu}_2 = -\hat{\sigma}_2 \hat{B}_2 + \hat{B}_1 \quad . \tag{A13}$$

Assume the existence of $B = \lim_{|I_0| \to \infty} \frac{1}{|I_0|} \sum_{i \in I_0} b_i$. It holds that $\hat{B}_2 \xrightarrow{p} \mu_2/\sigma_2 + B/\sigma_2$ and $\hat{B}_1 \xrightarrow{p} B$. Hence, we obtain from (A13):

$$\hat{\mu}_2 = -\hat{\sigma}_2 \hat{B}_2 + \hat{B}_1 \xrightarrow{p} -\sigma_2(\mu_2/\sigma_2 + B/\sigma_2) + B = \mu_2 \quad . \tag{A14}$$

Appendix B.2. Consistency for Multiplicative DIF Effects f_i with Condition (II)

From Section 2.3.1 (Equation (9)), we know that $\hat{a}_{i1} = a_{i1} = a_i/f_i$ and $\hat{a}_{i2} = a_{i2}\sigma_2 = \sigma_2 a_i f_i$. Then, we obtain:

$$\hat{T}_2 - \hat{T}_1 = \log \sigma_2 + \frac{1}{|I_0|} \sum_{i \in I_0} \log f_i \quad . \tag{A15}$$

With $\mathrm{E}(\log f_i) = 0$, we obtain from (A15):

$$\mathrm{E}(\hat{T}_2 - \hat{T}_1) = \log \sigma_2 \quad . \tag{A16}$$

This proves $\hat{\sigma}_2 \xrightarrow{p} \sigma_2$ using the same reasoning as in (A9).

The derivations for $\hat{B}_g$ ($g = 1, 2$) are the same as in Appendix B.1. Due to the consistency of $\hat{\sigma}_2$, we also obtain the consistency of $\hat{\mu}_2$ as in (A14).

Appendix C. Proof of Proposition 2

Appendix C.1. Consistency for Additive DIF Effects f_i with Condition (I)

Define:

$$\hat{U}_g = \frac{1}{|I_0|} \sum_{i \in I_0} \hat{a}_{ig} \quad \text{for } g = 1, 2 \quad . \tag{A17}$$

The expected values for $\hat{U}_1$ and $\hat{U}_2$ are given in as:

$$\mathrm{E}(\hat{U}_2) = \frac{1}{|I_0|}\sum_{i\in I_0} \mathrm{E}((\sigma_2(a_i+f_i))) = \sigma_2 \frac{1}{|I_0|}\sum_{i\in I_0} a_i \quad \text{and} \tag{A18}$$

$$\mathrm{E}(\hat{U}_1) = \frac{1}{|I_0|}\sum_{i\in I_0} a_i \quad . \tag{A19}$$

Using the notation $A = \lim_{|I_0|\to\infty} \frac{1}{|I_0|}\sum_{i\in I_0} a_i$, we obtain $\hat{U}_1 \xrightarrow{p} A$ and $\hat{U}_2 \xrightarrow{p} \sigma_2 A$. By the continuous mapping theorem ([147], p. 7), we obtain:

$$\hat{\sigma}_2 = \frac{\hat{U}_2}{\hat{U}_1} \xrightarrow{p} \sigma_2 \quad . \tag{A20}$$

The steps for deriving the consistency of $\hat{\mu}_2$ are the same as in Appendix B.1.

Appendix C.2. Consistency for Multiplicative DIF Effects f_i with Condition (IO)

For multiplicative DIF effects, we have:

$$\hat{U}_2 = \frac{1}{|I_0|}\sum_{i\in I_0} \hat{a}_i\sigma_2 f_i = \sigma_2 \frac{1}{|I_0|}\sum_{i\in I_0} \hat{a}_i \exp(\log f_i) \quad \text{and} \tag{A21}$$

$$\hat{U}_1 = \frac{1}{|I_0|}\sum_{i\in I_0} \hat{a}_i / f_i = \sigma_2 \frac{1}{|I_0|}\sum_{i\in I_0} \hat{a}_i \exp(-\log f_i) \quad . \tag{A22}$$

Because f_i has a symmetric distribution, it follows that $\alpha \equiv \mathrm{E}(\exp(\log f_i)) = \mathrm{E}(\exp(-\log f_i))$. Then, we have:

$$\mathrm{E}(\hat{U}_2) = \sigma_2 \alpha \frac{1}{|I_0|}\sum_{i\in I_0} \hat{a}_i \quad \text{and} \quad \mathrm{E}(\hat{U}_1) = \alpha \frac{1}{|I_0|}\sum_{i\in I_0} \hat{a}_i \quad . \tag{A23}$$

Assuming the existence of $A = \lim_{|I_0|\to\infty} \frac{1}{|I_0|}\sum_{i\in I_0} \hat{a}_i$, one obtains $\hat{U}_2 \xrightarrow{p} \sigma_2\alpha A$ and $\hat{U}_1 \xrightarrow{p} \alpha A$. By the continuous mapping theorem ([147], p. 7), we obtain:

$$\hat{\sigma}_2 = \frac{\hat{U}_2}{\hat{U}_1} \xrightarrow{p} \frac{\sigma_2 \alpha A}{\alpha A} = \sigma_2 \quad . \tag{A24}$$

As in Appendix C.1, the derivation for the consistency of $\hat{\mu}_2$ does not require new calculations.

Appendix D. Estimates in Haberman Linking

For the HAB linking with logarithmized item loadings, the linking function for the standard deviation is given as:

$$\begin{aligned} H_{1,\log}(\sigma_2, \boldsymbol{a}) &= \sum_{i\in I_0\cup I_1} (\log \hat{a}_{i1} - \log a_i)^2 \\ &\quad + \sum_{i\in I_0\cup I_2} (\log \hat{a}_{i2} - \log a_i - \log \sigma_2)^2 \quad . \end{aligned} \tag{A25}$$

Taking the derivative of $H_{1,\log}$ with respect to $\log a_i$ for $i \in I_0$ and setting it to zero provide (up to multiplication with a constant):

$$(\log \hat{a}_{i1} - \log a_i) + (\log \hat{a}_{i2} - \log a_i - \log \sigma_2) = 0 \quad . \tag{A26}$$

Similarly, setting the derivative of $H_{1,\log}$ with respect to $\log \sigma_2$ to zero provides:

$$\sum_{i \in I_0} (\log \hat{a}_{i2} - \log a_i - \log \sigma_2) = 0 \tag{A27}$$

Summing the equations of all $|I_0|$ items of (A26) and subtracting (A27) provide:

$$\sum_{i \in I_0} (\log \hat{a}_{i1} - \log a_i) = 0 \quad . \tag{A28}$$

With an estimate $\log \hat{\sigma}_2$ of $\log \sigma_2$, we obtain from (A26):

$$\log \hat{a}_i = \frac{1}{2}(\log \hat{a}_{i1} + \log \hat{a}_{i2} - \log \hat{\sigma}_2) \quad . \tag{A29}$$

Plugging (A29) into (A27) provides:

$$\log \hat{\sigma}_2 = \frac{1}{|I_0|} \sum_{i \in I_0} \log \hat{a}_{i2} - \frac{1}{|I_0|} \sum_{i \in I_0} \log \hat{a}_{i1} \quad . \tag{A30}$$

Hence, it holds that:

$$\hat{\sigma}_2 = \exp\left(\frac{1}{|I_0|} \sum_{i \in I_0} \log \hat{a}_{i2} - \frac{1}{|I_0|} \sum_{i \in I_0} \log \hat{a}_{i1} \right) \quad . \tag{A31}$$

For Haberman linking with untransformed item discriminations (HAB-nolog), the linking function for the standard deviation of the second group is given as:

$$H_{1,\text{nolog}}(\sigma_2, \boldsymbol{a}) = \sum_{i \in I_0 \cup I_1} (\hat{a}_{i1} - a_i - 1)^2 + \sum_{i \in I_0 \cup I_2} (\hat{a}_{i2} - a_i - \sigma_2)^2 \quad . \tag{A32}$$

Taking the derivative of $H_{1,\text{nolog}}$ with respect to a_i for $i \in I_0$ and setting it to zero provide:

$$(\hat{a}_{i1} - a_i - 1) + (\hat{a}_{i2} - a_i - \sigma_2) = 0 \quad . \tag{A33}$$

Estimates $\hat{a}_i$ are then given by:

$$\hat{a}_i = \frac{1}{2}(\hat{a}_{i1} + \hat{a}_{i2} - (1 + \sigma_2)) \quad . \tag{A34}$$

Setting the derivative of $H_{1,\text{nolog}}$ with respect to σ_2 to zero provides:

$$\sum_{i \in I_0} (\hat{a}_{i2} - a_i - \sigma_2) = 0 \quad . \tag{A35}$$

Substituting (A34) into (A35) provides:

$$\hat{\sigma}_2 = 1 + \frac{1}{|I_0|} \sum_{i \in I_0} \hat{a}_{i2} - \frac{1}{|I_0|} \sum_{i \in I_0} \hat{a}_{i1} \quad . \tag{A36}$$

Appendix E. Estimates in Invariance Alignment

We now derive closed expressions for the estimates from IA2 (see Section 3.4). To estimate σ_2, define:

$$f(\sigma_2) = \sum_{i \in I_0} \left(\hat{a}_{i1} - \frac{\hat{a}_{i2}}{\sigma_2} \right)^2 \quad . \tag{A37}$$

Then, it follows that:

$$f(\sigma_2) = \sum_{i \in I_0} \left(\hat{a}_{i1}^2 + \frac{\hat{a}_{i2}^2}{\sigma_2^2} - 2\frac{\hat{a}_{i1}\hat{a}_{i2}}{\sigma_2} \quad . \right) \tag{A38}$$

The derivative of f in (A38) is:

$$\frac{\partial f}{\partial \sigma_2} = \sum_{i \in I_0} \left(-2\frac{\hat{a}_{i2}^2}{\sigma_2^3} + 2\frac{\hat{a}_{i1}\hat{a}_{i2}}{\sigma_2^2} \right) \quad . \tag{A39}$$

To find the minimum, setting the derivative to zero provides:

$$\sum_{i \in I_0} \left(-\hat{a}_{i2}^2 + \hat{a}_{i1}\hat{a}_{i2}\sigma_2 \right) = 0 \tag{A40}$$

and we have:

$$\hat{\sigma}_2 = \frac{\sum_{i \in I_0} \hat{a}_{i2}^2}{\sum_{i \in I_0} \hat{a}_{i1}\hat{a}_{i2}} \quad . \tag{A41}$$

To estimate μ_2, define:

$$g(\mu_2) = \sum_{i \in I_0} \left(\hat{\nu}_{i1} - \hat{\nu}_{i2} + \mu_2 \frac{\hat{a}_{i2}}{\hat{\sigma}_2} \right)^2 \quad . \tag{A42}$$

We obtain:

$$g(\mu_2) = \sum_{i \in I_0} \left((\hat{\nu}_{i1} - \hat{\nu}_{i2})^2 + \mu_2^2 \frac{\hat{a}_{i2}^2}{\hat{\sigma}_2^2} - 2\mu_2 \frac{\hat{a}_{i2}}{\hat{\sigma}_2}(\hat{\nu}_{i1} - \hat{\nu}_{i2}) \right) \quad . \tag{A43}$$

Setting the derivative of g with respect to μ_2 to zero provides:

$$\sum_{i \in I_0} \left(2\mu_2 \frac{\hat{a}_{i2}^2}{\hat{\sigma}_2^2} - 2\frac{\hat{a}_{i2}}{\hat{\sigma}_2}(\hat{\nu}_{i1} - \hat{\nu}_{i2}) \right) = 0 \quad . \tag{A44}$$

Then, we obtain using (A41):

$$\hat{\mu}_2 = \frac{\sum_{i \in I_0} \frac{\hat{a}_{i2}}{\hat{\sigma}_2}(\hat{\nu}_{i1} - \hat{\nu}_{i2})}{\sum_{i \in I_0} \frac{\hat{a}_{i2}^2}{\hat{\sigma}_2^2}} = \hat{\sigma}_2 \frac{\sum_{i \in I_0} \hat{a}_{i2}(\hat{\nu}_{i1} - \hat{\nu}_{i2})}{\sum_{i \in I_0} \hat{a}_{i2}^2} = \frac{\sum_{i \in I_0} \hat{a}_{i2}(\hat{\nu}_{i1} - \hat{\nu}_{i2})}{\sum_{i \in I_0} \hat{a}_{i1}\hat{a}_{i2}} \quad . \tag{A45}$$

Appendix F. Item Parameters Used in the Simulation Study

In Table A1, the item parameters used in the simulation study are shown. Item discriminations a_i had a mean of 1.00 (SD = 0.28, Min = 0.50, Max = 1.42), and item difficulties b_i had an average of 0.00 (SD = 1.00, Min = −1.62, Max = 1.39)

Table A1. Item parameters used in the simulation study.

Item	a_i	b_i
1	0.95	−0.97
2	0.88	0.59
3	0.75	0.75
4	1.29	−0.79
5	1.28	1.23
6	1.29	−1.10
7	1.25	−0.67
8	0.97	0.20
9	0.73	1.26
10	1.27	0.05
11	1.42	1.22
12	0.75	−0.01
13	0.50	0.20
14	0.81	1.39
15	1.12	0.61
16	0.78	−1.00
17	1.30	−1.58
18	0.70	−1.62
19	1.29	1.06
20	0.74	−0.81

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

References

1. Cai, L.; Choi, K.; Hansen, M.; Harrell, L. Item response theory. *Annu. Rev. Stat. Appl.* **2016**, *3*, 297–321. [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. 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.
4. Battauz, M. Regularized estimation of the four-parameter logistic model. *Psych* **2020**, *2*, 269–278. [CrossRef]
5. Bürkner, P.C. Analysing standard progressive matrices (SPM-LS) with Bayesian item response models. *J. Intell.* **2020**, *8*, 5. [CrossRef] [PubMed]
6. Chang, H.H.; Wang, C.; Zhang, S. Statistical applications in educational measurement. *Annu. Rev. Stat. Appl.* **2021**, *8*, 439–461. [CrossRef]
7. Genge, E. LC and LC-IRT models in the identification of Polish households with similar perception of financial position. *Sustainability* **2021**, *13*, 4130. [CrossRef]
8. Jefmański, B.; Sagan, A. Item response theory models for the fuzzy TOPSIS in the analysis of survey data. *Symmetry* **2021**, *13*, 223. [CrossRef]
9. Karwowski, M.; Milerski, B. Who supports Polish educational reforms? Exploring actors' and observers' attitudes. *Educ. Sci.* **2021**, *11*, 120. [CrossRef]
10. Medová, J.; Páleníková, K.; Rybanskỳ, L.; Naštická, Z. Undergraduate students' solutions of modeling problems in algorithmic graph theory. *Mathematics* **2019**, *7*, 572. [CrossRef]
11. Mousavi, A.; Cui, Y. The effect of person misfit on item parameter estimation and classification accuracy: A simulation study. *Educ. Sci.* **2020**, *10*, 324. [CrossRef]
12. Palma-Vasquez, C.; Carrasco, D.; Hernando-Rodriguez, J.C. Mental health of teachers who have teleworked due to COVID-19. *Eur. J. Investig. Health Psychol. Educ.* **2021**, *11*, 515–528. [CrossRef]
13. Storme, M.; Myszkowski, N.; Baron, S.; Bernard, D. Same test, better scores: Boosting the reliability of short online intelligence recruitment tests with nested logit item response theory models. *J. Intell.* **2019**, *7*, 17. [CrossRef]
14. Tsutsumi, E.; Kinoshita, R.; Ueno, M. Deep item response theory as a novel test theory based on deep learning. *Electronics* **2021**, *10*, 1020. [CrossRef]
15. 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.
16. Kolen, M.J.; Brennan, R.L. *Test Equating, Scaling, and Linking*; Springer: New York, NY, USA, 2014. [CrossRef]
17. Lietz, P.; Cresswell, J.C.; Rust, K.F.; Adams, R.J. (Eds.) *Implementation of Large-Scale Education Assessments*; Wiley: New York, NY, USA, 2017. [CrossRef]
18. Maehler, D.B.; Rammstedt, B. (Eds.) *Large-Scale Cognitive Assessment*; Springer: New York, NY, USA, 2020. [CrossRef]

19. Wagemaker, H. International large-scale assessments: From research to policy. In *A Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*; Rutkowski, L., von Davier, M., Rutkowski, D., Eds.; Chapman Hall/CRC Press: London, UK, 2014; pp. 11–36. [CrossRef]
20. van der Linden, W.J. Unidimensional Logistic Response Models. In *Handbook of Item Response Theory, Volume One: Models*; CRC Press: Boca Raton, FL, USA, 2016; pp. 11–30. [CrossRef]
21. Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*; Danish Institute for Educational Research: Copenhagen, Denmark, 1960.
22. von Davier, A.A.; Carstensen, C.H.; von Davier, M. *Linking Competencies in Educational Settings and Measuring Growth*; (Research Report No. RR-06-12); Educational Testing Service: Princeton, NJ, USA, 2006. [CrossRef]
23. von Davier, A.A.; Holland, P.W.; Thayer, D.T. *The Kernel Method of Test Equating*; Springer: New York, NY, USA, 2004. [CrossRef]
24. Bolsinova, M.; Maris, G. Can IRT solve the missing data problem in test equating? *Front. Psychol.* **2016**, *6*, 1956. [CrossRef] [PubMed]
25. Liou, M.; Cheng, P.E. Equipercentile equating via data-imputation techniques. *Psychometrika* **1995**, *60*, 119–136. [CrossRef]
26. Meredith, W. Measurement invariance, factor analysis and factorial invariance. *Psychometrika* **1993**, *58*, 525–543. [CrossRef]
27. Millsap, R.E. *Statistical Approaches to Measurement Invariance*; Routledge: New York, NY, USA, 2011. [CrossRef]
28. van de Vijver, F.J.R. (Ed.) *Invariance Analyses in Large-Scale Studies*; OECD: Paris, France, 2019. [CrossRef]
29. Mellenbergh, G.J. Item bias and item response theory. *Int. J. Educ. Res.* **1989**, *13*, 127–143. [CrossRef]
30. Millsap, R.E.; Everson, H.T. Methodology review: Statistical approaches for assessing measurement bias. *Appl. Psychol. Meas.* **1993**, *17*, 297–334. [CrossRef]
31. Osterlind, S.J.; Everson, H.T. *Differential Item Functioning*; Sage Publications: New York, NY, USA, 2009. [CrossRef]
32. Penfield, R.D.; Camilli, G. Differential item functioning and item bias. In *Handbook of Statistics, Volume 26: Psychometrics*; Rao, C.R., Sinharay, S., Eds.; Elesvier: Amsterdam, The Netherlands, 2007; pp. 125–167. [CrossRef]
33. Uyar, S.; Kelecioglu, H.; Dogan, N. Comparing differential item functioning based on manifest groups and latent classes. *Educ. Sci. Theory Pract.* **2017**, *17*, 1977–2000. [CrossRef]
34. Lee, S.Y.; Hong, A.J. Psychometric investigation of the cultural intelligence scale using the Rasch measurement model in South Korea. *Sustainability* **2021**, *13*, 3139. [CrossRef]
35. Mylona, I.; Aletras, V.; Ziakas, N.; Tsinopoulos, I. Rasch validation of the VF-14 scale of vision-specific functioning in Greek patients. *Int. J. Environ. Res. Public Health* **2021**, *18*, 4254. [CrossRef]
36. Pichette, F.; Béland, S.; Leśniewska, J. Detection of gender-biased items in the peabody picture vocabulary test. *Languages* **2019**, *4*, 27. [CrossRef]
37. Shibaev, V.; Grigoriev, A.; Valueva, E.; Karlin, A. Differential item functioning on Raven's SPM+ amongst two convenience samples of Yakuts and Russian. *Psych* **2020**, *2*, 44–51. [CrossRef]
38. Silvia, P.J.; Rodriguez, R.M. Time to renovate the humor styles questionnaire? An item response theory analysis of the HSQ. *Behav. Sci.* **2020**, *10*, 173. [CrossRef]
39. Hanson, B.A. Uniform DIF and DIF defined by differences in item response functions. *J. Educ. Behav. Stat.* **1998**, *23*, 244–253. [CrossRef]
40. Teresi, J.A.; Ramirez, M.; Lai, J.S.; Silver, S. Occurrences and sources of differential item functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life and general health. *Psychol. Sci.* **2008**, *50*, 538–612.
41. Buchholz, J.; Hartig, J. Measurement invariance testing in questionnaires: A comparison of three multigroup-CFA and IRT-based approaches. *Psych. Test Assess. Model.* **2020**, *62*, 29–53.
42. Chalmers, R.P. Extended mixed-effects item response models with the MH-RM algorithm. *J. Educ. Meas.* **2015**, *52*, 200–222. [CrossRef]
43. De Boeck, P.; Wilson, M. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*; Springer: New York, NY, USA, 2004. [CrossRef]
44. De Boeck, P. Random item IRT models. *Psychometrika* **2008**, *73*, 533–559. [CrossRef]
45. de Jong, M.G.; Steenkamp, J.B.E.M.; Fox, J.P. Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *J. Consum. Res.* **2007**, *34*, 260–278. [CrossRef]
46. Doran, H.; Bates, D.; Bliese, P.; Dowling, M. Estimating the multilevel Rasch model: With the lme4 package. *J. Stat. Softw.* **2007**, *20*, 1–18. [CrossRef]
47. Fox, J.P.; Verhagen, A.J. Random item effects modeling for cross-national survey data. In *Cross-Cultural Analysis: Methods and Applications*; Davidov, E., Schmidt, P., Billiet, J., Eds.; Routledge: London, UK, 2010; pp. 461–482. [CrossRef]
48. Van den Noortgate, W.; De Boeck, P. Assessing and explaining differential item functioning using logistic mixed models. *J. Educ. Behav. Stat.* **2005**, *30*, 443–464. [CrossRef]
49. Muthén, B.; Asparouhov, T. Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychol. Methods* **2012**, *17*, 313–335. [CrossRef] [PubMed]
50. van de Schoot, R.; Kluytmans, A.; Tummers, L.; Lugtig, P.; Hox, J.; Muthén, B. Facing off with scylla and charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Front. Psychol.* **2013**, *4*, 770. [CrossRef] [PubMed]

51. Bechger, T.M.; Maris, G. A statistical test for differential item pair functioning. *Psychometrika* **2015**, *80*, 317–340. [CrossRef]
52. Camilli, G. The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In *Differential Item Functioning: Theory and Practice*; Holland, P.W., Wainer, H., Eds.; Erlbaum: Hillsdale, NJ, USA, 1993; pp. 397–417.
53. Doebler, A. Looking at DIF from a new perspective: A structure-based approach acknowledging inherent indefinability. *Appl. Psychol. Meas.* **2019**, *43*, 303–321. [CrossRef] [PubMed]
54. 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. *Psych. Test Assess. Model.* **2020**, *62*, 233–279.
55. Frederickx, S.; Tuerlinckx, F.; De Boeck, P.; Magis, D. RIM: A random item mixture model to detect differential item functioning. *J. Educ. Meas.* **2010**, *47*, 432–457. [CrossRef]
56. Byrne, B.M.; Shavelson, R.J.; Muthén, B. Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychol. Bull.* **1989**, *105*, 456–466. [CrossRef]
57. Magis, D.; Tuerlinckx, F.; De Boeck, P. Detection of differential item functioning using the lasso approach. *J. Educ. Behav. Stat.* **2015**, *40*, 111–135. [CrossRef]
58. Tutz, G.; Schauberger, G. A penalty approach to differential item functioning in Rasch models. *Psychometrika* **2015**, *80*, 21–43. [CrossRef]
59. Soares, T.M.; Gonçalves, F.B.; Gamerman, D. An integrated Bayesian model for DIF analysis. *J. Educ. Behav. Stat.* **2009**, *34*, 348–377. [CrossRef]
60. Kopf, J.; Zeileis, A.; Strobl, C. Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educ. Psychol. Meas.* **2015**, *75*, 22–56. [CrossRef] [PubMed]
61. Magis, D.; Béland, S.; Tuerlinckx, F.; De Boeck, P. A general framework and an R package for the detection of dichotomous differential item functioning. *Behav. Res. Methods* **2010**, *42*, 847–862. [CrossRef]
62. Teresi, J.A.; Ramirez, M.; Jones, R.N.; Choi, S.; Crane, P.K. Modifying measures based on differential item functioning (DIF) impact analyses. *J. Aging Health* **2012**, *24*, 1044–1076. [CrossRef]
63. DeMars, C.E. Alignment as an alternative to anchor purification in DIF analyses. *Struct. Equ. Model.* **2020**, *27*, 56–72. [CrossRef]
64. Lai, M.H.C.; Liu, Y.; Tse, W.W.Y. Adjusting for partial invariance in latent parameter estimation: Comparing forward specification search and approximate invariance methods. *Behav. Res. Methods* **2021**. [CrossRef] [PubMed]
65. 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.* **2021**. [CrossRef]
66. 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]
67. Oliveri, M.E.; von Davier, M. Investigation of model fit and score scale comparability in international assessments. *Psych. Test Assess. Model.* **2011**, *53*, 315–333.
68. Oliveri, M.E.; von Davier, M. Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *Int. J. Test.* **2014**, *14*, 1–21. [CrossRef]
69. OECD. *PISA 2015. Technical Report*; OECD: Paris, France, 2017.
70. von Davier, M.; Yamamoto, K.; Shin, H.J.; Chen, H.; Khorramdel, L.; Weeks, J.; Davis, S.; Kong, N.; Kandathil, M. Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assess. Educ.* **2019**, *26*, 466–488. [CrossRef]
71. Robitzsch, A. L_p loss functions in invariance alignment and Haberman linking with few or many groups. *Stats* **2020**, *3*, 246–283. [CrossRef]
72. 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]
73. El Masri, Y.H.; Andrich, D. The trade-off between model fit, invariance, and validity: The case of PISA science assessments. *Appl. Meas. Educ.* **2020**, *33*, 174–188. [CrossRef]
74. Shealy, R.; Stout, W. A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika* **1993**, *58*, 159–194. [CrossRef]
75. Zwitser, R.J.; Glaser, S.S.F.; Maris, G. Monitoring countries in a changing world: A new look at DIF in international surveys. *Psychometrika* **2017**, *82*, 210–232. [CrossRef]
76. 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]
77. Bock, R.D.; Aitkin, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **1981**, *46*, 443–459. [CrossRef]
78. von Davier, M.; Sinharay, S. Analytics in international large-scale assessments: Item response theory and population models. In *A Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*; Rutkowski, L., von Davier, M., Rutkowski, D., Eds.; Chapman Hall/CRC Press: London, UK, 2014; pp. 155–174. [CrossRef]
79. Robitzsch, A. A note on a computationally efficient implementation of the EM algorithm in item response models. *Quant. Comput. Methods Behav. Sci.* **2021**, *1*, e3783. [CrossRef]

80. González, J.; Wiberg, M. *Applying Test Equating Methods. Using R*; Springer: New York, NY, USA, 2017. [CrossRef]
81. Lee, W.C.; Lee, G. IRT linking and equating. In *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test*; Irwing, P., Booth, T., Hughes, D.J., Eds.; Wiley: New York, NY, USA, 2018; pp. 639–673. [CrossRef]
82. Sansivieri, V.; Wiberg, M.; Matteucci, M. A review of test equating methods with a special focus on IRT-based approaches. *Statistica* **2017**, *77*, 329–352. [CrossRef]
83. Haberman, S.J. *Linking Parameter Estimates Derived from an Item Response Model through Separate Calibrations*; (Research Report No. RR-09-40); Educational Testing Service: Princeton, NJ, USA, 2009. [CrossRef]
84. Battauz, M. Multiple equating of separate IRT calibrations. *Psychometrika* **2017**, *82*, 610–636. [CrossRef] [PubMed]
85. Asparouhov, T.; Muthén, B. Multiple-group factor analysis alignment. *Struct. Equ. Model.* **2014**, *21*, 495–508. [CrossRef]
86. Muthén, B.; Asparouhov, T. IRT studies of many groups: The alignment method. *Front. Psychol.* **2014**, *5*, 978. [CrossRef]
87. Muthén, B.; Asparouhov, T. Recent methods for the study of measurement invariance with many groups: Alignment and random effects. *Sociol. Methods Res.* **2018**, *47*, 637–664. [CrossRef]
88. Pokropek, A.; Davidov, E.; Schmidt, P. A Monte Carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. *Struct. Equ. Model.* **2019**, *26*, 724–744. [CrossRef]
89. Pokropek, A.; Lüdtke, O.; Robitzsch, A. An extension of the invariance alignment method for scale linking. *Psych. Test Assess. Model.* **2020**, *62*, 303–334.
90. Haebara, T. Equating logistic ability scales by a weighted least squares method. *Jpn. Psychol. Res.* **1980**, *22*, 144–149. [CrossRef]
91. Kim, S.; Kolen, M.J. Effects on scale linking of different definitions of criterion functions for the IRT characteristic curve methods. *J. Educ. Behav. Stat.* **2007**, *32*, 371–397. [CrossRef]
92. Weeks, J.P. plink: An R package for linking mixed-format tests using IRT-based methods. *J. Stat. Softw.* **2010**, *35*, 1–33. [CrossRef]
93. Arai, S.; Mayekawa, S.i. A comparison of equating methods and linking designs for developing an item pool under item response theory. *Behaviormetrika* **2011**, *38*, 1–16. [CrossRef]
94. Robitzsch, A. Robust Haebara linking for many groups: Performance in the case of uniform DIF. *Psych* **2020**, *2*, 155–173. [CrossRef]
95. OECD. *PISA 2006. Technical Report*; OECD: Paris, France, 2009.
96. 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: Newton, MA, USA, 2017.
97. Foy, P.; Yin, L. Scaling the TIMSS 2015 achievement data. In *Methods and Procedures in TIMSS 2015*; Martin, M.O., Mullis, I.V., Hooper, M., Eds.; IEA: Newton, MA, USA, 2016.
98. Foy, P.; Fishbein, B.; von Davier, M.; Yin, L. Implementing the TIMSS 2019 scaling methodology. In *Methods and Procedures: TIMSS 2019 Technical Report*; Martin, M.O., von Davier, M., Mullis, I.V., Eds.; IEA: Newton, MA, USA, 2020.
99. Gebhardt, E.; Adams, R.J. The influence of equating methodology on reported trends in PISA. *J. Appl. Meas.* **2007**, *8*, 305–322.
100. Fishbein, B.; Martin, M.O.; Mullis, I.V.S.; Foy, P. The TIMSS 2019 item equivalence study: Examining mode effects for computer-based assessment and implications for measuring trends. *Large-Scale Assess. Educ.* **2018**, *6*, 11. [CrossRef]
101. 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.
102. Kim, S.H.; Cohen, A.S. A comparison of linking and concurrent calibration under item response theory. *Appl. Psychol. Meas.* **1998**, *22*, 131–143. [CrossRef]
103. Hanson, B.A.; Béguin, A.A. Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Appl. Psychol. Meas.* **2002**, *26*, 3–24. [CrossRef]
104. Andersson, B. Asymptotic variance of linking coefficient estimators for polytomous IRT models. *Appl. Psychol. Meas.* **2018**, *42*, 192–205. [CrossRef]
105. Demirus, K.; Gelbal, S. The study of the effect of anchor items showing or not showing differantial item functioning to test equating using various methods. *J. Meas. Eval. Educ. Psychol.* **2016**, *7*, 182–201. [CrossRef]
106. Gübes, N.; Uyar, S. Comparing performance of different equating methods in presence and absence of DIF Items in anchor test. *Int. J. Progress. Educ.* **2020**, *16*, 111–122. [CrossRef]
107. He, Y.; Cui, Z. Evaluating robust scale transformation methods with multiple outlying common items under IRT true score equating. *Appl. Psychol. Meas.* **2020**, *44*, 296–310. [CrossRef]
108. Inal, H.; Anil, D. Investigation of group invariance in test equating under different simulation conditions. *Eurasian J. Educ. Res.*, *18*, 67–86. [CrossRef]
109. Kabasakal, K.A.; Kelecioğlu, H. Effect of differential item functioning on test equating. *Educ. Sci. Theory Pract.* **2015**, *15*, 1229–1246. [CrossRef]
110. Tulek, O.K.; Kose, I.A. Comparison of different forms of a test with or without items that exhibit DIF. *Eurasian J. Educ. Res.* **2019**, *19*, 167–182. [CrossRef]
111. Pohl, S.; Schulze, D. Assessing group comparisons or change over time under measurement non-invariance: The cluster approach for nonuniform DIF. *Psych. Test Assess. Model.* **2020**, *62*, 281–303.
112. Yurtçu, M.; Güzeller, C.O. Investigation of equating error in tests with differential item functioning. *Int. J. Assess. Tool. Educ.* **2018**, *5*, 50–57. [CrossRef]

113. R Core Team. *R: A Language and Environment for Statistical Computing*; R Core Team: Vienna, Austria, 2020. Available online: https://www.R-project.org/ (accessed on 20 August 2020).
114. Robitzsch, A.; Kiefer, T.; Wu, M. TAM: Test Analysis Modules; R Package Version 3.7-6; 2021. Available online: https://CRAN.R-project.org/package=TAM (accessed on 25 June 2021).
115. Robitzsch, A. Sirt: Supplementary Item Response Theory Models; R Package Version 3.9-4; 2020. Available online: https://CRAN.R-project.org/package=sirt (accessed on 17 February 2020).
116. Bates, D.; Mächler, M.; Bolker, B.; Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **2015**, *67*, 1–48. [CrossRef]
117. OECD. *PISA 2009. Technical Report*; OECD: Paris, France, 2012.
118. Falk, C.F.; Cai, L. Semiparametric item response functions in the context of guessing. *J. Educ. Meas.* **2016**, *53*, 229–247. [CrossRef]
119. Feuerstahler, L.M. Metric transformations and the filtered monotonic polynomial item response model. *Psychometrika* **2019**, *84*, 105–123. [CrossRef] [PubMed]
120. Feuerstahler, L. Flexible item response modeling in R with the flexmet package. *Psych* **2021**, *3*, 447–478. [CrossRef]
121. Ramsay, J.O.; Winsberg, S. Maximum marginal likelihood estimation for semiparametric item analysis. *Psychometrika* **1991**, *56*, 365–379. [CrossRef]
122. Rossi, N.; Wang, X.; Ramsay, J.O. Nonparametric item response function estimates with the EM algorithm. *J. Educ. Behav. Stat.* **2002**, *27*, 291–317. [CrossRef]
123. Anderson, D.; Kahn, J.D.; Tindal, G. Exploring the robustness of a unidimensional item response theory model with empirically multidimensional data. *Appl. Meas. Educ.* **2017**, *30*, 163–177. [CrossRef]
124. Martineau, J.A. Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *J. Educ. Behav. Stat.* **2006**, *31*, 35–62. [CrossRef]
125. Köhler, C.; Hartig, J. Practical significance of item misfit in educational assessments. *Appl. Psychol. Meas.* **2017**, *41*, 388–400. [CrossRef] [PubMed]
126. Sinharay, S.; Haberman, S.J. How often is the misfit of item response theory models practically significant? *Educ. Meas.* **2014**, *33*, 23–35. [CrossRef]
127. Zhao, Y.; Hambleton, R.K. Practical consequences of item response theory model misfit in the context of test equating with mixed-format test data. *Front. Psychol.* **2017**, *8*, 484. [CrossRef] [PubMed]
128. Bolt, D.M.; Deng, S.; Lee, S. IRT model misspecification and measurement of growth in vertical scaling. *J. Educ. Meas.* **2014**, *51*, 141–162. [CrossRef]
129. Guo, H.; Liu, J.; Dorans, N.; Feigenbaum, M. *Multiple Linking in Equating and Random Scale Drift*; (Research Report No. RR-11-46); Educational Testing Service: Princeton, NJ, USA, 2011. [CrossRef]
130. Puhan, G. Detecting and correcting scale drift in test equating: An illustration from a large scale testing program. *Appl. Meas. Educ.* **2008**, *22*, 79–103. [CrossRef]
131. Battauz, M. IRT test equating in complex linkage plans. *Psychometrika* **2013**, *78*, 464–480. [CrossRef] [PubMed]
132. Battauz, M. Factors affecting the variability of IRT equating coefficients. *Stat. Neerl.* **2015**, *69*, 85–101. [CrossRef]
133. Battauz, M. equateIRT: An R package for IRT test equating. *J. Stat. Softw.* **2015**, *68*, 1–22. [CrossRef]
134. Briggs, D.C.; Weeks, J.P. The sensitivity of value-added modeling to the creation of a vertical score scale. *Educ. Financ. Policy* **2009**, *4*, 384–414. [CrossRef]
135. Bjermo, J.; Miller, F. Efficient estimation of mean ability growth using vertical scaling. *Appl. Meas. Educ.* **2021**. [CrossRef]
136. Fischer, L.; Rohm, T.; Carstensen, C.H.; Gnambs, T. Linking of Rasch-scaled tests: Consequences of limited item pools and model misfit. *Front. Psychol.* **2021**, *12*, 633896. [CrossRef] [PubMed]
137. Pohl, S.; Haberkorn, K.; Carstensen, C.H. Measuring competencies across the lifespan-challenges of linking test scores. In *Dependent Data in Social Sciences Research*; Stemmler, M., von Eye, A., Wiedermann, W., Eds.; Springer: Cham, Switzerland, 2015; pp. 281–308. [CrossRef]
138. Tong, Y.; Kolen, M.J. Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Appl. Meas. Educ.* **2007**, *20*, 227–253. [CrossRef]
139. Barrett, M.D.; van der Linden, W.J. Estimating linking functions for response model parameters. *J. Educ. Behav. Stat.* **2019**, *44*, 180–209. [CrossRef]
140. 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]
141. Ogasawara, H. Standard errors of item response theory equating/linking by response function methods. *Appl. Psychol. Meas.* **2001**, *25*, 53–67. [CrossRef]
142. Haberman, S.J.; Lee, Y.H.; Qian, J. *Jackknifing Techniques for Evaluation of Equating Accuracy*; (Research Report No. RR-09-02); Educational Testing Service: Princeton, NJ, USA, 2009. [CrossRef]
143. Michaelides, M.P. A review of the effects on IRT item parameter estimates with a focus on misbehaving common items in test equating. *Front. Psychol.* **2010**, *1*, 167. [CrossRef]
144. Monseur, C.; Berezner, A. The computation of equating errors in international surveys in education. *J. Appl. Meas.* **2007**, *8*, 323–335. [PubMed]

145. Monseur, C.; Sibberns, H.; Hastedt, D. Linking errors in trend estimation for international surveys in education. *IERI Monogr. Ser.* **2008**, *1*, 113–122.
146. Xu, X.; von Davier, M. *Linking Errors in Trend Estimation in Large-Scale Surveys: A Case Study*; (Research Report No. RR-10-10); Educational Testing Service: Princeton, NJ, USA, 2010. [CrossRef]
147. Van der Vaart, A.W. *Asymptotic Statistics*; Cambridge University Press: Cambridge, UK, 1998; Volume 3. [CrossRef]