foundations

MDPI

Article

Bias-Corrected Root Mean Square Deviation Estimators

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

The root mean square deviation (RMSD) is a widely used item fit statistic in item response models. However, the sample RMSD is known to exhibit positive bias in small samples. To address this, seven alternative bias-corrected RMSD estimators are proposed and evaluated in a simulation study involving items with uniform differential item functioning (DIF). The results demonstrate that the proposed estimators effectively reduce the bias of the original RMSD statistic. Their performance is compared, and the most favorable estimators are highlighted for empirical research. Finally, the application of the various RMSD statistics is illustrated using PISA 2006 reading data.

Keywords: item response model; 2PL model; differential item functioning; RMSD; item fit

1. Introduction

Item response theory (IRT) models [1–6] are multivariate statistical models for vectors of discrete random variables. They are widely used in the social sciences, particularly in educational large-scale assessment (LSA; [7]) studies, where cognitive tasks are administered.

This article considers dichotomous (binary) random variables. Let $\boldsymbol{X} = (X_1, \ldots, X_I)$ denote a vector of I items, with $X_i \in \{0, 1\}$. A unidimensional IRT model [8] specifies the probability distribution of $\boldsymbol{X}$ as

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

where ϕ is the normal density with mean μ and SD σ. The latent variable θ, often referred to as a trait or ability, has parameters collected in $\boldsymbol{\delta} = (\mu, \sigma)$. The vector $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_I)$ contains the item parameters for the parametric item response functions (IRFs) $P_i(\theta; \boldsymbol{\gamma}_i) = \mathsf{P}(X_i = 1 \mid \theta)$. The latent ability variable θ can be viewed as a unidimensional summary of the high-dimensional contingency table of item responses $\boldsymbol{X}$ (see [9]). Larger values of θ (typically positive) are associated with more able persons who solve more items, whereas smaller values of θ (typically negative) indicate less able persons.

The two-parameter logistic (2PL; [10]) model is among the most widely used IRT models. Its IRF is given by

$$P_i(\theta; \boldsymbol{\gamma}_i) = \Psi\big(a_i(\theta - b_i)\big), \tag{2}$$

where a_i and b_i denote item discrimination and difficulty, respectively, and $\Psi(x) = (1 + \exp(-x))^{-1}$ is the logistic function. The item parameter vector is denoted by $\boldsymbol{\gamma}_i = (a_i, b_i)$. The 2PL model accommodates differences in how strongly items relate to the ability variable θ through variation in the discrimination parameter a_i, with larger

check for updates

Academic Editor: Wei-Shih Du

Received: 17 October 2025
Revised: 19 November 2025
Accepted: 24 November 2025
Published: 28 November 2025

Citation: Robitzsch, A. Bias-Corrected Root Mean Square Deviation Estimators. *Foundations* **2025**, *5*, 36. https://doi.org/10.3390/foundations5040036

Copyright: © 2025 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

positive values indicating stronger associations and values near zero indicating weaker associations. Variation in the marginal probability of answering an item correctly is mainly captured by the difficulty parameter b_i, where positive values correspond to more difficult items and negative values correspond to easier items.

For a sample of N individuals with independent and identically distributed observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ of $\boldsymbol{X}$, the parameters of the IRT model in (1) are consistently estimated via marginal maximum likelihood (MML; [11]), commonly implemented using the EM algorithm [12].

In LSA applications such as programme for international student assessment (PISA; [13,14]), item response functions (IRFs) are typically modeled parametrically as in (1). These studies involve multiple countries, and it is generally assumed that item parameters are invariant across countries. In practice, this assumption may be violated, as certain items can systematically advantage or disadvantage specific countries. This phenomenon is known as differential item functioning (DIF; [15,16]), although alternative terms such as measurement bias or item bias are also used [17]. The presence of DIF may bias group differences, which motivates the search for items with DIF to ensure that comparisons are less distorted [18]. However, DIF may also cancel out across the test (i.e., balanced DIF; [19–21]), a situation in which removing DIF items from the test might be required.

As a result, the assumed IRF represents a (slight) misspecification of the true IRT model (1) for a given country (or group, henceforth). The multivariate random vector $\boldsymbol{X}$ can then be expressed as

$$\mathsf{P}(\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\delta}, \boldsymbol{\gamma}) \simeq \int \prod_{i=1}^{I} \left[P_i^*(\theta; \boldsymbol{\gamma}_i)^{x_i} (1 - P_i^*(\theta; \boldsymbol{\gamma}_i))^{1-x_i} \right] \phi(\theta; \mu, \sigma) \, \mathrm{d}\theta, \tag{3}$$

where P_i^* denotes the assumed IRF and P_i represents the IRF in the data-generating model for the group. In practice, the approximation of P_i by P_i^* is generally expected to introduce only minimal distortion in estimating the distribution parameters μ and σ.

Assessing the adequacy of parametric IRFs (i.e., item fit; [22–26]) is a central issue in psychometrics. The discrepancy between the true IRF P_i and the assumed parametric IRF P_i^* should be quantified using an appropriate effect size measure, ideally accompanied by statistical inference. Of particular interest is the identification of misfitting items i for which the assumed IRFs P_i^* deviate substantially from P_i.

The present study focuses the root mean square deviation (RMSD; [27–36]). The motivation for examining RMSD lies in its widespread use in current PISA studies.

A weighted RMSD statistic is defined following Joo et al. [37]. Let ω_i denote a weighting function for item i such that it integrates to 1, that is, $\int \omega_i(\theta) \, \mathrm{d}\theta = 1$ (see [37]). The weighting function ω_i may be item-specific or common across items. The primary objective is to quantify the discrepancy between a data-generating IRF P_i and the model-assumed IRF P_i^*. The RMSD statistic applies weights to the deviations $P_i(\theta) - P_i^*(\theta)$ to provide a summarized measure of model-data discrepancy. The weighted RMSD statistic summarizes squared IRF differences by calculating

$$\mathsf{RMSD}_{\omega_i, i} = \sqrt{\int \left[P_i(\theta) - P_i^*(\theta) \right]^2 \omega_i(\theta) \, \mathrm{d}\theta}. \tag{4}$$

If the weighting function ω is chosen as the normal density with group mean μ and group SD σ, and is the same across all items, the RMSD statistic is referred to as the distribution-weighted RMSD. If the weighting function ω_i is item-specific, chosen as the normal density with mean b_i (the assumed item difficulty in the model) and SD of 1, it is referred to as the difficulty-weighted RMSD.

The assumed IRF P_i^* typically depends on item parameters that are either known (i.e., fixed) or estimated. In LSA studies, the item parameters used in P_i^* often originate from an international scaling step in which item parameters are obtained by pooling all country-specific datasets into a single total dataset. The resulting international item parameters are then treated as fixed in the subsequent country-specific scaling step. Differences between country-specific IRFs P_i and the model-assumed IRF P_i^* typically arise from country DIF [38]; that is, the functioning of items differs slightly across countries. For instance, some items become easier in a particular country relative to the international average, whereas others become more difficult. The RMSD statistic (4) is therefore used to detect items exhibiting country DIF.

As another application, P_i^* in (4) can represent the assumed IRF in the 2PL model. If the IRT model is treated as a working model, the true IRF P_i differs from the model-assumed IRF P_i^*. For example, the true IRF might include a guessing parameter [39] or could take the form of any monotone function of the ability variable θ (see [40,41]). In this context, the RMSD statistic can be employed to detect misfitting items.

The RMSD definition (4) is formulated at the population level and does not involve sample data. In empirical research, model fit is assessed in datasets with limited sample sizes. Therefore, it is essential to investigate estimators of the RMSD statistic that perform well in small-sample settings.

A sample-based version of the RMSD statistic is now defined. Let $\widehat{P}_i$ denote the observed IRF, a sample-based estimate of P_i, evaluated at a theta point θ_t ($t = 1, \ldots, T$) as

$$\widehat{p}_{it} = \widehat{P}_i(\theta_t) = \frac{\sum_{n=1}^{N} h_{nt} x_{ni}}{\sum_{n=1}^{N} h_{nt}}, \tag{5}$$

where h_{nt} represents the posterior distribution of person n at grid point θ_t. The posterior distribution is typically obtained by fitting the IRT model via MML [11], so the quantities in (5) can be computed directly from standard software output. A discrete evaluation of the weighting function w_i is then defined as

$$\omega_{it} = \frac{\omega_i(\theta_t)}{\sum_{s=1}^{T} \omega_i(\theta_s)} \text{ for } t = 1, \ldots, T. \tag{6}$$

The sample-based RMSD is then defined as

$$\widehat{\mathrm{RMSD}}_{\omega_i, i} = \sqrt{\sum_{t=1}^{T} \left[\widehat{p}_{it} - p_{it}^*\right]^2 \omega_{it}}, \tag{7}$$

where $p_{it}^* = P_i(\theta_t)^*$.

It has been shown in [33] that the RMSD estimator (7) is positively biased in small samples. Consequently, it is desirable to develop bias-corrected RMSD estimators with improved properties. This paper proposes seven alternative bias-corrected RMSD estimators, derived based on asymptotic bias considerations. The performance of these estimators is evaluated in a simulation study in which items exhibit DIF.

The rest of the article is organized as follows. Section 2 introduces the newly proposed bias-corrected RMSD estimators. Section 3 reports results from a simulation study examining the performance of these RMSD statistics under uniform DIF. Section 4 presents an empirical example using PISA 2006 reading data. Finally, Section 5 closes with a discussion.

2. Derivation of Bias-Corrected RMSD Estimators

This section discusses alternative bias-corrected estimators of the RMSD statistic. The population RMSD value is defined as

$$\mathcal{T} = \mathcal{T}(\boldsymbol{p}_i) = \sqrt{\sum_{t=1}^{T}(p_{it} - p_{it}^*)^2 \omega_{it}}. \tag{8}$$

This statistic depends on the true data-generating item response probabilities p_{it}, which are unknown and must be inferred to compute the RMSD statistic. The quantities p_{it}^* and ω_{it} are treated as fixed in the following discussion.

The originally proposed RMSD statistic $\hat{T}_0$ is defined as

$$\hat{T}_0 = \hat{T}_0(\widehat{\boldsymbol{p}}_i) = \sqrt{\sum_{t=1}^{T}(\widehat{p}_{it} - p_{it}^*)^2 \omega_{it}}, \tag{9}$$

and depends on the estimated item response probabilities $\widehat{\boldsymbol{p}}_i$. The substitution of $\boldsymbol{p}_i$ with $\widehat{\boldsymbol{p}}_i$ introduces sampling error, which leads to a positive bias because these errors appear in the squared terms of (9) (see [33,42]).

A bias correction approach relies on quantifying the sampling variances of the estimates $\widehat{p}_{it}$. Let $\boldsymbol{V}_i$ denote the variance matrix $\mathsf{Var}(\widehat{\boldsymbol{p}}_i)$. This matrix can be estimated using the framework of M-estimation [43], as the estimates $\widehat{p}_{it}$ satisfy the estimating equations

$$\sum_{n=1}^{N} h_{nt}(x_{ni} - p_{it}) = 0 \text{ for } t = 1, \ldots, T, \tag{10}$$

where h_{nt} represents the posterior probability of subject n at the theta grid point θ_t. The variance of $\widehat{p}_{it}$ is given by

$$v_t = \mathsf{Var}(\widehat{p}_{it}) = \frac{\sum_{n=1}^{N} h_{nt}^2 (x_{ni} - p_{it})^2}{\left(\sum_{n=1}^{N} h_{nt}\right)^2}, \tag{11}$$

and the covariance between $\widehat{p}_{is}$ and $\widehat{p}_{it}$ is expressed as

$$v_{st} = \mathsf{Cov}(\widehat{p}_{is}, \widehat{p}_{it}) = \frac{\sum_{n=1}^{N} h_{ns} h_{nt} (x_{ni} - p_{is})(x_{ni} - p_{it})}{\left(\sum_{n=1}^{N} h_{ns}\right)\left(\sum_{n=1}^{N} h_{nt}\right)}. \tag{12}$$

For distant θ points θ_s and θ_t, the numerator in (12) shows that the covariance v_{st} is approximately zero, since $h_{ns}h_{nt} \approx 0$ when the posterior distribution of subject n is concentrated around a specific θ value. Note that the formulas (11) and (12) also appeared in [44].

The bias in the squared RMSD, $\hat{T}_0^2$, is first considered. Note that

$$\mathsf{E}\left(\hat{T}_0^2\right) = \mathsf{E}\left(\sum_{t=1}^{T}(\widehat{p}_{it} - p_{it}^*)^2 \omega_{it}\right) = \mathcal{T}^2 + \mathsf{E}\left(\sum_{t=1}^{T}(\widehat{p}_{it} - p_{it})^2 \omega_{it}\right), \tag{13}$$

where $\mathsf{E}(\widehat{p}_{it} - p_{it}) = 0$. This directly implies

$$\mathsf{E}\left(\sum_{t=1}^{T}(\widehat{p}_{it} - p_{it}^*)(p_{it} - p_{it}^*)\omega_{it}\right) = \sum_{t=1}^{T}(p_{it} - p_{it}^*)\omega_{it}\mathsf{E}(\widehat{p}_{it} - p_{it}^*) = 0, \tag{14}$$

which proves (13).

The second term in (13) represents the positive bias in $\hat{T}_0^2$ and is given by

$$\mathcal{B} = \mathsf{Bias}(\hat{T}_0^2) = \mathsf{E}\left(\sum_{t=1}^{T}(\widehat{p}_{it} - p_{it})^2\omega_{it}\right) = \sum_{t=1}^{T} v_t\omega_t. \tag{15}$$

This result motivates the bias-corrected RMSD estimator

$$\hat{T}_1 = \sqrt{(\hat{T}_0^2 - \mathcal{B})_+}, \tag{16}$$

where $(x)_+ = \max(x, 0)$. This approach was also proposed in [33], but without incorporating the more appropriate variance estimation given in (11) and (12).

However, Jensen's inequality [43,45] implies that

$$\mathsf{E}\left(\sqrt{Z}\right) < \sqrt{\mathsf{E}(Z)} \tag{17}$$

for any positive random variable Z. Applying (17) to $\hat{T}_1$ and disregarding negative values yields

$$\mathsf{E}(\hat{T}_1) < \sqrt{\mathsf{E}\hat{T}_1^2} = \sqrt{\mathsf{E}(\hat{T}_0^2 - \mathcal{B})} = \sqrt{\mathcal{T}^2} = \mathcal{T}. \tag{18}$$

Thus, $\hat{T}_1$ exhibits a negative bias. A bias correction applied to the squared RMSD (i.e., $\hat{T}_0^2$) removes bias in $\hat{T}_1^2$ for $\mathcal{T}^2$, but introduces bias in $\hat{T}_1$. Therefore, it is more appropriate to derive a bias correction directly for $\hat{T}_0$ that operates on the square-root scale.

A second type of bias correction can be derived using a Taylor expansion of $\hat{T}_0$ around $\widehat{\boldsymbol{p}}_i$ with respect to the true IRF $\boldsymbol{p}_i$. Because the RMSD statistic is defined as the square root of a positively valued statistic that exhibits bias, the correction relies on the quadratic Taylor expansion of the square root function given by

$$\sqrt{x} \simeq \sqrt{x_0} + \frac{1}{2\sqrt{x_0}}(x - x_0) - \frac{1}{8x_0^{3/2}}(x - x_0)^2 \tag{19}$$

for positive x and x_0 with x close to x_0. Letting $y = \sqrt{x}$ and $y_0 = \sqrt{x_0}$ yields the approximation

$$y \simeq y_0 + \frac{1}{2y_0}(y^2 - y_0^2) - \frac{1}{8y_0^3}\left(y^2 - y_0^2\right)^2. \tag{20}$$

In (20), replacing y by the sample RMSD $\hat{T}_0$ and y_0 by the population RMSD $\mathcal{T}$ gives

$$\hat{T}_0 \simeq \mathcal{T} + \frac{1}{2\mathcal{T}}(\hat{T}_0^2 - \mathcal{T}^2) - \frac{1}{8\mathcal{T}^3}\left(\hat{T}_0^2 - \mathcal{T}^2\right)^2, \tag{21}$$

where $\mathcal{T}^2$ is the population MSD statistic. Taking expectations in (21) results in

$$\mathsf{E}(\hat{T}_0 - \mathcal{T}) \simeq \frac{1}{2\mathcal{T}}\mathsf{E}(\hat{T}_0^2 - \mathcal{T}^2) - \frac{1}{8\mathcal{T}^3}\mathsf{E}\left(\hat{T}_0^2 - \mathcal{T}^2\right)^2. \tag{22}$$

Assuming that $\mathsf{E}\left(\hat{T}_0^2 - \mathcal{T}^2\right)^2$ is approximately equal to $\mathsf{Var}(\hat{T}_0^2)$, the bias of $\hat{T}_0$ can then be obtained from (22) as

$$\mathsf{Bias}(\hat{T}_0) \simeq \frac{1}{2\mathcal{T}}\mathsf{Bias}(\hat{T}_0^2) - \frac{1}{8\mathcal{T}^3}\mathsf{E}\left(\hat{T}_0^2 - \mathcal{T}^2\right)^2. \tag{23}$$

The bias approximation (23) is used to derive bias-corrected estimators for $\hat{T}_0$. At first, only the linear term from the Taylor expansion is employed; that is, the bias component involving $\hat{T}_0^2$ (i.e., $\mathsf{Bias}(\hat{T}_0^2)$) from (23) which can be readily estimated as $\mathcal{B}$ in (15), while

ignoring the term that involves $\mathsf{E}\big(\hat{T}_0^2 - \mathcal{T}^2\big)^2$. The quantity $\mathcal{T}$ in this term can be estimated using $\hat{T}_0$ or $\hat{T}_1$, yielding the following bias-corrected RMSD estimators:

$$\hat{T}_2 = \left(\hat{T}_0 - \frac{\mathcal{B}}{2\hat{T}_0}\right)_+ \text{ and} \tag{24}$$

$$\hat{T}_3 = \left(\hat{T}_0 - \frac{\mathcal{B}}{2\hat{T}_1}\right)_+. \tag{25}$$

Because $\hat{T}_1$ is smaller than $\hat{T}_0$, the bias-corrected estimator $\hat{T}_2$ is always at least as large as $\hat{T}_3$.

Further bias-corrected estimators follow by retaining the second term in (23) that involves $\mathsf{E}\big(\hat{T}_0^2 - \mathcal{T}^2\big)^2$. First, note that

$$\mathsf{E}\Big(\hat{T}_0^2 - \mathcal{T}^2\Big)^2 = \Big[\mathsf{Bias}(\hat{T}_0^2)\Big]^2 + \mathsf{Var}(\hat{T}_0^2) = \mathcal{B}^2 + \mathsf{Var}(\hat{T}_0^2). \tag{26}$$

The term $\mathcal{B}^2$ can be regarded as small relative to $\mathsf{Var}(\hat{T}_0^2)$, which motivates the approximation of $\mathsf{E}\big(\hat{T}_0^2 - \mathcal{T}^2\big)^2$ by $\mathsf{Var}(\hat{T}_0^2)$. An approximate expression for $\mathsf{Var}(\hat{T}_0^2)$ is

$$\mathsf{Var}(\hat{T}_0^2) = 4\sum_{s=1}^{T}\sum_{t=1}^{T}(p_{is} - p_{is}^*)(p_{it} - p_{it}^*)v_{st}\omega_{is}\omega_{it}. \tag{27}$$

An natural estimator of this variance is given by

$$\mathcal{C} = \sum_{s=1}^{T}\sum_{t=1}^{T}(\widehat{p}_{is} - p_{is}^*)(\widehat{p}_{it} - p_{it}^*)v_{st}\omega_{is}\omega_{it}, \tag{28}$$

but the substitution of p_{it} with $\widehat{p}_{it}$ introduces an additional variance component that arises solely from sampling variability. This additional component contributes bias to the estimator and is approximately given by

$$\mathcal{D} = \sum_{s=1}^{T}\sum_{t=1}^{T}v_{st}^2\omega_{is}\omega_{it}. \tag{29}$$

The estimators $\hat{T}_4$ and $\hat{T}_5$ replace $\mathcal{T}$ by $\hat{T}_0$ and differ in the quantity used to substitute for $\mathsf{E}\big(\hat{T}_0^2 - \mathcal{T}^2\big)^2$. The statistic $\hat{T}_5$ corrects for the positive bias in the variance estimate that affects $\hat{T}_4$:

$$\hat{T}_4 = \left(\hat{T}_0 - \frac{\mathcal{B}}{2\hat{T}_0} + \frac{\mathcal{C}}{2\hat{T}_0^3}\right)_+ \text{ and} \tag{30}$$

$$\hat{T}_5 = \left(\hat{T}_0 - \frac{\mathcal{B}}{2\hat{T}_0} + \frac{\mathcal{C} - \mathcal{D}}{2\hat{T}_0^3}\right)_+. \tag{31}$$

Finally, the last two bias-corrected RMSD estimators are derived by substituting $\hat{T}_1$ for $\hat{T}_0$ in (30) and (31), resulting in

$$\hat{T}_6 = \left(\hat{T}_0 - \frac{\mathcal{B}}{2\hat{T}_1} + \frac{\mathcal{C}}{2\hat{T}_1^3}\right)_+ \text{ and} \tag{32}$$

$$\hat{T}_7 = \left(\hat{T}_0 - \frac{\mathcal{B}}{2\hat{T}_1} + \frac{\mathcal{C} - \mathcal{D}}{2\hat{T}_1^3}\right)_+. \tag{33}$$

In practice, divisions by zero values of $\hat{T}_1$ must be prevented when computing the bias-corrected estimators. In such cases, the bias-corrected RMSD estimates $\hat{T}_3$, $\hat{T}_6$, and $\hat{T}_7$ are set to zero.

In the simulation study in Section 3 and the empirical example in Section 4, the distribution-weighted RMSD statistics are denoted by $\hat{R}_d$, and the difficulty-weighted RMSD statistics by $\hat{U}_d$ ($d = 0, 1, \ldots, 7$).

3. Simulation Study

3.1. Method

The 2PL IRT model was applied for both data generation and analysis. The latent trait θ followed a normal distribution with mean 0 and standard deviation 1. The simulation study comprised $I = 40$ items. Ten base items were defined with discrimination parameters $a_i = 1$ for all items, and difficulty parameters b_i fixed at -2.00, -1.56, -1.11, -0.67, -0.22, 0.22, 0.67, 1.11, 1.56, and 2.00. These ten items were each duplicated four times to yield a total test length of 40 items. Two of the 40 items were simulated to exhibit DIF. Items j and $10 - j + 1$, for $j = 1$, $j = 3$, or $j = 5$, were assigned uniform DIF in the difficulty parameters b_i with values δ and $-\delta$, respectively. The DIF effect size δ was set to -0.6 and 0.6, representing large DIF magnitudes [16,42].

The sample size N was varied across 125, 250, 500, 1000, 2000, and 4000, reflecting typical conditions in large-scale assessment studies [7].

For each of the 6 (sample size N) $\times$ 3 (type of DIF items) $\times$ 2 (DIF effect size δ) $= 36$ simulation conditions, 1500 replications were performed. Item parameters were fixed at the values of the base items when fitting the 2PL model. While all item parameters remained fixed, the distribution parameters μ and σ of the latent trait were freely estimated. The presence of DIF was ignored during the scaling step. The distribution-weighted RMSD statistics $\hat{R}_d$ and difficulty-weighted RMSD statistics $\hat{U}_d$ ($d = 0, 1, \ldots, 7$) were computed (see Section 2).

A true RMSD value was defined using the IRF from the data-generating model that included the DIF items. Table 1 presents the RMSD values for items exhibiting DIF effects. The results indicate that the distribution-weighted RMSD varies with item difficulty b_i for a fixed DIF effect δ_i, whereas the difficulty-weighted RMSD statistic remains unaffected. For items without DIF effects (48 out of 50 items), the RMSD values were zero.

Table 1. Simulation Study: Population values of the root mean square deviation (RMSD) statistic as a function of item difficulty b_i and DIF effect size δ_i.

δ_i	b_i	dist	diff
-0.6	-0.22	0.122	0.125
	-1.11	0.100	0.125
	-2.00	0.069	0.125
0.6	-0.22	0.126	0.125
	-1.11	0.117	0.125
	-2.00	0.091	0.125

Note. dist = distribution-weighted RMSD; diff = difficulty-weighted RMSD.

The bias and root mean square error (RMSE) of the RMSD statistics were calculated with respect to the true RMSD values. A relative RMSE was defined by dividing each RMSD statistic by the originally proposed RMSD statistic $\hat{R}_0$ or $\hat{U}_0$, and multiplying by 100. In addition, the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles were computed for the RMSD statistics.

All analyses were conducted using the open-source statistical software R (Version 4.4.1; [46]). The 2PL model was estimated with the `sirt::xxirt()` function from

the R package sirt (Version 4.2-133; [47]). Custom R functions were developed to compute the various RMSD statistics. These functions, along with the replication materials, are available at https://osf.io/amd8k (accessed on 17 October 2025). Percentile plots were generated using the R package ggplot2 (Version 4.0.0; [48]).

3.2. Results

Table 2 presents the bias and relative RMSE for the distribution-weighted RMSD statistics of items with DIF. Substantial positive biases were observed for the RMSD statistic $\hat{R}_0$ at the smaller sample sizes ($N = 125$ and $N = 250$), which were markedly reduced by the bias-corrected RMSD statistics. The average absolute bias was $M = 0.004$ for $\hat{R}_0$ and $M = 0.001$ for $\hat{R}_4$, $\hat{R}_5$, and $\hat{R}_6$. The positive bias of $\hat{R}_0$ appeared relatively unaffected by item difficulty b_i. Across all simulation conditions, the RMSE was higher for the bias-corrected RMSD statistics than for the original statistic $\hat{R}_0$. The lowest average RMSE values were obtained for $\hat{R}_4$ ($M = 103.5$) and $\hat{R}_6$ ($M = 103.9$), whereas $\hat{R}_3$ showed the largest average RMSE value of $M = 111.3$.

Table 2. Simulation Study: Bias and relative root mean square error (RMSE) of distribution-weighted RMSD statistics $\hat{R}_d$ ($d = 0, \ldots, 7$) for items with differential item functioning (DIF) as a function of item difficulty b_i , DIF effect size δ_i and sample size N

			Bias								**Relative RMSE**							
b_i	δ_i	N	$\hat{R}_0$	$\hat{R}_1$	$\hat{R}_2$	$\hat{R}_3$	$\hat{R}_4$	$\hat{R}_5$	$\hat{R}_6$	$\hat{R}_7$	$\hat{R}_0$	$\hat{R}_1$	$\hat{R}_2$	$\hat{R}_3$	$\hat{R}_4$	$\hat{R}_5$	$\hat{R}_6$	$\hat{R}_7$
−0.22	−0.6	125	**0.016**	−0.005	−0.002	−0.007	0.003	0.002	0.004	0.000	100	**115.7**	107.3	**122.2**	104.1	108.3	106.7	**115.9**
		250	0.006	−0.004	−0.003	−0.004	0.000	−0.001	0.000	−0.001	100	108.7	106.5	**111.3**	103.4	104.6	102.8	106.0
		500	0.003	−0.002	−0.002	−0.002	0.000	−0.001	0.000	−0.001	100	104.0	103.6	104.4	101.8	102.1	101.8	102.1
		1000	0.001	−0.001	−0.001	−0.001	0.000	0.000	0.000	0.000	100	102.3	102.2	102.4	101.1	101.2	101.1	101.2
		2000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	100	100.6	100.6	100.6	100.2	100.2	100.2	100.2
		4000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	100	100.7	100.7	100.7	100.4	100.4	100.4	100.4
	0.6	125	**0.015**	−0.006	−0.003	−0.009	0.002	0.000	0.003	−0.001	100	**115.2**	107.8	**122.9**	104.4	108.3	104.7	**114.5**
		250	0.007	−0.004	−0.003	−0.004	0.000	0.000	0.000	−0.001	100	109.4	106.6	**112.3**	103.7	105.1	103.5	106.9
		500	0.003	−0.002	−0.002	−0.002	0.000	0.000	0.000	0.000	100	103.5	103.2	103.9	101.6	101.8	101.6	101.9
		1000	0.001	−0.001	−0.001	−0.001	0.000	0.000	0.000	0.000	100	102.0	101.9	102.0	100.9	101.0	100.9	101.0
		2000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	100	100.8	100.8	100.9	100.4	100.4	100.4	100.4
		4000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	100	100.6	100.6	100.6	100.3	100.3	100.3	100.3
−1.11	−0.6	125	**0.014**	−0.007	−0.004	**−0.011**	0.001	−0.002	0.002	−0.004	100	**122.5**	**111.4**	**133.4**	107.3	**113.5**	108.9	**125.6**
		250	0.006	−0.005	−0.004	−0.006	−0.001	−0.002	−0.001	−0.002	100	**114.4**	110.0	**119.7**	105.9	108.5	104.0	**112.0**
		500	0.002	−0.003	−0.003	−0.003	−0.001	−0.001	−0.001	−0.001	100	106.7	106.0	107.6	103.4	104.0	103.2	104.2
		1000	0.000	−0.003	−0.003	−0.003	−0.002	−0.002	−0.002	−0.002	100	105.3	105.1	105.5	103.1	103.3	103.1	103.3
		2000	−0.001	−0.002	−0.002	−0.002	−0.002	−0.002	−0.002	−0.002	100	103.5	103.5	103.6	102.1	102.2	102.1	102.2
		4000	−0.001	−0.002	−0.002	−0.002	−0.002	−0.002	−0.002	−0.002	100	103.1	103.1	103.2	102.0	102.0	102.0	102.0
	0.6	125	**0.013**	−0.009	−0.006	**−0.013**	0.000	−0.002	0.001	−0.004	100	**120.6**	**110.9**	**129.8**	106.9	**111.9**	108.0	**119.9**
		250	0.006	−0.005	−0.004	−0.006	−0.001	−0.002	−0.001	−0.002	100	**112.2**	108.3	**116.2**	104.9	106.7	104.0	108.7
		500	0.003	−0.003	−0.003	−0.003	−0.001	−0.001	−0.001	−0.001	100	105.1	104.6	105.7	102.5	102.8	102.4	102.9
		1000	0.001	−0.001	−0.001	−0.001	0.000	0.000	0.000	0.000	100	102.4	102.3	102.5	101.2	101.3	101.2	101.3
		2000	0.000	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001	100	101.8	101.8	101.8	101.0	101.0	101.0	101.0
		4000	0.000	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001	100	101.2	101.2	101.2	100.7	100.7	100.7	100.7
−2.00	−0.6	125	**0.015**	−0.006	−0.002	−0.010	0.002	−0.001	0.002	−0.006	100	**126.2**	**111.6**	**136.0**	108.4	**117.1**	**118.5**	**136.2**
		250	0.006	−0.006	−0.004	−0.008	−0.001	−0.002	−0.001	−0.004	100	**126.7**	**116.1**	**137.0**	**111.2**	**117.6**	**112.6**	**131.4**
		500	0.002	−0.004	−0.003	−0.004	−0.002	−0.002	−0.002	−0.002	100	**115.5**	**112.0**	**119.7**	107.7	**110.2**	107.2	**113.7**
		1000	0.001	−0.002	−0.002	−0.003	−0.001	−0.002	−0.001	−0.002	100	108.0	107.4	108.9	104.7	105.3	104.6	105.6
		2000	−0.001	−0.002	−0.002	−0.002	−0.002	−0.002	−0.002	−0.002	100	105.8	105.7	106.0	103.6	103.7	103.6	103.8
		4000	−0.001	−0.002	−0.002	−0.002	−0.002	−0.002	−0.002	−0.002	100	104.7	104.6	104.7	103.0	103.1	103.0	103.1
	0.6	125	**0.016**	−0.010	−0.005	**−0.014**	0.001	−0.002	0.003	−0.006	100	**121.7**	107.7	**133.6**	104.5	**110.9**	**110.1**	**125.1**
		250	0.007	−0.006	−0.005	−0.008	−0.001	−0.002	−0.001	−0.003	100	**116.8**	**110.4**	**123.9**	106.4	**110.1**	105.5	**114.4**
		500	0.003	−0.004	−0.003	−0.004	−0.001	−0.002	−0.001	−0.002	100	108.4	107.0	**110.2**	104.1	105.1	103.7	105.7
		1000	0.000	−0.003	−0.003	−0.003	−0.002	−0.002	−0.002	−0.002	100	105.9	105.6	106.2	103.3	103.6	103.3	103.6
		2000	0.000	−0.002	−0.002	−0.002	−0.002	−0.002	−0.002	−0.002	100	103.9	103.8	103.9	102.3	102.4	102.3	102.4
		4000	−0.001	−0.002	−0.002	−0.002	−0.002	−0.002	−0.002	−0.002	100	103.2	103.2	103.2	102.0	102.0	102.0	102.0

Note. The RMSD statistic $\hat{R}_0$ was used as the reference to compute the relative RMSE. Absolute bias values larger than 0.01 and RMSE values larger than 110.0 are printed in bold font.

Figure 1 shows percentile plots for the distribution-weighted RMSD statistics $\hat{R}_0$ and $\hat{R}_7$ for items with DIF ($\delta_i = 0.6$). The results indicate that sampling variability decreased substantially with increasing sample size. However, the variability of both $\hat{R}_0$ and $\hat{R}_7$ remained considerable at the smallest sample size ($N = 125$). For the easy item with $b_i = -2.00$, the bias-corrected RMSD statistic was frequently estimated as 0 when $N = 125$. The figure also illustrates that the median of $\hat{R}_0$ slightly decreased with increasing sample size and eventually converged to the true RMSD value, whereas the median of the bias-corrected RMSD $\hat{R}_7$ increased with increasing sample size.

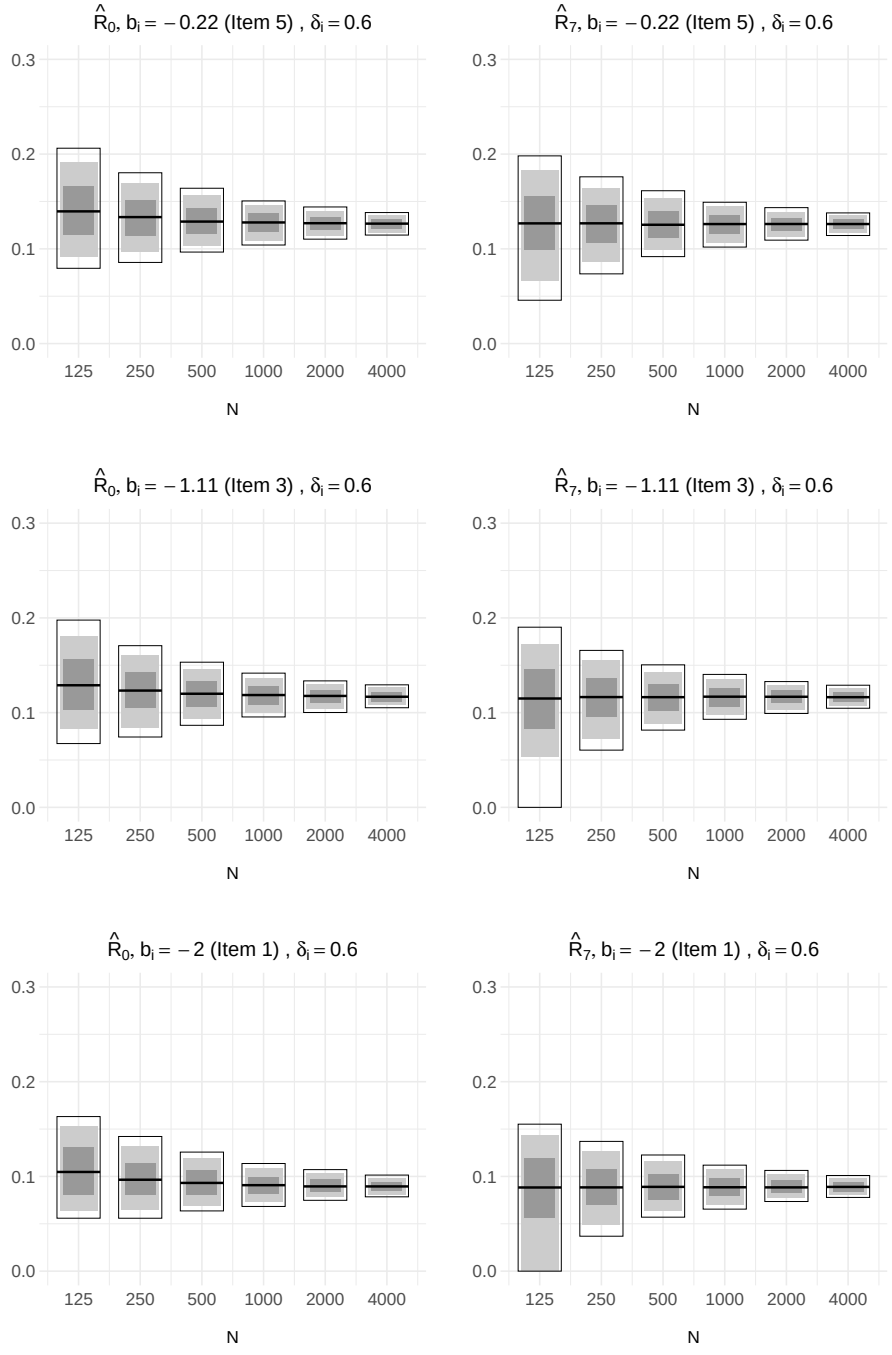


Figure 1. Simulation Study: Percentile plots for distribution-weighted RMSD statistics $\hat{R}_0$ and $\hat{R}_7$ for items with differential item functioning (DIF) as a function of item difficulty b_i and sample size N Note. Outlined in black: 5–95th percentile; light shading: 10–90th; dark gray: 25–75th. Medians are horizontal lines.

Table 3 presents the bias and relative RMSE for the difficulty-weighted RMSD statistics of items with DIF. The positive bias of the difficulty-weighted RMSD statistic $\hat{U}_0$ was more pronounced than the distribution-weighted RMSD statistic $\hat{R}_0$, particularly for easy or difficult items whose item difficulties b_i substantially differed from 0. Like for the $\hat{R}_d$ statistics, the corresponding bias-corrected statistics $\hat{U}_5$ ($M = 0.006$) and $\hat{U}_7$ had the lowest average absolute biases ($M = 0.006$), while the average absolute bias for $\hat{U}_0$ was $M = 0.017$. Notably, there was even a positive bias of 0.024 for $\hat{U}_0$ for $b_i = -2.00$ and $\delta_i = -0.60$ for the relatively large sample size of $N = 1000$. However, bias decreased with increasing sample sizes, but large sample sizes are required to obtain precise estimates of the difficulty-weighted RMSD statistic $\hat{R}_0$.

Table 3. Simulation Study: Bias and relative root mean square error (RMSE) of difficulty-weighted RMSD statistics $\hat{U}_d$ ($d = 0, \ldots, 7$) for items with differential item functioning (DIF) as a function of item difficulty b_i , DIF effect size δ_i and sample size N.

			Bias								**Relative RMSE**							
b_i	δ_i	N	$\hat{U}_0$	$\hat{U}_1$	$\hat{U}_2$	$\hat{U}_3$	$\hat{U}_4$	$\hat{U}_5$	$\hat{U}_6$	$\hat{U}_7$	$\hat{U}_0$	$\hat{U}_1$	$\hat{U}_2$	$\hat{U}_3$	$\hat{U}_4$	$\hat{U}_5$	$\hat{U}_6$	$\hat{U}_7$
−0.22	−0.6	125	**0.018**	−0.004	−0.001	−0.007	0.004	0.002	0.005	0.000	100	**112.6**	104.6	**119.4**	102.2	106.2	104.4	**114.0**
		250	0.007	−0.004	−0.003	−0.005	0.000	−0.001	0.000	−0.001	100	107.9	105.5	**111.0**	102.6	104.0	101.9	106.1
		500	0.003	−0.003	−0.002	−0.003	−0.001	−0.001	−0.001	−0.001	100	104.1	103.7	104.6	101.8	102.1	101.8	102.2
		1000	0.001	−0.002	−0.002	−0.002	−0.001	−0.001	−0.001	−0.001	100	102.5	102.4	102.6	101.3	101.3	101.3	101.3
		2000	0.001	−0.001	−0.001	−0.001	0.000	0.000	0.000	0.000	100	101.0	101.0	101.0	100.5	100.5	100.5	100.5
		4000	0.000	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001	100	101.1	101.1	101.2	100.7	100.7	100.7	100.7
	0.6	125	**0.015**	−0.005	−0.003	−0.008	0.002	0.001	0.003	−0.001	100	**116.1**	108.0	**123.5**	104.7	108.9	106.1	**116.6**
		250	0.006	−0.004	−0.003	−0.004	0.000	−0.001	0.000	−0.001	100	**110.1**	107.3	**113.2**	104.3	105.8	104.1	108.0
		500	0.003	−0.002	−0.002	−0.002	0.000	0.000	0.000	0.000	100	104.0	103.7	104.4	102.0	102.2	101.9	102.3
		1000	0.001	−0.001	−0.001	−0.001	0.000	0.000	0.000	0.000	100	102.1	102.1	102.2	101.1	101.1	101.1	101.1
		2000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	100	100.9	100.9	100.9	100.4	100.4	100.4	100.4
		4000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	100	100.7	100.7	100.7	100.4	100.4	100.4	100.4
−1.11	−0.6	125	**0.041**	−0.002	0.006	−0.010	**0.018**	0.008	**0.019**	−0.001	100	107.7	100.2	**113.2**	100.7	106.4	108.2	**116.6**
		250	**0.019**	−0.007	−0.003	**−0.011**	0.005	0.000	0.007	−0.005	100	**110.4**	103.1	**118.0**	102.3	108.3	104.3	**120.4**
		500	0.010	−0.004	−0.003	−0.006	0.001	0.000	0.002	−0.001	100	103.7	101.4	107.0	101.2	103.2	100.7	107.7
		1000	0.003	−0.005	−0.004	−0.005	−0.002	−0.002	−0.002	−0.003	100	103.6	102.9	104.3	101.9	102.7	102.1	103.3
		2000	−0.001	−0.005	−0.004	−0.005	−0.003	−0.003	−0.003	−0.003	100	104.5	104.3	104.8	103.0	103.2	103.0	103.4
		4000	−0.002	−0.004	−0.004	−0.004	−0.004	−0.004	−0.004	−0.004	100	105.3	105.2	105.4	103.7	103.8	103.7	103.8
	0.6	125	**0.024**	−0.009	−0.003	**−0.014**	0.004	−0.002	0.005	**−0.011**	100	**123.2**	**111.0**	**133.2**	107.6	**117.8**	**115.1**	**137.6**
		250	**0.012**	−0.007	−0.004	−0.009	0.000	−0.002	0.000	−0.006	100	**121.8**	**112.1**	**131.1**	108.2	**116.0**	**110.7**	**133.1**
		500	0.004	−0.005	−0.005	−0.006	−0.002	−0.003	−0.002	−0.003	100	**112.0**	109.8	**115.1**	106.8	109.1	106.9	**113.6**
		1000	0.002	−0.003	−0.003	−0.003	−0.001	−0.001	−0.001	−0.001	100	106.6	106.2	107.1	104.3	104.9	104.6	105.3
		2000	0.000	−0.002	−0.002	−0.002	−0.001	−0.001	−0.001	−0.001	100	104.5	104.3	104.6	103.1	103.3	103.2	103.4
		4000	−0.001	−0.002	−0.002	−0.002	−0.001	−0.001	−0.001	−0.001	100	103.2	103.1	103.2	102.2	102.3	102.2	102.3
−2.00	−0.6	125	**0.123**	**0.064**	**0.076**	**0.056**	**0.093**	**0.073**	**0.084**	**0.061**	100	99.8	97.8	100.6	99.2	100.9	103.4	103.6
		250	**0.075**	**0.019**	**0.031**	**0.011**	**0.049**	**0.030**	**0.043**	**0.018**	100	103.6	99.3	105.9	100.6	104.6	107.7	**110.2**
		500	**0.043**	0.000	0.008	−0.008	**0.022**	0.009	**0.022**	−0.001	100	107.3	101.4	**111.5**	103.1	108.8	109.9	**117.5**
		1000	**0.024**	−0.004	0.001	−0.009	0.010	0.003	**0.013**	−0.003	100	106.7	101.1	**112.3**	102.9	108.2	106.4	**119.1**
		2000	0.010	−0.006	−0.005	−0.008	0.001	−0.002	0.002	−0.003	100	104.1	101.7	107.4	102.9	105.8	103.2	**112.3**
		4000	0.002	−0.007	−0.007	−0.008	−0.004	−0.004	−0.003	−0.005	100	104.3	103.5	105.3	103.6	104.8	103.9	106.0
	0.6	125	**0.072**	**0.019**	**0.029**	**0.013**	**0.042**	**0.026**	**0.036**	**0.017**	100	103.9	98.7	106.0	98.9	104.4	107.2	109.1
		250	**0.041**	−0.002	0.007	−0.008	**0.017**	0.005	**0.014**	−0.004	100	**115.9**	105.7	**121.3**	104.3	**114.8**	**115.9**	**125.8**
		500	**0.019**	**−0.012**	−0.006	**−0.018**	0.002	−0.007	0.001	**−0.017**	100	**132.3**	**116.0**	**143.4**	**112.1**	**128.8**	**122.6**	**152.0**
		1000	0.008	−0.009	−0.006	**−0.011**	−0.002	−0.005	−0.001	**−0.011**	100	**127.0**	**116.5**	**139.0**	**112.2**	**124.1**	**113.6**	**149.7**
		2000	0.002	−0.007	−0.006	−0.007	−0.004	−0.005	−0.004	−0.006	100	**116.9**	**114.6**	**119.9**	**111.5**	**115.5**	**112.9**	**122.6**
		4000	−0.002	−0.007	−0.007	−0.007	−0.006	−0.006	−0.006	−0.006	100	**114.9**	**114.1**	**115.8**	**111.2**	**112.8**	**111.9**	**114.2**

Note. The RMSD statistic $\hat{U}_0$ was used as the reference to compute the relative RMSE. Absolute bias values larger than 0.01 and RMSE values larger than 110.0 are printed in bold font.

Figure 2 shows percentile plots for the difficulty-weighted RMSD statistics $\hat{U}_0$ and $\hat{U}_7$ for items with DIF ($\delta_i = 0.6$). Substantial variability was observed in the RMSD statistics, particularly for the smaller sample sizes ($N = 125$ and $N = 250$). For items with moderate difficulty (i.e., $b_i = -0.22$ and $b_i = -1.11$), the median RMSD values remained relatively stable across sample sizes.

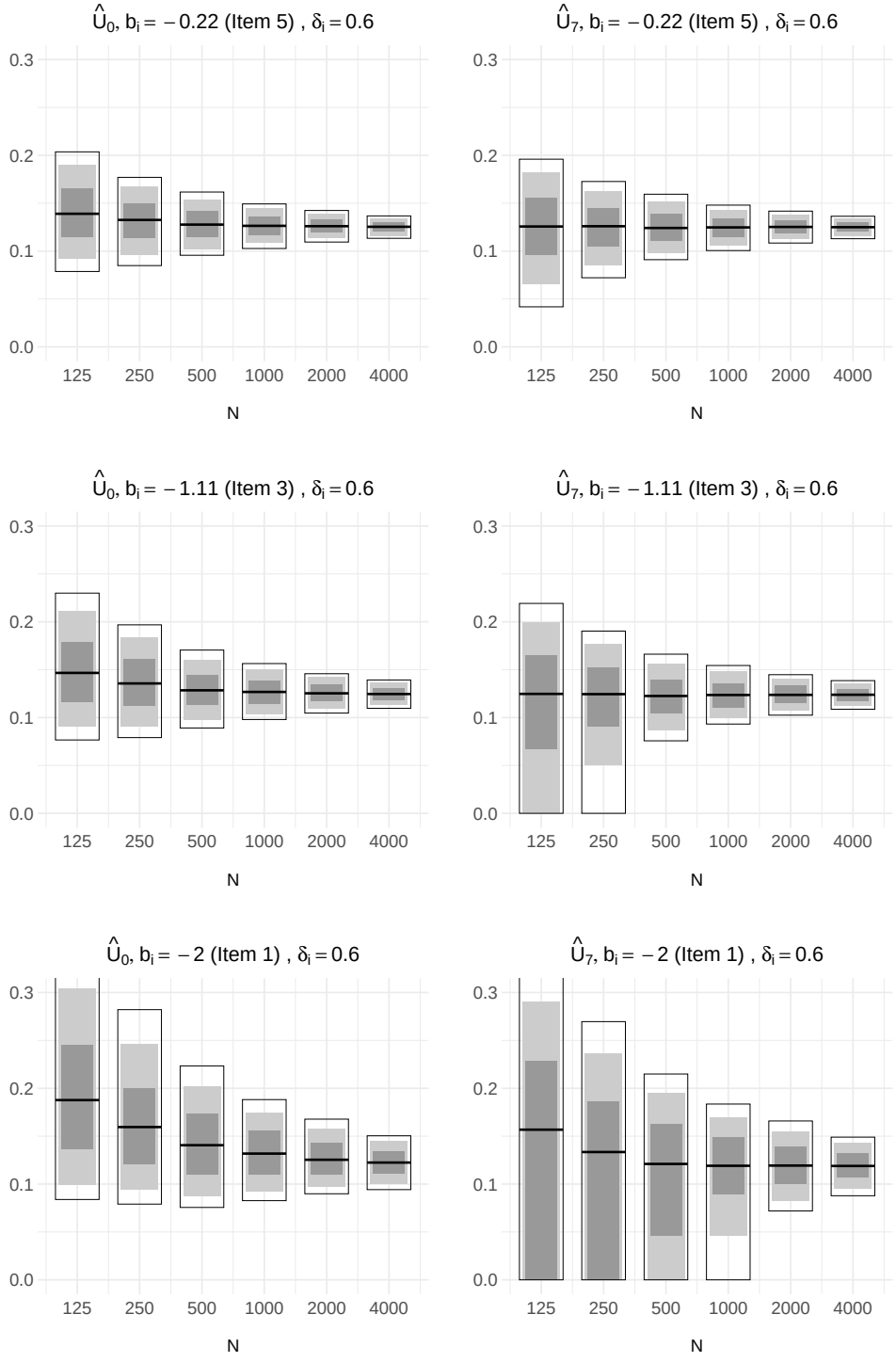


Figure 2. Simulation Study: Percentile plots for difficulty-weighted RMSD statistics $\hat{U}_0$ and $\hat{U}_7$ for items with differential item functioning (DIF) as a function of item difficulty b_i and sample size N Note. Outlined in black: 5–95th percentile; light shading: 10–90th; dark gray: 25–75th. Medians are horizontal lines.

Table 4 presents the bias and relative RMSE for the distribution-weighted RMSD statistics of items without DIF. The condition was selected in which the third and eight item had DIF effects of 0.6 and -0.6, respectively. The true RMSD value was 0 for items without DIF. The positive bias of all RMSD statistics was evident, but much more pronounced for the original RMSD statistic $\hat{R}_0$ with an average absolute bias of $M = 0.032$. The smallest

average absolute bias was observed of $\hat{R}_3$ ($M = 0.008$). This bias-corrected RMSD statistic had also the smallest average RMSE ($M = 44.6$).

Table 4. Simulation Study: Bias and relative root mean square error (RMSE) of distribution-weighted RMSD statistics $\hat{R}_d$ ($d = 0, \ldots, 7$) for items without differential item functioning (DIF) as a function of item difficulty b_i and sample size N.

		Bias								**Relative RMSE**							
b_i	N	$\hat{R}_0$	$\hat{R}_1$	$\hat{R}_2$	$\hat{R}_3$	$\hat{R}_4$	$\hat{R}_5$	$\hat{R}_6$	$\hat{R}_7$	$\hat{R}_0$	$\hat{R}_1$	$\hat{R}_2$	$\hat{R}_3$	$\hat{R}_4$	$\hat{R}_5$	$\hat{R}_6$	$\hat{R}_7$
−0.22	125	**0.068**	**0.022**	**0.033**	**0.017**	**0.039**	**0.032**	**0.033**	**0.021**	100	53.6	61.9	47.2	69.0	63.1	72.4	54.7
	250	**0.049**	**0.015**	**0.023**	**0.011**	**0.027**	**0.023**	**0.024**	**0.014**	100	51.2	60.2	44.2	67.5	61.5	72.0	52.7
	500	**0.034**	**0.011**	**0.016**	0.008	**0.019**	**0.016**	**0.017**	0.010	100	50.7	59.8	43.4	67.0	61.2	71.6	52.7
	1000	**0.024**	0.007	**0.011**	0.005	**0.013**	**0.011**	**0.012**	0.007	100	49.1	58.9	41.5	66.3	60.3	71.3	51.3
	2000	**0.017**	0.005	0.008	0.004	0.009	0.008	0.008	0.005	100	49.9	59.3	42.6	66.4	60.7	70.8	52.0
	4000	**0.012**	0.004	0.006	0.002	0.007	0.006	0.006	0.003	100	49.3	59.0	41.8	66.1	60.3	70.6	51.2
−0.67	125	**0.067**	**0.022**	**0.032**	**0.016**	**0.038**	**0.031**	**0.033**	**0.020**	100	52.1	60.9	45.1	68.3	62.0	72.7	53.1
	250	**0.047**	**0.014**	**0.022**	0.010	**0.026**	**0.021**	**0.022**	**0.013**	100	50.8	59.7	43.7	67.1	61.0	71.7	51.8
	500	**0.034**	0.010	**0.016**	0.008	**0.019**	**0.016**	**0.017**	0.010	100	51.3	60.4	44.5	67.5	61.7	72.2	53.6
	1000	**0.024**	0.007	**0.011**	0.005	**0.013**	**0.011**	**0.011**	0.007	100	48.9	58.7	41.5	66.2	60.1	70.7	51.5
	2000	**0.017**	0.005	0.008	0.004	0.009	0.007	0.008	0.005	100	50.3	59.2	43.3	66.4	60.7	70.3	52.1
	4000	**0.012**	0.004	0.006	0.003	0.007	0.005	0.005	0.003	100	49.6	59.0	42.2	66.3	60.4	70.1	51.6
−1.56	125	**0.062**	**0.021**	**0.030**	**0.016**	**0.036**	**0.029**	**0.031**	**0.019**	100	54.8	62.7	48.3	69.7	63.5	73.5	54.9
	250	**0.044**	**0.015**	**0.021**	**0.011**	**0.025**	**0.021**	**0.022**	**0.014**	100	53.6	61.8	47.1	69.1	63.0	72.8	55.2
	500	**0.032**	0.010	**0.015**	0.007	**0.018**	**0.015**	**0.016**	0.010	100	52.3	61.0	45.3	68.4	62.4	72.6	54.0
	1000	**0.022**	0.006	0.010	0.004	**0.012**	0.010	0.010	0.006	100	49.7	58.9	42.8	66.6	60.3	71.3	52.7
	2000	**0.016**	0.005	0.007	0.003	0.009	0.007	0.007	0.004	100	49.8	59.0	43.0	66.6	60.3	70.4	52.6
	4000	**0.011**	0.003	0.005	0.003	0.006	0.005	0.006	0.003	100	51.1	60.0	44.0	67.5	61.6	72.8	54.2
−2.00	125	**0.058**	**0.021**	**0.029**	**0.016**	**0.035**	**0.028**	**0.030**	**0.019**	100	56.6	64.0	50.7	71.0	64.8	74.3	57.1
	250	**0.042**	**0.014**	**0.021**	**0.011**	**0.024**	**0.020**	**0.021**	**0.013**	100	54.5	62.5	48.3	69.8	63.6	73.8	56.1
	500	**0.030**	0.010	**0.014**	0.007	**0.017**	**0.014**	**0.014**	0.009	100	53.4	61.5	47.1	69.1	62.9	72.8	56.0
	1000	**0.021**	0.006	0.009	0.004	**0.011**	0.009	0.010	0.006	100	49.6	58.7	42.3	66.8	60.1	71.7	52.9
	2000	**0.015**	0.005	0.007	0.004	0.009	0.007	0.008	0.005	100	52.7	61.1	46.1	68.8	62.8	73.4	55.9
	4000	**0.011**	0.003	0.005	0.002	0.006	0.005	0.005	0.003	100	51.3	60.0	44.6	67.8	61.6	72.5	55.5

Note. The RMSD statistic $\hat{R}_0$ was used as the reference to compute the relative RMSE. Absolute bias values larger than 0.01 and RMSE values larger than 110.0 are printed in bold font.

Figure 3 shows percentile plots for the distribution-weighted RMSD statistics $\hat{R}_0$ and $\hat{R}_7$ for items without DIF. The positive bias as evident for $\hat{R}_0$, while the median values of the $\hat{R}_7$ statistic was 0 which conincided with the true RMSD value. As expected, the sampling variability of $\hat{R}_0$ and $\hat{R}_7$ for non-DIF items was also reduced with increasing sample sizes.

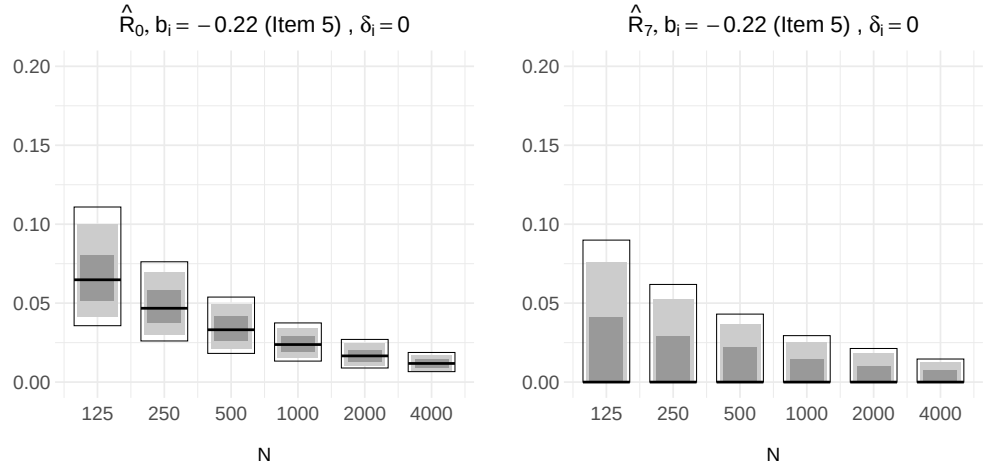


Figure 3. *Cont.*

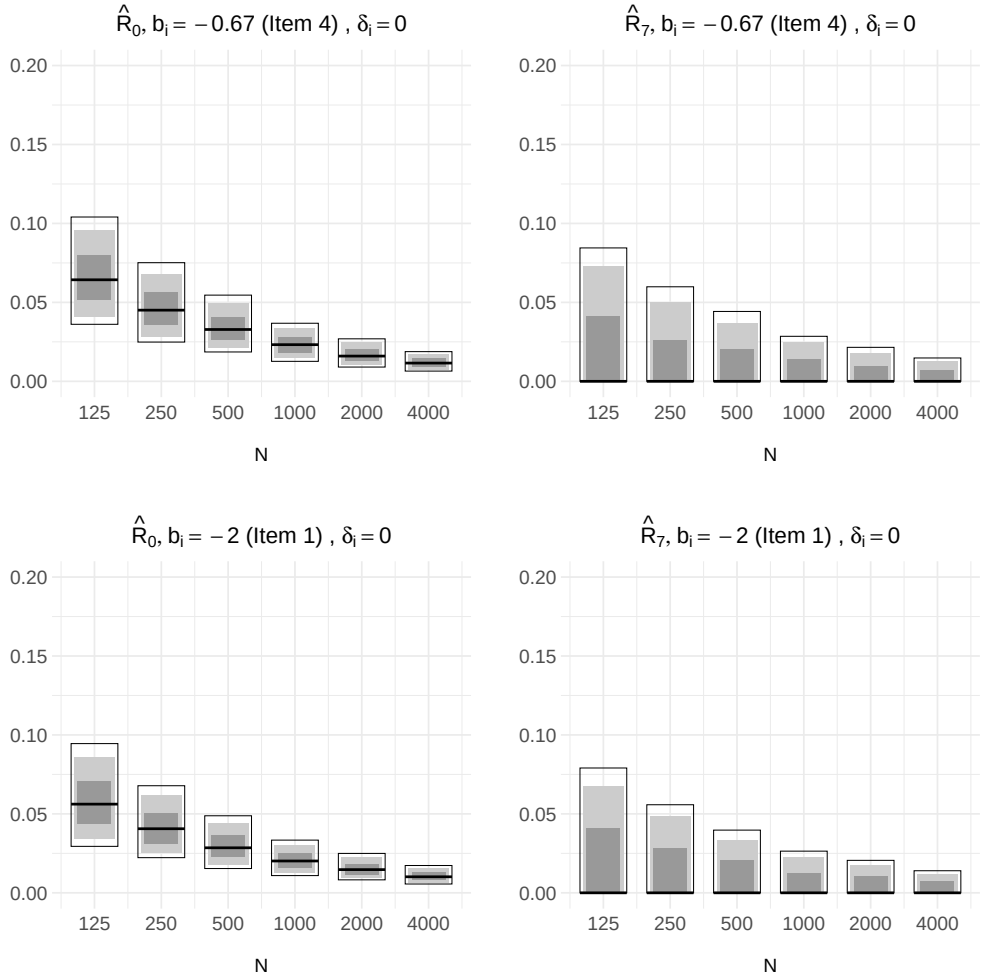


Figure 3. Simulation Study: Percentile plots for distribution-weighted RMSD statistics $\hat{R}_0$ and $\hat{R}_7$ for items without differential item functioning (DIF) as a function of item difficulty b_i and sample size N Note. Outlined in black: 5–95th percentile; light shading: 10–90th; dark gray: 25–75th. Medians are horizontal lines.

Table 5 presents the bias and relative RMSE for the difficulty-weighted RMSD statistics of items without DIF. The positive bias of the difficulty-weighted RMSD statistics $\hat{U}_d$ was most pronounced for smaller sample sizes and for items with more extreme difficulty levels. Similar to the distribution-weighted RMSD statistics, $\hat{U}_3$ showed the lowest average absolute bias ($M = 0.024$), followed by $\hat{U}_1$ ($M = 0.028$). The lowest average RMSE value was also observed for $\hat{U}_3$ ($M = 56.6$).

Figure 4 shows percentile plots for the difficulty-weighted RMSD statistics $\hat{U}_0$ and $\hat{U}_7$ for items without DIF. Across nearly all conditions, the median of the bias-corrected RMSD statistic $\hat{U}_7$ was 0 and matched the true RMSD value. Furthermore, the upper percentiles of the $\hat{U}_7$ statistic were not greater than those of the $\hat{U}_0$ statistic, indicating that using $\hat{U}_7$ instead of $\hat{U}_0$ for DIF assessment in non-DIF items does not introduce additional risk.

In summary, the original difficulty-weighted RMSD statistic exhibited larger positive biases and substantially greater variability than the original distribution-weighted RMSD statistic. The proposed bias-corrected RMSD statistics effectively reduced bias, although they introduced a small to moderate increase in variability. Notably, for items without DIF, the bias-corrected RMSD statistics had median values of 0 in most conditions, indicating favorable performance for DIF detection.

The bias-corrected RMSD statistics that applied the bias correction directly to RMSD using a Taylor approximation performed better than those based on a bias correction of RMSD^2 followed by taking its (positive) square root.

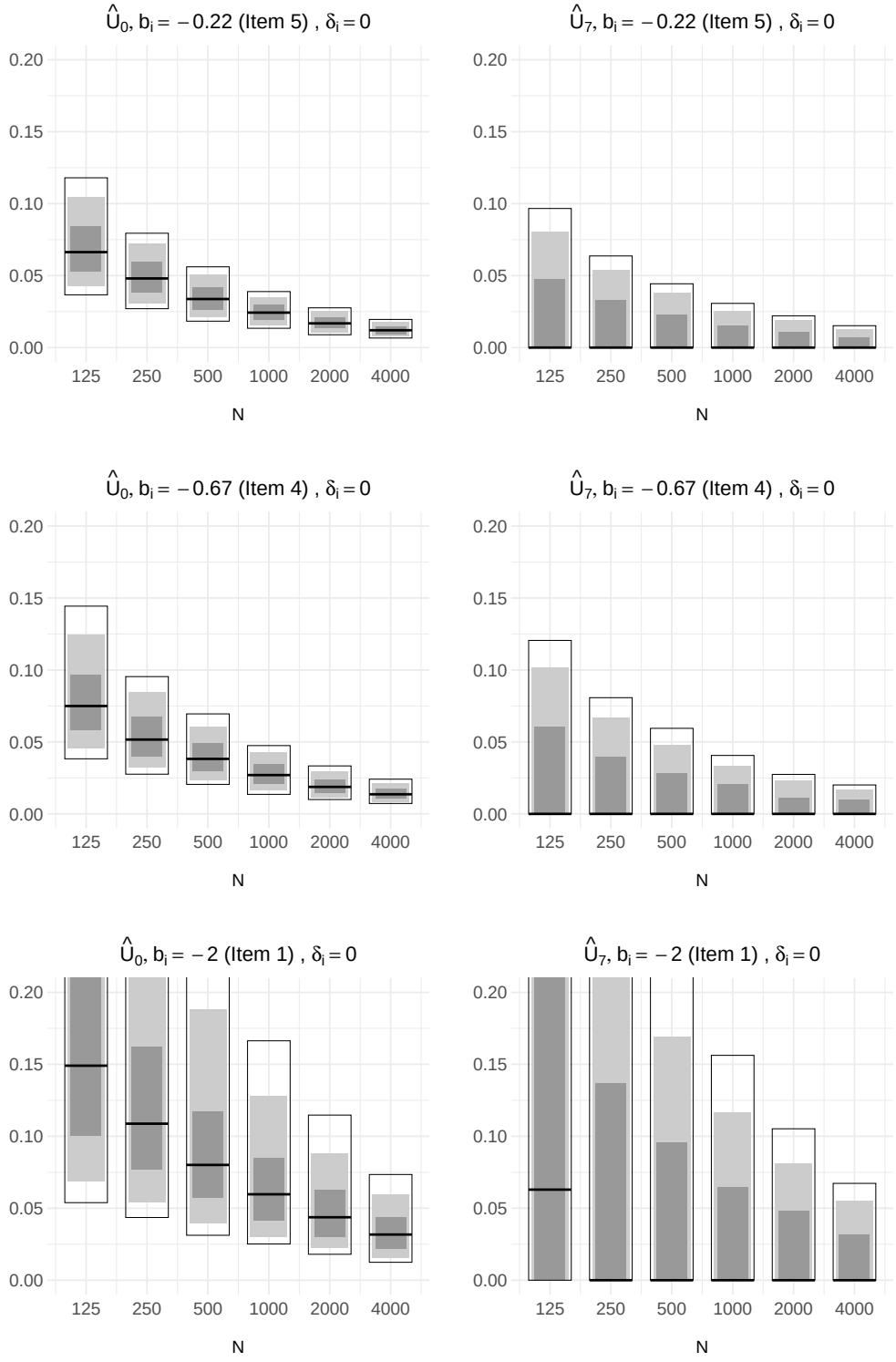


Figure 4. Simulation Study: Percentile plots for difficulty-weighted RMSD statistics $\hat{U}_0$ and $\hat{U}_7$ for items without differential item functioning (DIF) as a function of item difficulty b_i and sample size N Note. Outlined in black: 5–95th percentile; light shading: 10–90th; dark gray: 25–75th. Medians are horizontal lines.

Table 5. Simulation Study: Bias and relative root mean square error (RMSE) of difficulty-weighted RMSD statistics $\hat{U}_d$ ($d = 0, \ldots, 7$) for items without differential item functioning (DIF) as a function of item difficulty b_i and sample size N.

		Bias								Relative RMSE							
b_i	N	$\hat{U}_0$	$\hat{U}_1$	$\hat{U}_2$	$\hat{U}_3$	$\hat{U}_4$	$\hat{U}_5$	$\hat{U}_6$	$\hat{U}_7$	$\hat{U}_0$	$\hat{U}_1$	$\hat{U}_2$	$\hat{U}_3$	$\hat{U}_4$	$\hat{U}_5$	$\hat{U}_6$	$\hat{U}_7$
−0.22	125	**0.071**	**0.024**	**0.035**	**0.018**	**0.041**	**0.034**	**0.036**	**0.023**	100	55.2	63.0	48.9	70.3	64.4	74.4	57.1
	250	**0.050**	**0.016**	**0.024**	**0.012**	**0.029**	**0.024**	**0.025**	**0.015**	100	51.9	60.8	44.9	68.3	62.1	72.6	54.1
	500	**0.035**	**0.011**	**0.017**	0.008	**0.020**	**0.016**	**0.017**	0.010	100	51.3	60.3	44.2	67.7	61.8	72.1	53.5
	1000	**0.025**	0.007	**0.012**	0.005	**0.014**	**0.011**	**0.012**	0.007	100	49.2	59.1	41.7	66.6	60.4	71.9	51.6
	2000	**0.017**	0.005	0.008	0.004	0.010	0.008	0.008	0.005	100	50.3	59.6	43.2	67.0	61.1	72.2	53.1
	4000	**0.012**	0.004	0.006	0.003	0.007	0.006	0.006	0.003	100	49.7	59.3	42.3	66.6	60.7	71.8	51.5
−0.67	125	**0.081**	**0.031**	**0.043**	**0.025**	**0.051**	**0.041**	**0.045**	**0.030**	100	61.0	67.4	56.0	75.4	68.9	79.2	64.0
	250	**0.056**	**0.019**	**0.027**	**0.015**	**0.033**	**0.026**	**0.028**	**0.019**	100	55.8	63.2	50.1	71.8	64.8	76.1	60.1
	500	**0.041**	**0.014**	**0.020**	**0.011**	**0.024**	**0.019**	**0.021**	**0.014**	100	55.8	63.4	49.8	72.0	65.4	76.6	60.3
	1000	**0.028**	0.009	**0.013**	0.006	**0.016**	**0.013**	**0.014**	0.009	100	52.1	60.8	45.4	69.9	62.8	75.5	58.8
	2000	**0.020**	0.006	0.009	0.004	**0.011**	0.009	0.009	0.006	100	52.0	60.5	45.6	69.1	62.3	73.5	56.7
	4000	**0.014**	0.004	0.007	0.003	0.008	0.007	0.007	0.005	100	52.1	60.7	45.3	69.3	62.8	75.2	57.6
−1.56	125	**0.136**	**0.077**	**0.088**	**0.070**	**0.102**	**0.086**	**0.095**	**0.076**	100	78.1	80.8	75.9	87.3	82.2	88.7	80.4
	250	**0.097**	**0.048**	**0.057**	**0.043**	**0.069**	**0.056**	**0.063**	**0.049**	100	73.4	76.8	70.6	85.4	79.0	87.3	77.9
	500	**0.071**	**0.032**	**0.039**	**0.027**	**0.049**	**0.038**	**0.044**	**0.034**	100	69.2	73.4	65.5	84.0	76.5	86.7	76.5
	1000	**0.048**	**0.018**	**0.024**	**0.015**	**0.031**	**0.023**	**0.027**	**0.020**	100	61.2	67.1	56.5	79.9	70.2	83.0	70.1
	2000	**0.035**	**0.013**	**0.017**	0.010	**0.023**	**0.017**	**0.020**	**0.014**	100	58.4	65.3	53.0	79.2	68.7	82.1	69.6
	4000	**0.025**	0.008	**0.012**	0.007	**0.016**	**0.011**	**0.013**	0.010	100	56.6	63.5	51.2	77.0	66.8	80.2	67.2
−2.00	125	**0.185**	**0.123**	**0.133**	**0.117**	**0.150**	**0.131**	**0.139**	**0.122**	100	85.6	87.0	84.4	91.7	88.2	92.0	87.4
	250	**0.135**	**0.079**	**0.088**	**0.073**	**0.105**	**0.087**	**0.094**	**0.080**	100	81.3	83.3	79.6	90.0	85.1	90.3	84.6
	500	**0.098**	**0.050**	**0.059**	**0.046**	**0.073**	**0.058**	**0.064**	**0.052**	100	76.2	79.1	73.9	88.5	81.8	89.0	81.8
	1000	**0.071**	**0.032**	**0.039**	**0.028**	**0.052**	**0.039**	**0.044**	**0.036**	100	71.4	75.1	68.3	87.1	78.7	88.2	80.2
	2000	**0.051**	**0.022**	**0.027**	**0.019**	**0.036**	**0.027**	**0.031**	**0.025**	100	67.5	72.0	63.9	85.4	76.1	86.6	78.0
	4000	**0.035**	**0.013**	**0.017**	**0.011**	**0.024**	**0.017**	**0.020**	**0.016**	100	62.7	68.2	58.5	83.1	72.5	84.3	75.2

Note. The RMSD statistic $\hat{U}_0$ was used as the reference to compute the relative RMSE. Absolute bias values larger than 0.01 and RMSE values larger than 110.0 are printed in bold font.

4. Empirical Example

4.1. Method

In this section, alternative RMSD statistics are illustrated using a subdataset of the PISA 2006 dataset [49] for the reading domain. The complete PISA 2006 dataset is publicly available at https://www.oecd.org/en/data/datasets/pisa-2006-database.html (accessed on 17 October 2025). Processed data and the derived subdataset are available in the `data.pisa2006Read` dataset from the R package sirt (Version 4.2-133; [47]). The analysis focuses on item responses from Booklet 1 to examine RMSD statistics for sample sizes of approximately 200 to 300 students per item, reflecting conditions typical of PISA field trials.

This analysis included 22 countries with valid responses on all 28 reading items from Booklet 1. Sample sizes per country ranged from 287 to 1738 (with a mean of $M = 630.7$, a median of $Med = 381$, and an SD of $SD = 460.3$), yielding a total sample size of 13,876 students. In the original `data.pisa2006Read` dataset, student weights were normalized within each country to represent a hypothetical sample size of 5000, ensuring equal contributions across countries. Because students were randomly assigned to booklets, the total student weights were approximately balanced across countries.

Among the 28 reading items, several were originally scored polytomously but were recoded dichotomously in this analysis, with only the highest category coded as correct. The remaining items were treated as dichotomous, in line with the officially reported PISA analysis.

In the first step of the analysis, joint item parameters of the 2PL model, assumed to be invariant across countries, were estimated. The 2PL model was fitted as a single-group IRT model using student weights to ensure that all 22 countries contributed approximately

equally to the analysis. The estimated item discriminations a_i and item difficulties b_i are reported in Table 6. Item discriminations ranged from 0.59 to 1.95 ($M = 1.35$, $SD = 0.33$), and item difficulties ranged from -1.92 to 3.23 ($M = -0.19$, $SD = 1.14$).

Table 6. Empirical Example, PISA 2006 Reading, Booklet 1, Estonia: International item parameters and estimated RMSD statistics (standard errors in parentheses).

Item	a_i	b_i	$\hat{R}_0$	$\hat{R}_1$	$\hat{R}_3$	$\hat{R}_7$	$\hat{U}_0$	$\hat{U}_1$	$\hat{U}_3$	$\hat{U}_7$
R055Q01	1.33	−1.74	0.025 (0.018)	0.000 (0.020)	0.000 (0.020)	0.000 (0.019)	0.033 (0.070)	0.000 (0.081)	0.000 (0.083)	0.000 (0.081)
R055Q02	1.23	−0.05	0.051 (0.025)	0.026 (0.026)	0.014 (0.026)	0.033 (0.026)	0.056 (0.027)	0.031 (0.028)	0.020 (0.028)	0.038 (0.028)
R055Q03	1.53	−0.60	0.079 (0.025)	0.070 (0.026)	0.069 (0.026)	0.071 (0.026)	**0.084 (0.049)**	0.071 (0.051)	0.070 (0.051)	0.072 (0.050)
R055Q05	1.95	−0.94	0.059 (0.022)	0.050 (0.023)	0.049 (0.023)	0.052 (0.023)	**0.081 (0.035)**	0.068 (0.036)	0.066 (0.036)	0.067 (0.035)
R067Q01	1.39	−1.92	0.045 (0.018)	0.032 (0.021)	0.029 (0.021)	0.032 (0.021)	0.110 (0.124)	0.059 (0.151)	0.037 (0.157)	0.000 (0.154)
R067Q04	0.91	0.74	**0.171 (0.025)**	**0.166 (0.026)**	**0.166 (0.026)**	**0.167 (0.026)**	**0.181 (0.035)**	**0.172 (0.036)**	**0.172 (0.036)**	**0.175 (0.036)**
R067Q05	1.10	−0.24	**0.097 (0.025)**	**0.084 (0.027)**	**0.082 (0.027)**	**0.087 (0.027)**	**0.081 (0.036)**	0.061 (0.039)	0.058 (0.039)	0.061 (0.039)
R102Q04A	1.54	0.75	**0.097 (0.026)**	**0.087 (0.027)**	**0.086 (0.027)**	**0.091 (0.026)**	**0.090 (0.040)**	**0.081 (0.041)**	**0.081 (0.041)**	**0.085 (0.040)**
R102Q05	1.31	0.23	0.026 (0.028)	0.000 (0.029)	0.000 (0.029)	0.000 (0.029)	0.027 (0.030)	0.000 (0.032)	0.000 (0.032)	0.000 (0.031)
R102Q07	1.49	−1.57	0.021 (0.018)	0.000 (0.018)	0.000 (0.018)	0.000 (0.018)	0.032 (0.024)	0.000 (0.025)	0.000 (0.025)	0.000 (0.024)
R104Q01	1.60	−1.41	0.040 (0.017)	0.010 (0.018)	0.000 (0.018)	0.000 (0.018)	0.147 (0.024)	0.033 (0.025)	0.000 (0.025)	0.000 (0.025)
R104Q02	0.59	1.22	0.052 (0.024)	0.015 (0.027)	0.000 (0.027)	0.052 (0.027)	0.043 (0.054)	0.000 (0.059)	0.000 (0.060)	0.000 (0.058)
R104Q05	1.12	3.23	0.063 (0.013)	0.056 (0.015)	0.055 (0.016)	0.059 (0.016)	**0.241 (0.011)**	**0.163 (0.012)**	**0.144 (0.012)**	**0.177 (0.012)**
R111Q01	1.34	−0.73	0.048 (0.025)	0.024 (0.026)	0.012 (0.026)	0.020 (0.026)	0.064 (0.027)	0.040 (0.028)	0.033 (0.028)	0.023 (0.028)
R111Q02B	0.98	1.86	0.046 (0.021)	0.019 (0.025)	0.000 (0.026)	0.008 (0.025)	0.080 (0.055)	0.000 (0.060)	0.000 (0.061)	0.000 (0.057)
R111Q06B	1.44	0.51	**0.133 (0.025)**	**0.127 (0.026)**	**0.127 (0.026)**	**0.129 (0.026)**	**0.125 (0.029)**	**0.118 (0.030)**	**0.118 (0.030)**	**0.121 (0.030)**
R219Q01E	1.49	−0.32	0.072 (0.030)	0.059 (0.031)	0.058 (0.031)	0.062 (0.030)	**0.085 (0.041)**	0.074 (0.043)	0.073 (0.043)	0.077 (0.042)
R219Q01T	1.70	−0.81	**0.096 (0.027)**	**0.090 (0.028)**	**0.089 (0.028)**	**0.091 (0.027)**	**0.101 (0.056)**	**0.082 (0.059)**	0.080 (0.059)	**0.081 (0.058)**
R219Q02	1.43	−1.41	0.071 (0.019)	0.062 (0.020)	0.061 (0.020)	0.065 (0.020)	**0.126 (0.064)**	**0.104 (0.071)**	**0.102 (0.071)**	**0.106 (0.068)**
R220Q01	1.75	0.36	0.035 (0.028)	0.000 (0.029)	0.000 (0.029)	0.000 (0.029)	0.033 (0.030)	0.000 (0.031)	0.000 (0.031)	0.000 (0.030)
R220Q02B	1.45	−0.29	0.029 (0.027)	0.000 (0.028)	0.000 (0.028)	0.000 (0.028)	0.036 (0.039)	0.000 (0.040)	0.000 (0.040)	0.000 (0.040)
R220Q04	1.29	−0.31	0.069 (0.025)	0.054 (0.026)	0.052 (0.026)	0.055 (0.026)	0.071 (0.036)	0.051 (0.038)	0.047 (0.038)	0.050 (0.038)
R220Q05	1.80	−1.02	0.061 (0.022)	0.044 (0.022)	0.042 (0.022)	0.045 (0.022)	**0.145 (0.072)**	**0.090 (0.077)**	**0.074 (0.077)**	**0.126 (0.075)**
R220Q06	1.14	−0.51	**0.097 (0.031)**	**0.086 (0.032)**	**0.085 (0.032)**	**0.088 (0.032)**	**0.118 (0.031)**	**0.103 (0.033)**	**0.102 (0.033)**	**0.104 (0.034)**
R227Q01	0.66	0.00	**0.125 (0.024)**	**0.116 (0.027)**	**0.115 (0.028)**	**0.118 (0.027)**	**0.128 (0.029)**	**0.119 (0.033)**	**0.119 (0.033)**	**0.122 (0.032)**
R227Q02T	0.96	0.89	**0.103 (0.023)**	**0.097 (0.025)**	**0.096 (0.026)**	**0.098 (0.025)**	**0.094 (0.044)**	**0.086 (0.047)**	**0.086 (0.049)**	**0.086 (0.048)**
R227Q03	1.53	−0.21	**0.098 (0.026)**	**0.090 (0.027)**	**0.090 (0.027)**	**0.093 (0.027)**	**0.099 (0.034)**	**0.088 (0.036)**	**0.088 (0.036)**	**0.092 (0.036)**
R227Q06	1.72	−0.93	**0.248 (0.024)**	**0.246 (0.025)**	**0.245 (0.025)**	**0.247 (0.025)**	**0.277 (0.033)**	**0.274 (0.034)**	**0.274 (0.034)**	**0.275 (0.034)**

Note. a_i = item discrimination; b_i = item difficulty; $\hat{R}_d$ = distribution-weighted RMSD statistics ($d = 0, 1, 3, 7$); $\hat{U}_d$ = difficulty-weighted RMSD statistics ($d = 0, 1, 3, 7$). RMSD values larger than 0.08 are printed in bold font.

In the second step, the 2PL model was fitted separately for each of the 22 countries, fixing the item parameters to the joint international estimates while estimating only the country mean μ and country SD σ.

Finally, the original and bias-corrected distribution-weighted and difficulty-weighted RMSD statistics were computed under the assumption of independent student sampling. Although this assumption is violated in practice, the clustered sampling design primarily affects means and SDs rather than item parameters. Furthermore, the same independence assumption is employed in PISA reporting for both the field trial and main study analyses.

Statistical inference for the RMSD statistics was obtained using a nonparametric bootstrap [50] of students within countries. The bootstrap procedure involved independent resampling of students with replacement, maintaining the original sample size for each country. During the computation of the RMSD statistics, bootstrap sampling was applied to the posterior distribution, resulting in varying $\widehat{p}_{it}$ values across bootstrap samples and, consequently, a distribution of RMSD values. Standard errors were estimated using

a normal approximation to the empirical distributions, and 95% symmetric confidence intervals were computed.

The computed confidence intervals for the RMSD statistics enabled testing a minimum effect size hypothesis [51], assessing whether a given RMSD statistic was significantly greater than a prespecified cutoff value, such as 0.05. Item misfit was identified when the lower bound of the confidence interval exceeded this cutoff.

Instead of relying on a fixed cutoff value for the RMSD statistic, a data-driven approach proposed by von Davier and Bezirhan [36] identifies items with large RMSD values as outliers, which are subsequently flagged if they exceed a specified outlier criterion.

4.2. Results

In the first analysis, results for the RMSD statistics are presented for Estonia, which included 287 students with item responses from Booklet 1. Table 6 reports the estimated distribution-weighted and difficulty-weighted RMSD statistics. In most cases, the distribution-weighted RMSD values $\hat{R}_d$ were smaller than their difficulty-weighted counterparts $\hat{U}_d$ ($d = 0, 1, 3, 7$). As expected, the bias-corrected RMSD statistics were smaller than the original statistics and were often estimated as zero.

For six items (R055Q03, R055Q05, R104Q05, R219Q01E, R219Q02, and R220Q05), the original difficulty-weighted RMSD statistic $\hat{U}_0$ exceeded the commonly used cutoff value of 0.08, whereas this was not the case for the widely applied original distribution-weighted RMSD statistic $\hat{R}_0$. The standard errors of the RMSD statistics were relatively large, indicating that statistical uncertainty should be considered when making item selection decisions.

The subsequent analyses examined RMSD statistics for items across all 22 countries. Table 7 presents the percentages of items flagged when the RMSD statistics exceeded a specified cutoff value, were significantly greater than a given cutoff, or were identified as outliers according to the method proposed by von Davier and Bezirhan.

Using a strict cutoff value of 0.05, 59.1% and 72.4% of items were classified as misfitting based on $\hat{R}_0$ and $\hat{U}_0$, respectively. These proportions decreased slightly when the corresponding bias-corrected RMSD statistics were applied. When a cutoff value of 0.12 was used, the flagging rates dropped to 10.1% for $\hat{R}_0$ and 19.8% for $\hat{U}_0$. Incorporating statistical inference into the decision process further reduced the flagging rates substantially.

The data-driven RMSD cutoffs proposed by von Davier and Bezirhan resulted in flagging rates between 2.9% and 6.8%, indicating that most items fit the 2PL model adequately. The estimated data-driven cutoffs were 0.152 (for $\hat{R}_0$), 0.165 (for $\hat{R}_1$), 0.177 (for $\hat{R}_3$), 0.152 (for $\hat{R}_6$), 0.172 (for $\hat{R}_7$), 0.191 (for $\hat{U}_0$), 0.209 (for $\hat{U}_1$), 0.223 (for $\hat{U}_3$), 0.199 (for $\hat{U}_6$), and 0.222 (for $\hat{U}_7$).

Table 7. Empirical Example, PISA 2006 Reading, Booklet 1: Percentage of flagged items across items and countries according to several criteria regarding RMSD statistics

RMSD	$\hat{R}_0$	$\hat{R}_1$	$\hat{R}_3$	$\hat{R}_6$	$\hat{R}_7$	$\hat{U}_0$	$\hat{U}_1$	$\hat{U}_3$	$\hat{U}_6$	$\hat{U}_7$
>0.05	59.1	48.1	47.4	52.8	50.3	72.4	55.4	53.2	63.6	56.8
>0.08	28.7	24.7	24.2	25.8	25.6	45.6	34.1	32.8	38.6	35.9
>0.12	10.1	9.3	9.3	9.7	9.6	19.8	15.3	14.9	18.3	17.4
>0.15	5.2	4.9	4.9	4.9	4.9	12.5	9.3	9.1	11.4	10.9
sign. > 0.05	24.2	20.8	20.8	21.6	21.6	26.3	21.4	21.3	23.1	22.4
sign. > 0.08	10.7	9.7	9.7	10.2	10.2	14.9	12.0	11.7	13.8	13.3
sign. > 0.12	4.9	4.1	4.1	4.2	4.2	7.3	6.3	6.2	6.8	6.7
sign. > 0.15	2.3	1.8	1.8	1.9	1.9	5.7	4.2	3.9	5.2	5.0
>DD cutoff	5.0	4.1	2.9	4.9	2.9	6.8	4.2	3.4	5.8	4.2

Note. $\hat{R}_d$ = distribution-weighted RMSD statistics ($d = 0, 1, 3, 6, 7$); $\hat{U}_d$ = difficulty-weighted RMSD statistics ($d = 0, 1, 3, 6, 7$); sign. = significantly; DD cutoff = data-driven RMSD cutoff according to von Davier and Bezirhan [36].

Figure 5 presents a percentile plot of the empirical distribution of RMSD statistics across all items and countries. The bias-corrected RMSD statistics ($\hat{R}_d$ and $\hat{U}_d$ for $d = 1, 3, 6, 7$) yielded slightly smaller values than the original RMSD statistics ($\hat{R}_0$ and $\hat{U}_0$). However, the reduction in magnitude due to bias correction did not substantially alter the overall distribution of the original RMSD statistics.

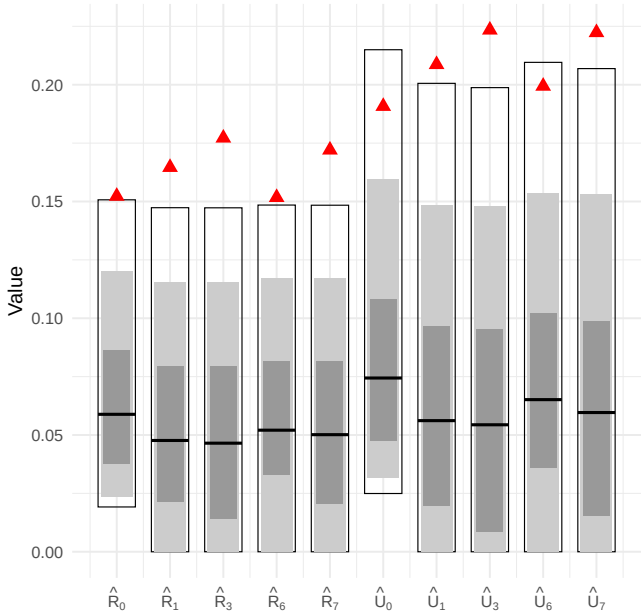


Figure 5. Empirical Example, PISA 2006 Reading, Booklet 1: Percentile plot of empirical distribution of RMSD statistics across all items and countries. Note. Outlined in black: 5–95th percentile; light shading: 10–90th; dark gray: 25–75th. Medians are horizontal lines. Red triangle: data-driven RMSD cutoff according to von Davier and Bezirhan [36]. $\hat{R}_d$ = distribution-weighted RMSD statistics ($d = 0, 1, 3, 6, 7$); $\hat{U}_d$ = difficulty-weighted RMSD statistics ($d = 0, 1, 3, 6, 7$).

5. Discussion

This paper proposed and examined seven alternative bias-corrected distribution-weighted and difficulty-weighted RMSD estimators. The key conclusion is that bias correction should be applied directly to the RMSD rather than to its squared form. The correction methods were based on either a first-order (linear) or a second-order (quadratic) Taylor expansion. All approaches effectively reduced the positive bias observed in the original RMSD statistic, though they slightly increased variance. Bias-corrected RMSD statistics that incorporated the quadratic terms of the Taylor expansion generally demonstrated higher precision. Notably, these bias-corrected RMSD statistics often had a median value of zero for items without DIF, corresponding to the population RMSD value of zero in such cases. In empirical research, the distribution-weighted bias-corrected RMSD statistics $\hat{R}_4$ and $\hat{R}_5$ and the difficulty-weighted bias-corrected RMSD statistics $\hat{U}_4$ and $\hat{U}_5$ are recommended, as they substantially reduce bias without substantially increasing variance.

The empirical example used PISA data from a single booklet. The results showed that the percentage of items flagged as misfitting by the RMSD statistic varied substantially depending on the chosen RMSD cutoff value. Moreover, the proportion of misfitting items was very low when using the data-driven cutoff proposed by [36].

The findings also indicated that sampling variance in the estimated item response functions, which are required for computing the RMSD statistic, becomes particularly critical in small samples, leading to positive bias in the RMSD. The proposed bias-correction methods are therefore especially relevant for small-scale applications aimed at detecting item misfit, such as field trials in PISA.

As an anonymous reviewer critically noted, the simulation study assumed equal item discriminations of 1, implying that the one-parameter logistic (1PL) model served as the data-generating model, whereas the 2PL model was fitted to the responses. This discrepancy can be viewed as a substantial limitation. The behavior of the different bias-correction approaches for the RMSD statistic may depend on variation in item discriminations. A more detailed investigation of this issue may be addressed in future research.

The study was limited to dichotomous items. Future research could extend the bias-corrected RMSD estimators to polytomous items. However, several alternative generalizations of the RMSD statistic exist for polytomous data, and it remains unclear which formulation should be preferred in applied research to enable a coherent generalization of the cutoff values used for dichotomous items.

The approach presented in this study is applicable to any weighted RMSD statistic [37]. It remains unclear whether the distribution-weighted RMSD statistic has notable disadvantages compared to the difficulty-weighted RMSD statistic. In the originally proposed distribution-weighted RMSD, misfitting items with extreme difficulties tend to receive lower RMSD values than misfitting items with moderate difficulty. When likelihood-based inference includes precision weighting in the estimation of group means, it is reasonable for item misfit to be partially influenced by the likelihood contribution. Items with low precision, which contribute weakly to the likelihood, should arguably not be strongly flagged in the item misfit analysis.

Funding: This research received no external funding.

Data Availability Statement: Replication material for the Simulation Study in Section 3 can be found at https://osf.io/amd8k (accessed on 17 October 2025). The dataset `data.pisa2006Read` used in the empirical example in Section 4 is available from the R package sirt (Version 4.2-133; https://doi.org/10.32614/CRAN.package.sirt; accessed on 27 September 2025).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

1PL	one-parameter logistic
2PL	two-parameter logistic
DIF	differential item functioning
IRF	item response function
IRT	item response theory
LSA	large-scale assessment
MML	marginal maximum likelihood
PISA	programme for international student assessment
RMSD	root mean square deviation
RMSE	root mean square error
SD	standard deviation

References

1. Baker, F.B.; Kim, S.H. *Item Response Theory: Parameter Estimation Techniques*; CRC Press: Boca Raton, FL, USA, 2004. [CrossRef]
2. Bock, R.D.; Moustaki, I. Item response theory in a general framework. *Handb. Stat.* **2007**, *26*, 469–513. [CrossRef]
3. Bock, R.D.; Gibbons, R.D. *Item Response Theory*; Wiley: Hoboken, NJ, USA, 2021. [CrossRef]
4. Chen, Y.; Li, X.; Liu, J.; Ying, Z. Item response theory – A statistical framework for educational and psychological measurement. *Stat. Sci.* **2025**, *40*, 167–194. [CrossRef]
5. Tutz, G. *A Short Guide to Item Response Theory Models*; Springer: Berlin/Heidelberg, Germany, 2025. [CrossRef]
6. 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.

7. 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]
8. 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]
9. von Davier, M. Why sum scores may not tell us all about test takers. *Newborn Infant Nurs. Rev.* **2010**, *10*, 27–36. [CrossRef]
10. 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.
11. 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]
12. Bock, R.D.; Aitkin, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **1981**, *46*, 443–459. [CrossRef]
13. OECD. *PISA 2018. Technical Report*; OECD: Paris, France, 2020. Available online: https://bit.ly/3zWbidA (accessed on 23 November 2025).
14. OECD. *PISA 2022. Technical Report*; OECD: Paris, France, 2024. Available online: https://rb.gy/crutc6 (accessed on 23 November 2025).
15. Holland, P.W.; Wainer, H. (Eds.) *Differential Item Functioning: Theory and Practice*; Lawrence Erlbaum: Hillsdale, NJ, USA, 1993. [CrossRef]
16. Penfield, R.D.; Camilli, G. Differential item functioning and item bias. *Handb. Stat.* **2007**, *26*, 125–167. [CrossRef]
17. Mellenbergh, G.J. Item bias and item response theory. *Int. J. Educ. Res.* **1989**, *13*, 127–143. [CrossRef]
18. Millsap, R.E. *Statistical Approaches to Measurement Invariance*; Routledge: New York, NY, USA, 2011. [CrossRef]
19. Robitzsch, A.; Lüdtke, O. Mean comparisons of many groups in the presence of DIF: An evaluation of linking and concurrent scaling approaches. *J. Educ. Behav. Stat.* **2022**, *47*, 36–68. [CrossRef]
20. Pohl, S.; Schulze, D.; Stets, E. Partial measurement invariance: Extending and evaluating the cluster approach for identifying anchor items. *Appl. Psychol. Meas.* **2021**, *45*, 477–493. [CrossRef]
21. Schulze, D.; Reuter, B.; Pohl, S. Measurement invariance: Dealing with the uncertainty in anchor item choice by model averaging. *Struct. Equ. Model.* **2022**, *22*, 521–530. [CrossRef]
22. Douglas, J.; Cohen, A. Nonparametric item response function estimation for assessing parametric model fit. *Appl. Psychol. Meas.* **2001**, *25*, 234–243. [CrossRef]
23. van Rijn, P.W.; Sinharay, S.; Haberman, S.J.; Johnson, M.S. Assessment of fit of item response theory models used in large-scale educational survey assessments. *Large-Scale Assess. Educ.* **2016**, *4*, 10. [CrossRef]
24. Sinharay, S.; Haberman, S.J. How often is the misfit of item response theory models practically significant? *Educ. Meas.* **2014**, *33*, 23–35. [CrossRef]
25. Sinharay, S.; Monroe, S. Assessment of fit of item response theory models: A critical review of the status quo and some future directions. *Brit. J. Math. Stat. Psychol.* **2025**, *78*, 711–733. [CrossRef] [PubMed]
26. Swaminathan, H.; Hambleton, R.K.; Rogers, H.J. Assessing the fit of item response theory models. *Handb. Stat.* **2007**, *26*, 683–718. [CrossRef]
27. Buchholz, J.; Hartig, J. Comparing attitudes across groups: An IRT-based item-fit statistic for the analysis of measurement invariance. *Appl. Psychol. Meas.* **2019**, *43*, 241–250. [CrossRef]
28. Buchholz, J.; Hartig, J. Measurement invariance testing in questionnaires: A comparison of three multigroup-CFA and IRT-based approaches. *Psychol. Test Assess. Model.* **2020**, *62*, 29–53. Available online: https://bit.ly/38kswHh (accessed on 23 November 2025).
29. Khorramdel, L.; Shin, H.J.; von Davier, M. GDM software mdltm including parallel EM algorithm. In *Handbook of Diagnostic Classification Models*; von Davier, M., Lee, Y.S., Eds.; Springer: Berlin/Heidelberg, Germany, 2019; pp. 603–628. [CrossRef]
30. Kunina-Habenicht, O.; Rupp, A.A.; Wilhelm, O. A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Stud. Educ. Eval.* **2009**, *35*, 64–70. [CrossRef]
31. 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]
32. Joo, S.; Ali, U.; Robin, F.; Shin, H.J. Impact of differential item functioning on group score reporting in the context of large-scale assessments. *Large-Scale Assess. Educ.* **2022**, *10*, 18. [CrossRef]
33. Robitzsch, A. Statistical properties of estimators of the RMSD item fit statistic. *Foundations* **2022**, *2*, 488–503. [CrossRef]
34. Sueiro, M.J.; Abad, F.J. Assessing goodness of fit in item response theory with nonparametric models: A comparison of posterior probabilities and kernel-smoothing approaches. *Educ. Psychol. Meas.* **2011**, *71*, 834–848. [CrossRef]
35. Tijmstra, J.; Bolsinova, M.; Liaw, Y.L.; Rutkowski, L.; Rutkowski, D. Sensitivity of the RMSD for detecting item-level misfit in low-performing countries. *J. Educ. Meas.* **2020**, *57*, 566–583. [CrossRef]

36. von Davier, M.; Bezirhan, U. A robust method for detecting item misfit in large scale assessments. *Educ. Psychol. Meas.* **2023**, *83*, 740–765. [CrossRef] [PubMed]
37. Joo, S.; Valdivia, M.; Svetina Valdivia, D.; Rutkowski, L. Alternatives to weighted item fit statistics for establishing measurement invariance in many groups. *J. Educ. Behav. Stat.* **2024**, *49*, 465–493. [CrossRef]
38. Glas, C.A.W.; Jehangir, M. Modeling country-specific differential functioning. In *A Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*; Rutkowski, L., von Davier, M., Rutkowski, D., Eds.; Chapman Hall/CRC Press: London, UK, 2013; pp. 97–115. [CrossRef]
39. Lord, F.M.; Novick, R. *Statistical Theories of Mental Test Scores*; Addison-Wesley: Reading, MA, USA, 1968.
40. Falk, C.F.; Cai, L. Semiparametric item response functions in the context of guessing. *J. Educ. Meas.* **2016**, *53*, 229–247. [CrossRef]
41. Feuerstahler, L.M. Metric transformations and the filtered monotonic polynomial item response model. *Psychometrika* **2019**, *84*, 105–123. [CrossRef]
42. Robitzsch, A.; Lüdtke, O. A review of different scaling approaches under full invariance, partial invariance, and noninvariance for cross-sectional country comparisons in large-scale assessments. *Psychol. Test Assess. Model.* **2020**, *62*, 233–279. Available online: https://bit.ly/3ezBB05 (accessed on 23 November 2025).
43. Boos, D.D.; Stefanski, L.A. *Essential Statistical Inference*; Springer: New York, NY, USA, 2013. [CrossRef]
44. Kondratek, B. Item-fit statistic based on posterior probabilities of membership in ability groups. *Appl. Psychol. Meas.* **2022**, *46*, 462–478. [CrossRef]
45. Held, L.; Sabanés Bové, D. *Applied Statistical Inference*; Springer: Berlin, Germany, 2014. [CrossRef]
46. 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).
47. Robitzsch, A. *sirt: Supplementary Item Response Theory Models*; R Package Version 4.2-133; R Core Team: Vienna, Austria, 2025.
48. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2016. [CrossRef]
49. OECD. *PISA 2006. Technical Report*; OECD: Paris, France, 2009. Available online: https://bit.ly/38jhdzp (accessed on 23 November 2025).
50. Davison, A.C.; Hinkley, D.V. *Bootstrap Methods and Their Application*; Cambridge University Press: Cambridge, UK, 1997. [CrossRef]
51. Grissom, R.J.; Kim, J.J. *Effect Sizes for Research: A Broad Practical Approach*; Lawrence Erlbaum: Mahwah, NJ, USA, 2005. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.