foundations

MDPI

Article

Comparing Different Specifications of Mean–Geometric Mean Linking

Alexander Robitzsch [1,2]

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

Abstract: Mean–geometric mean (MGM) linking compares group differences on a latent variable θ within the two-parameter logistic (2PL) item response theory model. This article investigates three specifications of MGM linking that differ in the weighting of item difficulty differences: unweighted (UW), discrimination-weighted (DW), and precision-weighted (PW). These methods are evaluated under conditions where random DIF effects are present in either item difficulties or item intercepts. The three estimators are analyzed both analytically and through a simulation study. The PW method outperforms the other two only in the absence of random DIF or in small samples when DIF is present. In larger samples, the UW method performs best when random DIF with homogeneous variances affects item difficulties, while the DW method achieves superior performance when such DIF is present in item intercepts. The analytical results and simulation findings consistently show that the PW method introduces bias in the estimated group mean when random DIF is present. Given that the effectiveness of MGM methods depends on the type of random DIF, the distribution of DIF effects was further examined using PISA 2006 reading data. The model comparisons indicate that random DIF with homogeneous variances in item intercepts provides a better fit than random DIF in item difficulties in the PISA 2006 reading dataset.

Keywords: item response model; 2PL model; mean–geometric mean linking; differential item functioning

MSC: 62H10; 62H25

Academic Editor: Jay Jahangiri

Received: 3 May 2025
Revised: 27 May 2025
Accepted: 4 June 2025
Published: 6 June 2025

Citation: Robitzsch, A. Comparing Different Specifications of Mean–Geometric Mean Linking. *Foundations* **2025**, *5*, 20. https://doi.org/10.3390/foundations5020020

1. Introduction

Item response theory (IRT) models [1–3] provide a statistical framework for modeling multivariate discrete outcomes. This work specifically addresses binary item responses and explores methods for comparing two populations using linking techniques. Consider a response vector $\boldsymbol{X} = (X_1, \ldots, X_I)$, where each variable $X_i \in \{0, 1\}$ represents a dichotomously scored item. A unidimensional IRT model [4] specifies the joint probability distribution $\mathsf{P}(\boldsymbol{X} = \boldsymbol{x})$ for response patterns $\boldsymbol{x} = (x_1, \ldots, x_I) \in \{0, 1\}^I$ through a parametric formulation:

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

where f denotes the normal density function, parameterized by the mean μ and the standard deviation (SD) σ. The distribution parameters μ and σ of the latent variable θ, often

referred to as a trait or ability variable, are collected in the vector $\delta = (\mu, \sigma)$. The vector $\gamma = (\gamma_1, \ldots, \gamma_I)$ collects the item parameters for the item response functions (IRFs) $P_i(\theta; \gamma_i) = \mathsf{P}(X_i = 1|\theta)$ for $i = 1, \ldots, I$. The IRF of the two-parameter logistic (2PL) model [5] is defined by

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

where a_i and b_i represent the item discrimination and the item difficulty b_i, respectively. The function $\Psi(x) = (1 + \exp(-x))^{-1}$ corresponds to the standard logistic distribution function. In this formulation, the item parameter vector is $\gamma_i = (a_i, b_i)$. Alternatively, the 2PL model can be reparametrized by replacing the difficulty parameter b_i with the intercept ν_i, resulting in

$$P_i(\theta; \gamma_i) = \Psi(a_i\theta + \nu_i). \tag{3}$$

The two 2PL parameterizations are related by the identity $\nu_i = -a_i b_i$ (see [6,7]). In this parametrization, the item parameter vector is $\gamma_i = (a_i, \nu_i)$.

Given a sample of N individuals with independent and identically distributed response vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ drawn from the distribution of $\boldsymbol{X}$, the parameters of the IRT model specified in (1) can be consistently estimated through marginal maximum likelihood (MML) methods [8,9].

IRT models are widely employed to compare the test performance of two groups by assessing differences in the parameters of the latent variable θ, as defined in the IRT framework of (1). This article specifically examines linking methods [10] based on the 2PL model.

In the first step of the linking approach, the 2PL model is estimated separately for each group, allowing for the presence of differential item functioning (DIF), where item behavior may vary across groups [11–13]. More specifically, item parameters are permitted to differ between groups, indicating that the groups may respond differently to an item even after accounting for overall differences in the θ variable. In the second step, differences in item parameters are used to estimate group differences in θ through a linking procedure [10,14,15].

This article evaluates the performance of mean–geometric mean (MGM; [7,10,16–19]) linking in the presence of DIF [13] in either item difficulties or item intercepts. The standard MGM method is based on the mean difference of log-transformed item discriminations for estimating the group SD and the mean difference of untransformed item difficulties for estimating the group mean. This study considers three specifications of MGM linking under random DIF [20–22] in item difficulties or intercepts. Prior research has shown that random DIF contributes to increased variance in the estimated linking parameters [23,24]. This concept is also referred to as a linking error in educational large-scale assessment studies [23,25–32].

The three MGM specifications considered here differ in the weighting of item difficulty differences when computing the group mean. To the best of the authors' knowledge, the performance of these MGM variants under random DIF has not yet been systematically examined. The performance of the MGM estimators is assessed analytically and through a simulation study.

The remainder of the article is organized as follows. Section 2 reviews the MGM specifications. Section 3 presents the results from a simulation study comparing their performance. Section 4 provides an empirical illustration using PISA data. Section 5 concludes with a discussion.

2. Mean–Geometric Mean Linking

2.1. Identified Item Parameters in Separate Scaling

The various specifications of the MGM method are based on item parameters from the 2PL model, estimated separately for each group. The following describes the identification of item parameters under the assumption that no DIF is present in item discriminations a_i or item difficulties b_i. In both groups, the latent variable θ is standardized by fixing its mean and SD to 0 and 1, respectively, allowing all item parameters to be estimated within each group. In the first group, the identified item parameters are given by $\widehat{a}_{i1} = a_i$, $\widehat{b}_{i1} = b_i$, and $\widehat{\nu}_{i1} = \nu_i = -a_i b_i$, where a_i and b_i represent the invariant item parameters in the 2PL model across groups.

In the second group, the latent variable θ is assumed to have a mean μ and SD σ. By fixing the θ mean to 0 and SD to 1, the identified item parameters from the separate 2PL model estimation in this group are given by

$$\widehat{a}_{i2} = \sigma a_i \text{ or } \widehat{\alpha}_{i2} = \log \widehat{a}_{i2} = \log \sigma + \log a_i, \tag{4}$$

$$\widehat{b}_{i2} = \sigma^{-1}(b_i - \mu) \text{ and} \tag{5}$$

$$\widehat{\nu}_{i2} = a_i \mu + \nu_i. \tag{6}$$

The MGM method aims to recover the parameters μ and σ using the group-specific item parameters $\widehat{a}_{ig}$ and $\widehat{b}_{ig}$ ($i = 1, \ldots, I$, $g = 1, 2$), obtained from separate estimations under the 2PL model.

2.2. Weighted Means

The different specifications of MGM linking for estimating μ are essentially different weighted means of item difficulty differences. The following briefly reviews the statistical properties of such weighted means. Let Y_i ($i = 1, \ldots, I$) denote normally distributed observations with mean μ and variances σ_i^2. A weighted mean with fixed weights w_i is defined as

$$\overline{Y}_w = \frac{\frac{1}{I}\sum_{i=1}^{I} w_i Y_i}{\frac{1}{I}\sum_{i=1}^{I} w_i}. \tag{7}$$

Note that the multiplication factor $1/I$ in (7) could be omitted; however, it is included to maintain consistency with later expressions in the various specifications of MGM linking. The expected value of $\overline{Y}_w$ is μ, and its variance is given by

$$\mathsf{Var}(\overline{Y}_w) = \frac{\sum_{i=1}^{I} w_i^2 \sigma_i^2}{\left(\sum_{i=1}^{I} w_i\right)^2}. \tag{8}$$

The minimal variance in (8) is attained when the observations Y_i are weighted by their precisions $1/\sigma_i^2$ (i.e., the inverses of their variances) (see [33]). A weighted mean using these weights is commonly referred to as a precision-weighted mean.

2.3. Random DIF in Item Difficulties or Item Intercepts

The occurrence of random DIF [20,34] can be characterized by whether DIF manifests in item difficulties [34] or item intercepts [35]. Assume that DIF arises only as deviations in item parameters for the second group relative to the first group. Further, assume invariant

item parameters. Let e_i and ϵ_i denote random DIF in item difficulties and item intercepts with zero means, respectively, under the assumptions

$$b_{i2} = b_i + e_i \quad \text{and} \quad \nu_{i2} = \nu_i + \epsilon_i. \tag{9}$$

The two DIF effects are related by

$$\epsilon_i = -a_i e_i \quad \text{or} \quad e_i = -\frac{\epsilon_i}{a_i}. \tag{10}$$

If the DIF effects e_i in item difficulties have variances τ_i^2, corresponding to random DIF, the variances of the DIF effects ϵ_i in item intercepts are given by

$$\mathsf{Var}(\epsilon_i) = a_i^2 \mathsf{Var}(e_i). \tag{11}$$

If the random DIF effects e_i have homogeneous variances τ^2, it follows from (11) that the corresponding DIF effects ϵ_i in item intercepts exhibit heterogeneous variances $a_i^2\tau^2$. Conversely, if the DIF effects ϵ_i in item intercepts have homogeneous variances τ^2, then the DIF effects e_i in item difficulties have heterogeneous variances τ^2/a_i^2.

The formulation of random DIF in terms of item difficulties or item intercepts is statistically equivalent when allowing for heterogeneous variances τ_i^2. However, empirical analyses may test whether random DIF with homogeneous variance is more plausible in item difficulties or in item intercepts. As will be shown later, the performance of the different MGM specifications depends on whether random DIF with homogeneous variances occurs in item difficulties or in item intercepts.

2.4. Estimation of σ in MGM Linking

In MGM linking, the SD σ is estimated using the means of log-transformed item discriminations. Specifically, the estimate $\widehat{\sigma}$ is computed as (see [10,16])

$$\widehat{\sigma} = \exp\left(\frac{1}{I} \sum_{i=1}^{I} \log \widehat{a}_{i2} - \frac{1}{I} \sum_{i=1}^{I} \log \widehat{a}_{i1} \right). \tag{12}$$

Since averages on the logarithmic scale are used, this method is referred to as log-mean-mean linking.

2.5. Estimation of μ in MGM Linking

The estimation of μ is now addressed. Three variants of weighted means of item difficulties are considered to derive the estimate $\widehat{\mu}$. In the formal treatment, assume that random DIF occurs in item difficulties, with $\mathsf{E}(e_i) = 0$ and $\mathsf{Var}(e_i) = \tau_i^2$. If random DIF in item difficulties exhibits homogeneous variance, then $\tau_i^2 = \tau^2$. Alternatively, if random DIF with homogeneous variances occurs in item intercepts, then $\tau_i^2 = \tau^2/a_i^2$.

In addition to random DIF, sampling errors affect the item parameter estimates. Let u_{ig} denote the sampling error in the estimated item difficulty ($i = 1, \ldots, I; g = 1, 2$). The estimated item difficulty in the first group is then given by

$$\widehat{b}_{i1} = b_i + u_{i1} \text{ with } \mathsf{E}(u_{i1}) = 0 \text{ and } \mathsf{Var}(u_{i1}) \simeq \frac{C_0 + C_1 b_i^2}{N a_i^2}, \tag{13}$$

where C_0 and C_1 are constants that depend on the dataset. The variance expression in (13) is supported by empirical evidence from simulation studies [9,36].

The estimated item difficulty in the second group satisfies

$$\widehat{b}_{i2} = \frac{1}{\sigma}(b_i + e_i - \mu) + u_{i2} \text{ with } \mathsf{E}(u_{i2}) = 0 \text{ and } \mathsf{Var}(u_{i2}) \simeq \frac{C_0 + C_1 \frac{1}{\sigma^2}(b_i + e_i - \mu)^2}{N a_i^2 \sigma^2}. \quad (14)$$

Here, u_{i2} represents the sampling error, while e_i denotes the random DIF effect. For a sufficiently large number of items, the estimated item difficulties can be treated as approximately independent across items [36].

2.5.1. Unweighted MGM Linking (UW)

The original variant of MGM linking for estimating μ is based on the difference in item difficulties, defined as

$$\widehat{\mu} = -\left(\frac{1}{I}\sum_{i=1}^{I}\widehat{\sigma}\widehat{b}_{i2} - \frac{1}{I}\sum_{i=1}^{I}\widehat{b}_{i1}\right) = -\frac{1}{I}\sum_{i=1}^{I}\left(\widehat{\sigma}\widehat{b}_{i2} - \widehat{b}_{i1}\right). \quad (15)$$

This estimator uses the previously calculated SD $\widehat{\sigma}$ from (12) and applies equal weights to the differences $\widehat{\sigma}\widehat{b}_{i2} - \widehat{b}_{i1}$ in item difficulties. For this reason, the estimator in (15) is referred to as unweighted MGM linking (UW).

To examine the statistical properties of $\widehat{\mu}$, the expression in (15) is rewritten using (13) and (14) as

$$\widehat{\mu} = \frac{\widehat{\sigma}}{\sigma}\mu - \frac{\widehat{\sigma} - \sigma}{\sigma} \cdot \frac{1}{I}\sum_{i=1}^{I} b_i - \frac{\widehat{\sigma}}{\sigma} \cdot \frac{1}{I}\sum_{i=1}^{I} e_i - \frac{1}{I}\sum_{i=1}^{I}(\widehat{\sigma}u_{i2} - u_{i1}). \quad (16)$$

Since $\mathsf{E}(\widehat{\sigma}) \to \sigma$ as $I \to \infty$, it follows that $\mathsf{E}(\widehat{\mu}) \to \mu$ under the assumptions $\mathsf{E}(e_i) = \mathsf{E}(u_{ig}) = 0$. Thus, a simplified form of $\widehat{\mu}$ is given by (16) as

$$\widehat{\mu} = \mu - \frac{1}{I}\sum_{i=1}^{I} e_i - \frac{1}{I}\sum_{i=1}^{I}(\widehat{\sigma}u_{i2} - u_{i1}). \quad (17)$$

To derive the variance of $\widehat{\mu}$, assume that $\mathsf{Var}(\widehat{\sigma}) \simeq 0$. Then, from (17),

$$\mathsf{Var}(\widehat{\mu}) = \frac{1}{I^2}\sum_{i=1}^{I}\tau_i^2 + \frac{1}{I^2}\sum_{i=1}^{I}\left[\sigma^2\mathsf{Var}(u_{i2}) + \mathsf{Var}(u_{i1})\right]. \quad (18)$$

Equation (18) shows that the variance of $\widehat{\mu}$ consists of two components: the variance due to random DIF effects e_i and the variance from sampling errors u_{ig}. Using (13) and (14), the expression can be further rephrased as

$$\mathsf{Var}(\widehat{\mu}) = \frac{1}{I^2}\sum_{i=1}^{I}\tau_i^2 + \frac{1}{NI^2}\sum_{i=1}^{I}\frac{C_0(\sigma^2 + 1) + C_1\left[(b_i + e_i - \mu)^2 + b_i^2\right]}{a_i^2}. \quad (19)$$

As the sample size increases, the contribution of the sampling error variance diminishes. However, the variance component due to random DIF remains nonzero even in the limit of infinite sample size.

2.5.2. Discrimination-Weighted MGM Linking (DW)

An alternative MGM linking estimate for μ relies on the identification of Equation (6). Following the rationale used in the invariance alignment [37,38] method, the absence of DIF effects yields the identity

$$\widehat{\nu}_{i2} - \widehat{\nu}_{i1} - \frac{\widehat{a}_{i2}}{\sigma}\mu = 0. \tag{20}$$

This identity motivates the estimation of μ as the minimizer of

$$H(\mu) = \sum_{i=1}^{I}\left(\widehat{\nu}_{i2} - \widehat{\nu}_{i1} - \frac{\widehat{a}_{i2}}{\widehat{\sigma}}\mu\right)^2, \tag{21}$$

which leads to the estimator

$$\widehat{\mu} = \widehat{\sigma}\frac{\frac{1}{I}\sum_{i=1}^{I}(\widehat{\nu}_{i2} - \widehat{\nu}_{i1})\widehat{a}_{i2}}{\frac{1}{I}\sum_{i=1}^{I}\widehat{a}_{i2}^2}. \tag{22}$$

The estimator $\widehat{\mu}$ in (22) can be further rewritten as (see [19])

$$\widehat{\mu} = \widehat{\sigma}\frac{\frac{1}{I}\sum_{i=1}^{I}\left(\widehat{a}_{i2}\widehat{b}_{i2} - \widehat{a}_{i1}\widehat{b}_{i1}\right)\widehat{a}_{i2}}{\frac{1}{I}\sum_{i=1}^{I}\widehat{a}_{i2}^2}. \tag{23}$$

To analyze the statistical properties of $\widehat{\mu}$ as defined in (23), simplifying assumptions are applied: $\mathsf{Var}(\widehat{\sigma}) \simeq 0$, $\widehat{a}_{i1} \simeq a_i$, and $\widehat{a}_{i2} \simeq \sigma a_i$. Under these assumptions, the estimator simplifies to

$$\widehat{\mu} = \frac{\sum_{i=1}^{I} a_i^2\left(\sigma\widehat{b}_{i2} - \widehat{b}_{i1}\right)}{\sum_{i=1}^{I} a_i^2}. \tag{24}$$

This expression reveals that $\widehat{\mu}$ is a weighted average of item difficulty differences, where the weights are proportional to the squared item discriminations. Consequently, the estimator in (24) is referred to as discrimination-based MGM linking (DW).

The estimator $\widehat{\mu}$ in (24) can be expressed in terms of the DIF effects e_i and the sampling errors u_{ig} as

$$\widehat{\mu} = \mu + \frac{\sum_{i=1}^{I} a_i^2 e_i}{\sum_{i=1}^{I} a_i^2} + \frac{\sum_{i=1}^{I} a_i^2(\sigma u_{i2} - u_{i1})}{\sum_{i=1}^{I} a_i^2}. \tag{25}$$

As with the UW estimator, this formulation yields an asymptotically unbiased estimate of μ, i.e., $\mathsf{E}(\widehat{\mu}) \to \mu$ as $I \to \infty$. The variance of $\widehat{\mu}$ is given by

$$\mathsf{Var}(\widehat{\mu}) = \frac{\sum_{i=1}^{I} a_i^4\tau_i^2}{\left(\sum_{i=1}^{I} a_i^2\right)^2} + \frac{\sum_{i=1}^{I} a_i^4\left[\sigma^2\mathsf{Var}(u_{i2}) + \mathsf{Var}(u_{i1})\right]}{\left(\sum_{i=1}^{I} a_i^2\right)^2}, \tag{26}$$

which can be further simplified using the expressions for $\mathsf{Var}(u_{ig})$ from (19) as

$$\mathsf{Var}(\widehat{\mu}) = \frac{\sum_{i=1}^{I} a_i^4 \tau_i^2}{\left(\sum_{i=1}^{I} a_i^2\right)^2} + \frac{1}{N} \cdot \frac{\sum_{i=1}^{I} a_i^2 \left\{ C_0(\sigma^2+1) + C_1\left[(b_i + e_i - \mu)^2 + b_i^2\right]\right\}}{\left(\sum_{i=1}^{I} a_i^2\right)^2}. \tag{27}$$

The left-hand variance component in (27) indicates that an optimal estimate of μ is obtained when the DIF effects e_i satisfy $\mathsf{Var}(e_i) = a_i^{-2}\tau^2$, corresponding to random DIF with homogeneous variances in item intercepts. In this case, the weighting by item discriminations enhances precision, as the sampling variance of estimated item difficulties is also proportional to a_i^{-2}. However, if random DIF with homogeneous variance occurs in item difficulties rather than in item intercepts, the discrimination-based weighting in DW may result in a higher variance compared to the equal weighting used in UW, particularly in large samples where the contribution from sampling error becomes negligible.

2.5.3. Precision-Weighted MGM Linking (PW)

The UW linking method assigns equal weights to the item difficulty differences $\widehat{\sigma}\widehat{b}_{i2} - \widehat{b}_{i1}$. As an alternative, these differences can be weighted by their precisions, that is, the inverse of their sampling variances [7,39]. This approach yields the estimator

$$\widehat{\mu} = -\frac{\sum_{i=1}^{I} \omega_i \left(\sigma \widehat{b}_{i2} - \widehat{b}_{i1}\right)}{\sum_{i=1}^{I} \omega_i}, \tag{28}$$

where the precision weights ω_i must be estimated. The variances $\mathsf{Var}(\widehat{b}_{i1})$ and $\mathsf{Var}(\widehat{b}_{i2})$ are obtained from the observed information matrix in the group-wise scaling models. Based on these variances, the weights ω_i are defined as

$$\omega_i = \left[\widehat{\sigma}^2 \mathsf{Var}(\widehat{b}_{i2}) + \mathsf{Var}(\widehat{b}_{i1})\right]^{-1}. \tag{29}$$

Using (13) and (14), the precision weights can be approximately determined as

$$\omega_i = \left[\widehat{\sigma}^2 \mathsf{Var}(u_{i2}) + \mathsf{Var}(u_{i1})\right]^{-1} = \frac{N a_i^2}{C_0(\sigma^2+1) + C_1\left[(b_i + e_i - \mu)^2 + b_i^2\right]}. \tag{30}$$

Importantly, (30) highlights that the estimated precision weights ω_i depend on the random DIF effects e_i. For small values of e_i, a linear Taylor approximation of (30) yields

$$\omega_i \simeq \omega_{i0} - \omega_{i1} e_i (b_i - \mu), \text{ where} \tag{31}$$

$$\omega_{i0} = N a_i^2 h_i,\ \omega_{i1} = 2 C_1 N a_i^2 h_i^2, \text{ and } h_i = \left\{ C_0(\sigma^2+1) + C_1\left[(b_i - \mu)^2 + b_i^2\right]\right\}^{-1}. \tag{32}$$

Note that ω_{i0} and ω_{i1} are independent of the random DIF effect e_i.

The estimator $\widehat{\mu}$ in (28) can be rephrased as

$$\widehat{\mu} \simeq \mu - \frac{\sum_{i=1}^{I} [\omega_{i0} - \omega_{i1} e_i (b_i - \mu)] e_i}{\sum_{i=1}^{I} (\omega_{i0} - \omega_{i1} e_i)} - \frac{\sum_{i=1}^{I} [\omega_{i0} - \omega_{i1} e_i (b_i - \mu)](u_{i2} - u_{i1})}{\sum_{i=1}^{I} (\omega_{i0} - \omega_{i1} e_i)}. \tag{33}$$

Assuming independence between e_i and u_{ig}, the expectation of $\widehat{\mu}$ as $I \to \infty$ is given by

$$\mathsf{E}(\widehat{\mu}) \to \mu - \frac{\sum_{i=1}^{I} \omega_{i1} \tau_i^2 (\mu - b_i)}{\sum_{i=1}^{I} \omega_{i0}} \text{ for } I \to \infty. \tag{34}$$

According to (34), the PW linking method may produce a negatively biased estimate of μ when μ is, on average, greater than b_i. However, in the absence of random DIF, the PW method does not exhibit bias in the estimation of μ.

The variance of the PW estimate can be derived analogously to that of the UW and DW estimators, although it offers limited additional insight. By construction, the PW linking method yields the smallest variance in the absence of DIF, as it employs optimal precision weights.

3. Simulation Study

In this Simulation Study, the performances of the three MGM linking specifications (i.e., UW, DW, and PW) outlined in Section 2.5 are compared.

3.1. Method

The data-generating model was based on the 2PL model applied to two groups. For the first group, the latent variable θ followed a standard normal distribution with a fixed mean of 0 and SD of 1. For the second group, θ was also normally distributed with a fixed mean $\mu = 0.3$ and SD $\sigma = 1.2$, which was consistent across all simulation conditions.

The simulation study used $I = 20$ or $I = 40$ items. Group-specific item parameters a_{ig} and b_{ig} for each item $i = 1, \ldots, I$ and for groups $g = 1, 2$ were derived from fixed base parameters and newly simulated random DIF effects in each replication. The item parameters were constructed using 10 base items. These base items were duplicated twice in the 20-item condition and four times in the 40-item condition. For the 10 base items, the base item discriminations a_{i0} were set to 0.6 for the first five items and 1.2 for the remaining five. Base item difficulties were assigned values of -1.4, -0.7, 0.0, 0.7, and 1.4 for the first five items, with the same sequence repeated for the remaining items. The complete set of item parameters is available at https://osf.io/xa4qz (accessed on 3 May 2025).

For the first group, item discriminations and item difficulties were set to the base item parameters. In the second group, DIF effects with a homogeneous variance were introduced either in item difficulties or item intercepts. A normally distributed random DIF effect with DIF SD τ was added to the corresponding item difficulty or item intercept. The DIF SD τ was chosen as 0, 0.25, or 0.5. Combined with the type of DIF effects (i.e., in item difficulties b_i or item intercepts ν_i), five different DIF conditions (i.e., no DIF, $\tau = 0.25$ and DIF in b_i, $\tau = 0.5$ and DIF in b_i, $\tau = 0.25$ and DIF in ν_i, and $\tau = 0.5$ and DIF in ν_i) were simulated. Item discriminations in the second group were kept identical to the base values, ensuring no DIF in discrimination parameters.

Per-group sample sizes of $N = 500$, 1000, 2000, and infinity (denoted as Inf) were selected to represent typical ranges encountered in medium- to large-scale testing scenarios involving the 2PL model [40]. For infinite sample sizes, no item responses were simulated. However, the item parameters used in MGM linking still included the random DIF effects in this case.

In each of the 4 (sample size N) $\times$ 2 (number of items I) $\times$ 5 (random DIF conditions) $=$ 40 simulation conditions, 7500 replications were conducted. The three MGM specifications—UW, DW, and PW—were applied to the simulated datasets. The bias, SD,

and root mean square error (RMSE) of the estimated mean $\widehat{\mu}$ were computed. The relative RMSE of the $\widehat{\mu}$ estimator was defined as the RMSE of a given method divided by the RMSE of the UW method, which served as the reference.

All analyses in this simulation study were performed using R (Version 4.4.1; [41]). The 2PL model was fitted using the `sirt::xxirt()` function from the R package sirt (Version 4.2-114; [42]). Dedicated functions were developed to estimate the different MGM models. Replication materials for this study can be accessed at https://osf.io/xa4qz (accessed on 3 May 2025).

3.2. Results

Table 1 presents the bias of the estimated group mean $\widehat{\mu}$ as a function of the number of items I and the sample size N. In the absence of DIF ($\tau = 0$), all three MGM methods yielded unbiased estimates. When DIF was present in either item difficulties b_i or item intercepts ν_i, the UW and DW methods continued to produce unbiased estimates. Consistent with the analytical findings in Section 2.5.3, the PW method exhibited bias under these conditions. The magnitude of the bias increased with larger DIF SD τ. Notably, the bias of the PW method did not diminish with increasing sample size N.

Table 1. Simulation Study: Bias of estimated mean $\widehat{\mu}$ as a function of the DIF SD τ, the type of DIF effects, the number of items I, and sample size N.

		τ = 0			τ = 0.25, DIF in b_i			τ = 0.5, DIF in b_i			τ = 0.25, DIF in ν_i			τ = 0.5, DIF in ν_i		
I	***N***	**UW**	**DW**	**PW**	**UW**	**DW**	**PW**	**UW**	**DW**	**PW**	**UW**	**DW**	**PW**	**UW**	**DW**	**PW**
20	500	0.003	−0.003	−0.007	0.002	−0.003	**−0.014**	0.009	0.001	**−0.026**	0.004	−0.002	**−0.013**	0.005	0.000	**−0.029**
	1000	0.002	−0.002	−0.004	0.002	−0.002	**−0.011**	0.001	−0.003	**−0.029**	0.003	−0.001	**−0.011**	0.000	−0.005	**−0.033**
	2000	0.001	0.000	−0.001	0.000	−0.001	−0.009	0.000	0.000	**−0.027**	−0.001	0.000	**−0.011**	−0.001	0.000	**−0.029**
	Inf	0.000	0.000	—	0.000	0.000	—	0.000	0.000	—	0.000	−0.001	—	−0.003	−0.001	—
40	500	0.005	0.003	−0.002	0.004	0.003	−0.008	0.004	0.006	**−0.024**	0.005	0.004	−0.008	0.005	0.009	**−0.022**
	1000	0.003	−0.001	−0.002	0.001	−0.002	−0.010	0.004	0.001	**−0.025**	0.001	−0.001	**−0.011**	−0.001	−0.004	**−0.031**
	2000	−0.003	0.001	−0.002	−0.003	0.001	−0.009	−0.004	0.001	**−0.027**	−0.002	0.002	−0.008	−0.004	0.002	**−0.027**
	Inf	0.000	0.000	—	0.001	0.001	—	0.000	0.001	—	−0.001	−0.001	—	0.002	0.001	—

Note. SD = standard deviation; Inf = infinite sample size; UW = unweighted mean–geometric mean linking; DW = discrimination-weighted mean–geometric mean linking; PW = precision-weighted mean–geometric mean linking; — = linking method not applied; Values of absolute bias larger than 0.010 are printed in bold font.

Table 2 reports the SD of the estimated group mean $\widehat{\mu}$ as a function of the number of items I and the sample size N. As expected, the SD decreased with increasing sample size and increased with higher DIF SD τ. The SD also declined with a larger number of items. In the no DIF condition ($\tau = 0$), the PW method produced estimates with the lowest SD, followed by the DW and UW methods. When random DIF was present in item difficulties b_i, PW resulted in the smallest SD for smaller sample sizes, whereas UW became more efficient than DW and PW as the sample size increased. A comparable pattern emerged for DIF in item intercepts ν_i, with the distinction that DW instead of UW yielded the smallest SD in larger samples.

Table 3 presents the relative RMSE of the estimated group mean $\widehat{\mu}$ as a function of the number of items I and the sample size N. The PW method exhibited the lowest RMSE in the no DIF condition and in DIF conditions with small sample sizes. In large samples, the UW method yielded the smallest RMSE when DIF was present in item difficulties b_i. In contrast, in conditions with DIF in item intercepts ν_i, the DW and PW methods outperformed UW, with DW showing a slight efficiency advantage over PW at larger sample sizes.

Table 2. Simulation Study: Standard deviation (SD) of estimated mean $\widehat{\mu}$ as a function of the DIF SD τ, the type of DIF effects, the number of items I, and sample size N.

		τ = 0			τ = 0.25, DIF in b_i			τ = 0.5, DIF in b_i			τ = 0.25, DIF in ν_i			τ = 0.5, DIF in ν_i		
I	***N***	**UW**	**DW**	**PW**	**UW**	**DW**	**PW**	**UW**	**DW**	**PW**	**UW**	**DW**	**PW**	**UW**	**DW**	**PW**
20	500	0.095	0.085	0.082	0.110	0.107	0.103	0.150	0.160	0.149	0.120	0.102	0.101	0.179	0.147	0.142
	1000	0.066	0.058	0.058	0.087	0.088	0.087	0.131	0.145	0.136	0.099	0.084	0.085	0.166	0.134	0.130
	2000	0.046	0.042	0.042	0.074	0.079	0.078	0.122	0.139	0.131	0.087	0.073	0.075	0.156	0.127	0.125
	Inf	0.000	0.000	—	0.056	0.066	—	0.112	0.130	—	0.072	0.058	—	0.147	0.118	—
40	500	0.083	0.079	0.077	0.090	0.091	0.087	0.116	0.124	0.114	0.099	0.089	0.087	0.135	0.117	0.109
	1000	0.058	0.054	0.054	0.071	0.071	0.069	0.098	0.106	0.099	0.078	0.069	0.069	0.120	0.100	0.097
	2000	0.040	0.038	0.038	0.056	0.060	0.059	0.089	0.103	0.095	0.065	0.057	0.057	0.111	0.092	0.089
	Inf	0.000	0.000	—	0.039	0.046	—	0.079	0.092	—	0.052	0.042	—	0.105	0.084	—

Note. Inf = infinite sample size; UW = unweighted mean–geometric mean linking; DW = discrimination-weighted mean–geometric mean linking; PW = precision-weighted mean–geometric mean linking; — = linking method not applied.

Table 3. Simulation Study: Relative root mean square error (RMSE) of estimated mean $\widehat{\mu}$ as a function of the DIF SD τ, the type of DIF effects, the number of items I, and sample size N.

		τ = 0			τ = 0.25, DIF in b_i			τ = 0.5, DIF in b_i			τ = 0.25, DIF in ν_i			τ = 0.5, DIF in ν_i		
I	***N***	**UW**	**DW**	**PW**	**UW**	**DW**	**PW**	**UW**	**DW**	**PW**	**UW**	**DW**	**PW**	**UW**	**DW**	**PW**
20	500	100	89.0	86.8	100	97.0	94.7	100	106.1	100.7	100	85.0	85.1	100	82.5	80.8
	1000	100	87.8	87.5	100	101.0	99.8	100	111.0	106.4	100	84.2	85.9	100	80.8	81.1
	2000	100	91.9	91.4	100	107.0	106.3	100	114.2	109.4	100	83.9	87.1	100	81.4	82.1
	Inf	—	—	—	100	116.7	—	100	116.1	—	100	80.7	—	100	79.8	—
40	500	100	95.3	92.9	100	100.3	96.3	100	106.7	99.9	100	90.2	87.8	100	86.2	82.2
	1000	100	93.0	93.1	100	100.0	98.8	100	109.0	104.3	100	87.7	88.9	100	83.4	84.3
	2000	100	96.0	94.5	100	107.6	106.1	100	115.2	111.4	100	86.7	87.7	100	83.2	84.0
	Inf	—	—	—	100	116.2	—	100	116.4	—	100	79.5	—	100	79.9	—

Note. SD = standard deviation; Inf = infinite sample size; — = linking method not applied; UW = unweighted mean–geometric mean linking; DW = discrimination-weighted mean–geometric mean linking; PW = precision-weighted mean–geometric mean linking; The UW method was used as the reference method to compute the relative RMSE.

Overall, the results of this Simulation Study showed that the performance of the UW, DW, and PW methods depended on the type of simulated DIF effects. The PW linking method yielded unbiased and efficient estimates only in the absence of random DIF. In DIF conditions with item difficulties affected, the UW method outperformed DW. Conversely, when DIF was simulated in item intercepts, the DW method was superior to UW.

4. Empirical Example: PISA 2006 Reading

The Simulation Study presented in Section 3 demonstrates that the performance of the three MGM methods depends on the presence and nature of random DIF in the data. To investigate whether random DIF occurs in item intercepts or item difficulties, the PISA 2006 dataset [43] for the reading domain was analyzed. This dataset includes participants from 26 selected countries (see Appendix A) that participated in the PISA 2006 study. The full PISA 2006 dataset is publicly accessible at https://www.oecd.org/en/data/datasets/pisa-2006-database.html as of (accessed on 3 May 2025).

Items in the reading domain were administered to a subset of students participating in the PISA 2006 study. The analysis included students who had been administered at least one item from the respective cognitive domain. In total, the analysis included 110,236 students, with sample sizes per country ranging from 2010 to 12,142 ($M = 4239.8$, $SD = 3046.7$).

A few of the 28 reading items were originally scored polytomous but were recoded into dichotomous scores for simplicity in this empirical example, with only the highest

category considered correct. The other items were handled as dichotomous, consistent with the original treatment in PISA.

Student (sampling) weights were applied in all analyses. To guarantee equal influence from each country, weights within each country were normalized to sum to 5000. It should be noted that the choice of 5000 is arbitrary; any constant value would serve equally well to balance contributions across countries.

In the first step, international item parameters were estimated by fitting the 2PL model to the weighted, combined dataset for each domain. These item parameters, along with other relevant information, are presented in Table 4. The average item discrimination was 1.402, suggesting a relatively well-discriminating test, while the average item difficulty was −0.163, indicating that the items were slightly easier relative to the ability of students in the total population.

Table 4. Empirical Example, PISA 2006 Reading: International item parameters and descriptive statistics of DIF effects for all 28 items.

				DIF Effects				
Item	**#CNT**	a_i	b_i	$\widehat{\tau}_{\text{obs}}$	$\widehat{\tau}_{\text{bc}}$	**Min**	**Max**	**p(SW)**
R055Q01	26	1.395	−1.486	0.218	0.210	−0.447	0.654	0.124
R055Q02	26	1.379	0.043	0.214	0.207	−0.394	0.411	0.522
R055Q03	26	1.620	−0.335	0.279	0.272	−0.445	0.496	0.095
R055Q05	26	2.117	−0.777	0.188	0.182	−0.353	0.644	0.002
R067Q01	26	1.228	−2.069	0.350	0.339	−0.664	1.022	0.033
R067Q04	26	0.832	0.723	0.710	0.694	−2.041	0.976	0.017
R067Q05	26	1.088	−0.307	0.526	0.513	−1.258	1.164	0.508
R102Q04A	25	1.460	0.669	0.383	0.373	−0.604	0.783	0.266
R102Q05	26	1.330	0.244	0.298	0.290	−0.611	0.435	0.073
R102Q07	24	1.418	−1.493	0.427	0.416	−0.680	0.821	0.083
R104Q01	26	1.628	−1.321	0.185	0.178	−0.267	0.445	0.122
R104Q02	26	0.583	1.337	0.685	0.664	−0.873	2.194	0.008
R104Q05	26	1.206	2.974	0.449	0.428	−0.674	1.066	0.530
R111Q01	26	1.365	−0.604	0.259	0.251	−0.400	0.558	0.366
R111Q02B	26	1.044	1.917	0.500	0.486	−0.858	1.027	0.738
R111Q06B	26	1.589	0.542	0.224	0.217	−0.589	0.307	0.124
R219Q01E	26	1.633	−0.250	0.295	0.287	−1.042	0.541	0.006
R219Q01T	26	1.861	−0.667	0.242	0.235	−0.522	0.478	0.986
R219Q02	26	1.534	−1.179	0.229	0.221	−0.423	0.346	0.451
R220Q01	26	1.762	0.308	0.211	0.205	−0.317	0.460	0.377
R220Q02B	25	1.521	−0.376	0.159	0.152	−0.221	0.338	0.143
R220Q04	26	1.302	−0.312	0.320	0.312	−0.546	0.373	0.003
R220Q05	26	1.977	−1.145	0.165	0.157	−0.370	0.286	0.297
R220Q06	26	1.167	−0.675	0.393	0.383	−0.500	0.688	0.014
R227Q01	26	0.778	−0.151	0.671	0.655	−1.550	1.275	0.827
R227Q02T	26	0.994	0.792	0.629	0.614	−0.995	1.437	0.738
R227Q03	26	1.665	−0.183	0.235	0.227	−0.650	0.484	0.557
R227Q06	26	1.766	−0.777	0.225	0.218	−0.314	0.555	0.021

Note . #CNT = number of countries per item; a_i = item discrimination in the 2PL model; b_i = item difficulty in the 2PL model; $\widehat{\tau}_{\text{obs}}$ = observed SD of DIF effects; $\widehat{\tau}_{\text{bc}}$ = bias-corrected estimate of SD of DIF effects; Min = smallest DIF effect per item across countries; Max = largest DIF effect per item across countries; p(SW) = *p*-value of Shapiro–Wilk test for normality of DIF effects.

In the second step, country means and country SDs were computed using the fixed international item parameters presented in Table 4. The means and SDs for the 26 countries, based on the original logit scale of the 2PL model, are reported in Table 5. The country means had an average of $M = 0.000$ (with $SD = 0.228$), while the country SDs had an average of $M = 0.973$ (with $SD = 0.078$).

In the third step, DIF effects e_i were determined for each country. The country mean and country SD were fixed at $\widehat{\mu}$ and $\widehat{\sigma}$, as obtained from the second step, while the international item parameters a_i and b_i were used. Specifically, the IRT model

$$\mathsf{P}(X_i = 1|\theta) = \Psi(a_i(\theta - b_i - e_i)) \text{ with } \mathsf{N}(\widehat{\mu}, \widehat{\sigma}^2) \tag{35}$$

was applied in each country, where N denotes the normal distribution. It is important to note that in (35), only DIF effects and their sampling variances were computed.

Using the data on estimated DIF effects, the distribution of DIF effects within each country was examined. As an initial descriptive step, the empirical SD of DIF effects e_i (i.e., $\widehat{\tau}_{\text{obs}}$) was calculated and is reported in Table 5. The average DIF SD was $M = 0.367$ with $SD = 0.091$, indicating considerable heterogeneity in DIF effects across countries.

Table 5. Empirical Example, PISA 2006 Reading: Descriptive statistics for countries and estimated SD of DIF effects.

CNT	N	I	M	SD	$\widehat{\tau}_{\text{obs}}$	$\widehat{\tau}_{\text{bc}}$	$\widehat{\tau}_{\nu,\text{bc}}$	ΔLL
AUS	7562	28	0.170	0.960	0.249	0.246	0.350	−1.46
AUT	2646	27	−0.037	1.033	0.272	0.265	0.310	3.86
BEL	4840	28	0.059	1.071	0.266	0.255	0.307	3.21
CAN	12,142	28	0.276	0.934	0.283	0.279	0.359	1.34
CHE	6578	28	0.023	0.958	0.327	0.320	0.377	3.87
CZE	3246	28	−0.168	1.130	0.335	0.326	0.393	3.25
DEU	2701	28	−0.039	1.140	0.522	0.500	0.445	11.56
DNK	2431	27	0.001	0.891	0.398	0.394	0.447	4.63
ESP	10,506	28	−0.351	0.815	0.413	0.408	0.479	3.90
EST	2630	28	−0.007	0.838	0.344	0.339	0.432	1.66
FIN	2536	28	0.516	0.854	0.330	0.326	0.378	4.26
FRA	2524	28	−0.010	0.984	0.332	0.320	0.405	1.88
GBR	7061	28	−0.016	0.985	0.340	0.336	0.447	0.48
GRC	2606	28	−0.431	0.952	0.490	0.479	0.510	6.62
HUN	2399	28	−0.148	0.918	0.320	0.306	0.371	3.23
IRL	2468	28	0.184	0.946	0.275	0.269	0.343	1.60
ISL	2010	28	−0.069	0.915	0.326	0.320	0.405	1.87
ITA	11,629	28	−0.285	0.984	0.350	0.340	0.422	2.55
JPN	3203	28	0.028	1.034	0.438	0.435	0.598	−0.51
KOR	2790	27	0.561	0.959	0.589	0.576	0.628	5.71
LUX	2443	27	−0.180	1.012	0.333	0.315	0.358	4.65
NLD	2666	28	0.092	1.017	0.429	0.425	0.516	2.98
NOR	2504	28	−0.107	1.018	0.453	0.439	0.461	7.10
POL	2968	28	0.068	1.000	0.306	0.302	0.411	−0.21
PRT	2773	28	−0.242	0.955	0.534	0.529	0.542	7.76
SWE	2374	28	0.107	1.004	0.288	0.283	0.327	4.39

Note. CNT = country label (see Appendix A); N = sample size per country; I = number of items per country; M = country mean; SD = country SD; $\widehat{\tau}_{\text{obs}}$ = observed SD of DIF effects in item difficulties b_i; $\widehat{\tau}_{\text{bc}}$ = bias-corrected SD estimate of DIF effects in item difficulties b_i; $\widehat{\tau}_{\nu,\text{bc}}$ = bias-corrected SD estimate of DIF effects in item intercepts ν_i; ΔLL = difference in log-likelihood values for models with DIF effects in item intercepts ν_i or item difficulties b_i. Positive ΔLL values indicate a better model fit for DIF effects in item intercepts.

To account for the contribution of sampling variance in the observed DIF effect estimates $\hat{e}_i$, maximum likelihood estimation was applied to the following model for DIF effects:

$$\hat{e}_i \sim \mathsf{N}(\kappa, \tau^2 + v_i^2) \text{ for } i = 1, \ldots, I. \tag{36}$$

Here, the parameters κ and the DIF SD τ were estimated, and v_i^2 denotes the estimated sampling variance of $\hat{e}_i$. Model (36) corresponds to a random-effects meta-analysis model with known error variances and was estimated using the `stats::optim()` function in R (Version: 4.4.1; [41]). The resulting τ estimates, referred to as $\widehat{\tau}_{\text{bc}}$ (i.e., bias-corrected

estimates), are also reported in Table 5 and were, as expected, slightly lower than the empirical values: $M = 0.359$, $SD = 0.089$. Note that model (36) assumes random DIF with homogeneous variances for DIF in item difficulties.

The model for DIF effects in (36) is contrasted with the alternative model

$$\hat{e}_i \sim \mathsf{N}\left(\kappa, \frac{\tau^2}{a_i^2} + v_i^2\right) \text{ for } i = 1, \ldots, I, \tag{37}$$

which represents DIF in item intercepts under the assumption of a homogeneous variance. Again, this model was fitted using the `stats::optim()` function in R [41]. The corresponding τ estimates, denoted as $\widehat{\tau}_{v,\mathrm{bc}}$, are reported in Table 5. On average, $\widehat{\tau}_{v,\mathrm{bc}}$ was slightly larger than $\widehat{\tau}_{\mathrm{bc}}$, with $M = 0.424$ and $SD = 0.083$. This result aligns with expectations, given that the average item discrimination clearly exceeded 1.

Because the competing models (36) and (37) were based on the same data (i.e., estimated DIF effects $\hat{e}_i$) and involved the same number of estimated parameters, their log-likelihood values can be directly compared to assess whether the assumption of DIF in item difficulties or item intercepts is more appropriate. The corresponding difference in log-likelihood, denoted as ΔLL, is reported in Table 5. Notably, the model assuming DIF in item intercepts provided a better fit in 23 out of 26 countries.

Table 4 also includes DIF SD estimates, $\widehat{\tau}_{\mathrm{obs}}$ and $\widehat{\tau}_{\mathrm{bc}}$, computed for individual items across countries to examine whether certain items were more susceptible to country-level DIF than others. Substantial variability in $\widehat{\tau}_{\mathrm{bc}}$ estimates across items was observed, with $M = 0.335$ and $SD = 0.165$. To evaluate the plausibility of the normality assumption for DIF effects, the Shapiro–Wilk test for normality was applied. The corresponding p values are listed as p(SW) in Table 4. A total of 8 out of 28 items showed statistically significant deviations from normality. Additionally, Figure 1 presents histograms of estimated DIF effects for nine selected items, along with the estimated DIF SD τ and the Shapiro–Wilk p value. While outliers in DIF effects were evident for items R104Q02 and R219Q01E, the overall pattern suggested unsystematic variation in DIF effects, supporting the plausibility of the normal distribution assumption for random DIF. In contrast, the commonly assumed partial invariance structure—where only a subset of items exhibit large absolute DIF effects while the majority show small or no DIF effects [44,45]—was clearly not supported by the data.

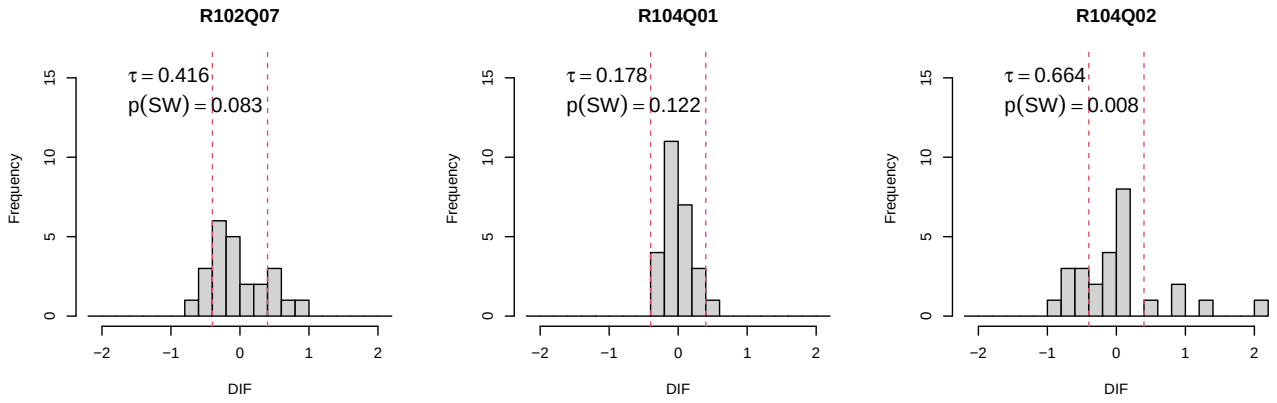

Figure 1. *Cont.*

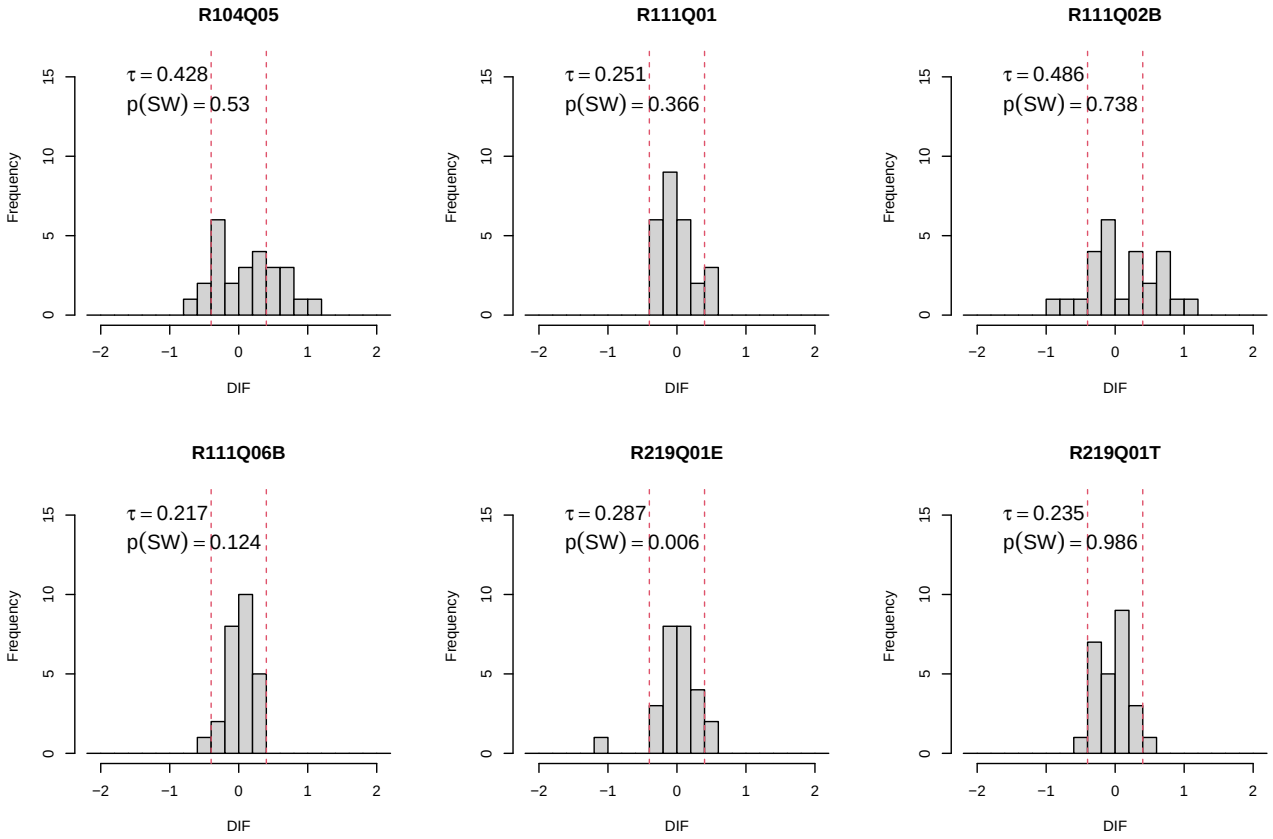

Figure 1. Empirical Example, PISA 2006 Reading: Histograms of estimated DIF effects for nine selected items (R102Q07, R104Q01, R104Q02, R104Q05, R111Q01, R111Q02B, R111Q06B, R219Q01E, and R219Q01T), along with estimated DIF SD τ and Shapiro–Wilk test for normality (p(SW)). DIF effects of -0.4 and 0.4 are displayed in a red vertical dashed line.

5. Discussion

This article examines three specifications of MGM linking that differ in the computation of the group mean $\widehat{\mu}$. The UW method assigns equal weight to all item difficulty differences; the DW method weights these differences by the squared item discrimination; and the PW method applies precision weights that account for the sampling error in item difficulty differences. The relative performance of the three methods depends on the data-generating model. When no random DIF is present, the PW method consistently outperforms the other two. In the presence of random DIF, the estimated group mean is influenced by both DIF and sampling error. Thus, the effectiveness of each method depends on the relative contribution of these two sources of variance.

When random DIF with homogeneous variances affects item difficulties, the UW method outperforms both DW and PW under large sample conditions. Conversely, if random DIF with homogeneous variances influences item intercepts, the DW method yields superior results among the three approaches.

Because the estimated precision weights in PW linking reflect the presence of DIF effects, a bias arises due to the covariance between these weights and the item difficulty differences since DIF affects both quantities. Therefore, if random DIF is suspected, the PW method is not recommended.

Given that the performance of the UW and DW methods depends on whether random DIF affects item difficulties or item intercepts, the PISA 2006 reading dataset was analyzed to investigate which assumption is more tenable. Model fit comparisons indicated that DIF effects were more prevalent in item intercepts than in item difficulties for the majority of countries. This empirical evidence suggests that the DW method may be preferable

when selecting an MGM specification in applied settings. Furthermore, as DIF in item intercepts appears to be the more tenable assumption in practice, future simulation studies may benefit from focusing on DIF effects in item intercepts rather than in item difficulties. Nevertheless, future research should examine whether this specific result from the PISA 2006 reading dataset generalizes to other empirical contexts.

However, weighting items based on item discrimination may not accurately reflect group differences, as the group difference should ideally assign equal weight to all items to preserve the intended test composition [46,47]. Nonetheless, this critique may not fully hold, as item weighting is already introduced when selecting the 2PL model—incorporating item discrimination—over the Rasch model [48], which applies equal weighting in the IRT framework.

Future research could examine the comparison between the UW and PW methods within the Rasch model. In this setting, MGM linking is replaced by mean–mean linking, as only the group means are aligned, and the group SDs are freely estimated. As noted by an anonymous reviewer, the Rasch model may exhibit special measurement properties compared to the 2PL model [49–51] (but see [52–54]), which can lead to its preference among practitioners [55–58]. Several established tools for detecting DIF have been developed within the Rasch framework [59,60]. The relative efficiency of the UW and PW methods depends on the relative magnitude of random DIF SD and sampling error. Importantly, the PW method is also expected to introduce bias in the estimated group mean under the Rasch model when random DIF effects are present.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the fact that this is a secondary data analysis, based on PISA data, for which there already is approval.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Replication material for the Simulation Study in Section 3 can be found at https://osf.io/xa4qz (accessed on 3 May 2025). The PISA 2006 dataset used in Section 4 can be downloaded from https://www.oecd.org/en/data/datasets/pisa-2006-database.html (accessed on 3 May 2025).

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

Abbreviations

The following abbreviations are used in this manuscript:

2PL	two-parameter logistic
DIF	differential item functioning
DW	discrimination-weighted mean–geometric mean linking
IRF	item response function
IRT	item response theory
MGM	mean–geometric mean
MML	marginal maximum likelihood
PW	precision-weighted mean–geometric mean linking
PISA	programme for international student assessment
RMSE	root mean square error
SD	standard deviation
UW	unweighted mean–geometric mean linking

Appendix A. Country Labels for the PISA 2006 Study

The country labels used in Table 5 are as follows: AUS = Australia; AUT = Austria; BEL = Belgium; CAN = Canada; CHE = Switzerland; CZE = Czech Republic; DEU = Germany; DNK = Denmark; ESP = Spain; EST = Estonia; FIN = Finland; FRA = France; GBR = United Kingdom; GRC = Greece; HUN = Hungary; IRL = Ireland; ISL = Iceland; ITA = Italy; JPN = Japan; KOR = Korea; LUX = Luxembourg; NLD = The Netherlands; NOR = Norway; POL = Poland; PRT = Portugal; SWE = Sweden.

References

1. Bock, R.D.; Gibbons, R.D. *Item Response Theory*; Wiley: Hoboken, NJ, USA, 2021. [CrossRef]
2. Reckase, M.D. *Multidimensional Item Response Theory Models*; Springer: New York, NY, USA, 2009. [CrossRef]
3. Yen, W.M.; Fitzpatrick, A.R. Item response theory. In *Educational Measurement*; Brennan, R.L., Ed.; Praeger Publishers: Westport, CT, USA, 2006; pp. 111–154.
4. van der Linden, W.J. Unidimensional logistic response models. In *Handbook of Item Response Theory, Volume 1: Models*; van der Linden, W.J., Ed.; CRC Press: Boca Raton, FL, USA, 2016; pp. 11–30. [CrossRef]
5. Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In *Statistical Theories of Mental Test Scores*; Lord, F.M., Novick, M.R., Eds.; MIT Press: Reading, MA, USA, 1968; pp. 397–479.
6. Kamata, A.; Bauer, D.J. A note on the relation between factor analytic and item response theory models. *Struct. Equ. Model.* **2008**, *15*, 136–153. [CrossRef]
7. van der Linden, W.J.; Barrett, M.D. Linking item response model parameters. *Psychometrika* **2016**, *81*, 650–673. [CrossRef] [PubMed]
8. Bock, R.D.; Aitkin, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **1981**, *46*, 443–459. [CrossRef]
9. Glas, C.A.W. Maximum-likelihood estimation. In *Handbook of Item Response Theory, Volume 2: Statistical Tools*; van der Linden, W.J., Ed.; CRC Press: Boca Raton, FL, USA, 2016; pp. 197–216. [CrossRef]
10. Kolen, M.J.; Brennan, R.L. *Test Equating, Scaling, and Linking*; Springer: New York, NY, USA, 2014. [CrossRef]
11. Holland, P.W.; Wainer, H., (Eds.) *Differential Item Functioning: Theory and Practice*; Lawrence Erlbaum: Hillsdale, NJ, USA, 1993. [CrossRef]
12. Millsap, R.E. *Statistical Approaches to Measurement Invariance*; Routledge: New York, NY, USA, 2011. [CrossRef]
13. Penfield, R.D.; Camilli, G. Differential item functioning and item bias. In *Handbook of Statistics, Vol. 26: Psychometrics*; Rao, C.R., Sinharay, S., Eds.; Elsevier: Amsterdam, The Netherlands, 2007; pp. 125–167. [CrossRef]
14. Lee, W.C.; Lee, G. IRT linking and equating. In *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test*; Irwing, P., Booth, T., Hughes, D.J., Eds.; Wiley: New York, NY, USA, 2018; pp. 639–673. [CrossRef]
15. Sansivieri, V.; Wiberg, M.; Matteucci, M. A review of test equating methods with a special focus on IRT-based approaches. *Statistica* **2017**, *77*, 329–352. [CrossRef]
16. Mislevy, R.J.; Bock, R.D. *BILOG 3. Item Analysis and Test Scoring with Binary Logistic Models*; Software Manual; Scientific Software International: Chicago, IL, USA, 1990.
17. Haberman, S.J. *Linking Parameter Estimates Derived from an Item Response Model Through Separate Calibrations*; (Research Report No. RR-09-40); Educational Testing Service: Princeton, NJ, USA, 2009. [CrossRef]
18. Battauz, M. equateIRT: An R package for IRT test equating. *J. Stat. Softw.* **2015**, *68*, 1–22. [CrossRef]
19. Robitzsch, A. Extensions to mean–geometric mean linking. *Mathematics* **2025**, *13*, 35. [CrossRef]
20. De Boeck, P. Random item IRT models. *Psychometrika* **2008**, *73*, 533–559. [CrossRef]
21. Fox, J.P.; Verhagen, A.J. Random item effects modeling for cross-national survey data. In *Cross-Cultural Analysis: Methods and Applications*; Davidov, E., Schmidt, P., Billiet, J., Eds.; Routledge: London, UK, 2010; pp. 461–482. [CrossRef]
22. de Jong, M.G.; Steenkamp, J.B.E.M.; Fox, J.P. Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *J. Consum. Res.* **2007**, *34*, 260–278. [CrossRef]
23. Monseur, C.; Berezner, A. The computation of equating errors in international surveys in education. *J. Appl. Meas.* **2007**, *8*, 323–335. https://bit.ly/2WDPeqD.
24. Michaelides, M.P.; Haertel, E.H. Selection of common items as an unrecognized source of variability in test equating: A bootstrap approximation assuming random sampling of common items. *Appl. Meas. Educ.* **2014**, *27*, 46–57. [CrossRef]
25. Monseur, C.; Sibberns, H.; Hastedt, D. Linking errors in trend estimation for international surveys in education. *IERI Monogr. Ser.* **2008**, *1*, 113–122. https://bit.ly/38aTVeZ.

26. Robitzsch, A.; Lüdtke, O. Linking errors in international large-scale assessments: Calculation of standard errors for trend estimation. *Assess. Educ.* **2019**, *26*, 444–465. [CrossRef]
27. Robitzsch, A. Linking error in the 2PL model. *J* **2023**, *6*, 58–84. [CrossRef]
28. Robitzsch, A. Estimation of standard error, linking error, and total error for robust and nonrobust linking methods in the two-parameter logistic model. *Stats* **2024**, *7*, 592–612. [CrossRef]
29. Robitzsch, A.; Lüdtke, O. An examination of the linking error currently used in PISA. *Meas. Interdiscip. Res. Persp.* **2024**, *22*, 61–77. [CrossRef]
30. Sachse, K.A.; Roppelt, A.; Haag, N. A comparison of linking methods for estimating national trends in international comparative large-scale assessments in the presence of cross-national DIF. *J. Educ. Meas.* **2016**, *53*, 152–171. [CrossRef]
31. Sachse, K.A.; Haag, N. Standard errors for national trends in international large-scale assessments in the case of cross-national differential item functioning. *Appl. Meas. Educ.* **2017**, *30*, 102–116. [CrossRef]
32. Wu, M. Measurement, sampling, and equating errors in large-scale assessments. *Educ. Meas.* **2010**, *29*, 15–27. [CrossRef]
33. Lohr, S.L. *Sampling: Design and Analysis*; Chapman and Hall/CRC: London, UK, 2021. [CrossRef]
34. Robitzsch, A. A comparison of linking methods for two groups for the two-parameter logistic item response model in the presence and absence of random differential item functioning. *Foundations* **2021**, *1*, 116–144. [CrossRef]
35. Chen, Y.; Li, C.; Ouyang, J.; Xu, G. DIF statistical inference without knowing anchoring items. *Psychometrika* **2023**, *88*, 1097–1122. [CrossRef]
36. Yuan, K.H.; Cheng, Y.; Patton, J. Information matrices and standard errors for MLEs of item parameters in IRT. *Psychometrika* **2014**, *79*, 232–254. [CrossRef] [PubMed]
37. Asparouhov, T.; Muthén, B. Multiple-group factor analysis alignment. *Struct. Equ. Model.* **2014**, *21*, 495–508. [CrossRef]
38. Muthén, B.; Asparouhov, T. IRT studies of many groups: The alignment method. *Front. Psychol.* **2014**, *5*, 978. [CrossRef]
39. Barrett, M.D.; van der Linden, W.J. Estimating linking functions for response model parameters. *J. Educ. Behav. Stat.* **2019**, *44*, 180–209. [CrossRef]
40. Rutkowski, L.; von Davier, M.; Rutkowski, D. (Eds.) *A Handbook of International Large-scale Assessment: Background, Technical Issues, and Methods of Data Analysis*; Chapman Hall/CRC Press: London, UK, 2013. [CrossRef]
41. R Core Team. *R: A Language and Environment for Statistical Computing*; R Core Team: Vienna, Austria, 2024. Available online: https://www.R-project.org (accessed on 15 June 2024).
42. Robitzsch, A. *sirt: Supplementary Item Response Theory Models*, R Package Version 4.2-114. 2025. Available online: https://github.com/alexanderrobitzsch/sirt (accessed on 7 April 2025).
43. OECD. *PISA 2006 Technical Report*; OECD: Paris, France, 2009. Available online: https://bit.ly/38jhdzp (accessed on 3 May 2025).
44. von Davier, M.; Yamamoto, K.; Shin, H.J.; Chen, H.; Khorramdel, L.; Weeks, J.; Davis, S.; Kong, N.; Kandathil, M. Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assess. Educ.* **2019**, *26*, 466–488. [CrossRef]
45. Joo, S.H.; Khorramdel, L.; Yamamoto, K.; Shin, H.J.; Robin, F. Evaluating item fit statistic thresholds in PISA: Analysis of cross-country comparability of cognitive items. *Educ. Meas.* **2021**, *40*, 37–48. [CrossRef]
46. Brennan, R.L. Misconceptions at the intersection of measurement theory and practice. *Educ. Meas.* **1998**, *17*, 5–9. [CrossRef]
47. Camilli, G. IRT scoring and test blueprint fidelity. *Appl. Psychol. Meas.* **2018**, *42*, 393–400. [CrossRef]
48. Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*; Danish Institute for Educational Research: Copenhagen, Denmark, 1960.
49. Heine, J.H.; Heene, M. Measurement and mind: Unveiling the self-delusion of metrification in psychology. *Meas. Interdiscip. Res. Persp.* 2024, *Epub ahead of print*. [CrossRef]
50. Salzberger, T. The illusion of measurement: Rasch versus 2-PL. *Rasch Meas. Trans.* **2002**, *16*, 882. Available online: https://tinyurl.com/25wzmzb5 (accessed on 3 May 2025).
51. Linacre, J.M. Understanding Rasch measurement: Estimation methods for Rasch measures. *J. Outcome Meas.* **1999**, *3*, 382–405. Available online: https://bit.ly/2UV6Eht (accessed on 3 May 2025).
52. Ballou, D. Test scaling and value-added measurement. *Educ. Financ. Policy* **2009**, *4*, 351–383. [CrossRef]
53. van der Linden, W.J. Fundamental measurement and the fundamentals of Rasch measurement. In *Objective Measurement: Theory Into Practice (Vol. 2)*; Wilson, M., Ed.; Ablex Publishing Corporation: Hillsdale, NJ, USA, 1994; pp. 3–24.
54. Robitzsch, A.; Lüdtke, O. Some thoughts on analytical choices in the scaling model for test scores in international large-scale assessment studies. *Meas. Instrum. Soc. Sci.* **2022**, *4*, 9. [CrossRef]
55. Andrich, D.; Marais, I. *A Course in Rasch Measurement Theory*; Springer: New York, NY, USA, 2019. [CrossRef]
56. Engelhard, G. *Invariant Measurement*; Routledge: New York, NY, USA, 2012. [CrossRef]
57. Melin, J.; Bonn, S.E.; Pendrill, L.; Lagerros, Y.T. A questionnaire for assessing user satisfaction with mobile health apps: Development using Rasch measurement theory. *JMIR mHealth uHealth* **2020**, *8*, e15909. [CrossRef]

58. Wu, M.; Tam, H.P.; Jen, T.H. *Educational Measurement for Applied Researchers*; Springer: Singapore, 2016. [CrossRef]
59. Tennant, A.; Pallant, J.F. DIF matters: A practical approach to test if differential item functioning makes a difference. *Rasch Meas. Trans.* **2007**, *20*, 1082–1084. Available online: https://rb.gy/wbiku0 (accessed on 3 May 2025).
60. Melin, J.; Cano, S.; Flöel, A.; Göschel, L.; Pendrill, L. The role of entropy in construct specification equations (CSE) to improve the validity of memory tests: Extension to word lists. *Entropy* **2022**, *24*, 934. [CrossRef]