Scilight

Applied Mathematics and Statistics
https://www.sciltp.com/journals/ams

Article

Check for updates

Bias-Corrected RMSD Item Fit Statistic via SIMEX

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

How To Cite: Robitzsch, A. Bias-Corrected RMSD Item Fit Statistic via SIMEX. *Applied Mathematics and Statistics* **2026**, *3*(1), 10. https://doi.org/10.53941/ams.2026.100010

Received: 4 May 2026
Revised: 11 June 2026
Accepted: 15 June 2026
Published: 18 June 2026

Abstract: This study evaluates the simulation extrapolation (SIMEX) method as a bias-correction approach for the distribution-weighted and difficulty-weighted root mean square deviation (RMSD) item fit statistics. The results indicate that SIMEX reduces the positive bias of the original RMSD statistic and can be applied in the context of differential item functioning (DIF) analysis. Although the SIMEX-based RMSD statistics showed slightly greater bias than previously proposed analytic corrections, they yielded lower RMSE for items with DIF. For items without DIF, the analytic bias-correction methods performed better with respect to both bias and root mean square error (RMSE). An empirical example further showed that the SIMEX-based and analytically bias-corrected RMSD statistics produced very similar estimates.

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

1. Introduction

Item response theory (IRT) models [1–4] are multivariate statistical models for discrete multivariate random variables. They are widely used in the social sciences, particularly in educational large-scale assessment (LSA; [5]) studies involving cognitive tasks.

The present article focuses on dichotomous (i.e., binary) random variables. Let $\boldsymbol{X} = (X_1, \ldots, X_I)$ denote a vector of I items, with $X_i \in \{0, 1\}$. A standard classical unidimensional IRT model [6] specifies the probability distribution of $\boldsymbol{X}$ as

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

where ϕ denotes the normal density with mean μ and standard deviation (SD) σ. The latent variable θ, often interpreted as a trait or ability, has distribution parameters $\boldsymbol{\delta} = (\mu, \sigma)$. The vector $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_I)$ contains the item parameters for the parametric item response functions (IRFs) $P_i(\theta; \boldsymbol{\gamma}_i) = \mathsf{P}(X_i = 1 | \theta)$. The latent variable θ may be viewed as a unidimensional summary of the contingency table of item responses $\boldsymbol{X}$ (see [7]). Larger values of θ are associated with more able persons who solve more items, whereas smaller values are associated with less able persons.

The general IRT model in (1) encompasses a range of specific models that differ in the functional form of the IRFs. In the present article, the two-parameter logistic (2PL; [8]) model is considered, for which the 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 for the 2PL model is given by $\boldsymbol{\gamma}_i = (a_i, b_i)$.

The parameters of the IRT model (1) can be consistently estimated by marginal maximum likelihood (MML; [9]), which is commonly implemented using the EM algorithm [10] on the basis of a sample of N individuals with independent and identically distributed observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ of $\boldsymbol{X}$.

In empirical applications, IRFs are typically modeled parametrically as in (1). A common assumption is that item parameters are invariant across groups whose differences in θ are of interest. In practice, however, this assumption may be violated when certain items systematically advantage or disadvantage specific groups. This phenomenon is known as differential item functioning (DIF; [11,12]), although related terms such as measurement bias or item bias are also used [13]. Because DIF can bias the estimation of group differences, identifying such items is important for obtaining less distorted comparisons [14].

When DIF is present, the assumed IRF constitutes a slight misspecification of the true IRT model (1) for a given group. The multivariate random vector $\boldsymbol{X}$ can then be expressed as

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

where P_i^* denotes the assumed IRF and P_i the IRF in the data-generating model for the group. Approximating P_i by P_i^* may distort the estimation of the distribution parameters μ and σ and thereby induce bias in these parameters, under the assumption that the true parameters are defined based on the IRFs P_i^*. In the misspecified IRT model (3), the distribution parameter estimates converge, as the sample size tends to infinity, to pseudo-true parameters that minimize the Kullback-Leibler distance between the specified and the true IRT model (see [15–17]).

Assessing the adequacy of parametric IRFs, that is, item fit [18–22], is therefore a central task in psychometrics. This requires a measure that quantifies the discrepancy between the true IRF P_i and the assumed IRF P_i^* and helps identify items for which this discrepancy is substantial.

To quantify item-level discrepancies between the true and assumed IRFs, the present study focuses on the root mean square deviation (RMSD; [23–35]) statistic, a widely used index in psychometrics. Let ω_i denote a weighting function for item i such that $\int \omega_i(\theta)\,\mathrm{d}\theta = 1$ (see [36]). The weighting function may be item-specific or common across items. The weighted RMSD statistic summarizes the discrepancy between the data-generating IRF P_i and the model-assumed IRF P_i^* as

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

If ω is chosen as the normal density with group mean μ and group SD σ and is the same across items, the statistic is referred to as the distribution-weighted RMSD. If ω_i is item-specific and chosen as the normal density with mean b_i and SD 1, the statistic is referred to as the difficulty-weighted RMSD [36,37]. The assumed IRF P_i^* typically depends on item parameters that are either known or estimated.

The RMSD in Equation (4) is defined at the population level. For empirical applications, a sample-based version is required. 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} denotes the estimated individual posterior distribution of person n at grid point θ_t. The posterior distribution is typically obtained by fitting the IRT model via MML [9]. A discrete evaluation of the weighting function ω_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 defined as

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

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

As shown in [30], the estimator (7) is positively biased in small samples. Bias-corrected RMSD estimators have therefore been proposed for improved performance under such conditions [31]. In the statistical literature, the

simulation extrapolation (SIMEX; [38]) method has been developed as a data-driven approach to bias correction under measurement error. In the present context, the difference between the population IRFs p_{it} and the sample-based IRFs $\widehat{p}_{it}$ is treated as measurement error, and SIMEX is applied as an alternative bias-correction method for the RMSD statistic. The central aim of the present article is to compare this SIMEX-based approach with analytic bias-correction methods for the RMSD statistic.

These considerations motivate a systematic investigation of SIMEX-based and analytic bias-correction methods for the RMSD statistic. The remainder of the article is organized as follows. Section 2 reviews analytic bias-correction methods for the RMSD statistic. Section 3 introduces the application of SIMEX to the RMSD statistic. Results from the simulation study are presented in Section 4. An empirical example illustrating the application of the different RMSD estimators is provided in Section 5. The article concludes with a discussion in Section 6.

2. Analytic Approach to Bias Correction in the RMSD Statistic

In this section, analytic bias-corrected RMSD estimators are reviewed [31]. The population RMSD is defined as

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

and depends on the true item response probabilities p_{it}, which are not directly observable. Throughout this section, p_{it}^* and ω_{it} are treated as fixed. In practice, $\boldsymbol{p}_i$ is replaced by the estimated item response probabilities $\widehat{\boldsymbol{p}}_i$, yielding the sample-based RMSD statistic

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

Because $\hat{T}_0$ is based on $\widehat{\boldsymbol{p}}_i$ rather than $\boldsymbol{p}_i$, it is subject to sampling error. Since the deviations are squared in (9), this induces positive bias (see [29,30]).

To quantify this bias, the sampling variance $\boldsymbol{V}_i = \mathsf{Var}(\widehat{\boldsymbol{p}}_i)$ of $\widehat{\boldsymbol{p}}_i$ is derived using M-estimation theory [39]. The estimating equations for $\widehat{p}_{it}$ are

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

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

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

and, for $s \neq t$, the covariance between $\widehat{p}_{is}$ and $\widehat{p}_{it}$ is

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

The formulas in (11) and (12) also appeared in [40].

These expressions imply that

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

where $\mathcal{B}$ denotes the positive bias in $\hat{T}_0^2$ induced by sampling error in $\widehat{\boldsymbol{p}}_i - \boldsymbol{p}_i$. A first bias-corrected estimator is

$$\hat{T}_1 = \sqrt{\left(\hat{T}_0^2 - \mathcal{B} \right)_+}, \tag{14}$$

where $(x)_+ = \max(x, 0)$. A related approach was proposed in [30], although without using the variance estimation in (11) and (12).

A more accurate correction was derived in [31] using a second-order Taylor expansion:

$$\mathsf{Bias}(\hat{T}_0) \simeq \frac{1}{2\mathcal{T}}\mathsf{Bias}(\hat{T}_0^2) - \frac{1}{8\mathcal{T}^3}\mathsf{E}\left(\hat{T}_0^2 - \mathcal{T}^2\right)^2. \tag{15}$$

Retaining only the linear term in (15) yields

$$\hat{T}_3 = \left(\hat{T}_0 - \frac{\mathcal{B}}{2\hat{T}_1}\right)_+. \tag{16}$$

Including the quadratic term leads to

$$\hat{T}_7 = \left(\hat{T}_0 - \frac{\mathcal{B}}{2\hat{T}_1} + \frac{\mathcal{C} - \mathcal{D}}{2\hat{T}_1^3}\right)_+, \text{ where} \tag{17}$$

$$\mathcal{C} = \sum_{s=1}^{T}\sum_{t=1}^{T}(\widehat{p}_{is} - p_{is}^*)(\widehat{p}_{it} - p_{it}^*)v_{st}\omega_{is}\omega_{it} \text{ and } \mathcal{D} = \sum_{s=1}^{T}\sum_{t=1}^{T} v_{st}^2\omega_{is}\omega_{it}. \tag{18}$$

The next section introduces a data-driven SIMEX approach for bias correction of the RMSD statistic.

3. SIMEX Approach to Bias Correction in the RMSD Statistic

Section 2 showed that the RMSD statistic $\hat{T}_0$ is biased because it is computed from estimated IRFs $\widehat{\boldsymbol{p}}_i$ rather than true IRFs $\boldsymbol{p}_i$. That section also derived the variance matrix $\boldsymbol{V}_i$ of $\widehat{\boldsymbol{p}}_i$, which can be estimated directly from the output of a fitted IRT model. Let $\boldsymbol{e}_i = \widehat{\boldsymbol{p}}_i - \boldsymbol{p}_i$ denote the vector of estimation errors. Then $\boldsymbol{e}_i$ has mean zero and variance matrix $\boldsymbol{V}_i$, so the bias of the RMSD statistic is determined by the sampling variance in $\boldsymbol{e}_i$.

In the present context, the central idea of SIMEX is to generate simulated deviations $\boldsymbol{e}_i^* = (e_{i1}^*, \ldots, e_{iT}^*)$ with variance matrix $\lambda\boldsymbol{V}_i$, thereby adding controlled amounts of sampling variance. The RMSD statistic T_0^* under the additional noise $\boldsymbol{e}_i^*$ is then computed by replacing $\widehat{\boldsymbol{p}}_i$ with $\widehat{\boldsymbol{p}}_i + \boldsymbol{e}_i^*$ in (9), which yields (see [32] for an application of the same approach to confidence interval computation)

$$T_0^* = \sqrt{\sum_{t=1}^{T}\left(\widehat{p}_{it} + e_{it}^* - p_{it}^*\right)^2\omega_{it}}. \tag{19}$$

The resulting RMSD estimates from (19) are then used to infer bias through extrapolation [38, 41–43]. Specifically, S replicated datasets of deviations are generated, yielding an estimated RMSD statistic for item i, denoted by $\hat{\beta}_s(\lambda)$ for $s = 1, \ldots, S$. This statistic is evaluated on a grid of λ values 0.5, 1.0, 1.5, and 2.0 [44,45]. SIMEX typically employs the quadratic extrapolation model [46]

$$\hat{\beta}(\lambda) = \mathsf{E}(\hat{\beta}_s(\lambda)) \approx \alpha_0 + \alpha_1\lambda + \alpha_2\lambda^2. \tag{20}$$

Because $\hat{\beta}(\lambda)$ typically increases with λ, the SIMEX-based RMSD statistic is obtained by extrapolating to $\lambda = -1$, which corresponds to the hypothetical absence of sampling variance in $\boldsymbol{e}_i$ and thus in $\widehat{\boldsymbol{p}}_i$. Under quadratic extrapolation, the resulting estimator is $\hat{\beta}(-1) = \alpha_0 - \alpha_1 + \alpha_2$.

An advantage of SIMEX is that it provides a fully data-driven bias-correction method. By contrast, the analytic corrections reviewed in Section 2 rely on asymptotic arguments. The comparative performance of the analytic methods and SIMEX is examined in the simulation study presented in Section 4.

4. Simulation Study

4.1. Method

The 2PL IRT model was used for both data generation and analysis. The latent trait θ was assumed to follow a normal distribution with mean 0 and SD 1. The simulation study considered a test with $I = 50$ items. Ten base items were specified with discrimination parameters $a_i = 1$ for all items and item difficulty parameters b_i set to -1.8, -1.4, -1.0, -0.6, -0.2, 0.2, 0.6, 1.0, 1.4, and 1.8. Each of these ten items was duplicated five times, yielding a total test length of 50 items.

Exactly one of the 50 items was simulated to exhibit DIF. In the simulation, a constant DIF effect of $\delta = 0.6$ was specified for item j when $j = 1$, $j = 3$, $j = 5$, $j = 6$, $j = 8$, or $j = 10$. The DIF effect size $\delta = 0.6$ was

selected to represent a moderate to large DIF magnitude [12].

The sample size N was varied across 125, 250, 500, and 1000 to reflect typical conditions in educational assessment studies [4,5]. Across the 4 (sample size N) $\times$ 6 (selected DIF item) $=$ 24 simulation conditions, 4000 replications were conducted per condition. Item parameters were fixed at the values of the base items when fitting the 2PL model. However, the distribution parameters μ and σ of the latent trait θ were freely estimated. The presence of DIF was ignored during this scaling step.

Distribution-weighted and difficulty-weighted RMSD measures were computed, including the originally defined statistic $\hat{T}_0$ as well as the bias-corrected statistics $\hat{T}_3$ and $\hat{T}_7$ (see Section 2). In addition, the SIMEX-based RMSD statistics were computed using 1000 simulated draws. To reduce the Monte Carlo error of the SIMEX method, a quasi-Monte Carlo approach based on a deterministic distribution designed to reduce simulation error was employed. Specifically, the 1000 multidimensional data points from a multidimensional uniform distribution with independent components were generated using a Sobol sequence [47], implemented by the function `qrng::sobol()` in R (Version 4.5.3; [48]) within the qrng package (Version 0.0-11; [49]). These multivariate data points were then transformed coordinate-wise with the inverse standard normal distribution function to approximate a multivariate normal distribution.

A true RMSD value was defined from the IRF and the item parameters of the data-generating model. Under this definition, the item with DIF had a positive true RMSD value, whereas the remaining 49 items without DIF had a true RMSD value of 0. The bias and root mean square error (RMSE) of the RMSD statistics were calculated relative to the true RMSD values. A relative RMSE was defined by dividing each RMSD statistic by the value of the RMSD estimate $\hat{T}_7$ and multiplying the result by 100.

All analyses were conducted using the open-source statistical software R (Version 4.5.3; [48]). The 2PL model was estimated with the `sirt::xxirt()` function from the R package sirt (Version 4.2-133; [50]). Dedicated R functions were developed to compute the different RMSD statistics. These functions, together with the replication materials, are available at https://osf.io/8scpw (accessed on 4 May 2026).

4.2. Results

Table 1 presents the performance of the four RMSD estimators for items with DIF across sample sizes and item difficulty levels. Overall, bias decreased as sample size increased for both the distribution-weighted and difficulty-weighted RMSD statistics, and all methods were approximately unbiased at $N = 1000$. For the distribution-weighted RMSD, the original statistic $\hat{T}_0$ tended to exhibit a small positive bias, whereas the SIMEX-based estimator and the two bias-corrected estimators generally reduced this bias, although they often yielded slightly negative values. Nevertheless, $\hat{T}_0$ showed the smallest relative RMSE across all conditions, followed by the SIMEX-based estimator. The estimator $\hat{T}_3$ had a larger RMSE than $\hat{T}_7$ for the distribution-weighted RMSD, but in some conditions it had a smaller RMSE than $\hat{T}_7$ for the difficulty-weighted RMSD statistic.

A similar pattern was observed for the difficulty-weighted RMSD, although biases were larger for the most extreme item difficulties. In particular, for very easy and very difficult DIF items (e.g., $b_i = -1.8$ and $b_i = 1.8$), $\hat{T}_0$ exhibited substantial positive bias at small sample sizes, whereas SIMEX and the bias-corrected estimators reduced this bias markedly. The largest bias was observed for the most difficult item, for which the difficulty-weighted $\hat{T}_0$ statistic had a bias of 0.106 at $N = 125$, compared with 0.072 for SIMEX, 0.040 for $\hat{T}_3$, and 0.044 for $\hat{T}_7$. For items of moderate difficulty, the methods performed similarly, and bias was generally very small for $N = 500$ or 1000. These findings are consistent with the definition of the RMSD statistics: the true values of the difficulty-weighted RMSD statistic are less influenced by item difficulty, whereas the distribution-weighted RMSD statistic takes lower values for items with more extreme difficulties.

Table 2 presents the bias and relative RMSE of the RMSD estimators for items without DIF. In this setting, all true RMSD values were zero. Across item difficulties and sample sizes, the original RMSD statistic $\hat{T}_0$ showed the largest positive bias for both the distribution-weighted and difficulty-weighted indices. The SIMEX-based estimator reduced this bias, whereas the bias-corrected estimators $\hat{T}_3$ and $\hat{T}_7$ showed the smallest bias overall. Bias generally decreased with increasing sample size, but remained non-negligibly positive for $\hat{T}_0$ and the SIMEX-based estimator. For example, for the easiest first item ($b_i = -1.8$), the distribution-weighted bias declined from 0.064 for $\hat{T}_0$ at $N = 125$ to 0.023 at $N = 1000$, whereas the corresponding bias for $\hat{T}_3$ decreased from 0.017 to 0.005. For the difficulty-weighted statistic, bias declined from 0.167 to 0.066 for $\hat{T}_0$ and from 0.098 to 0.025 for $\hat{T}_3$. In terms of relative RMSE, $\hat{T}_3$ outperformed $\hat{T}_7$ for non-DIF items. The SIMEX-based estimator had smaller RMSE values than the original statistic, but was inferior to the bias-corrected estimators $\hat{T}_3$ and $\hat{T}_7$.

Table 1. Simulation Study: Bias and relative root mean square error (RMSE) of estimated distribution-weighted and difficulty-weighted RMSD statistics for items with differential item functioning (DIF) as a function of item difficulty b_i and sample size N.

		Distribution-Weighted RMSD								**Difficulty-Weighted RMSD**							
		Bias				**Relative RMSE**				**Bias**				**Relative RMSE**			
Item (b_i)	N	$\hat{T}_0$	**SI**	$\hat{T}_3$	$\hat{T}_7$	$\hat{T}_0$	**SI**	$\hat{T}_3$	$\hat{T}_7$	$\hat{T}_0$	**SI**	$\hat{T}_3$	$\hat{T}_7$	$\hat{T}_0$	**SI**	$\hat{T}_3$	$\hat{T}_7$
1 (−1.8)	125	**0.016**	−0.001	**−0.016**	−0.007	79.2	84.2	107.3	100	**0.063**	**0.032**	0.004	0.006	88.8	86.2	96.8	100
	250	0.008	−0.002	−0.008	−0.003	86.3	92.5	107.2	100	**0.036**	**0.013**	−0.010	−0.009	76.3	78.3	95.5	100
	500	0.003	−0.002	−0.004	−0.002	94.1	98.6	103.9	100	**0.018**	0.003	**−0.013**	**−0.012**	67.9	73.0	93.3	100
	1000	0.001	−0.002	−0.003	−0.002	96.5	99.7	102.3	100	0.008	−0.002	−0.008	−0.007	71.7	78.6	92.5	100
3 (−1.0)	125	**0.016**	−0.001	**−0.013**	−0.005	83.1	88.2	108.3	100	**0.026**	0.005	**−0.012**	−0.009	75.4	78.0	97.7	100
	250	0.006	−0.003	−0.008	−0.003	90.6	96.8	106.6	100	**0.011**	−0.001	−0.010	−0.007	76.9	83.1	99.1	100
	500	0.002	−0.003	−0.005	−0.003	95.2	99.5	103.4	100	0.004	−0.002	−0.005	−0.003	88.7	94.9	101.7	100
	1000	−0.001	−0.003	−0.004	−0.003	96.1	99.8	102.3	100	0.000	−0.003	−0.005	−0.004	92.8	99.0	102.3	100
5 (−0.2)	125	**0.015**	−0.001	**−0.012**	−0.004	84.6	89.9	108.4	100	**0.015**	0.000	**−0.011**	−0.005	82.7	87.8	106.7	100
	250	0.005	−0.003	−0.007	−0.004	91.7	97.8	106.7	100	0.005	−0.003	−0.007	−0.003	90.9	97.1	106.1	100
	500	0.001	−0.003	−0.005	−0.003	95.2	99.6	103.3	100	0.001	−0.003	−0.005	−0.003	95.0	99.5	103.2	100
	1000	−0.001	−0.003	−0.004	−0.003	96.5	99.9	102.1	100	−0.001	−0.003	−0.004	−0.003	96.4	99.9	102.1	100
6 (0.4)	125	**0.014**	−0.001	**−0.012**	−0.005	83.1	89.0	107.7	100	**0.017**	0.000	**−0.012**	−0.004	84.8	88.5	106.6	100
	250	0.005	−0.003	−0.008	−0.004	90.6	97.3	106.6	100	0.006	−0.003	−0.008	−0.004	90.6	96.3	106.1	100
	500	0.001	−0.003	−0.005	−0.003	95.0	99.5	103.4	100	0.002	−0.003	−0.005	−0.003	95.4	99.4	103.2	100
	1000	−0.001	−0.003	−0.004	−0.003	96.1	99.8	102.2	100	−0.001	−0.003	−0.004	−0.003	96.1	99.8	102.2	100
8 (1.2)	125	**0.014**	−0.001	**−0.013**	−0.007	76.9	83.2	104.2	100	**0.039**	**0.014**	−0.010	−0.002	86.1	85.0	97.3	100
	250	0.006	−0.002	−0.007	−0.004	87.2	94.4	106.4	100	**0.021**	0.004	−0.009	−0.003	85.1	86.1	98.6	100
	500	0.002	−0.003	−0.005	−0.003	93.8	99.2	103.9	100	0.009	−0.001	−0.006	−0.002	92.8	94.9	100.6	100
	1000	−0.001	−0.003	−0.004	−0.003	95.7	99.7	102.4	100	0.003	−0.002	−0.005	−0.003	96.4	99.0	101.3	100
10 (1.8)	125	**0.016**	0.001	**−0.011**	−0.008	72.5	77.3	100.3	100	**0.106**	**0.072**	**0.040**	**0.044**	95.9	93.8	97.1	100
	250	0.007	−0.002	−0.008	−0.005	77.4	85.6	103.9	100	**0.064**	**0.035**	0.005	**0.011**	90.0	89.4	95.9	100
	500	0.003	−0.002	−0.004	−0.002	89.7	96.6	104.8	100	**0.037**	**0.016**	−0.007	−0.001	85.2	86.1	94.9	100
	1000	0.000	−0.002	−0.003	−0.002	94.4	99.4	103.1	100	**0.018**	0.005	−0.008	−0.003	85.4	87.1	94.2	100

Note. $\hat{T}_0$ = original RMSD statistic; SI = SIMEX-based RMSD statistic; $\hat{T}_3$ = bias-corrected RMSD statistic as defined by (16); $\hat{T}_7$ = bias-corrected RMSD statistic as defined by (17); Values of absolute bias larger than 0.010 are printed in bold font.

Table 2. Simulation Study: Bias and relative root mean square error (RMSE) of estimated distribution-weighted and difficulty-weighted RMSD statistics for items without differential item functioning (DIF) as a functionof item difficulty b_i and sample size N.

		Distribution-Weighted RMSD								**Difficulty-Weighted RMSD**							
		Bias				**Relative RMSE**				**Bias**				**Relative RMSE**			
Item (b_i)	N	$\hat{T}_0$	**SI**	$\hat{T}_3$	$\hat{T}_7$	$\hat{T}_0$	**SI**	$\hat{T}_3$	$\hat{T}_7$	$\hat{T}_0$	**SI**	$\hat{T}_3$	$\hat{T}_7$	$\hat{T}_0$	**SI**	$\hat{T}_3$	$\hat{T}_7$
1 (−1.8)	125	**0.064**	**0.038**	**0.017**	**0.020**	182.7	126.0	89.7	100	**0.167**	**0.127**	**0.098**	**0.102**	119.3	104.7	96.2	100
	250	**0.045**	**0.026**	**0.011**	**0.013**	189.2	126.9	86.7	100	**0.125**	**0.090**	**0.064**	**0.070**	122.2	105.0	94.0	100
	500	**0.033**	**0.019**	0.008	0.010	188.4	126.1	85.6	100	**0.093**	**0.065**	**0.042**	**0.049**	125.8	105.1	89.9	100
	1000	**0.023**	**0.013**	0.005	0.007	190.2	125.7	82.8	100	**0.066**	**0.043**	**0.025**	**0.031**	131.4	105.2	84.7	100
2 (−1.4)	125	**0.068**	**0.040**	**0.018**	**0.021**	183.5	125.7	88.5	100	**0.131**	**0.093**	**0.064**	**0.069**	128.0	106.9	94.2	100
	250	**0.048**	**0.027**	**0.011**	**0.014**	191.5	128.0	86.4	100	**0.093**	**0.062**	**0.038**	**0.043**	134.0	107.9	90.9	100
	500	**0.034**	**0.020**	0.008	0.010	189.8	126.5	84.4	100	**0.067**	**0.043**	**0.024**	**0.029**	139.4	108.4	86.5	100
	1000	**0.025**	**0.014**	0.005	0.007	191.9	126.7	83.4	100	**0.048**	**0.030**	**0.015**	**0.020**	144.1	108.2	82.1	100
4 (−0.6)	125	**0.072**	**0.042**	**0.017**	**0.021**	191.7	128.8	86.5	100	**0.083**	**0.051**	**0.025**	**0.030**	163.6	117.6	87.7	100
	250	**0.050**	**0.028**	0.010	**0.013**	204.6	133.2	84.7	100	**0.058**	**0.034**	**0.015**	**0.019**	173.3	119.4	83.3	100
	500	**0.036**	**0.020**	0.008	0.010	197.2	129.1	83.7	100	**0.042**	**0.024**	0.010	**0.013**	175.8	120.0	82.4	100
	1000	**0.026**	**0.014**	0.005	0.007	201.2	130.6	83.3	100	**0.030**	**0.017**	0.007	0.009	181.8	121.8	80.7	100
5 (−0.2)	125	**0.072**	**0.042**	**0.017**	**0.021**	194.1	129.6	86.2	100	**0.075**	**0.044**	**0.019**	**0.023**	184.8	125.7	86.5	100
	250	**0.051**	**0.029**	**0.011**	**0.015**	197.2	130.0	84.9	100	**0.053**	**0.030**	**0.012**	**0.016**	191.9	127.9	84.8	100
	500	**0.036**	**0.020**	0.008	0.010	200.0	130.1	84.0	100	**0.037**	**0.021**	0.008	0.010	196.0	128.3	83.5	100
	1000	**0.026**	**0.014**	0.005	0.007	203.9	132.0	83.7	100	**0.026**	**0.015**	0.005	0.007	199.4	129.8	83.3	100

Note. $\hat{T}_0$ = original RMSD statistic; SI = SIMEX-based RMSD statistic; $\hat{T}_3$ = bias-corrected RMSD statistic as defined by (16); $\hat{T}_7$ = bias-corrected RMSD statistic as defined by (17); Values of absolute bias larger than 0.010 are printed in bold font.

For empirical applications, the simulation results indicate that the original RMSD statistics should be interpreted with caution, particularly in small samples and for very easy or very difficult items, as they tend to overestimate RMSD. Bias-corrected approaches substantially improve performance. Among these, SIMEX appears to be a practically attractive option because it consistently reduces both bias and RMSE. However, the analytically bias-corrected RMSD estimators may be preferable in most conditions with respect to bias and in some conditions with respect to RMSE.

5. Empirical Example

The dataset `MathExamp14W` used in this Empirical Example is available in the R package psychotools (Version 0.7-6; [51]). It comprises item response data from 729 students on 13 items from a written introductory mathematics exam [52]. Gender was used as the grouping variable, with male students in Group 1 ($N = 403$) and female students in Group 2 ($N = 326$). The same RMSD estimators as in the Simulation Study in Section 4 were computed.

A multiple-group 2PL IRT model was estimated using the R package TAM (Version 4.3-25; [53]). The estimated group mean difference in θ was 0.322. The SD of θ was fixed to 1 in the male group and estimated as 1.234 in the female group.

Table 3 reports descriptive statistics and the estimated distribution-weighted and difficulty-weighted RMSD values for this empirical example. Item difficulties ranged from relatively low values for elasticity ($b_i = -1.084$), hesse ($b_i = -1.040$), and interest ($b_i = -0.965$) to a high value for payflow ($b_i = 2.271$). Item discrimination parameters ranged from 0.575 for quad to 1.691 for hesse. Differences in item p-values between male and female students were generally small, although somewhat larger differences were observed for the items annuity, matrix, payflow, and planning. The mean item p-value difference was 0.044 (SD = 0.039), favoring female students, and ranged from -0.032 for quad to 0.101 for annuity.

Across items, the original RMSD statistic $\hat{T}_0$ generally yielded the largest values, followed by the SIMEX-based estimates, whereas the bias-corrected statistics $\hat{T}_3$ and $\hat{T}_7$ were often close to zero, suggesting that DIF was frequently absent. This pattern is consistent with the simulation results, which indicated that the original RMSD statistic tends to be positively biased, whereas SIMEX and the analytic corrections reduce this upward bias.

Based on the estimated RMSD statistics, the strongest evidence of potential item-level misfit or DIF was observed for payflow, planning, and quad. For payflow, the distribution-weighted RMSD remained elevated across all estimators ($\hat{T}_0 = 0.048$, SIMEX $= 0.042$, $\hat{T}_3 = 0.040$, $\hat{T}_7 = 0.041$), and the corresponding difficulty-weighted values were the largest in the table ($\hat{T}_0 = 0.085$, SIMEX $= 0.074$, $\hat{T}_3 = 0.064$, $\hat{T}_7 = 0.067$). Planning also showed comparatively large RMSD values, particularly for the distribution-weighted statistic (0.052, 0.043, 0.037, and 0.040, respectively), whereas quad showed moderate values for the original and SIMEX estimators but notably smaller values after analytic correction. By contrast, several items, such as deriv, implicit, and lagrange, had corrected RMSD estimates of zero or near zero, indicating little evidence of meaningful DIF after accounting for estimation bias.

Overall, Table 3 indicates that most items showed little evidence of DIF once bias correction was applied, whereas a small subset of items—especially payflow and, to a lesser extent, planning—remained potentially noteworthy. Consistent with the simulation findings, interpretation should rely primarily on the bias-corrected or SIMEX-adjusted statistics rather than on the original RMSD values alone, particularly because the uncorrected statistic may overstate apparent DIF.

Table 3. Empirical Example: Descriptive statistics and estimated distribution-weighted and difficultyweighted RMSD statistics.

					Distribution-weighted RMSD				Difficulty-weighted RMSD			
Item	p_{i0}	p_{i1}	a_i	b_i	$\hat{T}_0$	**SI**	$\hat{T}_3$	$\hat{T}_7$	$\hat{T}_0$	**SI**	$\hat{T}_3$	$\hat{T}_7$
annuity	0.608	0.709	1.178	−0.576	0.029	0.017	0	0	0.040	0.031	0.021	0.025
deriv	0.705	0.715	1.095	−0.905	0.017	0	0	0	0.024	0.001	0	0
elasticity	0.732	0.779	1.205	−1.084	0.020	0.006	0	0	0.030	0.010	0	0
equations	0.375	0.442	1.128	0.573	0.028	0.017	0	0	0.037	0.027	0.018	0.010
hesse	0.772	0.801	1.691	−1.040	0.016	0.003	0	0	0.021	0.004	0	0
implicit	0.610	0.675	1.395	−0.454	0.008	0	0	0	0.010	0	0	0
integral	0.474	0.525	0.944	0.148	0.022	0.005	0	0	0.022	0.005	0	0
interest	0.710	0.709	1.005	−0.965	0.032	0.021	0	0	0.038	0.024	0	0
lagrange	0.404	0.429	0.703	0.685	0.018	0	0	0	0.016	0	0	0
matrix	0.603	0.696	1.610	−0.437	0.020	0.008	0	0	0.021	0.008	0	0
payflow	0.136	0.221	0.856	2.271	0.048	0.042	0.040	0.041	0.085	0.074	0.064	0.067
planning	0.400	0.429	0.859	0.620	0.052	0.043	0.037	0.040	0.043	0.031	0.019	0.012
quad	0.541	0.509	0.575	−0.066	0.042	0.029	0.011	0.024	0.042	0.029	0.010	0.024

Note. p_{i0}, p_{i1} = item p-values for male (Group 0) and female (Group 1) students; a_i = estimated item discrimination; b_i = estimated item difficulty; $\hat{T}_0$ = original RMSD statistic; SI = SIMEX-based RMSD statistic; $\hat{T}_3$ = bias-corrected RMSD statistic as defined by (16); $\hat{T}_7$ = bias-corrected RMSD statistic as defined by (17).

6. Discussion

The present study examined the SIMEX method as a bias-correction approach for the distribution-weighted and difficulty-weighted RMSD item fit statistics. Overall, the results showed that SIMEX was effective in reducing the positive bias of the original RMSD statistic, and its practical use was demonstrated in the context of differential item functioning (DIF) analysis. However, the SIMEX-based RMSD statistics generally exhibited slightly greater bias than previously proposed analytic bias-correction methods.

A more nuanced pattern emerged when estimator performance was evaluated across DIF conditions. For items exhibiting DIF, the SIMEX-based estimators produced lower root mean square error (RMSE) than the analytically bias-corrected RMSD estimators, suggesting that SIMEX may offer an advantage when the goal is to improve overall estimation accuracy under item misfit. In contrast, for items without DIF, where the true RMSD values were zero, the analytic bias-correction methods generally outperformed SIMEX with respect to both bias and RMSE.

The SIMEX approach is also expected to perform well when multiple items are affected by DIF, although the simulation study considered only the case in which a single item showed DIF. More generally, the proportion of DIF items influences the expected values of the RMSD statistics for all items, even in infinite sample sizes. This applies to both DIF and non-DIF items. As the proportion of DIF items increases, the RMSD statistics for non-DIF items tend to move further away from zero, whereas those for DIF items tend to be smaller than when the proportion of DIF items is lower [30].

As SIMEX relies on numerical simulations, it is computationally more demanding than analytical approaches to bias correction of the RMSD statistic. However, SIMEX is applied only to the computation of the RMSD statistics after the IRT model has been fitted. In general, fitting the IRT model is computationally more demanding than repeatedly computing the RMSD statistic in SIMEX. Hence, the computational burden of the SIMEX approach remains quite low.

The SIMEX approach to bias correction of the RMSD statistic can be extended to polytomous IRT models such as the generalized partial credit model [54] or the graded response model [55]. In this case, the discretized IRF must be estimated separately for each category. More specifically, the sampling error associated with the nonparametric IRF estimates $\widehat{\boldsymbol{p}}_{ik}$ for item i and category k is required. A variance matrix for all estimates collected in the vector $\widehat{\boldsymbol{p}}_i = (\widehat{\boldsymbol{p}}_{i1}, \ldots, \widehat{\boldsymbol{p}}_{iK})$ is then derived, and SIMEX simulates noisy versions of this estimate by adding a simulated vector $\boldsymbol{e}_i^*$. The SIMEX estimate of the RMSD statistic obtained by quadratic extrapolation is then derived exactly as in the case of dichotomous items.

The results from the Simulation Study suggest that bias correction of the RMSD statistic can be important, particularly for sample sizes less than or equal to 250. Such sample sizes occur in field trials of educational large-scale assessment studies such as programme for international student assessment (PISA; [56]), which typically report the RMSD item fit statistic. Therefore, the use of bias-corrected RMSD estimates, whether analytical or SIMEX-based, may be advisable in empirical research. Of course, bias-corrected RMSD estimates exhibit greater variability and may slightly reduce power, but they may also decrease false positive (i.e., type-I error) rates.

These findings suggest that the SIMEX approach may be especially valuable in situations where analytic bias-correction formulas are not yet available, such as for polytomous items. More broadly, the results indicate that SIMEX has the potential to serve as a general framework for reducing finite-sample bias in item fit effect size measures. Nevertheless, when analytically derived bias-correction methods are available, they may remain the preferred option because of their stronger performance in conditions where no item misfit is present.

7. Conclusions

The SIMEX approach has been proposed as an alternative procedure for correcting bias in the RMSD item-fit statistic. Overall, it exhibited favorable performance and produced results comparable to those obtained using analytical bias-correction methods. The SIMEX approach is particularly advantageous in settings where analytical derivations of the bias are unavailable, as it is based entirely on computational procedures.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Data Availability Statement

Replication materials for Simulation Study, described in Section 4, are available at https://osf.io/8scpw (accessed on 4 May 2026). The dataset used in Section 5 is publicly available through the R package psychotools at https://doi.org/10.32614/CRAN.package.psychotools (accessed on 4 May 2026).

Conflicts of Interest

The author declares no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the author used ChatGPT-5.4 to support language editing and enhance readability. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Abbreviations

2PL	two-parameter logistic
DIF	differential item functioning
IRF	item response function
IRT	item response theory
MML	marginal maximum likelihood
RMSE	root mean square error
SD	standard deviation
SIMEX	simulation extrapolation

References

1. Bock, R.D.; Moustaki, I. 15 Item Response Theory in a General Framework. In *Handbook of Statistics*; Rao, C.R.; Sinharay, S., Eds.; Wiley: Hoboken, NY, USA, 2007; pp. 469–513. https://doi.org/10.1016/S0169-7161(06)26015-2.
2. Bock, R.D.; Gibbons, R.D. *Item Response Theory*; Wiley: Hoboken, NY, USA, 2021. https://doi.org/10.1002/9781119716723.
3. Chen, Y.; Li, X.; Liu, J.; et al. Item Response Theory—A Statistical Framework for Educational and Psychological Measurement. *Stat. Sci.* **2025**, *40*, 167–194. https://doi.org/10.1214/23-STS896.
4. Yen, W.M.; Fitzpatrick, A.R. Item Response Theory. In *Educational Measurement*; Brennan, R.L., Ed.; Praeger Publishers: London, UK, 2006; pp. 111–154.
5. 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. https://doi.org/10.1201/b16061.
6. van der Linden, W.J. Unidimensional Logistic Response Models. In *Handbook of Item Response Theory*; van der Linden, W.J., Ed.; CRC Press: Boca Raton, UK, 2016; pp. 11–30. https://doi.org/10.1201/9781315119144.
7. von Davier, M. Why Sum Scores May Not Tell Us All about Test Takers. *Newborn Infant Nurs. Rev.* **2010**, *10*, 27–36. https://doi.org/10.1053/j.nainr.2009.12.011.
8. 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.
9. Glas, C.A.W. Maximum-Likelihood Estimation. In *Handbook of Item Response Theory*; van der Linden, W.J., Ed.; CRC Press: Boca Raton, FL, USA, 2016; pp. 197–216. https://doi.org/10.1201/b19166-11.
10. Bock, R.D.; Aitkin, M. Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm. *Psychometrika* **1981**, *46*, 443–459. https://doi.org/10.1007/BF02293801.
11. Holland, P.W.; Wainer, H. (Eds.) *Differential Item Functioning: Theory and Practice*; Lawrence Erlbaum: Hillsdale, NJ, USA, 1993. https://doi.org/10.4324/9780203357811.
12. Penfield, R.D.; Camilli, G. Differential Item Functioning and Item Bias. In *Handbook of Statistics*; Rao, C.R., Sinharay, S., Eds.; Wiley: Hoboken, NY, USA, 2007; pp. 125–167. https://doi.org/10.1016/S0169-7161(06)26005-X.
13. Mellenbergh, G.J. Item Bias and Item Response Theory. *Int. J. Educ. Res.* **1989**, *13*, 127–143. https://doi.org/10.1016/0883-0355(89)90002-5.
14. Millsap, R.E. *Statistical Approaches to Measurement Invariance*; Routledge: New York, NY, USA, 2011. https://doi.org/10.4324/9780203821961.
15. Held, L.; Sabanés Bové, D. *Applied Statistical Inference*; Springer: Berlin, Germany, 2014. https://doi.org/10.1007/978-3-642-37887-4.

16. Van der Vaart, A.W. *Asymptotic Statistics*; Cambridge University Press: Cambridge, UK, 1998. https://doi.org/10.1017/CBO9780511802256.
17. White, H. Maximum Likelihood Estimation of Misspecified Models. *Econometrica* **1982**, *50*, 1–25. https://doi.org/10.2307/1912526.
18. Douglas, J.; Cohen, A. Nonparametric Item Response Function Estimation for Assessing Parametric Model Fit. *Appl. Psychol. Meas.* **2001**, *25*, 234–243. https://doi.org/10.1177/01466210122032046.
19. van Rijn, P.W.; Sinharay, S.; Haberman, S.J.; et al. Assessment of Fit of Item Response Theory Models Used in Large-Scale Educational Survey Assessments. *Large-Scale Assess. Educ.* **2016**, *4*, 10. https://doi.org/10.1186/s40536-016-0025-3.
20. Sinharay, S.; Haberman, S.J. How Often Is the Misfit of Item Response Theory Models Practically Significant? *Educ. Meas.* **2014**, *33*, 23–35. https://doi.org/10.1111/emip.12024.
21. Sinharay, S.; Monroe, S. Assessment of Fit of Item Response Theory Models: A Critical Review of the Status Quo and Some Future Directions. *Br. J. Math. Stat. Psychol.* **2025**, *78*, 711–733. https://doi.org/10.1111/bmsp.12378.
22. Swaminathan, H.; Hambleton, R.K.; Rogers, H.J. 21 Assessing the Fit of Item Response Theory Models. In *Handbook of Statistics*; Rao, C.R., Sinharay, S., Eds.; Wiley: Hoboken, NY, USA, 2007; pp. 683–718. https://doi.org/10.1016/S0169-7161(06)26021-8.
23. 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. https://doi.org/10.1177/0146621617748323.
24. 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.
25. 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: Cham, Switzerland, 2019; pp. 603–628. https://doi.org/10.1007/978-3-030-05584-4_30.
26. 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. https://doi.org/10.1016/j.stueduc.2009.10.003.
27. Joo, S.H.; Khorramdel, L.; Yamamoto, K.; et al. Evaluating Item Fit Statistic Thresholds in PISA: Analysis of Cross-Country Comparability of Cognitive Items. *Educ. Meas.* **2021**, *40*, 37–48. https://doi.org/10.1111/emip.12404.
28. Joo, S.; Ali, U.; Robin, F.; et al. Impact of Differential Item Functioning on Group Score Reporting in the Context of Large-Scale Assessments. *Large-Scale Assess. Educ.* **2022**, *10*, 18. https://doi.org/10.1186/s40536-022-00135-7.
29. 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.
30. Robitzsch, A. Statistical Properties of Estimators of the RMSD Item Fit Statistic. *Foundations* **2022**, *2*, 488–503. https://doi.org/10.3390/foundations2020032.
31. Robitzsch, A. Bias-Corrected Root Mean Square Deviation Estimators. *Foundations* **2025**, *5*, 36. https://doi.org/10.3390/foundations5040036.
32. Robitzsch, A. Confidence Interval Estimation for RMSD and MD Item Fit Statistics. *Univ. J. Appl. Math.* **2026**, *14*, 9–21. https://doi.org/10.13189/ujam.2026.140102.
33. 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. https://doi.org/10.1177/0013164410393238.
34. Tijmstra, J.; Bolsinova, M.; Liaw, Y.L.; et al. Sensitivity of the RMSD for Detecting Item-Level Misfit in Low-Performing Countries. *J. Educ. Meas.* **2020**, *57*, 566–583. https://doi.org/10.1111/jedm.12263.
35. von Davier, M.; Bezirhan, U. A Robust Method for Detecting Item Misfit in Large Scale Assessments. *Educ. Psychol. Meas.* **2023**, *83*, 740–765. https://doi.org/10.1177/00131644221105819.
36. Joo, S.; Valdivia, M.; Svetina Valdivia, D.; et al. Alternatives to Weighted Item Fit Statistics for Establishing Measurement Invariance in Many Groups. *J. Educ. Behav. Stat.* **2024**, *49*, 465–493. https://doi.org/10.3102/10769986231183326.
37. Robitzsch, A. Comparing Weighted RMSD, Weighted MD, Infit, and Outfit Item Fit Statistics Under Uniform Differential Item Functioning. *Mathematics* **2025**, *13*, 3752. https://doi.org/10.3390/math13233752.
38. Carroll, R.J.; Küchenhoff, H.; Lombard, F.; et al. Asymptotics for the SIMEX Estimator in Nonlinear Measurement Error Models. *J. Am. Stat. Assoc.* **1996**, *91*, 242–250. https://doi.org/10.2307/2291401.
39. Boos, D.D.; Stefanski, L.A. *Essential Statistical Inference*; Springer: New York, NY, USA, 2013. https://doi.org/10.1007/978-1-4614-4818-1.
40. Kondratek, B. Item-Fit Statistic Based on Posterior Probabilities of Membership in Ability Groups. *Appl. Psychol. Meas.* **2022**, *46*, 462–478. https://doi.org/10.1177/01466216221108061.
41. Cook, J.R.; Stefanski, L.A. Simulation-Extrapolation Estimation in Parametric Measurement Error Models. *J. Am. Stat. Assoc.* **1994**, *89*, 1314–1328. https://doi.org/10.2307/2290994.
42. Stefanski, L.A.; Cook, J.R. Simulation-Extrapolation: The Measurement Error Jackknife. *J. Am. Stat. Assoc.* **1995**, *90*,

1247–1256. https://doi.org/10.2307/2291515.

43. Sevilimedu, V.; Yu, L. Simulation Extrapolation Method for Measurement Error: A Review. *Stat. Methods Med. Res.* **2022**, *31*, 1617–1636. https://doi.org/10.1177/09622802221102619.
44. Carroll, R.J.; Ruppert, D.; Stefanski, L.A.; et al. *Measurement Error in Nonlinear Models: A Modern Perspective*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2006. https://doi.org/10.1201/9781420010138.
45. Lederer, W.; Küchenhoff, H. A Short Introduction to the SIMEX and MCSIMEX. *R News* **2006**, *6*, 26–31.
46. Buonaccorsi, J.P. *Measurement Error: Models, Methods, and Applications*; CRC Press: Boca Raton, FL, USA, 2010. https://doi.org/10.1201/9781420066586.
47. Jäckel, P. *Monte Carlo Methods in Finance*; Wiley: New York, NY, USA, 2002.
48. R Core Team. R: A Language and Environment for Statistical Computing, 2026. Available online: https://www.R-project.org (accessed on 12 March 2026).
49. Hofert, M.; Lemieux, C. qrng: (Randomized) Quasi-Random Number Generators, 2026. R package Version 0.0-11. https://cran.r-project.org/web/packages/qrng/index.html (accessed on 22 January 2026).
50. Robitzsch, A. sirt: Supplementary Item Response Theory Models, 2025. R Package Version 4.2-133. https://cran.r-project.org/web/packages/sirt/index.html (accessed on 27 September 2025).
51. Zeileis, A.; Strobl, C.; Wickelmaier, F.; et al. psychotools: Psychometric Modeling Infrastructure, 2026. R Package Version 0.7-6. https://cran.r-project.org/web/packages/psychotools/index.html (accessed on 11 February 2026).
52. Zeileis, A. Examining Exams Using Rasch Models and Assessment of Measurement Invariance. *Austrian J. Stat.* **2025**, *54*, 9–26. https://doi.org/10.17713/ajs.v54i3.2055.
53. Robitzsch, A.; Kiefer, T.; Wu, M. TAM: Test Analysis Modules, 2025. R Package Version 4.3-25. https://cran.r-project.org/web/packages/TAM/index.html (accessed on 28 August 2025).
54. Muraki, E. A Generalized Partial Credit Model: Application of an EM Algorithm. *Appl. Psychol. Meas.* **1992**, *16*, 159–176. https://doi.org/10.1177/014662169201600206.
55. Forero, C.G.; Maydeu-Olivares, A. Estimation of IRT Graded Response Models: Limited Versus Full Information Methods. *Psychol. Methods* **2009**, *14*, 275–299. https://doi.org/10.1037/a0015825.
56. OECD. *PISA 2022. Technical Report*; OECD: Paris, France, 2024.