Psychological Test and Assessment Modeling, Volume 62, 2020 (2), 233-279

# A review of different scaling approaches under full invariance, partial invariance, and noninvariance for cross-sectional country comparisons in large-scale assessments

*Alexander Robitzsch*[1,2] *& Oliver Lüdtke*[1,2]

## Abstract

One of the primary goals of international large-scale assessments (ILSAs) in education is the comparison of country means in student achievement. The present article introduces a framework for discussing differential item functioning (DIF) for country comparisons in ILSAs. Three different linking methods are compared: concurrent calibration based on full invariance, concurrent calibration based on partial invariance using the MD or RMSD statistics, and separate calibration with subsequent nonrobust and robust linking approaches. Furthermore, we show analytically the bias in country means of different linking methods in the presence of DIF. In a simulation study, we show that partial invariance and robust linking approaches provide less biased country mean estimates than the full invariance approach in the case of biased items. Some guidelines are derived for the selection of cutoff values for the MD and RMSD statistics in the partial invariance approach.

Keywords: international large-scale assessments, linking, differential item functioning, multiple groups, RMSD statistic

[1] *Correspondence concerning this article should be addressed to:* Alexander Robitzsch, PhD, Leibniz Institute for Science and Mathematics Education (IPN), Olshausenstr. 62, 24118 Kiel, Germany; email: robitzsch@leibniz-ipn.de

[2] Centre for International Student Assessment, Germany

One of the major goals of international large-scale assessments (ILSAs) in education is the comparison of country means in student achievement. For example, beginning in 2000, every three years, the Programme for International Student Assessment (PISA) provides international comparisons of student performance for a large group of countries (72 countries in PISA 2015; OECD, 2017). These comparisons are reported for three content areas (reading, mathematics, and science; OECD, 2017) and receive considerable attention from educational researchers, policymakers, and the media. A major methodological challenge of these comparisons is that the items of the achievement tests can function differently (differential item functioning; DIF) in specific countries, and this can result in item parameters that are not invariant across countries. This could be the case, for example, if an item is relatively easier or more difficult for a specific country than at the international level (Camilli, 2006; Holland & Wainer, 1993). The existence and distribution of noninvariant item parameters across countries has been studied extensively in the ILSA literature (Kreiner & Christensen, 2014; Oliveri & von Davier, 2014, 2017), and it has been argued that ignoring country DIF in the calibration of item parameters has the potential to bias the estimation of country-specific means and standard deviations (Kankaras & Moors, 2014).

In the present article, we distinguish three different strategies for conducting cross-national comparisons in the presence of country DIF. In the first approach, DIF effects are completely ignored and common item parameters are estimated in a multiple-group item response theory (IRT) model. In this approach, country-specific item parameters are treated as if they were completely invariant across groups (*full invariance* approach) when estimating country-specific means and standard deviations. In the second approach, items with noninvariant parameters across countries are identified by using country-specific item fit statistics and cutoffs as screening criteria. In the current operational procedure of PISA, the mean deviation (MD) and root mean square deviation (RMSD) statistics are used to select items with country DIF (OECD, 2018; Oliveri & von Davier, 2011). Based on this screening process, country-specific means and standard deviations are estimated in a multiple-group IRT model in which the parameters of items with no country DIF are constrained to be equal across countries, and the parameters of items with country DIF are allowed to be country-specific (*partial invariance* approach). In the third approach, no invariance assumptions are made for the group-specific item parameters, and all country DIF effects are modeled when estimating country-specific means and standard deviations (*noninvariance* approach). More specifically, item parameters are calibrated separately and allowed to vary across groups. Linking methods for multiple groups (e.g., Haberman or Haebara linking) are then used to obtain country means by linking the group-specific item parameters on a common metric.

Although there is some evidence from secondary analyses of ILSAs that the results of country comparisons are quite robust against the choice of different statistical analysis models (Jerrim, Parker, Choi, Chmielewski, Sälzer, & Shure, 2018), simulation studies that compare the performance of the different approaches for treating country DIF are scarce. In the present study, we report the results of a comprehensive simulation study in which we manipulated the number of items, the sample sizes, and the amount and type of DIF effects (i.e., DIF variance and distribution) in order to investigate the accuracy of the three different approaches when comparing country means in the presence of DIF effects. Our main goal was to better understand the conditions under which the partial invariance approach – which is currently used in the PISA study – results in more accurate estimates of country means than the full invariance and noninvariance approaches. We focus on the 1PL model and address the extension to more complex IRT models (e.g., 2PL) in the Discussion section.

## Differential item functioning for multiple groups

In the following, we discuss the concept of differential item functioning (DIF; Holland & Wainer, 1993; Millsap, 2011) for multiple countries (also referred to as groups). Let $g = 1,\ldots,G$ denote a collection of groups to which a test consisting of $I$ items is administered. It is assumed that a unidimensional item response model holds in each group with group-specific item response functions (IRF) $P_{ig}(\theta)$, indicating the probability of a correct item response conditional on ability $\theta$. For the sake of simplicity, we only consider the 1PL model (i.e., the Rasch model; Rasch, 1960), which is given as

$$P_{ig}(\theta) = \Psi\left(\theta - b_i - e_{ig}\right),\ \theta \sim \mathrm{N}\left(\mu_g, \sigma_g^2\right)\ , \tag{1}$$

where $b_i$ is the common item difficulty for item $i$ ($i = 1,\ldots,I$) and $e_{ig}$ are group-specific item difficulty deviations with nonzero values indicating differential item functioning; $\Psi$ denotes the logistic distribution function, and it is assumed that the abilities are normally distributed in group $g$. It is well known that the model given in Equation 1 is not identified and identification constraints among item parameters are needed to estimate the means $\mu_g$ and the standard deviations $\sigma_g$ of the ability distributions for the $G$ groups and the DIF effects $e_{ig}$ (Bechger & Maris, 2015; Doebler, 2019; Soares, Goncalvez, & Gamerman, 2009; Strobl, Kopf, Hartmann, & Zeileis, 2018). In order to illustrate this identification problem, it is instructive to reparametrize Equation 1 as follows

$$P_{ig}(\theta) = \Psi(\theta - \underbrace{(b_i - \mu_g + e_{ig})}_{=b_{ig}}) = \Psi(\theta - b_{ig})\ ,\ \theta \sim \mathrm{N}\left(0, \sigma_g^2\right)\ , \tag{2}$$

where the group-specific item parameters $b_{ig}$ are identified without constraints and are composed into the common item difficulties $b_i$ of item $i$, the means of the ability distributions $\mu_g$, and the DIF effects $e_{ig}$. However, given the $I \cdot G$ group-specific item parameters $b_{ig}$, further constraints on the $I \cdot G$ DIF effects $e_{ig}$ are needed to identify the $G$ group means $\mu_g$[3].To resolve the identification issue, the items for each group are partitioned into two distinct sets. More specifically, it is assumed that for each group $g$ a subset of *anchor items* $\mathcal{J}_{A,g} \subset \mathcal{J} = \{1,\ldots,I\}$ exists such that $\sum_{i \in \mathcal{J}_{A,g}} e_{ig} = 0$ for $g = 1,\ldots,G$ (also

[3] The identification issue occurs because $\mu_g$ is only identified up to group-specific constants $c_g$. Identified parameters can be written as $b_{ig} = b_i - \mu_g + e_{ig} = b_i - (\mu_g + c_g) + (e_{ig} + c_g)$, which shows that group means $\mu_g$ and DIF effects $e_{ig}$ cannot be simultaneously computed without further constraints.

referred to as equal mean difficulty constraint; Kopf, Zeileis, & Strobl, 2015). The set of *biased items* is defined as $\mathcal{J}_{B,g} = \mathcal{J} \setminus \mathcal{J}_{A,g}$ (Camilli, 2006). It is important to emphasize that DIF effects of biased items can differ from zero on average, i.e., $\sum_{i \in \mathcal{J}_{B,g}} e_{ig} \neq 0$. In many simulation studies that investigate the consequences of DIF, the DIF effects $e_{ig}$ of anchor items are chosen to be "small" (or even zero) compared to DIF effects of biased items. In this case, it is plausible to consider DIF effects of biased items as outliers (De Boeck, 2008; Magis & De Boeck, 2011). Furthermore, we denote a test to have *balanced DIF* if for all groups, the DIF effects sum to zero (within each group $g$). Because the DIF effects of anchor items sum to zero by definition, balanced DIF is equivalent to the condition that DIF effects of biased items sum to zero (i.e., $\sum_{i \in \mathcal{J}_{B,g}} e_{ig} = 0$ for $g = 1,\ldots,G$). A test has *unbalanced DIF* if there exists at least one group $g$ for which $\sum_{i \in \mathcal{J}_{B,g}} e_{ig} \neq 0$ holds.

One central argument in the DIF literature is that items with DIF effects bias the estimated group means and should, therefore, not be included in group comparisons (e.g., OECD, 2017, for arguments in PISA). Typically, biased estimates of group means can be expected in the case of unbalanced DIF. The reasoning behind this argument – and the DIF concept – is illustrated in three small fictitious data examples with three items and four groups.

In the first dataset, we assume that the four true group means are −0.2, −0.2, −0.1, and 0.5, and that the common item difficulties for the three items are given as −1, 0, and 1 (see Table 1). We further assume that three out of 12 possible DIF effects are different from zero: item I1 is easier in group G2 (i.e., $e_{12} = -1$), item I2 is more difficult in group G3 (i.e., $e_{23} = 1$), and item I3 is more difficult in group G4 (i.e., $e_{34} = 0.5$). All three items I1, I2, and I3, serve as anchor items for the first group G1, while the remaining three groups have exactly one biased item (group G2: item I1; group G3: item I2; group G4: item I3). It can be seen that the set of anchor items differs among groups. Moreover, there is unbalanced DIF because the DIF effects of biased items do not sum to zero.

In practice, one has to infer these true parameters from the "observed data" which are the 12 group-specific parameters $b_{ig}$ that are identified from the item responses and are given in the upper right part of Table 1. Thus, the main goal is to determine the four group means $\mu_g$ and three common item parameters $b_i$ from the group-specific item parameters $b_{ig}$, which are given by (see Equation 2)[4]

$$b_{ig} = b_i - \mu_g + e_{ig} . \tag{3}$$

[4] For reasons of identification, the sum of group means is chosen to be zero, i.e., $\sum_g \mu_g = 0$ .

**Table 1:**
Illustrative Dataset 1: True DIF Effects, Identified Parameters, Estimated Group Means, and Estimated DIF Effects from Two-Way ANOVA Estimation (Using OLS and Robust Estimation)

| | | **True DIF Effects ($e_{ig}$)** | | | | **Identified Parameters ($b_{ig}$)** | | | |
|---|---|---|---|---|---|---|---|---|---|
| Item | $b_i$ | G1 | G2 | G3 | G4 | G1 | G2 | G3 | G4 |
| I1 | −1 | 0 | **−1** | 0 | 0 | −0.8 | −1.8 | −0.9 | −1.5 |
| I2 | 0 | 0 | 0 | **1** | 0 | 0.2 | 0.2 | 1.1 | −0.5 |
| I3 | 1 | 0 | 0 | 0 | **0.5** | 1.2 | 1.2 | 1.1 | 1.0 |
| | $\mu_g$ | −0.2 | −0.2 | −0.1 | 0.5 | – | – | – | – |
| | | **Estimated DIF Effects (OLS)** | | | | **Estimated DIF Effects (Robust)** | | | |
| | I1 | **0.29** | **−0.37** | −0.04 | **0.12** | 0.00 | −1.00 | 0.00 | 0.00 |
| | I2 | **−0.21** | **0.13** | **0.46** | **−0.38** | 0.00 | 0.00 | 1.00 | 0.00 |
| | I3 | −0.08 | **0.25** | **−0.42** | **0.25** | 0.00 | 0.00 | 0.00 | 0.50 |
| | $\hat{\mu}_g$ | −0.16 | **0.18** | **−0.39** | **0.37** | −0.20 | −0.20 | −0.10 | 0.50 |

*Note.* $b_i$ = common item parameter; $\mu_g$ = true group mean; $\hat{\mu}_g$ = estimated group mean; $e_{ig}$ = DIF effects; DIF effects of biased items are printed in bold. Estimated DIF effects and estimated group means are printed in bold if the absolute bias exceeds 0.10.

This is a two-way analysis of variance (ANOVA) without repeated measurements with main effects $\mathbf{b} = (b_1,\ldots,b_I)$ and $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_G)$, and interaction effects $e_{ig}$. Equation 3 corresponds to a linear regression in which DIF effects $e_{ig}$ are residuals. As pointed out by Camilli (1993) and van der Linden (1994) DIF effects can be interpreted as interaction effects in a two-way ANOVA. The parameters of the two-way ANOVA can be determined by minimizing an estimation function $H(\boldsymbol{\mu},\mathbf{b}) = \sum_{ig}\rho(b_{ig} - b_i + \mu_g) = \sum_{ig}\rho(e_{ig})$, where $\rho$ is a differentiable loss function (Fox, 2016; see also Davies, 2014, Ch. 6, for different loss functions). Using the estimation function $H$ implies estimation constraints on the residuals, namely, $\frac{\partial H}{\partial \mu_g} = \sum_i \rho'(e_{ig}) = 0$, where $\rho'$ denotes the first derivative of $\rho$. For unbiased estimation of group means, it is vital that the estimation constraints on the residuals implied by the estimation function $H$ coincide with the identification constraints on the DIF effects in the data-generating model (i.e., DIF effects sum to zero in the group-specific sets of anchor items). Applying ordinary least squares (OLS) estimation of the ANOVA model (using the loss function $\rho(x) = x^2/2$) implies that the DIF effects $e_{ig}$ sum to zero within the groups[5]. If there exist biased items (which imply unbalanced DIF), this estimation

[5] In OLS estimation, it holds that $\rho'(x) = x$, which results in $\sum_i \rho'(e_{ig}) = \sum_i e_{ig} = 0$.

constraint violates the identification constraint for DIF effects in the data-generating model. Alternatively, if DIF effects of biased items are interpreted as outliers, robust regression methods can be used for estimating the two-way ANOVA parameters. One robust regression approach is the bisquare loss function, which depends on a tuning constant $k$ and is defined as $\rho(x) = k^2/6 \cdot \left[1-\left(1-(x/k)^2\right)^3\right]$ for $|x| < k$ and $\rho(x) = k^2/6$ for $|x| \geq k$ (Fox, 2016).

In the left panel of Figure 1, the two loss functions are depicted. As can be seen, the bisquare loss function shows similar behavior as the least squares loss function for observations near to zero, but strongly differs for values substantially different from zero. In the robust approach using the bisquare loss function, residuals larger in absolute value than $k$ do not contribute to the loss function, and large DIF effects will be downweighted in the estimation (see right panel in Figure 1). In contrast, all observations are equally weighted in OLS estimation.

As can be seen in the left lower part of Table 1, when using least squares estimation, all items show DIF effects, and the estimated group means are biased. This is not surprising as the group comparisons also rely on the items with DIF. For example, the group mean of the second group, G2, is estimated as 0.18 while the true value is −0.20. This can be explained by the fact that item I1 in group G2 (with DIF effect $e_{12} = -1$) is included in the sum constraint and thus affects the estimation of G2's group mean. In the right lower

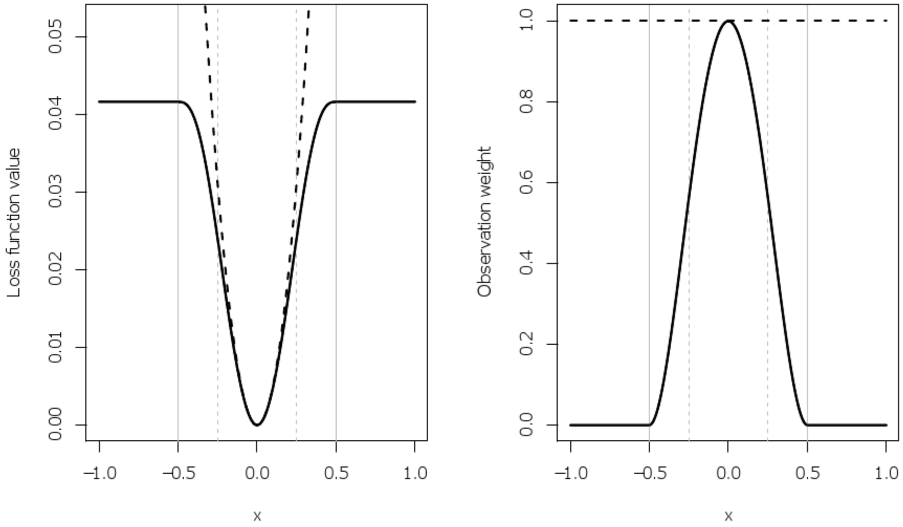

**Figure 1:**
Properties of least squares loss function (dashed line) and bisquare loss function (solid line). Left panel: loss functions. Right panel: observation weights induced by the two loss functions.

part of Table 1, results of robust bisquare regression using $k = .5$ are displayed. It can be seen that large residuals (i.e., biased items with DIF effects) are automatically detected as outliers and do not bias group means. Hence, by applying robust regression, an estimation constraint is only posed on those items in a group which are not considered to be outliers (i.e., they are detected as anchor items).

In practice, it is not realistic to assume that the majority of DIF effects is exactly zero (i.e., $e_{ig} = 0$) as most items will at least slightly differ in their functioning across countries in ILSAs (Grisay, Gonzales, & Monseur, 2009; Kankaras & Moors, 2014; Robitzsch & Lüdtke, 2019; Sachse, Roppelt, & Haag, 2016). Thus, it is reasonable to assume that there are small DIF effects for anchor items that sum to zero (Weeks, von Davier, & Yamamoto, 2014; Strobl et al., 2018). In the upper part of Table 2, we present a second illustrative dataset, which includes the same set of group-specific biased items but also group-specific anchor items with small DIF effects that add to zero within groups.

It is important to emphasize that in this scenario, not all anchor items need to show DIF (i.e., $e_{ig} = 0$), even though it would be possible as long as the DIF effects sum to zero within each group. As in the first dataset, the presence of biased items results in unbalanced DIF, and therefore estimates of the group means are biased if OLS estimation is employed (see lower left part of Table 2). Furthermore, the estimated DIF effects differ from the true DIF effects. This finding can be explained by the fact that OLS estimation implies a zero-sum constraint for all items within a group (biased items and anchor items), while the data-generating model involves the zero-sum constraint only for the set

**Table 2:**
Illustrative Dataset 2: True DIF Effects, Identified Parameters, Estimated Group Means, and Estimated DIF Effects from Two-Way ANOVA Estimation (Using OLS and Robust Estimation)

| | | **True DIF Effects ($e_{ig}$)** | | | | **Identified Parameters ($b_{ig}$)** | | | |
|---|---|---|---|---|---|---|---|---|---|
| Item | $b_i$ | G1 | G2 | G3 | G4 | G1 | G2 | G3 | G4 |
| I1 | −1 | 0 | **−1** | 0.1 | −0.1 | −0.8 | −1.8 | −0.8 | −1.6 |
| I2 | 0 | −0.1 | 0 | **1** | 0.1 | 0.1 | 0.2 | 1.1 | −0.4 |
| I3 | 1 | 0.1 | 0 | −0.1 | **0.5** | 1.3 | 1.2 | 1 | 1 |
| | $\mu_g$ | −0.2 | −0.2 | −0.1 | 0.5 | – | – | – | – |
| | | **Estimated DIF Effects (OLS)** | | | | **Estimated DIF Effects (Robust)** | | | |
| | I1 | **0.29** | **−0.37** | 0.06 | 0.02 | 0.00 | −1.00 | 0.10 | −0.10 |
| | I2 | **−0.31** | **0.13** | **0.46** | **−0.28** | −0.10 | 0.00 | 1.00 | 0.10 |
| | I3 | 0.02 | **0.25** | **−0.52** | **0.25** | 0.10 | 0.00 | −0.10 | 0.50 |
| | $\hat{\mu}_g$ | −0.16 | **0.18** | **−0.39** | **0.37** | −0.20 | −0.20 | −0.10 | 0.50 |

*Note.* $b_i$ = common item parameter; $\mu_g$ = true group mean; $\hat{\mu}_g$ = estimated group mean; $e_{ig}$ = DIF effects; DIF effects of biased items are printed in bold. Estimated DIF effects and estimated group means are printed in bold if the absolute bias exceeds 0.10.

of anchor items. However, when using robust estimation (lower right part of Table 2), estimated group means are again unbiased, and DIF effects of biased items are correctly detected as outliers.

In the third illustrative dataset (upper part of Table 3), it is assumed that all items are anchor items (i.e., show DIF effects that sum to zero in each group). Hence, there is balanced DIF, and OLS estimation provides unbiased estimates of group means because the estimation constraint of OLS (i.e., DIF effects are assumed to sum to zero within each group) coincides with the identification constraint for the DIF effects in the data-generating model (see lower left part of Table 3). In contrast, robust estimation, which treats large DIF effects as outliers, produces biased estimates of group means because the estimation constraint does not coincide with the identification constraints (see lower right part of Table 3). Thus, in the constellation of the third dataset, the removal of items with large DIF effects would have undesirable effects on the calculation of group means.

At a more conceptual level, it is crucial to understand that the decision about whether an item with a DIF effect is classified as an anchor item or a biased item is not a primarily statistical question (see Camilli, 1993; Penfield & Camilli, 2007; Zwitser, Glaser, & Maris, 2017, for this argument). It needs to be emphasized that comparisons involving group $g$ only rely on items from the set of anchor items $\mathcal{J}_{A,g}$ and that biased items from the set $\mathcal{J}_{B,g}$ do not contribute to the group comparisons. Therefore, removing items from the anchor item set based solely on statistical criteria could result in construct underrepresentation (i.e., removing items with DIF effects that are construct relevant; see

**Table 3:**
Illustrative Dataset 3: True DIF Effects, Identified Parameters, Estimated Group Means, and Estimated DIF Effects from Two-Way ANOVA Estimation (Using OLS and Robust Estimation)

| | | **True DIF Effects ($e_{ig}$)** | | | | **Identified Parameters ($b_{ig}$)** | | | |
|---|---|---|---|---|---|---|---|---|---|
| Item | $b_i$ | G1 | G2 | G3 | G4 | G1 | G2 | G3 | G4 |
| I1 | −1 | 0.3 | −0.2 | 0.1 | −0.2 | −0.5 | −1 | −0.8 | −1.7 |
| I2 | 0 | −0.3 | 0.6 | 0 | −0.3 | −0.1 | 0.8 | 0.1 | −0.8 |
| I3 | 1 | 0 | −0.4 | −0.1 | 0.5 | 1.2 | 0.8 | 1 | 1 |
| | $\mu_g$ | −0.2 | −0.2 | −0.1 | 0.5 | – | – | – | – |
| | | **Estimated DIF Effects (OLS)** | | | | **Estimated DIF Effects (Robust)** | | | |
| | I1 | 0.30 | −0.20 | 0.10 | −0.20 | **0.12** | **0.00** | −0.04 | **−0.06** |
| | I2 | −0.30 | 0.60 | 0.00 | −0.30 | −0.25 | **1.02** | 0.08 | **0.06** |
| | I3 | 0.00 | −0.40 | −0.10 | 0.50 | 0.03 | **0.00** | −0.04 | **0.84** |
| | $\hat{\mu}_g$ | −0.20 | −0.20 | −0.10 | 0.50 | **−0.38** | **−0.01** | **−0.25** | **0.63** |

*Note.* $b_i$ = common item parameter; $\mu_g$ = true group mean; $\hat{\mu}_g$ = estimated group mean; $e_{ig}$ = DIF effects; DIF effects of biased items are printed in bold. Estimated DIF effects and estimated group means are printed in bold if the absolute bias exceeds 0.10.

Camilli, 1993; Kane, 2006; Penfield & Camilli, 2007; Shealy & Stout, 1993). Construct underrepresentation could occur if relevant facets of a trait are not represented in the anchor item set (Camilli, 2006). For example, Wu (2010) mentioned significant DIF between Asian and Western countries for items on formal mathematics and items with real-life applications. In this example, the mean of DIF effects of a subset of items (so-called item bundles) differs from zero (Gierl et al., 2001). This finding could empirically occur if unidimensionality holds, or it could be caused by secondary dimensions that show differences in group means (Shealy & Stout, 1993). However, eliminating particular items with DIF effects involving formal mathematics from country comparisons in PISA could be seen as a severe threat to validity because the secondary dimension involving formal mathematics is interpreted as construct relevant. Thus, even if the categorization of items according to their DIF effects into sets of anchor items and biased items seems to be purely statistical, the conclusion that some items with large DIF effects (e.g., formal mathematics items in Asia) are construct relevant implies that these items must be included in the anchor item set. Consequently, if one is quite confident that all items are construct relevant, no items with DIF effects should be removed from linking. Hence, all items serve as anchor items and an identification constraint (i.e., $\sum_i e_{ig} = 0$) has to be assumed. Only those items with DIF effects that are identified as construct irrelevant constitute the set of biased items. In contrast, in a purely statistical approach, DIF effects are identified as outliers in a statistical model, and the corresponding items are subsequently treated as construct irrelevant for group comparisons.

## Approaches for multiple-group comparisons

In the following, we discuss different approaches for comparing group means in the presence of DIF in the 1PL model. At a conceptual level, we distinguish three strategies for calibrating the item parameters, which differ with respect to the degree of invariance they assume for the item parameters (see also van de Vijver, 2019, for a recent overview). First, we discuss calibration approaches that assume *full invariance* of item parameters across groups. In this strategy, DIF effects are ignored and not modeled (i.e., DIF effects $e_{ig}$ are not included as parameters in the statistical model) when estimating group means and standard deviations. Second, we discuss approaches that rely on *partial invariance*. In this approach, it is assumed that group-specific item parameters only need to be included for a subset of DIF effects (Rutkowski & Rutkowski, 2018; von Davier et al., 2019). Typically, the set of items with modeled DIF effects (which is aimed to match the set of biased items) is allowed to vary from group to group, and some statistical technique is needed to determine the set of biased items for each group. In most applications, there is only a minority of items for which DIF effects are modeled, and all other DIF effects are set to zero, indicating that there are no DIF effects for most items in each group (i.e., the anchor items). Third, we discuss approaches that do not make any invariance assumptions for item parameters (complete *noninvariance*). In these approaches, all DIF effects are allowed, and the group-specific means and standard deviations are estimated under some identification constraint for the DIF effects (e.g., DIF effects sum to zero).

## Concurrent calibration under full invariance

In concurrent calibration under the assumption of full invariance of item parameters, maximum likelihood (ML) estimation is used to estimate a multiple-group model without including any DIF effects for item parameters. More specifically, the following log-likelihood function is maximized with respect to the unknown model parameters $(\boldsymbol{\mu},\boldsymbol{\sigma},\mathbf{b})$ for all item responses $\mathbf{x}_{pg}$ of person $p = 1,\ldots,N_g$ in group $g = 1,\ldots,G$:

$$l(\boldsymbol{\mu},\boldsymbol{\sigma},\mathbf{b}) = \sum_{g=1}^{G}\sum_{p=1}^{N_g} v_{pg} \log\left[\int\left\{\prod_{i=1}^{I} P_i(\theta;b_i)^{x_{pgi}}\left(1-P_i(\theta;b_i)\right)^{1-x_{pgi}}\right\} f_g\left(\theta;\mu_g,\sigma_g\right)\mathrm{d}\theta\right], \quad (4)$$

where abilities in group $g$ are assumed to be normally distributed (i.e., $\theta \sim \mathrm{N}\left(\mu_g,\sigma_g^2\right)$), and all group means and standard deviations are collected in vectors $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, respectively (marginal maximum likelihood estimation, MML; see Adams, Wu, & Carstensen, 2007; von Davier & Sinharay, 2014)[6]. In addition, invariant item parameters **b** across groups are assumed, and the IRF is given as $P_i(\theta;b_i) = \Psi(\theta - b_i)$. As can be seen, concurrent calibration under full invariance estimates all common item parameters as well as country-specific means and standard deviations in one step. It is typically used with sampling weights $v_{pg}$ to accommodate the multistage sampling design in ILSAs.

If the model is correctly specified, maximum likelihood estimates are consistent (White, 1982). However, in general, the model in Equation 4 is misspecified as the true IRFs $P_{ig}(\theta) = \Psi(\theta - b_i - e_{ig})$ involve group-specific DIF effects, which are ignored in concurrent calibration under full invariance. For misspecified models, it has been shown that the ML estimate $(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\sigma}},\hat{\mathbf{b}})$ is still consistent and converges to the maximizer of the Kullback-Leibler information (White, 1982; see also Kuha & Moustaki, 2015). In this sense, ignoring DIF effects in the estimation of Equation 4 provides group mean estimates that are the best approximation of the group-specific distributions of item responses with respect to the Kullback-Leibler information. In the case of a misspecified model, model-robust standard errors (also known as sandwich standard errors; White, 1982) have to be used for valid statistical inference. Given true data-generating parameters $(\boldsymbol{\mu},\boldsymbol{\sigma},\mathbf{b})$, ML estimates of misspecified models are typically biased $(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\sigma}},\hat{\mathbf{b}})$ even in the case of infinite sample sizes, that is, $\hat{\boldsymbol{\mu}} \neq \boldsymbol{\mu}$. Nevertheless, we can derive $\hat{\boldsymbol{\mu}}$ as a function of the true parameters of interest $(\boldsymbol{\mu},\boldsymbol{\sigma},\mathbf{b})$ and DIF effects **e** (see Kolenikov, 2011, for a similar technique for misspecified structural equation models). Assuming large group sample

[6] In Equation 4, the log-likelihood function is formulated only for binary data without a multi-matrix design. The extension to polytomous responses and multi-matrix designs would result in only slight changes in notation.

sizes (see Appendices B and C), the bias of the estimator of the group mean $\hat{\mu}_g$ can be approximated as a weighted combination of the DIF effects $e_{ig}$:

$$\hat{\mu}_g = \mu_g + \sum_{i=1}^{I} w_{ig} e_{ig} \ , \tag{5}$$

where the weights $w_{ig} = w_{ig}\left(\mu_g, \sigma_g, b_i\right)$ in ML estimation are primarily driven by the precision of the IRFs. As the IRFs of items with more extreme difficulties are less precisely estimated, DIF effects for items with extreme difficulties are down-weighted in Equation 5. Therefore, the bias of a group mean is mainly caused by items for which DIF effects $e_{ig}$ are large, and their item difficulties are located close to the center of the distribution, that is, $\left|\mu_g - b_i\right|$ is small.

Alternatively, concurrent calibration under full invariance can be conducted with limited information methods such as diagonally weighted least squares (DWLS; Cai & Moustaki, 2018; see Rutkowski & Svetina, 2017, for an application in ILSAs). Using a probit IRF, only item thresholds and tetrachoric correlations are needed as input for a two-stage estimation procedure. In a first step, item thresholds and pairwise tetrachoric correlations are calculated for each group. In the second step, these statistics are used for estimating the parameter vector $\left(\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{b}\right)$ (Cai & Moustaki, 2018). Interestingly, the bias derivation in Equation 5 still holds for limited information estimation (see Appendix C). In DWLS, weights $w_{ig}$ are given either by the precision of group-specific item thresholds or by user-defined values. In the special case of unweighted least squares estimation, weights are given as $w_{ig} = 1/I$. Typically, limited information methods are expected to be more robust in modeling violations than ML estimators (MacCallum, Browne, & Cudeck, 2007).

To sum up, it can be expected that the concurrent calibration approach under full invariance with fixed items only provides unbiased estimates in either the trivial case of no true DIF effects or if the constraint $\sum_i w_{ig} e_{ig} = 0$ is fulfilled in the data-generating model. If only anchor items and no biased items exist, the condition $\sum_i \frac{1}{I} \cdot e_{ig} = 0$ is fulfilled, but it can be expected that for the precision weighted mean of DIF effects it holds that $\sum_i w_{ig} e_{ig} \approx 0$ in the case of a test with many items. This property would imply approximately unbiased estimates[7].

[7] If items (or DIF effects of items in a group) would be treated as random (instead as fixed) and DIF effects would have a zero expected value (i.e., $E_i\left(e_{ig}\right) = 0$), one can show by using Equation 5 that group means are unbiased with the misspecified scaling model under full invariance. This finding has already been demonstrated in simulation studies (Sachse et al., 2016; Robitzsch & Lüdtke, 2019).

## Concurrent calibration under partial invariance using DIF statistics

In contrast to concurrent calibration under full invariance, which ignores DIF effects, concurrent calibration under partial invariance allows some of the item parameters with large DIF effects $\mathbf{e}$ to vary across groups. In this approach, country comparisons are based on a multiple-group IRT model in which, for some of the items, item-by-group interactions are specified (Glas & Jehangir, 2014; Oliveri & von Davier, 2011, 2014; von Davier et al., 2019). The decision about which item parameters obtain group-specific item parameters is usually based on DIF statistics that measure how strongly group-specific item parameters deviate from common item parameters. The basic idea is to identify the set of biased items $\mathcal{J}_{B,g}$ for each group that obtain unique item parameters while the DIF effects in the set of anchor items $\mathcal{J}_{A,g}$ are set to zero. More specifically, a DIF statistic $T_{ig}$ of interest is selected and item $i$ for group $g$ is declared to be in the DIF item set $\mathcal{J}_{B,g}$ if $|T_{ig}| > c$ for some chosen cutoff value $c$. In the partial invariance approach, the estimation of country means is typically conducted using an iterative procedure (OECD, 2017; von Davier et al., 2019). First, a multiple-group IRT model is estimated under the assumption of full invariance to obtain estimates for the country means and standard deviations as well as for the common item parameters $(\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{b})$. Based on these parameters, the DIF statistic $T_{ig}$ is calculated for every item in every group. In the next step, a multiple-group model is specified in which DIF effects $e_{ig}$ are freely estimated if $|T_{ig}| > c$ in the previous step. From this model, parameter estimates for $(\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{b}, \mathbf{e})$ are obtained where a subset of DIF effects in $\mathbf{e}$ is set to zero. Furthermore, the calculation of the DIF statistic can be repeated and it can be checked whether, for some of the items, the DIF statistics are still not acceptable. If this is the case, more DIF effects can be freed in the estimation and the process can be iterated until no item shows any nonacceptable DIF (e.g., Kopf et al., 2015, for a description of this purification process). A large variety of DIF statistics have been proposed in the literature (see Penfield & Camilli, 2007, for an overview). In the following, we will first discuss a DIF statistic that is based on item difficulties and was used in PISA until 2012 (OECD, 2014). Second, we will discuss the MD and RMSD statistics, which are currently used in the operational procedure of PISA (OECD; 2017; von Davier et al., 2019).

### *DIF in item difficulties*

DIF in a 1PL model (also referred to as uniform DIF, see Penfield & Camilli, 2007) can be directly quantified by the size of the DIF effects $e_{ig}$. Often, a group-specific 1PL model is fitted and item difficulties from this group-specific model are compared with the difficulties from a multiple-group 1PL model in which all item parameters are assumed to be equal across groups (i.e., full invariance assumption). In order to place the two sets of item parameters onto the same metric, either an identification constraint can be imposed on the two calibrations (i.e., sum of item difficulties equals zero) or a linking method can be used (e.g., mean-mean linking; Kolen & Brennan, 2014). Differences in item difficulties $e_{ig}$ are classified to be large if absolute values exceed .64. Moderate DIF is said to exist for values between .43 and .64 (see Penfield & Camilli, 2007).

*MD and RMSD statistics*

Since 2015, the mean deviation (MD) and root mean squared deviation (RMSD) statistics are used in PISA to identify items with DIF (Oliveri & von Davier, 2011, 2014). In these statistics, the distance between a group-specific IRF $P_{ig}$ and a reference IRF $P_i$ (which does not include group-specific item parameters) is measured in the probability metric:

$$\begin{aligned}\mathrm{MD}_{ig} &= \int \left(P_{ig}(\theta) - P_i(\theta)\right) f_g(\theta)\,\mathrm{d}\theta,\\ \mathrm{RMSD}_{ig} &= \sqrt{\int \left(P_{ig}(\theta) - P_i(\theta)\right)^2 f_g(\theta)\,\mathrm{d}\theta}.\end{aligned} \tag{6}$$

The MD statistic measures the difference between an observed group mean based on the IRF $P_{ig}$ of group $g$ and an expected group mean based on a common IRF $P_i$ (see also Glas & Jehangir, 2014). The RMSD statistic quantifies the distance between the groupwise IRF and the common IRF. It can be shown that $\mathrm{RMSD}_{ig} \geq |\mathrm{MD}_{ig}|$ (see Raju, van der Linden, & Fleer, 1995). Several benchmarks for interpreting DIF effects as large have been proposed for the MD and RMSD statistics: .055 (Buchholz & Hartig, 2019), .08 (Köhler, Robitzsch, & Hartig, 2019), .10 (Oliveri & von Davier, 2011, 2014), .12 (OECD, 2017, p. 151), .15 (OECD, 2017, p. 174; von Davier et al., 2019), and .20 (OECD, 2015, p. 30). Because the absolute value of the MD statistic is bounded by the RMSD statistic, we use established effect sizes for the RMSD statistic also for the MD statistic.

One desirable feature of the MD and RMSD statistics is that the magnitude of these statistics is influenced not only by the size of the DIF effect $e_{ig}$ but also by the ability distribution in the specific groups (Wainer, 1993). To further illustrate this feature of the MD and RMSD statistics, we derived closed formulas by using a probit approximation of the logistic IRF in the 1PL multiple-group model. For $P_{ig}(\theta) = \Psi(\theta - b_i - e_{ig})$, $P_i(\theta) = \Psi(\theta - b_i)$, and a normal distribution $\mathrm{N}\left(\mu_g, \sigma_g^2\right)$ for group $g$, it can be shown that (see Appendix D)

$$\mathrm{MD}_{ig} = \Phi\left(\frac{\mu_g - b_i}{\sqrt{D^2 + \sigma_g^2}}\right) - \Phi\left(\frac{\mu_g - b_i - e_{ig}}{\sqrt{D^2 + \sigma_g^2}}\right) \text{ for } e_{ig} > 0\,, \tag{7}$$

where $D = 1.701$ is the conversion factor from the logit to the probit link function. For $e_{ig} < 0$, the two terms in Equation 7 have to be exchanged. As can be seen, the magnitude of MD not only depends on the size of the DIF effect $e_{ig}$ but is also affected by the difference $\mu_g - b_i$. Thus, the same DIF effect $e_{ig}$ (measured in the logit metric) provides different MD statistics (measured in the probability metric), depending on the magnitude $|\mu_g - b_i|$. As a consequence, the MD statistic has a tendency to be smaller in very low- or very high-achieving countries as, in these countries, $|\mu_g - b_i|$ is expected to substantially differ from zero (see also DeMars, 2011). However, items with small absolute values of MD statistics are therefore expected to induce only a slight bias when ignoring DIF effects in the calibration (see weights $w_{ig}$ in Equation 5 for the full invariance approach). Hence, if only items with large |MD| statistics are declared to be non-invariant, only those items are eliminated from the set of anchor items that have the potential to induce large biases in the group comparisons.

The computation of RMSD is slightly more complicated because bivariate normal integrals are now needed:

$$\left(\mathrm{RMSD}_{ig}\right)^2 = \int_{\mu_g - b_i - e_{ig}}^{\mu_g - b_i} \int_{\mu_g - b_i - e_{ig}}^{\mu_g - b_i} \phi_2\left(\frac{u}{\sqrt{D^2+\sigma_g^2}}, \frac{v}{\sqrt{D^2+\sigma_g^2}}, \frac{\sigma_g^2}{D^2+\sigma_g^2}\right) \mathrm{d}u\mathrm{d}v \text{ for } e_{ig} > 0 \,, \quad (8)$$

where $\phi_2 = \phi_2(x,y,\rho)$ is the bivariate normal density function. The lower and upper bounds of the integrals have to be exchanged for a negative $e_{ig}$. The integral in Equation 8 can be interpreted as an integration over the two-dimensional rectangle $[\mu_g - b_i - e_{ig}, \mu_g - b_i]^2$. Again, it turns out that the RMSD depends on the size of the DIF effect $e_{ig}$ as well as on the difference $\mu_g - b_i$ (see Tijmstra, Liaw, Bolsinova, Rutkowski, & Rutkowski, 2019, for an empirical demonstration).

Another challenging aspect in the interpretation of the MD and RMSD statistics is that their definitions involve an unknown group-specific IRF and an unknown ability distribution (see Equation 6). Both need to be estimated from sample data and this can result in biased estimates of the population MD and RMSD statistics. More specifically, the IRF and the ability distribution are reconstructed from the output of the MML estimation of a multiple-group IRT model (assuming fully invariant item parameters) and are based on aggregating individual posterior distributions (Köhler et al., 2019; see also van Rijn, Sinharay, Haberman, & Johnson, 2016). The group-specific IRF $P_{ig}(\theta)$ is replaced by a sample-based empirical IRF $\hat{P}_{ig}(\theta)$ and the density $f_g$ is replaced by an empirical distribution $\hat{f}_g$, resulting in a sample-based definition of the RMSD

$$\widehat{\mathrm{RMSD}}_{ig} = \sqrt{\int \left(\hat{P}_{ig}(\theta) - \hat{P}_i(\theta)\right)^2 \hat{f}_g(\theta)\mathrm{d}\theta} \quad (9)$$

Given the large overall sample size in the multiple-group ILSA context, we can assume that the common IRF is reliably und unbiasedly estimated, that is, $\hat{P}_i(\theta) \approx P_i(\theta)$. However, it can be shown that the sample RMSD, in general, is a biased estimator of the population RMSD (see Appendix E and Köhler et al., 2019, for similar arguments and an empirical illustration)

$$E\left(\widehat{\mathrm{RMSD}}_{ig}\right)^2 = RMSD_{ig}^2 + B_{1,+}(n) + B_{2,+}\left(\left|\hat{f}_g - f_g\right|\right) + B_{3,\pm}\left(P_{ig} - P_i, \left|\hat{f}_g - f_g\right|\right) \quad (10)$$

Three main sources of bias for the empirical RMSD must be distinguished. First, as $\hat{P}_{ig}$ relies on a finite sample size $n$, some sampling variability always contributes to the estimation of the RMSD and can only be reduced when the sample size goes to infinity (but see Köhler et al., 2019, for attempts to correct the bias). Thus, the first biasing term $B_{1,+}(n)$ that is a function of the sample size (see Appendix E), is always positive and only vanishes in large samples. As a consequence, higher RMSD values can be expected in smaller samples. Second, DIF effects in all items of group $g$ can bias the estimates of individual posterior distributions, which, in turn, bias the group distribution $f_g$ (i.e., $\hat{f}_g \neq f_g$). If such a bias is present, the empirical IRF $\hat{P}_{ig}$ does not converge (samples sizes tending to infinity) to the true IRF $P_{ig}$, but rather to a function $P_{ig}^* \neq P_{ig}$. The positive second bias term $B_{2,+}\left(\left|\hat{f}_g - f_g\right|\right)$ reflects the distance $|P_{ig}^* - P_{ig}|$. The third bias term $B_{3,\pm}$ can be positive or negative and only vanishes if the group distribution is not biasedly estimated (i.e., $\hat{f}_g = f_g$) or in the case of no DIF effect (i.e., $P_{ig} = P_i$)[8]. Using the same proof strategy, it can also be shown that the bias of the MD statistic is only affected by a biased estimation of the group distribution (i.e., $\hat{f}_g \neq f_g$) and that the sampling fluctuation of $\hat{P}_{ig}$ does not bias the MD statistic.

Overall, the partial invariance approach has the potential to remove bias in estimated group means if the items with large DIF statistics belong to the set of biased items. In this case, the biased items are correctly removed from group comparisons[9]. However, if items from the set of anchor items show large DIF statistics, bias could be introduced by removing those anchor items from group mean comparisons. Furthermore, even if no bias is introduced by removing items from the anchor item set (e.g., items with large positive and large negative DIF effects are removed from the identification of group means), efficiency losses in group mean comparisons could be expected because group comparisons now rely on a smaller number of items (see Sachse et al., 2016; Robitzsch & Lüdtke, 2019).

## Linking with separate calibrations under full noninvariance

In the third approach, no invariance assumptions are made for the group-specific item parameters. In this approach, group comparisons are based on a two-step procedure that

[8] As it is shown in Appendix E (Equation A14), the third bias term $B_{3,\pm}$ is computed as a weighted integral of the product $\left(P_{ig}(\theta) - P_i(\theta)\right) \cdot B_g(\theta)$, where $B_g(\theta)$ denotes the bias in estimated abilities. If the product is mostly negative for θ values, then an underestimation of the RMSD statistic can be expected. This case could occur when a 2PL model is fitted and the data had been generated from a 3PL model.

[9] Note that some items in a group are removed from the computation of the respective group mean. Hence, these items are practically removed from linking (and the subsequent group comparisons). However, this has the consequence that in the case of more than two groups, the difference between group means does not involve a full set of common item parameters.

combines the separate calibration of item parameters within groups and linking methods (see Kolen & Brennan, 2014; Lee & Lee, 2018, for overviews). In a first step, an item response model is fitted separately for each group, resulting in item parameter estimates $\hat{\mathbf{b}}_g$ ($g = 1,\ldots,G$). Hence, item parameters are calibrated separately and allowed to vary across groups. In a second step, the parameters of the group-specific ability distributions (i.e., group means $\mu_g$) are obtained by placing the group-specific item parameters onto a common metric (Battauz, 2017). In the following, we discuss the Haberman and the Haebara linking methods, which are suited for linking multiple groups.

### *Haberman linking*

In Haberman linking, the group-specific item parameters $b_{ig}$ are used to simultaneously estimate common item parameters $b_i$ and group means $\mu_g$. In the 1PL model with DIF effects, it holds that $e_{ig} = b_{ig} - (b_i - \mu_g)$. Haberman (2009) proposed a regression approach to estimate group means $\boldsymbol{\mu}$ and common item parameters $\mathbf{b}$ by minimizing the variation of residuals $e_{ig}$

$$H(\boldsymbol{\mu},\mathbf{b}) = \sum_{g=1}^{G}\sum_{i=1}^{I}\rho\left(\hat{b}_{ig} - b_i + \mu_g\right), \qquad (11)$$

where ρ is a loss function (Fox, 2016), and the identification constraint $\mu_1 = 0$ is used. It should be emphasized that Equation 11 corresponds to a two-way ANOVA (see Equation 3 in section "Differential item functioning for multiple groups"). Haberman (2009) proposed the squared loss function $\rho(x) = x^2/2$, which results in a linear regression model estimated by OLS (i.e., $L_2$ regression). Because DIF effects are often characterized as outlying observations, robust loss functions should be preferred for the unbiased estimation of group means (Fox, 2016). Here, we use the bisquare loss function which depends on a tuning parameter $k$. Using this robust loss function, residuals larger in absolute value than $k$ do not contribute to the loss function (Fox, 2016)[10]. Alternative loss functions are $\rho(x) = |x|$ (median regression or $L_1$ regression) or $\rho(x) = \sqrt{|x|}$ (used in invariance alignment; see Muthen & Asparouhov, 2014). It can be shown that the estimation constraints implied by Haberman linking are given as[11] (see Appendix F)

$$\frac{\partial H}{\partial \mu_g} = \sum_{i=1}^{I}\rho'\left(e_{ig}\right) = \sum_{i=1}^{I} w_{ig} e_{ig} = 0 \text{ with } w_{ig} = \frac{\rho'\left(e_{ig}\right)}{e_{ig}}. \qquad (12)$$

Unbiased estimates of group means are provided if the identification constraints for DIF effects coincide with the estimation constraints that are given by Equation 12. Note that

[10] As pointed out by an anonymous reviewer, robust linking can also be conceptualized as a variant of partial invariance (see also He, Cui, & Osterlind, 2015, for such a view).

[11] In the case of a nondifferentiable loss function ρ, either the derivative can be interpreted as a subdifferential (Hastie et al., 2015) or the loss function is replaced by a differentiable approximating function.

for a squared loss function, it holds that $w_{ig} = 1$ and the condition $\sum_i e_{ig} = 0$ is obtained[12] for each group $g$. The weights for the bisquare loss function are given as $w_{ig} = \left(1 - \left(e_{ig} / k\right)^2\right)^2$ for $|e_{ig}| < k$ and zero for DIF effects $|e_{ig}| \geq k$. Hence, outlying DIF effects receive a value of zero and do not contribute to the linking.

*Haebara linking*

It is known that Haberman linking can be unstable in small sample sizes because item parameter estimates can be imprecisely estimated. However, estimates of IRFs can be quite stable even in the case of unstable item parameters (Ogasawara, 2002). The Haebara linking method relies on linking IRFs across groups (Kolen & Brennan, 2014) and, hence, has the potential to provide more stable group mean estimates than Haberman linking. A generalization of Haebara linking to multiple groups is based on distances of estimated IRFs and IRFs assuming common item parameters across groups. The estimation of group means $\boldsymbol{\mu}$ relies on the minimization of the function (using the constraint $\mu_1 = 0$)

$$H(\boldsymbol{\mu}, \mathbf{b}) = \sum_{g=1}^{G} \sum_{i=1}^{I} \int \rho\left(\Psi\left(\theta - \hat{b}_{ig}\right) - \Psi\left(\theta - b_i + \mu_g\right)\right) \omega(\theta) \mathrm{d}\theta \qquad (13)$$

with a loss function $\rho$ and a weighting function $\omega$. Haebara (1980) proposed a quadratic loss function $\rho(x) = x^2 / 2$. Alternatively, He and colleagues (He et al., 2015; He & Cui, 2019) proposed a robust version of Haebara linking using $\rho(x) = |x|$, which should be superior to a quadratic loss function if there are only a few outlying biased items. The following nonlinear estimation constraint is fulfilled (see Appendix F):

$$\frac{\partial H}{\partial \mu_g} = \sum_{i=1}^{I} \int \rho'\left(\Psi\left(\theta - b_i - e_{ig} + \mu_g\right) - \Psi\left(\theta - b_i + \mu_g\right)\right) \Psi'\left(\theta - b_i + \mu_g\right) \omega(\theta) \mathrm{d}\theta = 0 \qquad (14)$$

## Purpose

Until PISA 2012, concurrent calibration under full invariance had been used and items with DIF effects were only excluded from country comparisons if they could be explained by translation issues (Adams, 2003). Hence, the scaling model was – to some extent – misspecified because DIF effects were ignored. Beginning with PISA 2015, a concurrent calibration under partial invariance was established in which model refinement was based on the RMSD as DIF statistic (OECD, 2017). It was argued that this

[12] Concurrent calibration under full invariance estimated by unweighted least squares (which sets all weights in DWLS equal to one) for a probit version of the 1PL model is equivalent to Haberman linking with a quadratic loss function.

approach should lead to better model fit and to more stable and less biased estimates when a few item-by-country interactions were included to deal with the presence of DIF (Oliveri & von Davier, 2011, 2014; von Davier et al., 2019). However, to the best of our knowledge, simulation studies that evaluate the performance of this approach in the context of ILSAs are still lacking. In this article, we investigate the conditions under which concurrent calibration under partial invariance (using DIF statistics) is superior to a misspecified concurrent calibration that assumes full invariance and to separate calibration with a subsequent linking step. We expect that the performance of the different approaches depends on the type and amount of DIF effects and on the sample size.

A recent methodological case study that used the PISA 2015 data showed that concurrent calibration under full invariance and partial invariance resulted in very similar country means (Jerrim et al., 2018). This finding is consistent with research showing that concurrent calibration under full invariance did not result in biased estimates if there were DIF effects with some variation, but all items belonged to the anchor item set (Sachse et al., 2016; Robitzsch & Lüdtke, 2019)[13]. However, even in conditions with moderate sample sizes, concurrent calibration under full invariance was outperformed by a separate calibration approach with subsequent linking. These results are in line with findings from the linking literature that separate calibration approaches are superior to the concurrent calibration approach in the case of model violations (Kolen & Brennan, 2014; Kang & Petersen, 2012; Lee & Lee, 2018). In our simulation study, we expected the performance of the partial invariance approach to depend on the proportion and type of DIF effects of biased items. In the case of balanced DIF (i.e., DIF effects of biased items cancel out), efficiency losses of estimated group means can be expected when items are removed from country comparisons in the partial invariance approach (Sachse et al., 2016; see also DeMars, 2020). Furthermore, in this case, it could be expected that concurrent calibration under full invariance and the linking approach are superior to the partial invariance approach. In the case of unbalanced DIF (i.e., DIF effects in biased items are of the same sign), DeMars (2020) showed that a robust linking approach provided less biased group mean estimates than scaling approaches that rely on DIF statistics. Thus, we expected that concurrent calibration under partial invariance would not outperform separate calibration approaches when appropriate robust linking methods were used in which items with large DIF effects were down-weighted. However, the question of whether concurrent calibration has some advantages in small samples remains open.

## Simulation study

The main goal of the simulation study was to evaluate different methods for comparing country means in the presence of country DIF. To this end, we assumed a 1PL model for $G = 20$ countries. For each country, abilities were normally distributed with mean $\mu_g$ and

[13] In these studies, DIF effects of the anchor items were treated as random. This means that DIF effects vanished on average. In the DIF definition of this paper, it is assumed that DIF effects are fixed and that these effects exactly sum to zero.

standard deviation $\sigma_g$. Across all conditions of the simulation, the country means and standard deviations were held fixed and ranged between −0.92 and 0.81 for means (with an average of 0.00), and 0.82 and 1.06 for standard deviations (with an average of 0.91). The population containing all students in all countries had a mean of zero and a standard deviation of one. Country-specific item parameters $\beta_{ig}$ were generated according to $\beta_{ig} = b_i + e_{ig}$, where $b_i$ is the common item parameter and $e_{ig}$ is the country-specific DIF effect. In each country, the sets of biased and anchor items were held fixed across conditions with a fixed proportion of biased items. For a fixed proportion $\pi_B$ of biased items, a discrete variable $Z_{ig}$ was defined for each item in each group, which had values of 0 (if the item was an anchor item), +1 (biased item with a positive DIF effect), and −1 (biased item with a negative DIF effect). Furthermore, standardized effects $\varepsilon_{ig}$ were specified which were nonzero for anchor items and zero for biased items. These effects fulfilled the conditions $\sum_i \varepsilon_{ig} = 0$ (i.e., DIF effects of anchor items sum to zero) and $\frac{1}{I(1-\pi_B)} \cdot \sum_i \varepsilon_{ig}^2 = 1$ (*SD* of DIF effects of anchor items equals one). In the case of balanced DIF, DIF effects were computed as $e_{ig} = \left(\left|Z_{ig}\right| - 1\right) SD_A \varepsilon_{ig} + Z_{ig}\delta$, where $SD_A$ was the prespecified standard deviation of DIF effects for the anchor items. Hence, half of the biased items received a DIF effect of δ and for the other half, the DIF effect was set to –δ. In the case of unbalanced DIF, all biased items within a country received a DIF effect of either δ or –δ. This property was implemented by defining a variable $D_g$ with equally frequent values +1 and −1 for each country. The DIF effects for unbalanced DIF were defined as $e_{ig} = \left(\left|Z_{ig}\right| - 1\right) SD_A \varepsilon_{ig} + \left|Z_{ig}\right| D_g \delta$. All data generating parameters can be downloaded from https://osf.io/53vqr/.

For each condition of the simulation design, 200 replications were generated. More specifically, we manipulated the following six factors in our simulation design: number of persons (*N* = 200, 500, and 1000), number of items (*I* = 20, and 40), proportion of biased items (0%, 10%, and 30%), DIF effect size for the biased items (δ = 0, .3, .6, and 1), standard deviation of DIF effects for anchor items ($SD_A$ = 0, .1, .2, and .3), and type of DIF effects for biased items (balanced vs. unbalanced). It needs to be mentioned that we did not implement a multi-matrix design, but we would not expect different findings by adopting such a design.

We used three different scaling strategies to obtain country means in each replication. First, we specified a multiple-group 1PL model with invariant item parameters across countries (full invariance approach; FI). Second, we implemented a partial invariance approach in which DIF statistics were used to identify items with DIF. Two different DIF statistics were applied: DIF in item difficulties in the logit metric, and the MD statistic. In the partial invariance approach based on item difficulties (PI-DIF), absolute values of .4 or .6 were used as cutoff values for declaring country-specific DIF. In the partial invariance approach based on the MD statistic (PI-MD), items in a country with absolute MD values larger than .05, .08, or .12 received country-specific parameters. Third, we used two linking approaches (Haberman method and Haebara method) in which no invariance assumptions were made for the item parameters. In both approaches, a robust version (RHAB and RHAE) or a nonrobust version (HAB and HAE) was used to link the item parameters that were obtained from separate calibrations within each country. For all analyses, the R software (R Core Team, 2019) and the R packages sirt (Robitzsch, 2019) and TAM (Robitzsch, Kiefer, & Wu, 2019) were used.

In each scaling strategy, the mean of the first country was set to zero and the standard deviation was set to one. For the country comparisons, country means were linearly transformed so that the total population of students across countries had a mean of zero and a standard deviation of one. We used two criteria to evaluate the different approaches: average absolute bias and average root mean square error (RMSE) across countries. Average absolute bias was determined by calculating the average of the absolute bias of each country mean (i.e., absolute deviation of estimated country mean from true country mean) across countries. Average absolute biases greater than .03 were considered as substantial because standard errors of country means in ILSAs are usually about that size (e.g., OECD, 2017). The average RMSE was calculated by averaging the RMSEs across countries; this indicates how stably the country means were estimated.

## Results

Table 4 shows the average absolute bias for the conditions with 40 items and a sample size of $N = 1{,}000$ in the case of balanced DIF (i.e., DIF effects of biased items sum to zero within each country). As can be seen, the FI as well as the HAB and HAE approaches, which did not remove items from country comparisons, produced approximately unbiased estimates of country means. Furthermore, the partial invariance approaches (PI-DIF and PI-MD) performed only slightly worse than the FI, HAB, and HAE approaches and substantial differences were only observed in conditions with a large standard deviation for the DIF effects of the anchor items (i.e., $SD_A = .3$). The robust linking approaches (RHAB and RHAE) performed similarly to the partial invariance approaches. However, the RHAB approach produced substantially biased country comparisons when the $SD_A$ for the DIF effects of the anchor items was large.

Table 5 shows the average RMSE for the same conditions (i.e., 40 items and a sample size of $N = 1{,}000$ for the balanced DIF condition). Overall, the results closely match the findings for the bias. First, country mean estimates that were produced by the approaches that were based on separate calibrations of the item parameters (HAB and HAE) were not less stable than the estimates of the full invariance approach (FI), and even outperformed the partial invariance approaches when the $SD_A$ was large (i.e., strong variation of DIF effects in the anchor item set). Second, the partial invariance approaches that allowed country-specific item parameters by using different cutoff values for DIF statistics only provided more stable estimates of country means than the FI approach when there were no DIF effects for anchor items (i.e., the $SD_A = 0$). In addition, the performance of the partial invariance approach depended on the choice of the specific cutoff value for the DIF statistic. For example, cutoff values of .05 and .08 for the MD statistic – which result in more country-specific item parameters – outperform a cutoff value of .12.

Table 6 shows the average absolute bias for the conditions with 40 items and a sample size of $N = 1{,}000$ in the case of unbalanced DIF (i.e., all DIF effects of biased items were either positive with a value of $\delta$ or negative with a value of $-\delta$ for each country). The country mean estimates produced by the FI and nonrobust linking approaches (HAB and HAE), which did not remove any items from the comparisons, were biased. This bias was substantially reduced with partial invariance approaches (PI-DIF and PI-MD) and robust linking approaches (RHAB and RHAE) if the DIF effect size $\delta$ of biased items was large relative to the standard deviation $SD_A$ of the DIF effects of anchor items (e.g., for $\delta = .6$, conditions with $SD_A \leq .2$). It is interesting that the robust linking methods based on separate calibration even outperformed partial invariance approaches based on concurrent calibration in many conditions. However, robust Haberman linking had worse performance in the presence of strong DIF effects for anchor items (i.e., $SD_A = .3$). It is also evident from Table 6 that the choice of cutoff values is crucial for partial invariance approaches. The partial invariance approach based on the MD statistic performed better with cutoff values of .05 or .08 than with .12. Overall, cutoff values have to be chosen that are smaller than the size of the absolute values of the DIF effects of biased items, in order to remove biased items from the comparison of country means.

Figure 2 shows the influence of sample size on the performance of the selected linking methods in the case of unbalanced DIF. It can be concluded that the general findings for $N = 1{,}000$ can also be transferred to smaller sample sizes of $N = 200$ and $N = 500$. Concurrent calibration under full invariance (FI) and nonrobust linking (HAE) were also more biased than partial invariance approaches and robust linking (RHAE) in smaller samples. If there were no DIF effects in anchor items ($SD_A = 0$), RHAE even outperformed partial invariance approaches based on the MD statistic (PI-MD), with cutoffs of .08 and .12. However, the PI-MD approach with a cutoff of .08 performed slightly better than the RHAE approach in the condition $SD_A = 0.2$. Importantly, approaches based on separate calibration (HAE, RHAE) were not less stable than concurrent calibration (FI, PI) even with a small sample size of $N = 200$. Hence, the different performance of linking methods with respect to average RMSE was mainly driven by average absolute bias.

## Empirical example: cross-sectional country comparisons for reading in PISA 2006

In order to illustrate the different approaches to estimating country means, we analyzed the data from the PISA 2006 assessment. In this reanalysis, we included all 26 OECD countries that participated in 2006 and focused on the reading domain, which was a minor domain in PISA 2006. Thus, reading items were only administered to a subset of the participating students, and we included only those students who received a test booklet with at least one reading item. This resulted in a total sample size of 110,236 students (ranging from 2,010 to 12,142 between countries). In total, 28 reading items nested within eight testlets were used in PISA 2006. As some items were polytomous, a partial credit model (PCM) was used in the analysis and the item difficulty of the PCM was used for linking. We specified eight different linking methods to obtain estimates of

**Table 4:**
Average Absolute Bias of Group Means as a Function of Proportion of DIF items, DIF Effect Size of Biased Items, and Standard Deviation of DIF Effects for Anchor Items for a Sample Size of $N = 1{,}000$, $I = 40$ Items, and Balanced DIF

| | | | **PI-DIF** | | **PI-MD** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| size | $SD_A$ | FI | .4 | .6 | .05 | .08 | .12 | HAB | HAE | RHAB | RHAE |
| *No biased items* | | | | | | | | | | | |
| 0 | 0 | .002 | .002 | .002 | .002 | .002 | .002 | .002 | .002 | .002 | .002 |
| 0 | .1 | .004 | .004 | .004 | .005 | .004 | .004 | .002 | .005 | .003 | .006 |
| 0 | .2 | .006 | .011 | .007 | .012 | .010 | .007 | .002 | .009 | .021 | .013 |
| 0 | .3 | .009 | .021 | .017 | .019 | .020 | .014 | .002 | .013 | **.069** | .025 |
| *10 % biased items* | | | | | | | | | | | |
| .3 | 0 | .003 | .004 | .003 | .003 | .003 | .003 | .002 | .004 | .002 | .002 |
| .3 | .1 | .004 | .004 | .004 | .005 | .003 | .004 | .001 | .005 | .004 | .006 |
| .3 | .2 | .006 | .012 | .007 | .014 | .009 | .008 | .002 | .009 | **.035** | .016 |
| .3 | .3 | .009 | .022 | .017 | .026 | .019 | .013 | .002 | .014 | **.073** | .028 |
| .6 | 0 | .007 | .002 | .010 | .002 | .004 | .006 | .002 | .008 | .002 | .002 |
| .6 | .1 | .007 | .003 | .010 | .003 | .005 | .006 | .002 | .009 | .004 | .005 |
| .6 | .2 | .008 | .013 | .011 | .014 | .012 | .008 | .002 | .011 | .028 | .016 |
| .6 | .3 | .010 | .023 | .023 | .027 | .022 | .016 | .002 | .014 | **.094** | .028 |
| 1 | 0 | .011 | .002 | .002 | .002 | .003 | .008 | .002 | .014 | .002 | .002 |
| 1 | .1 | .012 | .004 | .003 | .004 | .003 | .009 | .002 | .015 | .003 | .005 |
| 1 | .2 | .013 | .012 | .007 | .017 | .009 | .010 | .002 | .016 | .025 | .016 |
| 1 | .3 | .014 | .023 | .019 | **.031** | .022 | .017 | .001 | .018 | **.084** | .028 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *30 % biased items* | | | | | | | | | | | |
| .3 | 0 | .005 | .007 | .005 | .011 | .006 | .005 | .002 | .006 | .011 | .006 |
| .3 | .1 | .005 | .007 | .005 | .008 | .006 | .005 | .002 | .007 | .013 | .009 |
| .3 | .2 | .007 | .014 | .008 | .014 | .012 | .008 | .002 | .011 | **.040** | .019 |
| .3 | .3 | .009 | .024 | .020 | .018 | .024 | .018 | .002 | .013 | **.055** | **.031** |
| .6 | 0 | .007 | .011 | .027 | .003 | .017 | .023 | .002 | .010 | .002 | .005 |
| .6 | .1 | .008 | .009 | .027 | .008 | .014 | .023 | .002 | .012 | .005 | .010 |
| .6 | .2 | .008 | .016 | .023 | .018 | .019 | .022 | .002 | .013 | **.035** | .021 |
| .6 | .3 | .011 | .026 | .030 | .030 | .026 | .027 | .002 | .016 | **.071** | **.032** |
| 1 | 0 | .013 | .002 | .009 | .002 | .004 | .026 | .002 | .018 | .002 | .005 |
| 1 | .1 | .013 | .006 | .006 | .010 | .007 | .022 | .001 | .019 | .004 | .010 |
| 1 | .2 | .014 | .019 | .010 | .019 | .018 | .021 | .002 | .020 | .029 | .020 |
| 1 | .3 | .015 | **.037** | .025 | .029 | **.034** | **.032** | .002 | .023 | **.056** | **.033** |

*Note*. size = size of DIF effect for biased items; $SD_A$ = standard deviations of DIF effects for anchor items; FI = concurrent calibration (CC) assuming full invariance; PI-DIF = CC based on partial invariance with cutoffs for DIF statistic in logit metric; PI-MD = CC based on partial invariance with cutoffs for MD statistic; HAB = Haberman linking; HAE = Haebara linking; RHAB = robust Haberman linking; RHAE = robust Haebara linking. Values larger than .03 are printed in bold.

**Table 5:**
Average Root Mean Squared Error (RMSE) of Group Means as a Function of Proportion of DIF items, DIF Effect Size of Biased Items, and Standard Deviation of DIF Effects for Anchor Items for a Sample Size of $N$ = 1,000, $I$ = 40 Items, and Balanced DIF

|  |  |  | **PI-DIF** |  | **PI-MD** |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
| size | $SD_A$ | FI | .4 | .6 | .05 | .08 | .12 | HAB | HAE | RHAB | RHAE |
| *No biased items* |  |  |  |  |  |  |  |  |  |  |  |
| 0 | 0 | .030 | .030 | .030 | .030 | .030 | .030 | .030 | .030 | .030 | .031 |
| 0 | .1 | .031 | .031 | .031 | .032 | .031 | .031 | .031 | .031 | .031 | .033 |
| 0 | .2 | .032 | .036 | .032 | .037 | .035 | .032 | .031 | .033 | **.045** | .040 |
| 0 | .3 | .032 | **.043** | .039 | **.042** | **.043** | .036 | .031 | .034 | **.105** | **.047** |
| *10 % biased items* |  |  |  |  |  |  |  |  |  |  |  |
| .3 | 0 | .031 | .031 | .031 | .032 | .031 | .031 | .031 | .031 | .032 | .032 |
| .3 | .1 | .031 | .031 | .031 | .032 | .031 | .031 | .031 | .031 | .032 | .034 |
| .3 | .2 | .031 | .036 | .032 | .038 | .034 | .032 | .031 | .032 | **.058** | **.041** |
| .3 | .3 | .032 | **.044** | .038 | **.046** | **.043** | .036 | .030 | .034 | **.108** | **.052** |
| .6 | 0 | .032 | .031 | .036 | .031 | .032 | .034 | .031 | .032 | .031 | .032 |
| .6 | .1 | .031 | .030 | .035 | .031 | .032 | .032 | .030 | .031 | .031 | .033 |
| .6 | .2 | .031 | .036 | .036 | .038 | .035 | .033 | .030 | .032 | **.053** | **.041** |
| .6 | .3 | .032 | **.045** | **.044** | **.048** | **.045** | .039 | .030 | .034 | **.139** | **.052** |
| 1 | 0 | .033 | .031 | .031 | .031 | .031 | .033 | .031 | .034 | .031 | .032 |
| 1 | .1 | .033 | .031 | .030 | .031 | .031 | .033 | .030 | .034 | .031 | .033 |
| 1 | .2 | .033 | .035 | .031 | .039 | .034 | .034 | .030 | .035 | **.049** | **.041** |
| 1 | .3 | .034 | **.045** | .040 | **.050** | **.044** | .039 | .030 | .036 | **.118** | **.052** |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *30 % biased items* | | | | | | | | | | | |
| .3 | 0 | .031 | .033 | .031 | .036 | .032 | .031 | .031 | .031 | .036 | .033 |
| .3 | .1 | .030 | .032 | .030 | .035 | .031 | .030 | .030 | .031 | .039 | .036 |
| .3 | .2 | .032 | .038 | .032 | **.041** | .036 | .032 | .031 | .033 | **.076** | **.044** |
| .3 | .3 | .032 | **.045** | .040 | **.044** | **.044** | .039 | .030 | .034 | **.098** | **.053** |
| .6 | 0 | .031 | .035 | **.049** | .032 | .040 | **.044** | .030 | .032 | .031 | .032 |
| .6 | .1 | .032 | .036 | **.050** | .035 | .040 | **.045** | .031 | .034 | .034 | .037 |
| .6 | .2 | .032 | **.041** | **.047** | **.042** | **.043** | **.045** | .031 | .034 | **.071** | **.045** |
| .6 | .3 | .033 | **.051** | **.054** | **.052** | **.052** | **.049** | .030 | .036 | **.149** | **.054** |
| 1 | 0 | .034 | .031 | .035 | .032 | .032 | **.047** | .031 | .036 | .032 | .033 |
| 1 | .1 | .033 | .032 | .033 | .035 | .033 | **.044** | .030 | .036 | .032 | .036 |
| 1 | .2 | .034 | **.041** | .034 | **.042** | **.041** | **.045** | .031 | .038 | **.053** | **.044** |
| 1 | .3 | .035 | **.056** | **.048** | **.052** | **.055** | **.055** | .031 | .040 | **.099** | **.056** |

*Note*. size = size of DIF effect for biased items; $SD_A$ = standard deviations of DIF effects for anchor items; FI = concurrent calibration (CC) assuming full invariance; PI-DIF = CC based on partial invariance with cutoffs for DIF statistic in logit metric; PI-MD = CC based on partial invariance with cutoffs for MD statistic; HAB = Haberman linking; HAE = Haebara linking; RHAB = robust Haberman linking; RHAE = robust Haebara linking. Values larger than .040 are printed in bold.

**Table 6:**
Average Absolute Bias of Group Means as a Function of Proportion of DIF items, DIF Effect Size of Biased Items, and Standard Deviation of DIF Effects for Anchor Items for a Sample Size of $N = 1{,}000$, $I = 40$ Items, and Unbalanced DIF

| | | | **PI-DIF** | | **PI-MD** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| size | $SD_A$ | FI | .4 | .6 | .05 | .08 | .12 | HAB | HAE | RHAB | RHAE |
| *No biased items* | | | | | | | | | | | |
| 0 | 0 | .002 | .002 | .002 | .002 | .002 | .002 | .002 | .002 | .002 | .003 |
| 0 | .1 | .004 | .004 | .004 | .005 | .004 | .004 | .002 | .005 | .003 | .006 |
| 0 | .2 | .006 | .011 | .006 | .011 | .010 | .007 | .002 | .009 | .019 | .013 |
| 0 | .3 | .008 | .023 | .017 | .019 | .021 | .014 | .002 | .013 | **.069** | .025 |
| *10 % biased items* | | | | | | | | | | | |
| .3 | 0 | .028 | .027 | .028 | .016 | .028 | .028 | .030 | .028 | .016 | .011 |
| .3 | .1 | .027 | .025 | .027 | .019 | .026 | .027 | .029 | .026 | .019 | .018 |
| .3 | .2 | .029 | **.037** | .029 | **.038** | **.034** | .028 | **.031** | .029 | **.052** | **.037** |
| .3 | .3 | .029 | **.053** | **.039** | **.048** | **.050** | **.035** | .030 | .029 | **.099** | **.055** |
| .6 | 0 | **.054** | .005 | **.045** | .004 | .017 | **.051** | **.059** | **.054** | .003 | .010 |
| .6 | .1 | **.055** | .006 | **.045** | .012 | .017 | **.052** | **.060** | **.054** | .004 | .019 |
| .6 | .2 | **.055** | .024 | **.045** | **.039** | .027 | **.050** | **.059** | **.055** | .029 | **.037** |
| .6 | .3 | **.056** | **.053** | **.062** | **.054** | **.055** | **.059** | **.059** | **.056** | **.093** | **.059** |
| 1 | 0 | **.091** | .001 | .001 | .004 | .004 | .017 | **.100** | **.089** | .002 | .011 |
| 1 | .1 | **.091** | .006 | .003 | .020 | .006 | .018 | **.099** | **.089** | .004 | .018 |
| 1 | .2 | **.093** | **.034** | .009 | **.058** | .026 | .020 | **.101** | **.091** | .024 | **.039** |
| 1 | .3 | **.093** | **.067** | **.032** | **.081** | **.059** | **.034** | **.100** | **.091** | **.075** | **.058** |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *30 % biased items* | | | | | | | | | |
| .3 | 0 | **.087** | **.087** | **.087** | **.069** | **.087** | **.087** | **.088** | **.088** | **.062** | **.044** |
| .3 | .1 | **.088** | **.088** | **.088** | **.085** | **.088** | **.088** | **.089** | **.088** | **.082** | **.072** |
| .3 | .2 | **.087** | **.113** | **.088** | **.119** | **.106** | **.088** | **.089** | **.087** | **.149** | **.118** |
| .3 | .3 | **.087** | **.144** | **.110** | **.139** | **.140** | **.102** | **.090** | **.086** | **.211** | **.156** |
| .6 | 0 | **.177** | **.086** | **.173** | **.039** | **.112** | **.175** | **.179** | **.175** | .002 | **.046** |
| .6 | .1 | **.175** | **.093** | **.172** | **.069** | **.115** | **.174** | **.177** | **.173** | .006 | **.074** |
| .6 | .2 | **.174** | **.142** | **.180** | **.123** | **.156** | **.178** | **.177** | **.172** | **.137** | **.129** |
| .6 | .3 | **.175** | **.183** | **.222** | **.162** | **.196** | **.211** | **.178** | **.173** | **.227** | **.189** |
| 1 | 0 | **.286** | .022 | **.080** | **.077** | **.036** | **.136** | **.290** | **.280** | .002 | **.046** |
| 1 | .1 | **.286** | **.054** | **.081** | **.121** | **.061** | **.133** | **.291** | **.280** | .005 | **.076** |
| 1 | .2 | **.285** | **.120** | **.115** | **.183** | **.118** | **.155** | **.290** | **.279** | **.031** | **.131** |
| 1 | .3 | **.285** | **.182** | **.175** | **.226** | **.176** | **.207** | **.289** | **.278** | **.065** | **.191** |

*Note*. size = size of DIF effect for biased items; $SD_A$ = standard deviations of DIF effects for anchor items; FI = concurrent calibration (CC) assuming full invariance; PI-DIF = CC based on partial invariance with cutoffs for DIF statistic in logit metric; PI-MD = CC based on partial invariance with cutoffs for MD statistic; HAB = Haberman linking; HAE = Haebara linking; RHAB = robust Haberman linking; RHAE = robust Haebara linking. Values larger than .03 are printed in bold.

**Table 7:**
Country Means for the Reading Domain for PISA 2006 for 26 Selected OECD Countries

| | | | | | | | **PI-MD** | | **PI-DIF** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Country | $SD_{DIF}$ | $SD_{MD}$ | MD+ | MD− | Rg | FI | .12 | .08 | .5 | HAB | HAE | RHAB | RHAE |
| AUS | .36 | .06 | 0 | 0 | 8 | 515 | 515 | 516 | 517 | 519 | 516 | 523 | 520 |
| AUT | .29 | .06 | 1 | 1 | 4 | 494 | 494 | 495 | 493 | 496 | 496 | 497 | 495 |
| BEL | .26 | .04 | 0 | 0 | 3 | 504 | 504 | 504 | 505 | 503 | 505 | 505 | 506 |
| CAN | .36 | .07 | 1 | 2 | 9 | 530 | 534 | 533 | 537 | 528 | 528 | 535 | 529 |
| CHE | .36 | .07 | 0 | 2 | 9 | 499 | 504 | 505 | 499 | 501 | 502 | 508 | 503 |
| CZE | .36 | .06 | 0 | 1 | 6 | 483 | 486 | 480 | 480 | 485 | 484 | 486 | 483 |
| DEU | .43 | .07 | 1 | 1 | 10 | 494 | 495 | 500 | 498 | 492 | 495 | 502 | 500 |
| DNK | .38 | .10 | 2 | 3 | 7 | 499 | 502 | 501 | 506 | 503 | 502 | 503 | 505 |
| ESP | .46 | .07 | 0 | 3 | 10 | 465 | 473 | 467 | 468 | 468 | 466 | 475 | 470 |
| EST | .43 | .08 | 1 | 4 | 12 | 504 | 511 | 509 | 508 | 504 | 500 | 512 | 510 |
| FIN | .39 | .08 | 0 | 2 | 8 | 548 | 556 | 553 | 549 | 556 | 552 | 549 | 552 |
| FRA | .36 | .06 | 0 | 2 | 7 | 495 | 500 | 500 | 496 | 496 | 499 | 493 | 499 |
| GBR | .41 | .07 | 2 | 2 | 6 | 498 | 496 | 496 | 492 | 497 | 498 | 496 | 496 |
| GRC | .49 | .13 | 2 | 2 | 14 | 467 | 458 | 459 | 453 | 460 | 461 | 456 | 454 |
| HUN | .32 | .05 | 0 | 0 | 11 | 487 | 487 | 489 | 490 | 484 | 484 | 479 | 486 |
| IRL | .33 | .06 | 1 | 0 | 7 | 520 | 518 | 519 | 520 | 520 | 517 | 513 | 515 |
| ISL | .37 | .07 | 0 | 3 | 17 | 491 | 500 | 495 | 496 | 494 | 492 | 508 | 495 |
| ITA | .38 | .07 | 2 | 1 | 8 | 475 | 472 | 470 | 478 | 472 | 473 | 472 | 471 |
| JPN | .56 | .10 | 4 | 3 | 15 | 500 | 494 | 489 | 495 | 499 | 502 | 487 | 492 |
| KOR | .49 | .07 | 4 | 0 | 15 | 560 | 545 | 551 | 558 | 554 | 557 | 545 | 554 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LUX | .29 | .05 | 0 | 1 | 6 | 482 | 484 | 480 | 480 | 481 | 482 | 485 | 486 |
| NLD | .49 | .09 | 2 | 1 | 12 | 507 | 502 | 505 | 504 | 506 | 511 | 501 | 499 |
| NOR | .40 | .10 | 2 | 2 | 15 | 487 | 483 | 486 | 481 | 487 | 486 | 472 | 485 |
| POL | .37 | .07 | 1 | 1 | 7 | 510 | 509 | 512 | 514 | 509 | 509 | 516 | 513 |
| PRT | .49 | .10 | 2 | 2 | 5 | 476 | 472 | 475 | 472 | 476 | 474 | 475 | 471 |
| SWE | .31 | .06 | 1 | 0 | 6 | 511 | 507 | 510 | 509 | 510 | 510 | 509 | 513 |

*Note*. $SD_{DIF}$ = DIF standard deviation; $SD_{MD}$ = standard deviation of MD statistics; MD+ = number of items with MD statistic larger than .12; MD− = number of items with MD statistic smaller than −.12; Rg = range of country mean estimates among different linking methods; FI = concurrent calibration (CC) assuming full invariance; PI-DIF = CC based on partial invariance with cutoffs for DIF statistic in logit metric; PI-MD = CC based on partial invariance with cutoffs for MD statistic; HAB = Haberman linking; HAE = Haebara linking; RHAB = robust Haberman linking; RHAE = robust Haebara linking.

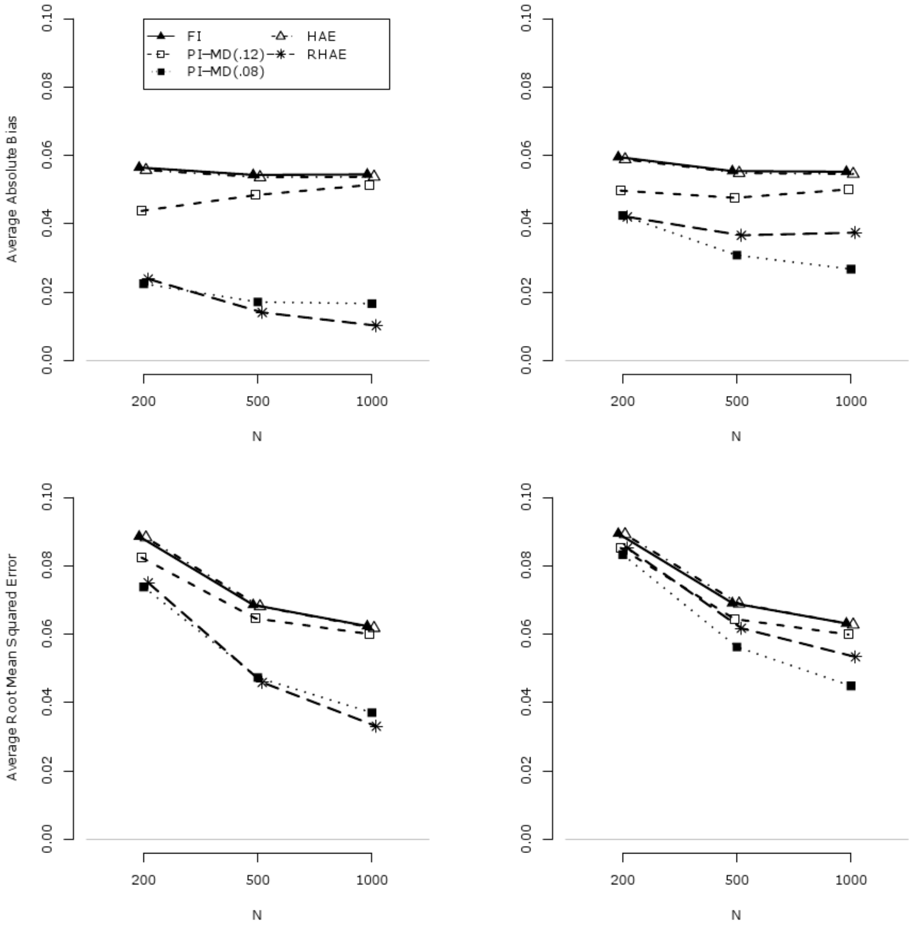

**Figure 2:**
Average absolute bias (upper panels) and average root mean squared error (lower panels) for unbalanced DIF for 10 % biased items with a DIF effect size of .6, $I = 40$ items, for a standard deviation of DIF effects for anchor items of $SD_A = 0$ (left panels) and $SD_A = 0.2$ (right panels) as a function of sample size.

country means: a full invariance approach (concurrent calibration with multiple groups), a partial invariance approach with DIF detection based on the MD statistic or the DIF effect size at the logit metric, and two nonrobust linking methods (Haberman and Haebara) as well as two robust linking methods (RHAB and RHAE). For all analyses, student weights within a country were normalized to a sum of 5,000, so that all countries contributed equally to the analyses. Finally, all estimated country means were linearly transformed such that the distribution containing all (weighted) students in all 26 countries had a mean of 500 (points) and a standard deviation of 100. Note that this transformation is not equivalent to the one used in officially published PISA data.

In Table 7, the country mean estimates obtained from the eight different linking methods are shown. Within a country, the range of country means differed between 3 and 17 points ($M = 9.1$) across the different methods. These differences between the linking methods can be explained by different amounts of country DIF. Furthermore, there was a strong cross-country correlation of .75 between the standard deviation of the DIF statistic in the logit metric and the standard deviation of the MD statistic. It is instructive to first focus on the comparison of country means based on the assumption of full invariance in a concurrent calibration approach that ignores DIF (similar to the PISA method used until 2012) and the partial invariance approach based on the MD statistic with a cutoff of .12 (similar to the PISA method starting from 2015, but note that PISA used the 2PL model instead of the 1PL model). About 9.4 % of all items across countries exceeded an absolute value of .12 for the MD statistic and there was an average absolute difference of 4.2 points between the two approaches, with a maximum discrepancy of 15 points (South Korea, KOR). As shown in Table 7, South Korea had four flagged DIF items with an MD statistic larger than .12, while there was no flagged DIF item with an MD statistic smaller than −.12. In the partial invariance approach, those four items that advantaged South Korea were excluded from the linking, which explains the drop of 15 points in the partial invariance approach (545 points) compared to the full invariance approach (560 points). A number of countries also had many more flagged DIF items with negative MD statistics than with positive MD statistics (Spain, ESP; Estonia, EST; Island, ISL). In these cases, items that were disadvantageous for a country were removed from the linking in the partial invariance approach. Overall, it can be concluded that the magnitude of the difference between the full and partial invariance approaches was similar to that of the standard errors caused by person sampling (about 3 points). Hence, the choice of a particular linking method is of practical relevance for at least some of the countries (but see Jerrim et al., 2018, for a similar analysis with the PISA 2015 data).

Table 8 shows the average absolute differences and correlations of the country mean estimates for the different linking methods. First, it needs to be emphasized that even a high correlation of .991 of the country means, which was provided by different methods (full invariance and Haberman linking), can result in a nonnegligible average absolute difference of 2.3 points (with a maximum of 8 points for Finland, FIN). Second, based on the pattern of correlations, the linking methods can be summarized into different groups that produced similar results. Concurrent calibration under full invariance and nonrobust linking methods (Haberman and Haebara) performed relatively similarly ($r$ = .991, .993, .995, respectively). This can be explained by the fact that neither approach removes items from the linking. Moreover, the partial invariance approaches based on the MD statistic with cutoffs of .12 and .08 performed similarly to the robust Haebara approach ($r$ = .990, .988, .991, respectively). Finally, when interpreting the results, it needs to be taken into account that the observed discrepancies in country means could be smaller for more recent PISA assessments as the number of items in a domain has been substantially increased in the recent PISA assessments.

**Table 8:**
Average Absolute Differences (Upper Diagonal) and Correlations (Lower Diagonal) for Different Linking Methods for Reading Domain in PISA 2006 for 26 Selected OECD Countries

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1: FI | | 4.2 | 3.5 | 3.6 | **2.3** | **2.2** | 6.6 | 4.3 |
| 2: PI-MD .12 | .968 | | **2.6** | 3.7 | 3.2 | 3.6 | 3.7 | **3.0** |
| 3: PI-MD .08 | .981 | **.990** | | 3.0 | **2.8** | **3.1** | 4.5 | **2.5** |
| 4: PI-DIF .5 | .982 | .983 | .987 | | 3.7 | 4.0 | 5.2 | 3.2 |
| 5: HAB | **.991** | .985 | .987 | .982 | | **1.8** | 5.5 | 3.2 |
| 6: HAE | **.993** | .975 | .982 | .980 | **.995** | | 6.0 | 3.4 |
| 7: RHAB | .934 | .975 | .967 | .964 | .953 | .941 | | 4.1 |
| 8: RHAE | .975 | .988 | **.991** | .987 | .985 | .979 | .973 | |

*Note*. FI = concurrent calibration (CC) assuming full invariance; PI-DIF = CC based on partial invariance with cutoffs for DIF statistic in logit metric; PI-MD = CC based on partial invariance with cutoffs for MD statistic; HAB = Haberman linking; HAE = Haebara linking; RHAB = robust Haberman linking; RHAE = robust Haebara linking. Absolute differences smaller than 3.0 and correlations larger than .990 are printed in bold.

## Discussion

In this article, we compared different linking approaches for computing group means (or country means) in the presence of DIF effects. The main goal was to investigate the conditions under which the partial invariance approach produces more accurate estimates of country means than the full invariance and noninvariance approaches. First, we argued that it is instructive for the understanding of DIF to differentiate between two group-specific item sets: anchor items and biased items. It was emphasized that the unbiased identification of group means can only be conducted based on the group-specific set of anchor items. Importantly, both anchor items, as well as biased items, are allowed to possess DIF effects, but the DIF effects of anchor items average to zero within a group. Furthermore, we discussed three different approaches for comparing country means in the presence of DIF (full invariance, partial invariance, and linking based on separate calibration) and showed analytically that these approaches place different constraints on the DIF effects for anchor items and biased items. Through a simulation study, it was shown that the performance of the different linking approaches depended on the nature of the DIF effects of anchor and biased items. If the DIF effects of biased items were not balanced, the concurrent calibration approach based on full invariance as well as the nonrobust linking methods (Haberman, Haebara) provided biased country means, whereas the partial invariance approaches based on the DIF and MD statistics were able to reduce the bias in the country means. However, in several conditions of the simulation study, the cutoff value of .08 for the MD statistic performed better than a cutoff value of .12. Moreover, robust linking approaches (robust Haberman, robust Haebara) even outperformed partial invariance approaches in some conditions. Thus, it could be questioned whether the partial invariance approaches always produce more stable estimates of group means in the context of ILSAs than linking approaches based on separate calibration[14]. Interestingly, it has been shown that separate calibration with subsequent linking (i.e., noninvariance approach) can be reformulated as a concurrent scaling approach in a multiple-group IRT model with side conditions (von Davier & von Davier, 2007).

As it is true for all simulation studies, conclusions are limited to the conditions which were investigated in our study. First, we restricted ourselves to 20 or 40 items, while ILSAs include a much larger item pool in general. Further studies could investigate larger numbers of items or could treat items as random instead as fixed. Second, we assumed that the maximum proportion of biased items was 30%. We would believe that the test construction was not successful if more than 30% of the items would be biased items (Magis & De Boeck, 2011). Third, we only chose two extreme conditions of DIF for biased items, namely, the case of balanced DIF in which the DIF effects of biased items sum to zero and the case of unbalanced DIF in which all items either have a joint positive or negative biasing DIF effect $\delta$. In reality, DIF effects of biased items likely follow a distribution between these two extreme scenarios. These constellations should be investigated in future simulation studies, which should also provide more practical guidelines for choosing among different scaling approaches in the presence of biased items. Fourth, our simulation design intended to mimic data constellations that are characteristic for the assessment of cognitive constructs. However, the assessment of DIF for non-cognitive constructs (e.g., student motivation) in ILSAs has received increasing attention (Avvisati, Le Donné, & Paccagnella, 2019; Buchholz & Hartig, 2019; Cieciuch et al., 2019; He, Barrera-Pedemonte, & Buchholz, 2019; Rutkowski & Svetina, 2017; Zieger, Jerrim, & Sims, 2019). It would be interesting to extend the design of our simulation to conditions that are more realistic for non-cognitive constructs (for example, polytomous item responses with four categories, and scales with three to ten items).

When discussing the reasoning behind the DIF concept, we emphasized that the decision whether an item induces bias for country comparisons is not purely (or maybe even not primarily) a statistical question. As it was pointed out by Camilli (1993; see also Penfield & Camilli, 2007), DIF detection procedures should be accompanied by expert reviews of items showing DIF. Only those items should be removed from country comparisons for which it is defensible to argue that DIF was caused by construct irrelevant factors (see also El Masri & Andrich, 2020). Until PISA 2015, items with DIF effects were only declared as DIF items and omitted from scaling for a particular country if translation issues were confirmed (Adams, 2003; see also Kreiner & Christensen, 2014). This approach practically ignores DIF unless it could be demonstrated that DIF was construct irrelevant. On the other hand, a purely statistical approach (e.g., based on partial invariance used since PISA 2015 or robust linking) ignores the fact that DIF items could be construct relevant. Including country-specific item parameters for construct relevant DIF items has the potential to result in construct underrepresentation for country compari-

[14] Note that in separate calibration estimated item parameters rely on smaller sample sizes. This increased uncertainty in estimated item parameters could, however, affect other parameters involving the ability distribution (e.g., quantiles or regression coefficients; see, e.g., Tsutakawa & Johnson, 1990).

sons. However, within-country analyses are not affected by this potential construct underrepresentation because no items are removed from scaling.

As the discussion of the concept of DIF for the 1PL model made clear, the identification of country means requires non-testable identification constraints. Hence, no analytical reasoning or simulation studies can be used to prove that partial invariance or robust linking methods (robust Haberman and robust Haebara) outperform a full invariance approach that practically ignores DIF effects in scaling in practical applications like PISA. Unfortunately, identification constraints and their ambiguous consequences are often neglected in the psychometric literature. For example, it is often argued that at least partial invariance for item intercepts is needed to allow meaningful (or valid) comparisons of group means (e.g., van de Vijver, 2019) in order to avoid the issue of the comparison of apples and oranges. It is important to emphasize that the labels "meaningful" or "valid" are not clearly defined statistical terms and, hence, these labels are often confounded with the concept of statistical bias (see, for example, He et al., 2019). We showed analytically and through a simulation study that unbiased estimates of country means can be obtained if noninvariance does hold (i.e., all items have DIF effects, but are anchor items). Moreover, one critical aspect of the partial invariance approach (as well as other approaches that result in the inclusion of country-specific item parameters or a down-weighting of items in a country, such as robust linking approaches) is that comparisons of different groups do not rely on a full set of common item parameters as a subset of items receives unique item parameters which are not comparable across most groups. For example, the country mean comparison of Germany and Poland in PISA does not involve a full set of common item parameters for each country if the sets of country-specific noninvariant items – that receive country-specific item parameters – differ between the two countries. More critically, the determination of how a country comparison is conducted (i.e., which items are used as anchor items in a country) is – in the current operational use since PISA 2015 – defined by means of a model misfit in a psychometric model (see von Davier et al., 2019)[15]. Hence, we think that this property of the partial invariance approach does not resolve the problem of comparing apples and oranges because it is also a potential threat to validity (i.e., the construct is not properly represented; see also Kuha & Moustaki, 2015, and El Masri & Andrich, 2020). In contrast, approaches using full invariance or complete noninvariance (non-robust linking methods based on separate calibration; Haberman and Haebara methods) use the full set of common item parameters for country comparisons and are less affected by construct representation concerns. In more detail, the full invariance approach uses the full set of common item parameters. The non-robust linking approaches allow for country-specific

[15] Brennan (1998, p. 8) argues: "[…] it is inappropriate to allow a scaling model to be the sole determiner of how content and formats are weighted in arriving at scores. To do so is to delegate to statistical models a responsibility that, at a minimum, should be shared with test developers and policymakers." He proceeds: "the role of scaling in drawing inferences about test scores is one of the most neglected aspects of validation, and the notion that scaling is (or should be) solely a psychometric matter may be the single most widely held misconception about measurement". This underlines the importance of the alignment of the test specification and the statistical methodology for test analysis. It seems that, in large-scale assessments, these two goals are often disconnected.

item parameters but do not involve a down-weighting of items in particular countries as in robust linking approaches.

However, it could be debated whether the application of the partial invariance approach could threaten the validity of country comparisons (as pointed out by an anonymous reviewer who disagreed with this statement). In the partial invariance approach, a small subset of items receives country-specific item parameters in the scaling model. Hence, these unique items are removed from the likelihood estimation for estimating common international item parameters but still provide additional information for ability estimation at the level of countries. Therefore, one could argue that the inclusion of unique item parameters leads to a slightly decreased comparability of countries. However, this does not threaten the validity of cross-country comparisons as the full set of common item parameters is used in all countries, and the majority of these items receive common international item parameters.

Moreover, we think that it is informative to quantify the amount of noninvariance as an additional source of uncertainty in country comparisons. Linking errors are usually reported for trend estimates of countries in large-scale assessments (OECD, 2017; Robitzsch & Lüdtke, 2019; Wu, 2010). Because country DIF introduces decreased comparability of countries, appropriate linking errors reflecting the extent of noninvariance could also be defined for country means and country mean differences.

A significant limitation of our study is that we only considered the 1PL model. Since PISA 2015, a 2PL model is in operational use, and the generalizability of our findings to the 2PL model could be questioned. Two different cases of DIF have to be distinguished in the 2PL model. First, in the case of only uniform DIF, DIF is only allowed in item intercepts, and the IRFs are given as $P_{ig}(\theta) = \Psi\left(a_i\left(\theta - b_i - e_{ig}\right)\right)$. From an analytical point of view, the same arguments can be applied to the 2PL model in this case because the identification issues of the 1PL model (regarding the item difficulties) can be directly translated to the 2PL model. As only uniform DIF is present, the common item slopes $a_i$ can be estimated in the total sample comprising all countries. Moreover, it is shown in Appendix D that MD and RMSD formulas are quite similar when allowing common item slopes (by replacing group-specific standard deviations $\sigma_g$ by $a_i\sigma_g$ in the corresponding formulas). Therefore, it would be interesting whether our recommendation for a more stringent use of a cutoff of .05 or .08 (instead of .12) for the MD or RMSD statistic could also be generalized to the 2PL model in a future simulation study. Second, the case of nonuniform DIF (i.e., DIF is also present in item slopes) requires additional research, and we think that it is not apparent whether all of our findings can be generalized to this case. While one could use similar cutoff values for MD and RMSD statistics, the linking approaches need further consideration by the inclusion of item slope parameters (see Battauz, 2017, for discussion of different variants of Haberman linking in the 2PL model).

We want to point out that the partial invariance approach has some similarity to regularization based estimation approaches for DIF (Tutz & Schauberger, 2015; see also Hastie et al., 2015, for a broad overview of regularized estimation). Both estimation approaches result in a small subset of DIF effects, which are estimated to be nonzero, while most of the items receive DIF effects of zero. However, regularization techniques do not rely on

DIF statistics but require the selection of additional tuning parameters. Furthermore, the invariance alignment procedure for multiple-group IRT item response models (Muthen & Asparouhov, 2014) could be regarded as an alternative linking method. The alignment method employs a loss function (i.e., $\rho(x) = \sqrt{|x|}$ ) in the linking optimization function, which imposes sparsity on DIF effects. It would be interesting to compare the alignment method to the fused lasso regularization approach (Hastie et al., 2015), which uses a penalty function that is quite similar to the optimization function in the alignment procedure. In the near future, we expect that regularization methods are more commonly applied in the psychometric literature because many of the problems that had been tackled with ad-hoc multi-step approaches could be handled in a much more principled way with regularization techniques (see for recent examples Bauer, Belzak, & Cole, 2020; Huang, 2019; Schauberger & Mair, 2020).

## Conclusion

In this article, we investigated the performance of full invariance, partial invariance, and full noninvariance (nonrobust and robust linking) approaches for country comparisons in large-scale assessments depending on different assumptions about the nature of DIF effects in the 1PL model. We found that in the presence of biased items, and balanced DIF (i.e., DIF effects of biased items average to zero), the full invariance and nonrobust linking approaches provide (approximately) unbiased country means with similar variability. In this case, the partial invariance and robust linking approaches could introduce slight biases by incorrectly removing items from country comparisons. However, in the presence of biased items, and unbalanced DIF (i.e., DIF effects of biased items do not average to zero), the full invariance and nonrobust linking approaches were biased, whereas the partial invariance approach with an appropriate cutoff value (for the DIF statistic) as well as the robust linking approaches performed similarly and strongly reduced the bias. These findings clearly showed the potential of the partial invariance and the robust linking approaches to increase the accuracy of country mean comparisons in the case of unbalanced DIF. However, we would also argue that the decision of whether an item should be considered as an anchor item or biased item always needs to take into account whether the corresponding DIF effect is construct relevant or construct irrelevant. Thus, it is not possible to base the choice of a psychometric model solely on statistical grounds.

## References

Adams, R. J. (2003). Response to ‘Cautions on OECD’s recent educational survey (PISA)’. *Oxford Review of Education, 29*(3), 379–389. doi: 10.1080/03054980307445

Adams, R. J., Wu, M. L., & Carstensen, C. H. (2007). Application of multivariate Rasch models in international large-scale educational assessments. In M. von Davier & C. H.

Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 271–280). New York: Springer. doi: 10.1007/978-0-387-49839-3_17

Avvisati, F., Le Donné, N., Paccagnella, M. (2019). A meeting report: Cross-cultural comparability of questionnaire measures in large-scale international surveys. *Measurement Instruments for the Social Sciences, 1*:8. doi: 10.1186/s42409-019-0010-z

Battauz, M. (2017). Multiple equating of separate IRT calibrations. *Psychometrika, 82*(3), 610–636. doi: 10.1007/s11336-016-9517-x

Bauer, D. J., Belzak, W. C., & Cole, V. T. (2020). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. *Structural Equation Modeling, 27*(1), 43–55. doi: 10.1080/10705511.2019.1642754

Bechger, T. M., & Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika, 80*(2), 317–340. doi: 10.1007/s11336-014-9408-y

Brennan, R. L. (1998). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issues and Practice, 17*(1), 5–9. doi: 10.1111/j.1745-3992.1998.tb00615.x

Buchholz, J., & Hartig, J. (2019). Comparing attitudes across groups: An IRT-based item-fit statistic for the analysis of measurement invariance. *Applied Psychological Measurement, 43*(3), 241–250. doi: 10.1177/0146621617748323

Cai, L., & Moustaki, I. (2018). Estimation methods in latent variable models for categorical outcome variables. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test* (pp. 253–277). New York: Wiley. doi: 10.1002/9781118489772.ch9

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 397–417). Hillsdale, NJ: Erlbaum.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Eds.), *Educational measurement* (pp. 221–256). Westport: Praeger Publisher.

Cieciuch, J., Davidov, E., Schmidt, P., & Algesheimer, R. (2019). How to obtain comparable measures for cross-national comparisons. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie, 71*(Suppl 1), 157–186. doi: 10.1007/s11577-019-00598-7

Davies, P. L. (2014). *Data analysis and approximate models*. Boca Raton: CRC Press. doi: 10.1201/b17146

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*(4), 533-559. doi: 10.1007/S11336-008-9092-X

DeMars, C. E. (2011) An analytic comparison of effect sizes for differential item functioning, *Applied Measurement in Education, 24*(3), 189–209. doi: 10.1080/08957347.2011.580255

DeMars, C. E. (2020). Alignment as an alternative to anchor purification in DIF analyses. *Structural Equation Modeling, 27*(1), 56–72. doi: 10.1080/10705511.2019.1617151

Doebler, A. (2019). Looking at DIF from a new perspective: A structure-based approach acknowledging inherent indefinability. *Applied Psychological Measurement, 43*(4), 303–321. doi: 10.1177/0146621618795727

El Masri, Y. H., & Andrich, D. (2020): The trade-off between model fit, invariance, and validity: The case of PISA science assessments. *Applied Measurement in Education*. Advance online publication. doi: 10.1080/08957347.2020.1732384

Fox, J. (2016). *Applied regression analysis and generalized linear models*. Thousand Oaks: Sage.

Fox, J.-P., & Verhagen, A. J. (2010). Random item effects modeling for cross-national survey data. In E. Davidov, P. Schmidt & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 461–482). London: Routledge Academic.

Gierl, M. J., Bisanz, J., Bisanz, G. L., Boughton, K. A., & Khaliq, S. N. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20*(2), 26–36. doi: 10.1111/j.1745-3992.2001.tb00060.x

Glas, C. A. W., & Jehangir, K. (2014). Modeling country-specific differential functioning. In L. Rutkowski, M. von Davier & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 97–115). Boca Raton: CRC Press. doi: 10.1201/b16061

Grisay, A., Gonzalez, E., & Monseur, C. (2009). Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 2*, 63–83.

Haberman, S. J. (2009). *Linking parameter estimates derived from an item response model through separate calibrations* (Research Report RR-09-40). Princeton, NJ. Educational Testing Service. doi: 10.1002/j.2333-8504.2009.tb02197.x

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*(3), 144–149. doi: 10.4992/psycholres1954.22.144

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity. The lasso and generalizations*. Boca Raton: CRC Press. doi: 10.1201/b18401

He, J., Barrera-Pedemonte, F., & Buchholz, J. (2019). Cross-cultural comparability of non-cognitive constructs in TIMSS and PISA. *Assessment in Education: Principles, Policy & Practice, 26*(4), 369–385. doi: 10.1080/0969594X.2018.1469467

He, Y., & Cui, Z. (2019). Evaluating robust scale transformation methods with multiple outlying common items under IRT true score equating. *Applied Psychological Measurement*. Advance online publication. doi: 10.1177/0146621619886050

He, Y., Cui, Z., & Osterlind, S. J. (2015). New robust scale transformation methods in the presence of outlying common items. *Applied Psychological Measurement, 39*(8), 613–626. doi: 10.1177/0146621615587003

Holland, P. W., & Wainer, H. (Eds.) (1993). *Differential item functioning: Theory and practice*. Hillsdale, NJ: Erlbaum.

Huang, P. H. (2018). A penalized likelihood method for multi-group structural equation modelling. *British Journal of Mathematical and Statistical Psychology, 71*(3), 499–522. doi: 10.1111/bmsp.12130

Jerrim, J., Parker, P., Choi, A., Chmielewski, A. K., Sälzer, C., & Shure, N. (2018). How robust are cross-country comparisons of PISA scores to the scaling model used? *Educational Measurement: Issues and Practice, 37*(4), 28–39. doi: 10.1111/emip.12211

Kang, T., & Petersen, N. S. (2012). Linking item parameters to a base scale. *Asia Pacific Education Review, 13*(2), 311–321. doi: 10.1007/s12564-011-9197-2

Kane, M. T. (2006). Validation. In R. L. Brennan (Eds.), *Educational measurement* (pp. 17–64). Westport: Praeger Publisher.

Kankaraš, M., & Moors, G. (2014). Analysis of cross-cultural comparability of PISA 2009 scores. *Journal of Cross-Cultural Psychology, 45*(3), 381–399. doi: 10.1177/0022022113 511297

Köhler, C., Robitzsch, A., & Hartig, J. (2019). A bias corrected RMSD item fit statistic: An evaluation and comparison to alternatives. *Journal of Educational and Behavioral Statistics*. Advance online publication. doi: 10.3102/1076998619890566

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking*. New York: Springer. doi: 10.1007/978-1-4939-0317-7

Kolenikov, S. (2011). Biases of parameter estimates in misspecified structural equation models. *Sociological Methodology*, *41*(1), 119–157. doi: 10.1111/j.1467-9531.2011.01236.x

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement, 75*(1), 22–56. doi: 10.1177/0013164414529792

Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika, 79*(2), 210–231. doi: 10.1007/s11336-013-9347-z

Kuha, J., & Moustaki, I. (2015). Nonequivalence of measurement in latent variable modeling of multigroup data: A sensitivity analysis. *Psychological Methods, 20*(4), 523–536. doi: 10.1037/met0000031

Lee, W.-C., & Lee, G. (2018). IRT linking and equating. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test* (pp. 639–673). New York: Wiley. doi: 10.1002/9781118 489772.ch21

MacCallum, R. C., Browne, M. W., & Cai, L. (2007). Factor analysis models as approximations. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100* (pp. 153–175). Mahwah: Lawrence Erlbaum.

Magis, D., & De Boeck, P. (2011). Identification of differential item functioning in multiple-group settings: A multivariate outlier detection approach. *Multivariate Behavioral Research, 46*(5), 733–755. doi: 10.1080/00273171.2011.606757

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge. doi: 10.4324/9780203821961

Muthen, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology, 5*:978. doi: 10.3389/fpsyg.2014.00978

OECD (2014). *PISA 2012 technical report*. Paris: OECD Publishing.

OECD (2015). *PISA 2015 field trial analysis report. Outcomes of the cognitive assessment (JT03371930)*. Paris: OECD.

OECD (2017). *PISA 2015 technical report*. Paris: OECD Publishing.

Ogasawara, H. (2002). Stable response functions with unstable item parameter estimates. *Applied Psychological Measurement, 26*(3), 239–254. doi: 10.1177/0146621602026003 001

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling, 53*(3), 315–333.

Oliveri, M. E., & von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing, 14*(1), 1–21. doi: 10.1080/15305058.2013.825265

Oliveri, M. E., & von Davier, M. (2017). Analyzing the invariance of item parameters used to estimate trends in international large-scale assessments. In H. Jiao, & R. W. Lissitz (Eds.), *Test fairness in the new generation of large-scale assessment* (pp. 121–146). New York: Information Age Publishing.

Owen, D. (1980). A table of normal integrals. *Communications in Statistics: Simulation and Computation, 9*(4). 389–419. doi: 10.1080/03610918008812164

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26: Psychometrics* (pp. 125–167). Amsterdam: Elsevier. doi: 10.1016/S0169-7161(06)26005-X

R Core Team (2019). *R: A language and environment for statistical computing*. Vienna, Austria. https://www.R-project.org/

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*(4), 353–368. doi: 10.1177/014662169501900405

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Robitzsch, A. (2019). *sirt: Supplementary item response theory models*. R package version 3.6-21. http://CRAN.R-project.org/package=sirt

Robitzsch, A., Kiefer, T., & Wu, M. (2019). *TAM: Test analysis modules*. R package version 3.2-24. http://CRAN.R-project.org/package=TAM

Robitzsch, A., & Lüdtke, O. (2019). Linking errors in international large-scale assessments: calculation of standard errors for trend estimation. *Assessment in Education: Principles, Policy & Practice, 26*(4), 444–465. doi: 10.1080/0969594X.2018.1433633

Rutkowski, L., & Rutkowski, D. (2018). Improving the comparability and local usefulness of international assessments: A look back and a way forward. *Scandinavian Journal of Educational Research, 62*(3), 354–367. doi: 10.1080/00313831.2016.1261044

Rutkowski, L., & Svetina, D. (2017). Measurement invariance in international surveys: Categorical indicators and fit measure performance. *Applied Measurement in Education, 30*(1), 39–51. doi: 10.1080/08957347.2016.1243540

Sachse, K. A., Roppelt, A., & Haag, N. (2016). A comparison of linking methods for estimating national trends in international comparative large-scale assessments in the presence of cross-national DIF. *Journal of Educational Measurement, 53*(2)*,* 152–171. doi: 10.1111/jedm.12106

Schauberger, G., & Mair, P. (2020). A regularization approach for the detection of differential item functioning in generalized partial credit models. *Behavior Research Methods, 52*(1). 279–294. doi: 10.3758/s13428-019-01224-2

Shealy, R., & Stout, W. A. (1993). 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, 58*, 159–194. doi: 10.1007/BF02294572

Soares, T. M., Goncalves, F. B., & Gamerman, D. (2009). An integrated Bayesian model for DIF analysis. *Journal of Educational and Behavioral Statistics, 34*(3), 348–377. doi: 10.3102/1076998609332752

Strobl, C., Kopf, J., Hartmann, R., & Zeileis, A. (2018). *Anchor point selection: An approach for anchoring without anchor items*. Working Papers in Economics and Statistics, 2018-03. University of Innsbruck.

Tijmstra, J., Liaw, Y., Bolsinova, M., Rutkowski, L., & Rutkowski, D. (2019). Sensitivity of the RMSD for detecting item-level misfit in low-performing countries. *Journal of Educational Measurement*. Advance online publication. doi: 10.1111/jedm.12263

Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, 55*(2), 371−390. doi: 10.1007/BF02295293

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika, 80*(1), 21–43. doi: 10.1007/s11336-013-9377-6

van de Vijver, F. J. R. (2019). *Invariance analyses in large-scale studies*. Paris: OECD.

van der Linden, W. J. (1994). Fundamental measurement and the fundamentals of Rasch measurement. In M. Wilson (Ed.), *Objective measurement: Theory into practice. Vol. 2* (pp. 3–24). Hillsdale, NJ: Ablex Publishing Corporation.

van Rijn, P. W., Sinharay, S., Haberman, S. J., & Johnson, M. S. (2016). Assessment of fit of item response theory models used in large-scale educational survey assessments. *Large-scale Assessments in Education, 4*:10, 1–23. doi: 10.1186/s40536-016-0025-3

von Davier, M., & Sinharay, S. (2014). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 155–174). Boca Raton: CRC Press. doi: 10.1201/b16061

von Davier, M., & von Davier, A. A. (2007). A unified approach to IRT scale linking and scale transformations. *Methodology, 3*(3), 115–124. doi: 10.1027/1614-2241.3.3.115

von Davier, M., Yamamoto, K., Shin, H. J., Chen, H., Khorramdel, L., Weeks, J., Davis, S., Kong, N., & Kandathil, M. (2019). Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assessment in Education: Principles, Policy & Practice, 26*(4), 466–488. doi: 10.1080/0969594X.2019.1586642

Wainer, H. (1993). Model-based standardized measurement on an item's differential impact. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 123–135). Hillsdale, NJ: Erlbaum.

Weeks, J., von Davier, M., & Yamamoto, K. (2014). Design considerations for the program for international student assessment. In L. Rutkowski, M. von Davier & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 259–275). Boca Raton: CRC Press. doi: 10.1201/b16061

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica, 50*(1), 1–25. doi: 10.2307/1912526

Wu, M. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice*, *29*(4), 15–27. doi: 10.1111/j.1745-3992.2010.00190.x

Zieger, L. R., Jerrim, J., & Sims, S. (2019). Comparing teachers' job satisfaction across countries. A multiple-pairwise measurement invariance approach. *Educational Measurement: Issues and Practice, 38*(3), 75–85. doi: 10.1111/emip.12254

Zwitser, R. J., Glaser, S. S. F., & Maris, G. (2017). Monitoring countries in a changing world: A new look at DIF in international surveys. *Psychometrika, 82*(1), 210–232. doi: 10.1007/s11336-016-9543-8

# Appendix

## Appendix A: preliminaries

In this section, we introduce some notation and basic facts, which will be used in the following appendices. We assume that a multiple-group 1PL model holds with $P_{ig}(\theta) = \Psi(\theta - b_{ig})$ where the identified parameter $b_{ig}$ is given as $b_{ig} = b_i - \mu_g - e_{ig}$. The primary parameter of interest is the group mean $\mu_g$. First, note that the logistic IRF can quite well be approximated by the probit IRF by $\Psi(\theta - b) = \Phi(D^{-1}(\theta - b))$, where $D = 1.701$ and $\Phi$ denotes the standard normal distribution function. Assuming a normal distribution for ability $\theta$, rules involving Gaussian integrals (Owen, 1980) can be applied to compute marginal item probabilities, thresholds, or tetrachoric correlations that provide closed forms of quantities of interest in which the latent variable $\theta$ has been integrated out. For example, this advantage is used in limited information estimation methods of item response models that do not require numerical integration methods that involve integrals of latent variables (Cai & Moustaki, 2018). In the following, we use real-valued loss functions $\rho = \rho(e)$ for estimation (see Fox, 2016), which are symmetric, nonnegative, fulfilling $\rho(0) = 0$, and are monotonically nondecreasing in $|e|$.

## Appendix B: maximum likelihood estimation for concurrent calibration under full invariance

To derive the bias in the multiple-group concurrent calibration approach, we assume that the number of items tends to infinity. In this case, individual posterior distributions converge to a point estimate, namely the individual maximum likelihood estimate. Moreover, it is assumed that the common item parameter $b_i$ can be consistently estimated. Let $P_i(\theta) = \Psi(\theta - b_i)$ be the common IRF and $W_i(\theta) = P_i(\theta)(1 - P_i(\theta))$ be the corresponding information function. Now we compute the estimated ability $\theta_1$ for persons in group $g$ for which data has been generated with ability $\theta_0$. More concretely, we seek to determine a function $m_g$ for which $\theta_1 = m_g(\theta_0)$ holds. If abilities are unbiasedly estimated, $m_g$ is the identity function. However, in the presence of DIF effects, the misspecified multiple-group IRT model under full invariance has the potential to introduce bias. For a large number of items and a large number of persons with ability $\theta_0$, the estimated ability $\theta_1$ is given as the maximizer of the Kullback-Leibler information, which is given as

$$l(\theta) = \frac{1}{I}\sum_{i=1}^{I}\left[P_{ig}(\theta_0)\log P_i(\theta) + \left(1 - P_{ig}(\theta_0)\right)\log\left(1 - P_i(\theta)\right)\right]. \quad \text{(A1)}$$

Setting the first derivative of $l$ to zero, $\theta_1$ is given as the root of the nonlinear equation

$$l_1(\theta) = \frac{\partial l}{\partial \theta} = \frac{1}{I}\sum_{i=1}^{I}\left[P_{ig}(\theta_0) - P_i(\theta)\right] = 0. \quad \text{(A2)}$$

We now can use two Taylor approximations of $P_{ig}$ and $P_i$ in Equation A2 to determine $m_g$. First, we obtain $P_i(\theta) = \Psi(\theta - b_i) \approx P_i(\theta_0) + W_i(\theta_0)(\theta - \theta_0)$ by using a linear Taylor approximation with respect to θ. Second, we apply a linear Taylor approximation of $P_{ig}$ with respect to the DIF effect $e_{ig}$ (see Kolenikov, 2011, for a similar strategy), resulting in $P_{ig}(\theta) = \Psi(\theta - b_i - e_{ig}) \approx P_i(\theta) - W_i(\theta)e_{ig}$. Using the abbreviation $\tilde{W}_i(\theta) = W_i(\theta) / \sum_u W_u(\theta)$, we obtain $\theta_1$ by inserting the Taylor approximations in Equation A2:

$$\theta_1 = m_g(\theta_0) = \theta_0 - \sum_{i=1}^{I} \tilde{W}_i(\theta_0) e_{ig} = \theta_0 - B_g(\theta_0), \tag{A3}$$

where $B_g$ denotes the bias in estimated abilities. The estimated group mean for group $g$ is then given as $\hat{\mu}_g = \int m_g(\theta) f_g(\theta) \mathrm{d}\theta$ where $f_g$ is the density of group $g$. Using Equation A3, this can be further simplified to

$$\hat{\mu}_g = \mu_g + \sum_{i=1}^{I} w_{ig} e_{ig} \ , \tag{A4}$$

where the weights $w_{ig}$ are given as $w_{ig} = w_{ig}(\mu_g, \sigma_g, b_i) = -\int \tilde{W}_i(\theta) f_g(\theta) \mathrm{d}\theta$.

## Appendix C: limited information methods for concurrent calibration

We now derive the bias of estimated group means if weighted least squares estimation is used. The estimated group means $\hat{\mu}_g$ rely on estimated item thresholds $\hat{\tau}_{ig}$ and pairwise tetrachoric correlations $\hat{\rho}_{ijg}$. For specified weights $w_{ig}$ (assuming $\sum_i w_{ig} = 1$ without loss of generality) and $w_{ijg}$ in the weighted least squares estimation and known common item difficulties $b_i$ (obtained in a previous estimation step), estimated group means and standard deviations are given as the minimizers of

$$H(\mu_g, \sigma_g) = \sum_i w_{ig} (\hat{\tau}_{ig} - \tau_{ig})^2 + \sum_{i,j} w_{ijg} (\hat{\rho}_{ijg} - \rho_{ijg})^2 \ , \tag{A5}$$

where $\tau_{ig}$ and $\rho_{ijg}$ are model-implied item intercepts and tetrachoric correlations, respectively. In large samples, the model-implied and estimated tetrachoric correlations coincide, that is, $\hat{\rho}_{ijg} = \rho_{ijg} = \sigma_g^2 / (D^2 + \sigma_g^2)$. Hence, the group-specific standard deviation $\sigma_g$ can be consistently estimated from data and can be identified from tetrachoric correlations only. Using $f_g = (D^2 + \sigma_g^2)^{-1/2} \sigma_g$, we obtain $\tau_{ig} = f_g(\mu_g - b_i)$ and $\hat{\tau}_{ig} = f_g(\mu_g - b_i - e_{ig})$. The estimated group mean as the minimizer of Equation A5 is then given as

$$\hat{\mu}_g = \mu_g + \sum_{i=1}^{I} w_{ig} e_{ig} \ . \tag{A6}$$

Hence, estimated group means using weighted least squares are biased in general. Note that the general form in Equation A6 is the same as for concurrent calibration using ML, but with differently defined weights (see Equation A4).

## Appendix D: population MD and RMSD statistics

Using the definition of the population MD statistic and the probit approximation of the logistic IRF (see Appendix A), the MD statistic is computed as

$$\mathrm{MD}_{ig} = \int \left[ \Phi\left( D^{-1}\left( \theta - b_i - e_{ig} \right) \right) - \Phi\left( D^{-1}\left( \theta - b_i \right) \right) \right] \sigma_g^{-1} \, \phi\left( \left( \theta - \mu_g \right) / \sigma_g \right) \mathrm{d}\theta \ . \quad \text{(A7)}$$

Using the rule for change of variables in integration, we rewrite Equation A7 as

$$\mathrm{MD}_{ig} = \int \left[ \Phi\left( D^{-1}\left( \sigma_g u + \mu_g - b_i - e_{ig} \right) \right) - \Phi\left( D^{-1}\left( \sigma_g u + \mu_g - b_i \right) \right) \right] \phi(u) \mathrm{d}u \ . \quad \text{(A8)}$$

Applying Formula (10,010.8) of Owen (1980) to Equation A8 results in

$$\mathrm{MD}_{ig} = \Phi\left( \frac{\mu_g - b_i - e_{ig}}{\sqrt{D^2 + \sigma_g^2}} \right) - \Phi\left( \frac{\mu_g - b_i}{\sqrt{D^2 + \sigma_g^2}} \right) . \quad \text{(A9)}$$

The formula for the square of the RMSD statistic (in the following defined as MSD) relies on the bivariate normal distribution $\Phi_2$. It holds that

$$\mathrm{MSD}_{ig} = \int \left[ \Phi\left( D^{-1}\left( \sigma_g u + \mu_g - b_i - e_{ig} \right) \right) - \Phi\left( D^{-1}\left( \sigma_g u + \mu_g - b_i \right) \right) \right]^2 \phi(u) \mathrm{d}u \ . \quad \text{(A10)}$$

Expanding the terms in Equation A10 leads to a representation $\mathrm{MSD}_{ig} = T_1 + T_2 - 2T_3$. Using $T\left(x, y, \sigma_g\right) = \Phi_2\left( \frac{x}{\sqrt{D^2 + \sigma_g^2}}, \frac{y}{\sqrt{D^2 + \sigma_g^2}}, \frac{\sigma_g^2}{D^2 + \sigma_g^2} \right)$, it follows that $T_1 = T(\mu_g - b_i - e_{ig}, \mu_g - b_i - e_{ig}, \sigma_g)$, $T_2 = T(\mu_g - b_i, \mu_g - b_i, \sigma_g)$, and, $T_3 = T(\mu_g - b_i, \mu_g - b_i - e_{ig}, \sigma_g)$ by applying Formula (20,010.3) of Owen (1980). Finally, for $e_{ig} > 0$, MSD can be equivalently written as a two-dimensional integral with respect to the bivariate normal density $\phi_2$:

$$\mathrm{MSD}_{ig} = \int_{\mu_g - b_i - e_{ig}}^{\mu_g - b_i} \int_{\mu_g - b_i - e_{ig}}^{\mu_g - b_i} \phi_2\left( \frac{u}{\sqrt{D^2 + \sigma_g^2}}, \frac{v}{\sqrt{D^2 + \sigma_g^2}}, \frac{\sigma_g^2}{D^2 + \sigma_g^2} \right) \mathrm{d}u\mathrm{d}v \ \text{for} \ e_{ig} > 0 \ . \quad \text{(A11)}$$

The case of $e_{ig} < 0$ can be handled by using redefined quantities $\tilde{b}_i = b_i + e_{ig}$ and $\tilde{e}_{ig} = -e_{ig}$.

The formula of the MD statistic can similarly be derived for the case of non-uniform DIF in the 2PL model. Assume that $P_i(\theta) = \Psi(a_i\theta - b_i)$ is the joint IRF and $P_{ig}(\theta) = \Psi((a_i + u_{ig})\theta - b_i - e_{ig})$ is the group-specific IRF. The same computations as for Equation A9 provide the MD statistic

$$\mathrm{MD}_{ig} = \Phi\left(\frac{\mu_g - b_i - e_{ig}}{\sqrt{D^2 + (a_i + u_{ig})^2 \sigma_g^2}}\right) - \Phi\left(\frac{\mu_g - b_i}{\sqrt{D^2 + a_i^2\sigma_g^2}}\right). \tag{A12}$$

In the case of uniform DIF in the 2PL model, the DIF effect $u_{ig}$ is zero and Equation A12 coincides with Equation A9 if $\sigma_g^2$ is replaced by $a_i^2\sigma_g^2$. The formula for the RMSD statistic for the 2PL model with uniform DIF can be obtained by applying the same reasoning to Equation A11.

## Appendix E: empirical MD and RMSD statistics

In the following, we use the notation from the Section "MD and RMSD statistics". As in Appendix B, we assume a large number of items. Then, the empirical IRF $\hat{P}_{ig}(\theta)$ converges to an IRF $P_{ig}^*(\theta) = P_{ig}(m_g^{-1}(\theta))$, where $m_g(\theta) = \theta - B_g(\theta)$ denotes the transformation of true abilities into estimated abilities (see Appendix B)[16]. The estimated density for group *g* is given as $\hat{f}_g(\theta) = f_g(m_g^{-1}(\theta))\left|m_g'(\theta)\right|$. The square of the expected value of the estimated RMSD statistic can be approximately written (by expanding the terms in Equation 9, but neglecting terms involving cross-products of $\hat{P}_{ig} - P_{ig}^*$, and applying the rule of change of variables in integration) as

$$\begin{aligned} E\left(\widehat{\mathrm{RMSD}}_{ig}\right)^2 &\approx E\left[\int\left[\hat{P}_{ig}(\theta) - P_{ig}(\theta)\right]^2 f_g(\theta)\mathrm{d}\theta\right] \\ &+ \int\left[P_{ig}(\theta) - P_i(\theta + B_g(\theta))\right]^2 f_g(\theta)\mathrm{d}\theta\,. \end{aligned} \tag{A13}$$

[16] We assume that the transformation function $m_g$ is (piecewise) monotone and differentiable in order to define $m_g^{-1}$ appropriately.

The first term in Equation A12 denotes the positive bias $B_{1,+}(n) = E\left[\int\left[\hat{P}_{ig}(\theta) - P_{ig}(\theta)\right]^2 f_g(\theta)\mathrm{d}\theta\right]$ caused by the finite sampling of persons. Again, we can apply a Taylor approximation $P_i(\theta + B_g(\theta)) \approx P_i(\theta) + W_i(\theta)B_g(\theta)$ (see Appendix B). Inserting this simplification into Equation A13 and omitting cross-products results in

$$\begin{aligned} E\left(\widehat{\mathrm{RMSD}}_{ig}\right)^2 &\approx \mathrm{RMSD}_{ig}^2 + B_{1,+}(n) + \underbrace{\int\left[W_i(\theta)B_g(\theta)\right]^2 f_g(\theta)\mathrm{d}\theta}_{B_{2,+}\left(|\widehat{f}_g - f_g|\right)} \\ &+ \underbrace{\int\left[P_{ig}(\theta) - P_i(\theta)\right]W_i(\theta)B_g(\theta)f_g(\theta)\mathrm{d}\theta}_{B_{3,\pm}\left(P_{ig}-P_i, |\widehat{f}_g - f_g|\right)} . \end{aligned} \qquad \text{(A14)}$$

By making use of the same proof strategy, it can be shown that the bias in the MD statistic only depends on misestimated ability, that is, $|\widehat{f}_g - f_g| \neq 0$.

## Appendix F: estimation constraints for Haberman and Haebara linking methods

We first derive the estimating equation for Haberman linking. Taking the derivative of Equation 11 with respect to $\mu_g$ provides

$$\frac{\partial H}{\partial \mu_g} = \sum_i \rho'\left(\hat{b}_{ig} - b_i + \mu_g\right) = 0 \qquad \text{(A14)}$$

For large sample sizes, it holds that $\hat{b}_{ig} = b_{ig} = b_i + e_{ig} - \mu_g$. Inserting this quantity into Equation A14 results in $\sum_i \rho'\left(e_{ig}\right) = 0$ (see Equation 12). In the case of Haberman linking with OLS estimation (i.e., $\rho(x) = x^2/2$), the constraint $\sum_i e_{ig} = 0$ is fulfilled, and we obtain $\hat{\mu}_g = \mu_g + \sum_i e_{ig}$.

The condition for Haebara linking can be obtained similarly by taking the first derivative of $H$ in Equation 13 with respect to $\mu_g$ and changing the order of integration and differentiation. Again, the relation $\hat{b}_{ig} = b_i + e_{ig} - \mu_g$ is used to obtain Equation 14.