Joo *et al. Large-scale Assessments in Education* (2022) 10:18
https://doi.org/10.1186/s40536-022-00135-7

Large-scale Assessments in Education

RESEARCH Open Access

Impact of differential item functioning on group score reporting in the context of large-scale assessments

Sean Joo[1*], Usama Ali[2,4], Frederic Robin[2] and Hyo Jeong Shin[3]

*Correspondence:
Sean Joo
sjoo@ku.edu
[1]University of Kansas, Kansas, USA
[2]Educational Testing Service, New Jersey, USA
[3]Sogang University, Seoul, South Korea
[4]South Valley University, Qena, Egypt

Abstract

We investigated the potential impact of differential item functioning (DIF) on group-level mean and standard deviation estimates using empirical and simulated data in the context of large-scale assessment. For the empirical investigation, PISA 2018 cognitive domains (Reading, Mathematics, and Science) data were analyzed using Jackknife sampling to explore the impact of DIF on the country scores and their standard errors. We found that the countries that have a large number of DIF items tend to increase the difference of the country scores computed with and without the DIF adjustment. In addition, standard errors of the country score differences also increased with the number of DIF items. For the simulation study, we evaluated bias and root mean squared error (RMSE) of the group mean and standard deviation estimates using the multigroup item response theory (IRT) model to explore the extent to which DIF items create a bias of the group mean scores and how effectively the DIF adjustment corrects the bias under various conditions. We found that the DIF adjustment reduced the bias by 50% on average. The implications and limitations of the study are further discussed.

Keywords Large-scale assessment, Programme for International Student Assessment, Differential item functioning, Group score reporting, Jackknife sampling

Introduction

The core purpose of national and international large-scale assessments (LSAs), such as National Assessment of Educational Progress (NAEP), the Programme for International Student Assessment (PISA), and the Programme for the International Assessment of Adult Competencies (PIAAC) is a comparison of education qualities among regions, states, and nations. Such comparisons provide important insights for educational researchers and policymakers to evaluate the current educational system and students' academic progress over time (e.g., Cosgrove & Cartwright 2014; Neumann et al., 2010). To achieve this, constructing high-level scale comparability is a critical requirement. Scale comparability refers to the condition in which assessments are comparable across all country- or state-level groups and across assessment cycles, such that the group-level

scores are on the same metric (e.g., Mazzeo & von Davier 2014; Oliveri & von Davier, 2014).

In LSAs, item response theory (IRT) methodology has been implemented in the scaling procedure to establish a common metric allowing for comparability across participating groups and assessment cycles. IRT analysis also allows researchers to investigate psychometric properties of items (i.e., slopes and difficulties), reliability, and validity of the assessments in general. For example, both PISA and PIAAC incorporated a two-parameter logistic model (2PLM; Birnbaum 1968) and generalized partial credit model (GPCM; Muraki 1992) as measurement models for dichotomous and polytomous response items, respectively. Moreover, multigroup IRT scaling (a.k.a. concurrent calibration; Bock & Zimowski 1997; Kolen & Brennan, 2014) has been implemented since the PISA 2015 main survey to put the multiple country-by-language-by-cycle scores on the same metric (OECD, 2016). Specifically, items are calibrated simultaneously with the equality constraint on their parameters across the participating countries and economies and assessment cycles in the multigroup IRT model (von Davier et al., 2019). These estimated item parameters are also referred to as *international* or *common* item parameters that contribute to the scale comparability.

Recently, educational researchers and practitioners have raised a practical question regarding common item parameters in LSAs: Does item calibration with an equality constraint present a scaling approach that is too restrictive? (e.g., Rutkowski et al., 2010; Rutkowski & Svetina, 2014; Svetina & Rutkowski, 2014; Switzer et al., 2017). Wu (2010) also noted that LSAs generally include the various sources of error induced by measurement, sampling, and equating procedures, and the scaling procedure should identify the source and magnitude of error to increase the validity of the results. It has been explicitly argued that these sources of error tend to create item misfits from the set of common item parameters for a particular group or assessment cycle (Oliveri & von Davier, 2011; Oliveri & Davier, 2014). For example, in PISA, the target population is considerably diverse in the sense that each student's country, language, culture, ethnicity, socioeconomic status, and background are different within the assessment sample (de Jong et al., 2007; Kreiner & Christensen, 2014; Sachse et al., 2016). Several cross-country studies have also reported that the scale comparability is not easily retained, and country-, language- or culture-specific item parameter calibration should be carefully investigated (e.g., Ercikan 2002; Ercikan & Koh, 2005; Gierl & Khaliq, 2001).

For these reasons, the heterogeneity of item parameters in LSAs should be considered in the scaling process, and several studies have suggested a group-specific or assessment cycle-specific item parameter approach to increase the validity of the scores as well as to improve the measurement precision (Oliveri & von Davier, 2011; Oliveri & von Davier, 2014). In addition, for trend items, which refer to items that were previously administered in past assessment cycles, fixed item parameter calibration (FIPC) has been suggested and implemented in the operational scaling procedure. FIPC links the previously administered scales to the current assessment scales by "fixing" the estimated item parameters from the previous assessment cycles, including both international and country-specific item parameters. Using this approach, the linking error from cycle to cycle can be substantially reduced without losing the validity of the scales. A previous study demonstrated the benefits of the FIPC approach in the context of PISA (König et al., 2021).

Regardless of group- or assessment cycle-specific item parameters, the item misfit should be precisely addressed in the IRT scaling procedure to obtain accurate group performance proficiency estimates that are comparable across groups and cycles. The term, item misfit, is also referred to as the *lack of measurement invariance* (Meredith, 1993), *item bias* (Lord, 1980), *differential item functioning* (DIF; Holland & Thayer 1988), or *item-by-country interaction* (OECD, 2016) in educational measurement literature. Because the main scope of the study is closely related to this type of item misfit in the context of LSAs, we, henceforth, refer to the misfit of items as DIF throughout the paper. In addition, it is worthwhile note that in this study we considered the DIF as the fixed effect in the context of LSAs, which is consistent with the PISA operational approach. However, previous LSA studies also have considered DIF as the random effect and incorporated the hierarchical random effect model to examine the effect of DIF (De Jong et al., 2007; Fox & Verhagen, 2018).

DIF Adjustment and Group Score Estimation

To address DIF items in IRT scaling, the *unique* item parameter calibration approach has been proposed and implemented in operational settings (OECD, 2016, 2019). Specifically, in subsequent IRT scaling procedures, unique item parameters (group-specific or cycle-specific) are separately estimated for the items detected as DIF. This adjustment for DIF items has several advantages in terms of psychometric properties and scale comparability in the context of LSAs. First, estimating unique item parameters significantly improved the overall model fit (Joo et al., 2021; Oliveri & von Davier, 2011; Oliveri & von Davier, 2014; Rutkowski & Svetina, 2014, 2017). For example, Oliveri & von Davier (2011) applied various IRT models to the PISA 2006 cognitive domain data and compared the fitted models. They concluded that the multigroup 2PLM with partially unique item parameters was the best fitting model based on several model fit indices, including the Akaike Information Criterion (AIC; Akaike 1974) and the Bayesian Information Criterion (BIC; Schwarz 1978).

Second, the group-specific unique item parameter approach for addressing DIF items can reduce the bias in the group score estimates and increase the stability of group rankings (Rutkowski & Rutkowski, 2018). Note that the bias in group means depends on the interplay of the distribution of DIF effects and the chosen linking method (Robitzsch, 2021). In this study, we mainly focus on the bias of the group mean caused by the DIF distribution. Although the true group mean parameters and group rankings are unknown in real data applications, it has been shown via simulation studies that the group-specific unique item parameter estimation can produce accurate group mean parameter estimates. More specifically, group specific unique item parameters reduce the bias which is defined as the difference between the generating and estimated group mean parameters. For example, Rutkowski et al., (2016) conducted a simulation study that mimicked the PISA 2009 main survey design and investigated the country achievement estimates. They compared several approaches for computing country achievement estimates by varying the samples for item parameter calibration. Their results showed that the most restrictive sample, with common item parameters, produced bias in the country achievement estimates up to 12.49 on the PISA scale, and the less restrictive sample reduced the gap between the true and estimated country achievement estimates.

Purpose

Although several previous studies have shown the psychometric benefits of allowing unique item parameters for DIF items, it is yet unknown the degree to which the proportion of unique item parameters can be acceptable without harming the comparability of test results across participating groups and cycles. For example, in operational LSA, several groups are investigated further in IRT scaling when they require a large proportion of unique item parameters, indicating possible data quality or integrity issues that may affect group score comparability. We defined the group score comparability as the proportion of international item parameters (i.e., invariant item parameters) across country and language groups in this study (OECD, 2019). Although Rutkowski et al., (2016) reported that a less restrictive calibration sample showed a less biased result of the country performance scores, it is unknown whether allowing country-, language-, or cycle-specific unique item parameters can still produce comparable group score results. More importantly, a simulation study that manipulates a different level of DIF items is needed to systematically examine the extent to which the group scores are biased by the DIF items and how effectively the DIF adjustment corrects the bias.

Therefore, the purposes of the study are (a) to examine the issue with DIF items in the context of LSAs, (b) to quantify their impact to help researchers and practitioners interpret group-level score results more carefully, and (c) to provide an empirically-based recommendation for addressing the issues with DIF items. To achieve such purposes, we conducted two studies: In the first study, we compared the precision and reliability of country-level score estimates computed with and without the DIF adjustment using the PISA 2018 main survey data. We incorporated the Jackknife sampling method to obtain the country score difference and standard error estimates. In the second study, we conducted a simulation study to investigate the impact of DIF items and the adjustment on the group mean and standard deviation estimates using the multigroup IRT scaling approach.

Study 1

PISA 2018 main survey data

To empirically investigate the impact of the unique item parameter approach on the group scores in the context of LSAs, we analyzed the PISA 2018 cognitive domains (Reading, Mathematics, and Science) main survey data. In the PISA 2018 main survey, Reading was the major domain in that Reading items were administered to all students, and Mathematics and Science were the minor domains in that one of the domains, either Mathematics or Science, was administered. Depending on the participating countries, the PISA also have been administered in either computer-based assessment (CBA) or paper-based assessment (PBA) mode since the 2015 cycle. In 2018, 70 countries participated in CBA and nine countries participated in PBA. A total of 244 Reading items was administered for CBA countries, consisting of 172 new items and 72 trend items. For PBA countries 72 Reading trend items were administered. For Mathematics, 83 items were administered to both CBA and PBA countries, and for Science, 115 items were administered to CBA countries and 85 to PBA countries, all of which had been administered in the previous cycle. In this analysis, we considered all cognitive major and minor domain data and also included both CBA and PBA countries. Moreover, we considered country-by-language groups where countries that have multiple languages

were divided into multiple language groups. Thus, the total number of country-by-language groups comprised 85 for CBA countries and 12 for PBA countries. Note that the country-by-language group approach is consistent with the PISA 2018 operational procedure (OECD, 2019).

As typical for LSAs, PISA also incorporates the balanced incomplete block (BIB) design for test administration. In the BIB design, students are required to respond to only a subset of the total item pool. The BIB design is a commonly used test administration design, especially for LSAs, because large-scale surveys generally cover a broad range of content and information. Using the BIB design, unbiased group-level score estimates can be obtained without overwhelming participating students with a large number of items. However, because only a subset of items is administered to students, a large proportion of data is missing by design. In this analysis, these missing responses were excluded from the IRT scaling, and they do not contribute to the item parameter estimation. Finally, we used senate weights so that the sample size per country to be equal as 5,000 (OECD, 2019).

IRT scaling and group score estimation

To analyze PISA cognitive domain data, we conducted multigroup IRT calibration (Bock & Zimowski, 1997). We initially applied the equality constraint across all country-by-language groups. More specifically, for new items, item parameters of all groups were estimated to be the same across groups. For trend items, item parameters were fixed at the estimates from previous assessment cycles. Note that the fixed item parameters were concurrently calibrated from data collected from PISA cycles 2006 to 2015. This fixed item parameter calibration approach is commonly used in operational settings to put the current assessment cycle scales on the same metric. Data from each PISA cognitive domain (Reading, Mathematics, and Science) was separately calibrated with the IRT models, such as 2PLM for dichotomous responses and GPCM for polytomous responses. The multigroup 2PLM (Eq. 1) and GPCM (Eq. 2) probability functions are described as:

$$P\left(X_{ijg}=1|\theta_{ig}\right)=\frac{\exp\left[Da_j\left(\theta_{ig}-b_j\right)\right]}{1+\exp\left[Da_j\left(\theta_{ig}-b_j\right)\right]} \tag{1}$$

$$P\left(X_{ijg}=k|\theta_{ig}\right)=\frac{\exp\left[\sum_{r=0}^{k}Da_j\left(\theta_{ig}-b_j+t_{jr}\right)\right]}{\sum_{u=0}^{m_j}\exp\left[\sum_{r=0}^{u}Da_j\left(\theta_{ig}-b_j+t_{jr}\right)\right]} \tag{2}$$

where θ_{ig} is latent trait parameter for the i^{th} student for g^{th} group, a_j is discrimination parameter, b_j is difficulty parameter, and t_{jr} is category threshold parameter of the j^{th} item. D is the scaling constant for the logit link function, assumed to be 1.7. For the GPCM, m_j is the total number of categories − 1 for the j^{th} item (e.g., X_{ijg} = 0, 1, …, m_j), and the category threshold parameter has additional constraints $t_{j0}=0$ and $\sum_{r=1}^{k}t_{jr}=0$. In the multigroup structure, θ_{ig} is assumed to be distributed as $N(\mu_g, \sigma_g^2)$, where μ_g is the mean and σ_g^2 is the variance of the g^{th} group. The parameters of the multigroup 2PLM and GPCM were estimated using marginal maximum likelihood (MML) estimation with expectation-maximization (EM) algorithm (Bock & Aitkin, 1981).

To evaluate the group-level score accuracy and precision, we estimated the group-specific posterior mean and standard deviation estimates from the multigroup IRT model. The mean and standard deviation estimates were estimated for each country-by-language group, and the estimates were rescaled to the PISA scale using the transformation coefficients. The transformation coefficients consist of scaling (A) and centering (B) factors and can be used to make a linear transformation from the logit scale to the PISA scale for group scores.

$$PISA_g = A\mu_g + B \tag{3}$$

Each of the PISA cognitive domains has different transformation coefficients. In this study, we used the transformation coefficients that have been provided in the PISA 2015 Technical Report (OECD, 2016). For example, the scaling factor A was 131.58, and the centering factor B was 437.95 for the Reading domain. Similarly, for Mathematics and Science, respectively, the scaling factors A were 135.90 and 168.32, and the centering factors B were 514.18 and 494.54. The detailed description about computing the transformation coefficients is delineated in the PISA 2015 Technical Report (OECD, 2016). It is important to note that the rescaled group score estimates considered in this study are different than the typical LSA operational procedure. In operational settings, a latent regression model is generally used to address the heterogeneity of the group population distribution (Mislevy, 1984; Mislevy et al., 1992). Moreover, several plausible values (PV) are randomly drawn from the posterior distributions for individuals to compute the proficiency estimates for groups (von Davier et al., 2009). However, our preliminary studies found high Pearson correlation between rescaled PISA country mean scores and PV-based PISA country mean scores (above 0.95 across all domains). In addition, because the main purpose of the study is to investigate the impact of DIF on group score estimates, we used the direct estimates of the group scores from the multigroup IRT model and transformed the estimates to the PISA scale. This approach can also reduce possible confounding effects from the latent regression model and the PV procedure.

DIF detection and adjustment

After the initial multigroup IRT scaling was done, we evaluated item fit using the two quantities: mean deviation (MD) and root mean squared deviation (RMSD), for each item-by-group. The MD and RMSD for item j were computed as:

$$MD_{jg} = \int \left[P_{jg}^{obs}(\theta) - P_{jg}^{exp}(\theta)\right] f_g(\theta)\, d\theta \tag{4}$$

$$RMSD_{jg} = \sqrt{\int \left[P_{jg}^{obs}(\theta) - P_{jg}^{exp}(\theta)\right]^2 f_g(\theta)\, d\theta} \tag{5}$$

where $P_{jg}^{obs}(\theta)$ indicates the group-specific observed item characteristic curve (ICC) of item j, and $P_{jg}^{exp}(\theta)$ indicates the group-specific expected ICC of item j. $f_g(\theta)$ also represents the estimated group density function for group g. The integrals in Eqs. 4 and 5 are approximated with Gaussian-Hermite quadrature points ranging from −5 to 5 (von Davier, 2005).

To compute the observed ICC probability, we used the following definition:

$$P_{jg}^{obs}(\theta) = \sum_{i=1}^{N} \frac{x_{ijg} L_{ig}(\theta|\boldsymbol{X}) A_g(\theta)}{\sum_{q=1}^{Q} L_{ig}(\theta_q|\boldsymbol{X}) A_g(\theta_q)} \tag{6}$$

where x_{ijg} is the observed response from examinee *i* of group *g* for item *j*, $A_g(\theta)$ is the normalized group weight for group *g*, and $L_{ig}(\theta|\boldsymbol{X})$ is the likelihood function for examinee *i* of group *g*. The likelihood function is defined as:

$$L_{ig}(\theta|\boldsymbol{X}) = \prod_{j} P(X_{ijg} = x_{ijg}|\theta) \tag{7}$$

where $P(X_{ijg} = x_{ijg}|\theta)$ is the category response probability for x_{ijg}. To compute the expected ICC probability in Eqs. 4 and 5, we used the item response probability functions defined in Eqs. 1 and 2. Note that in this study, we computed the RMSD quantities based on sample statistics following the PISA operational scaling procedure (OECD, 2019). However, readers are referred to Köhler et al., (2020) for detailed descriptions and differences of the population and sample RMSD statistics.

The DIF item for each group was determined by using an RMSD cutoff of 0.12. That is, if the RMSD value for an item-by-group is greater than or equal to 0.12, then the item is flagged as DIF. Although various RMSD statistics and their cutoffs have been suggested (Robitzsch & Lüdtke, 2020, 2022), and a fixed RMSD cutoff could be unreasonable (Köhler et al., 2020; Robitzsch, 2022), in the current study, we used the conventional RMSD cutoff of 0.12 to detect DIF because the RMSD of 0.12 is currently used in the PISA and PIAAC operational scaling procedure for cognitive domains (OECD, 2016, 2019; Yamamoto et al., 2013). To be consistent with the operational procedure in LSAs and to increase the generalizability of the study, it is important to use the same RMSD cutoff value to detect DIF. In addition, the validity of the RMSD cutoff of 0.12 has been empirically evaluated in terms of scale comparability, overall model-data fit, and group score reliability (Joo et al., 2021).

To adjust the DIF in the multigroup IRT model, we re-estimated the unique item parameters in the subsequent scaling procedure. Specifically, item parameters detected as DIF were re-estimated for the DIF detected groups (i.e., DIF groups). In addition, we considered partially unique item parameters for the DIF groups. That is, if DIF items have the same direction of MD (positive or negative) for the DIF groups, the same unique item parameters were estimated across the DIF groups for the DIF item. Note that the partially unique item parameters approach has several advantages in the context of LSAs in that it could increase the scale comparability across the DIF groups and still hold partial invariance (Byrne et al., 1989). The DIF adjustment in the multigroup IRT model was iteratively conducted and continued until no DIF item-by-groups were detected.

Group score difference and Jackknife sampling

To investigate the impact of DIF on group score estimates, we compared the group scores with and without DIF adjustment. Specifically, we separately computed the rescaled PISA group (i.e., country-by-language group) scores from the initial IRT scaling, where no adjustment to misfit was considered, denoted as $\widehat{\mu}_{g0}$ for $g=1, \ldots G$, and the rescaled PISA group scores from the final IRT scaling, where adjustments to misfit took

place, denoted as $\widehat{\mu}_{gF}$. Then the rescaled PISA group scores difference $d\left(\widehat{\mu}_g\right)$ were computed as:

$$d\left(\widehat{\mu}_g\right) = \widehat{\mu}_{gF} - \widehat{\mu}_{g0} \tag{8}$$

We directly computed the group mean differences from the full invariance and partial invariance models. However, Robitzsch & Lüdtke (2022) recently proposed the adjusted and weighted group mean estimates and their statistical inference in the partial invariance approach. To estimate the standard error of the group mean difference estimate, we incorporated the Jackknife sampling method (Efron & Tibshirani, 1994). Using the Jackknife sampling approach, the sampling distribution of the group mean difference estimate can be formed and used to estimate the standard error of the group mean difference estimate. We first stratified the PISA item response data by respondents. More specifically, for each country-by-language group and senate weight, we stratified the samples, computed the statistic, and created the sampling distribution.

To describe the Jackknife sampling procedure more formally, suppose a stratum is denoted as $\boldsymbol{X}_s$, for s=1, ..., S. We then first subtracted the stratum from the total PISA data denoted as $\boldsymbol{X}_{-s}$. Note that the stratum was created and subtracted for each country-by-language group and senate weight, respectively, then combined to obtain $\boldsymbol{X}_{-s}$. Once $\boldsymbol{X}_{-s}$ is constructed, we applied the multigroup IRT scaling described in Eqs. 1 and 2 and computed the group score difference statistic using Eq. 8, denoted as $d_{-s}\left(\widehat{\mu}_g\right)$. This procedure was conducted iteratively and continued until the number of iterations reached S. To summarize and report the result, we computed the following statistics:

$$\widehat{d}\left(\widehat{\mu}_g\right) = \frac{1}{S}\sum_{s=1}^{S} d_{-s}\left(\widehat{\mu}_g\right) \tag{9}$$

$$SE\left[\widehat{d}\left(\widehat{\mu}_g\right)\right] = \sqrt{\frac{S}{S-1}\sum_{s=1}^{S}\left[d_{-s}\left(\widehat{\mu}_g\right) - \widehat{d}\left(\widehat{\mu}_g\right)\right]^2} \tag{10}$$

Note that the group score difference statistic was separately computed for each cognitive domain and each assessment type (CBA and PBA). To explore the results consistent with the PISA score reporting, we aggregated the results to the country-level groups, denoted as $\widehat{d}\left(\widehat{\mu}_c\right)$, and $SE\left[\widehat{d}\left(\widehat{\mu}_c\right)\right]$, c=1, ..., C by using senate weights:

$$\widehat{d}\left(\widehat{\mu}_c\right) = \sum_{l=1}^{L_c}\frac{w_{l(c)}}{\sum_{m=1}^{L_c} w_m}\widehat{d}\left(\widehat{\mu}_{l(c)}\right) \tag{11}$$

$$SE\left[\widehat{d}\left(\widehat{\mu}_c\right)\right] = \sum_{l=1}^{L_c}\frac{w_{l(c)}}{\sum_{m=1}^{L_c} w_m}SE\left[\widehat{d}\left(\widehat{\mu}_{l(c)}\right)\right] \tag{12}$$

where $\widehat{\mu}_{l(c)}$ is the group score and $w_{l(c)}$ is the senate weight of the l^{th} language group for the c^{th} country, and L_c indicates the total number of language groups for the c^{th} country.

Study 1 Results

Proportion of DIF items

Figure 1 shows the proportion of DIF items for each country across PISA scores. The results were summarized by each cognitive domain and assessment type. As shown in Fig. 1, the proportion of DIF items was mainly higher for high- or low-level performance countries. In addition, the Reading domain has the highest proportion of DIF items followed by Science and Mathematics. For the Reading domain, the proportion of DIF items ranged from 7 to 33% with a mean of 14% for CBA countries. The corresponding values ranged from 7 to 39% with a mean of 23% for PBA countries. For Mathematics, the proportion of DIF items ranged from 1 to 36% with a mean of 8% for CBA countries and from 1 to 42% with a mean of 14% for PBA countries. For Science, the proportion of DIF items ranged from 4 to 26% with a mean of 13% for CBA countries, and from 9 to 39% with a mean of 20% for PBA countries. Note that the proportions of DIF items are similar to the PISA 2018 scaling results provided in the technical report (OECD, 2019).

Country score difference

Figure 2 shows the country score difference from IRT scaling with and without DIF adjustment. The country score difference and their standard error estimates were obtained from the Jackknife sampling method. The left column of Fig. 2 shows the country score differences across the proportions of DIF items and the right column shows the standard error of the country score estimates.

As shown in Fig. 2 panels a, c, and e, the country score differences tend to increase as the proportion of DIF items increased. As expected, the country score differences were substantial for the Reading domain, given that the proportion of DIF items was relatively high. For Reading, the minimum and maximum differences were −10.88 points and 12.89 points, respectively, and the average difference was 1.36 points for CBA countries. For PBA countries, the minimum and maximum differences were −3.43 points and 3.90 points, and the average difference was 1.52 points. Similarly, for Mathematics, the minimum and maximum differences were −16.85 points and 5.48 points, and the average difference was −0.11 points for CBA countries. For PBA countries, the minimum and maximum differences were −16.25 points and 1.92 points, and the average difference was −3.20 points. Lastly, for Science, the minimum and maximum differences were −10.13 points and 5.33 points, and the average difference was 0.46 points for CBA countries. For PBA countries, the minimum and maximum differences were −7.67 points and 4.86 points, and the average difference was −0.41 points.

Because the positive and negative group score differences can cancel each other, we additionally explored the descriptive statistics for the absolute score differences. For CBA countries, the absolute score differences ranged from 0.01 to 12.89 with the average of 2.50 points for Reading, 0.04 to 16.85 with the average of 1.93 points for Mathematics, and 0.02 to 10.13 with the average of 1.84 points for Science. Similarly, for PBA countries, the absolute score differences ranged from 1.51 to 3.90 with the average of 2.62 points for Reading, 0.00 to 16.25 with the average of 4.13 points for Mathematics, and 0.67 to 7.67 with the average of 2.81 points for Science.

Finally, it was clearly shown that standard error estimates increased substantially as the proportion of DIF items increased. As shown in Fig. 2 (panels b, d, and f), the standard error estimates for all cognitive domains consistently increased. For CBA countries,

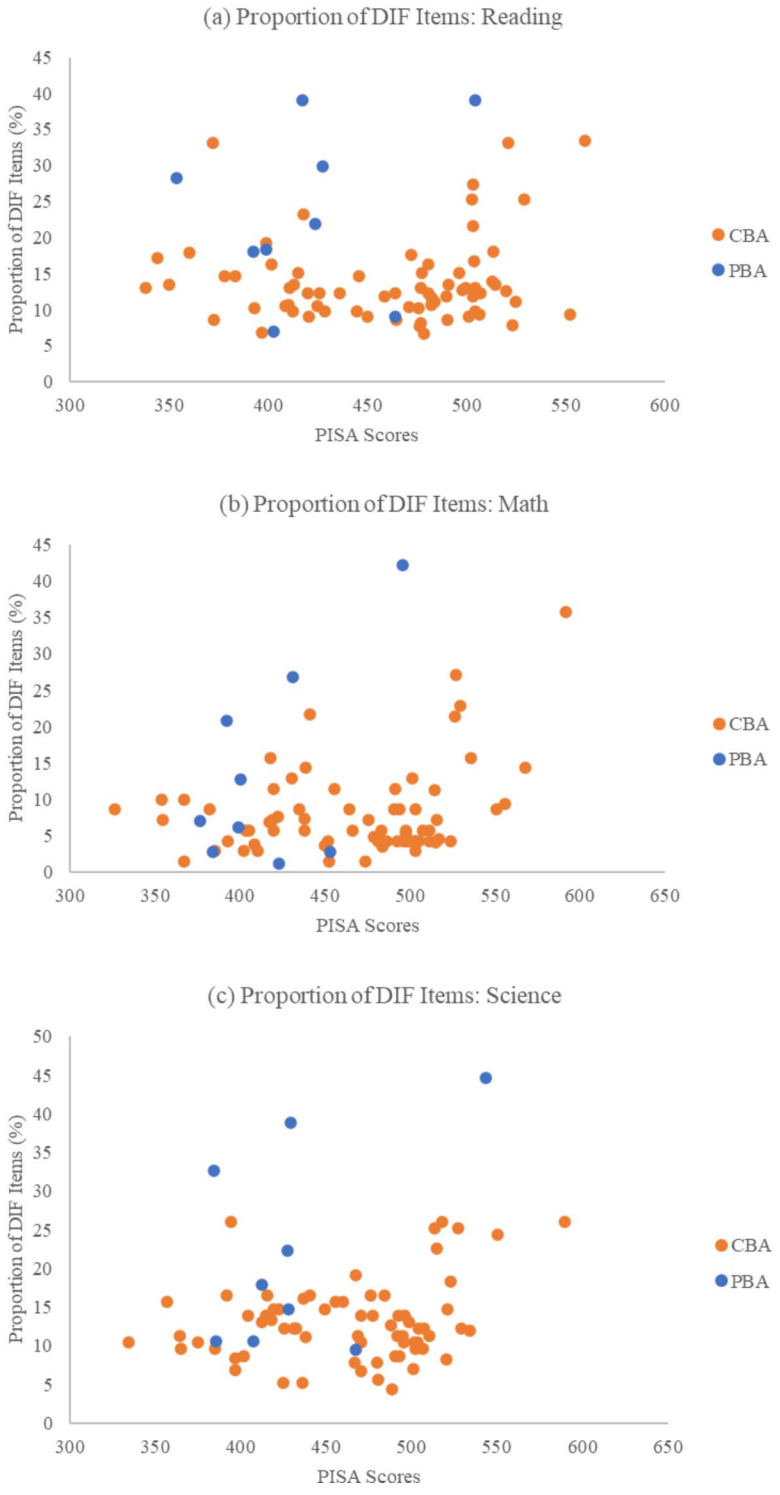

Fig. 1 The proportion of DIF items for CBA and PBA countries across PISA scores

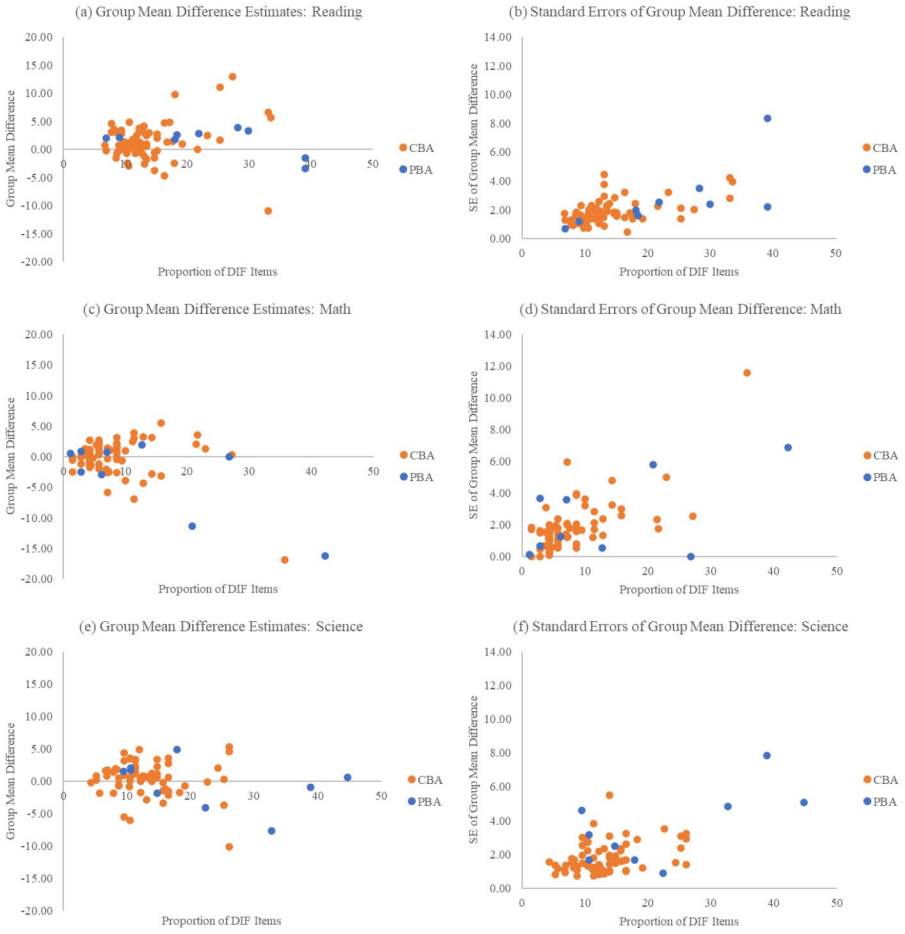

Fig. 2 PISA 2018 cognitive domain country score differences with and without DIF adjustment and their standard error estimates across the proportion of DIF items

the standard errors ranged from 0.44 to 4.47 with the average of 1.83 for Reading, 0.01 to 11.57 with the average of 1.91 for Mathematics, and 0.71 to 5.49 with the average of 1.79 for Science. Similarly, for PBA countries, the standard errors ranged from 0.69 to 8.35 with the average of 2.71 for Reading, 0.00 to 6.88 with the average of 2.51 for Mathematics, and 0.89 to 7.84 with the average of 3.58 for Science.

Study 2

Although we showed the impact of DIF items on group score estimates and their standard errors from the empirical PISA data, the extent to which DIF items cause the bias in the group score estimates and how the DIF adjustment addresses this bias is still unknown. Therefore, we conducted a simulation study in which data were generated with various levels of DIF in the context of LSAs.

Simulation design

For the simulation study design, we set the total number of items administered to each student to 40 and the number of item responses from each item to 500; we fixed the total number of groups to 10. In addition, we considered dichotomous response data only,

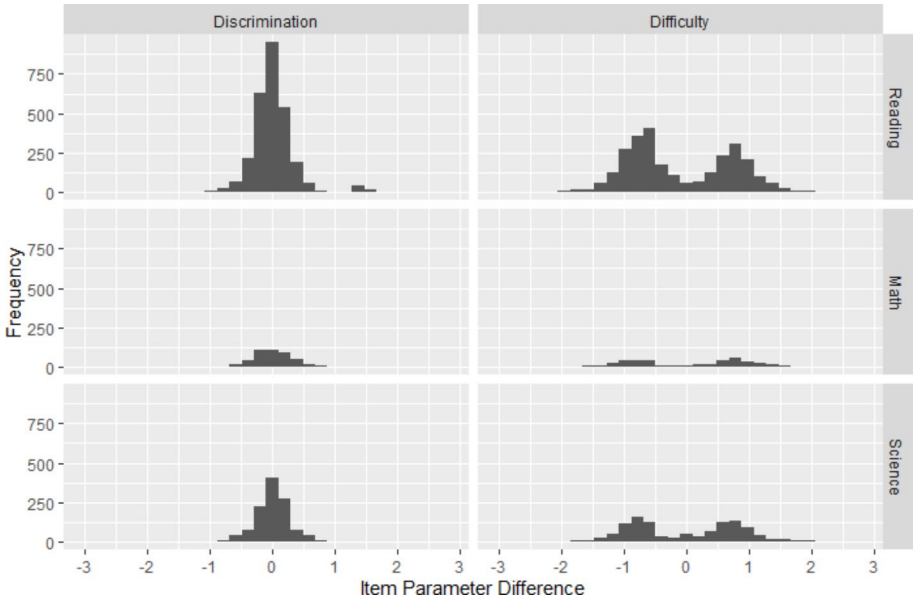

Fig. 3 The distribution of the PISA item parameter differences for DIF items

given that the majority of items in PISA consist of dichotomous response items. We varied four simulation conditions associated with how DIF items were generated:

1. Proportion of DIF items: 10%, 20%, and 40% of total items.
2. Type of DIF items: *a* parameter shift (nonuniform DIF) and *b* parameter shift (uniform DIF).
3. Size of DIF items: 0.3 (small) or 0.6 (large) for *a* parameter shift, and 0.5 (small) or 1.00 (large) for *b* parameter shift.
4. Direction of DIF items: positive and negative item parameter shift.

The proportions of DIF items we considered in the simulation conditions were consistent with the PISA 2018 cognitive domain data. As shown in Figs. 1 and 2, the proportion of DIF items ranged from approximately 5–40% across countries. In addition, we considered two types of DIF items in the study: uniform and nonuniform DIF. Previous studies have shown that different types of DIF items affect Type I error and DIF detection rates (e.g., Buchholz & Hartig 2019; Stark et al., 2006), and we expect that uniform DIF would have a larger impact on group score estimates than nonuniform DIF in the multigroup IRT model. We also considered an *a* parameter shift of 0.3 and a *b* parameter shift of 0.5 as small DIF size and an *a* parameter shift of 0.5 and a *b* parameter shift of 1.0 as large. To understand the range of DIF size more explicitly, we investigated the distribution of item parameter differences using the PISA 2018 data. Figure 3 shows the distributions of discrimination (*a*) and difficulty (*b*) parameter differences for DIF and nonDIF items across countries. As expected, the Reading domain has the highest difference in item parameters, followed by Science and Mathematics. The discrimination parameter difference ranged approximately from −1 to 1 and the difficulty parameter difference ranged from −2 to 2. The distribution for the discrimination parameter difference also showed a unimodal shape, whereas the corresponding distribution for the

difficulty parameter showed a bimodal distribution. Based on the item parameter difference distribution, we chose the DIF size values for the simulation study.

Data generation

The true item parameters were randomly drawn from uniform distributions with the ranges commonly observed in general LSAs. For discrimination parameters, the item parameters were randomly drawn from $U(0.75, 2.25)$, and for difficulty parameters, the item parameters were randomly drawn from $U(-2.00, 2.00)$. Note that the item parameters were randomly drawn for each replication to reduce the impact of item parameters on DIF items. To generate simulees for each group, we randomly generated latent trait parameters from the $N(\mu_g, 1)$, where μ_g is the true group mean parameter for group g. The true group mean parameters μ_g were also randomly generated from $U(-2, 2)$, which are in the range of commonly observed group scores in LSAs.

To generate DIF, we chose two groups (Group 2 and Group 3) as DIF groups. For the DIF groups, depending on the simulation conditions (e.g., proportions of DIF items, type of DIF items, size of DIF items, and direction of DIF items), we created item parameters that are different than the true item parameters (i.e., DIF item parameters). For example, for the condition where 40% of total items, uniform, large, and positive direction DIF were considered, we randomly selected 16 items of out 40 items and added the value of 1 to the true b parameters. Similarly, for the condition where 20% of total items, nonuniform, small, and negative direction DIF were considered, we randomly selected 8 items of out 40 items and subtracted the value of 0.3 from the true a parameters. We then generated the item response data using the DIF item parameters along with the DIF group simulees. For the DIF-free (nonDIF) groups, we used the nonDIF group-specific item parameters along with the group simulees to generate the item response data. Finally, the item response data for both DIF and nonDIF groups were combined to create the total dataset.

Analysis

The generated item responses were analyzed with two multigroup IRT models. Specifically, we first fitted a multigroup model with item parameter equality constraints across groups (denoted as DIF unadjusted model). We then identified the DIF items using RMSD and re-estimated the multigroup model with unique item parameters (denoted as DIF adjusted model). The group mean and standard deviation estimates from both models (i.e., DIF unadjusted and adjusted models) were also separately estimated. The estimated group mean estimates were then rescaled by multiplying 100 and adding 500 to be similar to the PISA scale scores. Similarly, the estimated group standard deviation estimates were also rescaled by multiplying 100. To evaluate the accuracy of the mean and standard deviation estimates, we computed bias and root mean squared error (RMSE) as follows:

$$Bias_g = \frac{\sum_{r=1}^{R} \widehat{\delta}_{g(r)} - \delta_{g(r)}}{R} \tag{13}$$

$$RMSE_g = \sqrt{\frac{\sum_{r=1}^{R} \left(\widehat{\delta}_{g(r)} - \delta_{g(r)}\right)^2}{R}} \tag{14}$$

where $\widehat{\delta}_{g(r)}$ is the estimated group mean or group standard deviation of the g^{th} group at the r^{th} replication, $\delta_{g(r)}$ is the true group mean or group standard deviation parameter for the g^{th} group at the r^{th} replication, and R is the total number of replications. For the current study, we set the total number of replications to 100. The bias and RMSE were computed for each group and separately averaged for the DIF groups and the nonDIF groups.

Study 2 Results

Table 1 shows bias and RMSE results for the group mean estimates from the DIF unadjusted and adjusted models across simulation conditions. Overall, the group mean bias was more substantial for the DIF groups than the nonDIF groups. More importantly, the DIF adjustment considerably reduced the bias and RMSE for the DIF groups. For the nonDIF groups, the bias ranged from −0.82 to 1.31 across the simulation conditions and the average bias from the DIF unadjusted and adjusted models were −0.16 and 0.13, respectively, for positive DIF and 0.14 and 0.01 for negative DIF. The bias for the nonDIF groups can be considered minimal based on the criteria provided by Hoogland & Boomsma (1998). For the DIF groups, the bias from the DIF unadjusted model ranged from −38.13 to 39.34, and the average bias was −8.35 for positive DIF and 8.43 for negative DIF. However, using the DIF adjusted model substantially reduced the bias by 50% on average. The bias from the DIF adjusted model ranged from −17.64 to 17.40, and the average bias was −3.61 for positive DIF and 3.52 for negative DIF. From the RMSE results, we found a similar pattern. RMSE of the nonDIF groups was consistent across the simulation conditions, whereas the corresponding value of the DIF groups ranged from 7.19 to 41.14 for the DIF unadjusted model. Similarly, using the DIF adjusted model substantially reduced the RMSE by 50%, ranging from 6.79 to 24.54 across the simulation conditions.

As the size and proportion of DIF increased, the bias and RMSE increased substantially, as expected. It is worthwhile to note that bias of the group mean estimates were more evident when DIF was created by shifting the b parameter (uniform DIF) than the a parameter (nonuniform DIF). When nonuniform DIF was considered, the highest bias was 1.93 across simulation conditions and fitted models. The bias and RMSE from nonuniform DIF were comparable to the results from the nonDIF groups. In addition, the direction of DIF also affected the direction of bias for the group mean estimates. For the positive DIF conditions, the direction of bias was negative, indicating that the group mean estimates were underestimated. For the negative DIF conditions, the direction of bias was positive indicating that the group mean estimates were overestimated.

Table 2 illustrates the bias and RMSE of the group standard deviation estimates for the nonDIF and DIF groups. The overall pattern of the results was similar to the group mean estimate results. For the nonDIF groups, the bias ranged from −1.48 to 0.20 and the average bias was −0.85 for positive DIF and −0.43 for negative DIF. In contrast, for the DIF groups, the bias was nonignorable, ranging from −10.68 to 2.20 using the DIF unadjusted model and −8.19 to 2.80 using the DIF adjusted model. The average bias of the DIF unadjusted and adjusted models were −1.26 and −0.94 for positive DIF and −3.92 and −2.99 for negative DIF. Overall, the bias showed negative values across the simulation conditions, indicating that the standard deviation estimates were underestimated

Table 1 *Bias and RMSE of Group Mean Estimates for DIF and nonDIF groups across the multigroup IRT models*

				Bias				RMSE			
				Unadjusted		Adjusted		Unadjusted		Adjusted	
Direction	Percent	Size	Type	NonDIF	DIF	NonDIF	DIF	NonDIF	DIF	NonDIF	DIF
Positive	10%	Small	Non	0.14	−0.17	−0.38	−0.21	7.66	7.82	7.57	7.19
			Uni	−0.16	−5.01	0.69	−2.36	7.72	9.22	7.19	8.70
		Large	Non	0.42	0.40	0.20	−0.68	7.45	8.13	7.35	7.60
			Uni	−0.21	−8.99	0.20	−0.44	8.00	12.99	8.21	8.11
	20%	Small	Non	0.20	0.36	0.72	1.14	7.89	8.35	7.98	8.72
			Uni	−0.04	−9.13	0.00	−7.67	7.33	12.79	7.89	11.46
		Large	Non	0.19	0.57	−0.01	0.00	7.55	7.65	7.95	8.22
			Uni	−0.99	−18.13	0.36	−0.81	7.65	20.90	7.71	7.27
	40%	Small	Non	0.29	−0.16	0.84	0.12	7.30	7.86	7.31	7.57
			Uni	−0.80	−20.18	0.19	−17.64	7.71	22.06	7.49	19.89
		Large	Non	−0.30	−1.67	−0.46	−1.67	7.57	9.30	7.31	8.66
			Uni	−0.70	−38.13	−0.82	−13.16	7.66	39.55	7.88	24.54
Negative	10%	Small	Non	0.37	0.42	0.74	0.82	7.34	7.89	7.69	6.79
			Uni	0.13	4.62	1.31	3.59	7.26	9.27	7.44	8.73
		Large	Non	0.68	1.37	−0.21	−0.39	7.62	9.04	7.05	8.12
			Uni	0.51	8.28	−0.07	0.74	7.78	12.18	7.62	7.67
	20%	Small	Non	−0.59	−0.40	−0.47	−0.80	7.49	7.19	7.53	7.08
			Uni	−0.59	9.66	−0.49	6.25	7.94	12.90	7.85	11.10
		Large	Non	0.15	−0.57	0.04	0.28	7.68	9.43	7.47	8.32
			Uni	0.58	19.04	0.23	1.53	7.25	21.66	7.64	7.68
	40%	Small	Non	0.98	0.87	−0.52	−0.06	7.86	8.81	7.90	8.43
			Uni	−0.63	19.13	−0.74	17.40	7.56	21.00	7.72	19.99
		Large	Non	−0.23	−0.58	0.23	1.93	7.65	14.66	7.79	10.33
			Uni	0.27	39.34	0.04	10.95	6.94	41.14	7.53	19.99

Note. Unadjusted = multigroup IRT model with the common item parameters without adjustment to DIF, Adjusted = multigroup IRT model with the unique item parameters adjusting for DIF, NonDIF = NonDIF groups, DIF = DIF groups, Non = Nonuniform DIF, Uni = Uniform DIF. Bias and RMSE were computed for the PISA scale scores

Table 2 *Bias and RMSE of Group Standard Deviation Estimates for DIF and nonDIF groups across the multigroup IRT models*

				Bias				RMSE			
				Unadjusted		Adjusted		Unadjusted		Adjusted	
Direction	Percent	Size	Type	NonDIF	DIF	NonDIF	DIF	NonDIF	DIF	NonDIF	DIF
Positive	10%	Small	Non	−0.91	−0.67	−1.48	−1.03	5.04	4.97	5.06	4.73
			Uni	−1.22	−1.53	−0.60	−1.16	4.97	5.42	4.95	5.29
		Large	Non	−0.79	−0.03	−1.11	−0.47	5.38	5.13	4.92	4.99
			Uni	−0.47	−2.17	−0.36	−0.84	5.16	6.95	4.75	5.28
	20%	Small	Non	−1.04	0.32	−0.78	0.84	5.08	4.68	5.35	5.60
			Uni	−1.38	−2.76	−0.09	−1.38	4.85	6.48	4.95	6.38
		Large	Non	−1.12	0.62	−0.76	1.47	5.12	5.11	4.99	5.35
			Uni	−0.98	−4.65	−0.60	−1.81	4.91	7.91	5.13	5.82
	40%	Small	Non	−1.48	1.13	−0.79	1.30	5.07	5.11	4.82	5.35
			Uni	0.20	−1.73	−1.04	−2.81	4.73	6.40	5.06	7.44
		Large	Non	−1.22	2.20	−0.74	2.80	5.36	6.25	5.40	5.74
			Uni	0.17	−5.83	−0.79	−8.19	5.31	10.16	5.28	13.95
Negative	10%	Small	Non	−0.42	−1.02	−1.07	−2.22	4.83	5.34	5.50	5.45
			Uni	−0.76	−0.98	0.07	−0.49	4.79	5.63	5.22	5.22
		Large	Non	−0.08	−3.03	−1.18	−2.50	4.94	6.03	4.72	5.38
			Uni	−0.22	−1.86	−1.08	−1.54	5.02	6.27	5.38	5.64
	20%	Small	Non	−0.61	−2.35	−0.12	−1.45	5.13	5.34	5.11	5.36
			Uni	−0.71	−2.24	−1.33	−2.52	5.12	6.25	5.28	6.40
		Large	Non	−0.02	−6.06	−0.06	−3.80	5.62	8.38	4.72	6.51
			Uni	−0.18	−3.79	−0.62	−1.66	5.21	7.70	5.10	5.71
	40%	Small	Non	−0.63	−4.45	−0.10	−3.82	5.05	7.12	5.34	6.86
			Uni	−0.43	−2.59	−1.00	−1.96	5.44	6.73	5.05	6.80
		Large	Non	−0.43	−10.68	0.32	−7.00	5.33	12.72	5.08	9.50
			Uni	−0.67	−8.01	−0.52	−6.96	5.10	10.78	4.54	12.24

Note. Unadjusted=multigroup IRT model with the common item parameters without adjustment to DIF, Adjusted=multigroup IRT model with the unique item parameters adjusting for DIF, NonDIF=NonDIF groups, DIF=DIF groups, Non=Nonuniform DIF, Uni=Uniform DIF. Bias and RMSE were computed for the PISA scale scores

regardless of the direction of DIF items. In addition, the bias was more substantial for the negative DIF than the positive DIF.

Discussion and Conclusion

In this study, we examined the impact of DIF items on group scores in the context of LSAs. Although much literature has previously discussed the benefits of the IRT calibration method for addressing DIF items in the multigroup structure (e.g., Oliveri & von Davier 2011, 2014, Rutkowski & Svetina, 2014; Rutkowski et al., 2016; von Davier et al., 2019), the degree to which the DIF adjustment affects the accuracy and precision of group performance estimates had not yet been empirically shown. To fill this gap, we conducted two studies. In the first study, we empirically showed the impact of the DIF adjustment on country score estimates using PISA 2018 main survey data. To precisely examine the effects of DIF adjustment, we incorporated Jackknife sampling to estimate the country score difference estimates and their standard errors. In the Jackknife sampling approach, we incorporated the DIF adjustment within the multigroup IRT scaling process for group comparisons. The multigroup IRT model with the DIF adjustment takes the uncertainty of items across countries and assessment cycles into account by simultaneously estimating international item parameters and fixing trend item parameters from previous assessment cycles (OECD, 2016, 2019; von Davier et al., 2019). This approach is comparable to the linking method using trend items in the presence of DIF (Robitzsch, 2021; Robitzsch & Lüdtke, 2019). In the second study, we conducted a simulation study to explore the consequence of DIF items and their adjustment on the group mean estimates directly obtained from the multigroup IRT models.

Based on the first study results, we found that the DIF items have a nonnegligible impact on the country scores and their standard error estimates for PISA 2018 cognitive domains. As the proportion of DIF items increased, the difference of the country score estimates obtained with and without the DIF adjustment considerably increased. The highest country score difference was −16.85 points on the PISA scale, observed in the Mathematics domain when the proportion of DIF items for the country was nearly 40%. Across countries, the Reading domain showed the largest score differences followed by Science and Mathematics, given that the proportion of DIF items was largest for Reading. In addition, we found that the standard error of country score differences increased as the proportion of DIF items increased, implying that the country score reliabilities can also be affected by DIF items. Consistent with the country score difference results, the standard error was highest for Reading followed by Mathematics and Science. Given that the proportion of DIF items per country in PISA 2018 data was as high as 40%, the results from the PISA data analysis provide the empirical evidence in which the DIF adjustment affects the country scores.

In the second study, we computed bias and RMSE of the group mean estimates from the two multigroup IRT models; the model with the constrained item parameters (DIF unadjusted model) and the model with the unique item parameters for the detected DIF items (DIF adjusted model). The data were generated by varying the proportion, size, type, and direction of DIF items, and we obtained the group mean estimates directly from the DIF adjusted and unadjusted models. We first found that the group mean estimates were underestimated when uniform DIF items were generated with a positive direction and overestimated when uniform DIF items were generated with a

negative direction. The group mean bias increased as the proportion and size of DIF items increased, and the bias was as high as 39.34 points on the PISA scale when 40% of the items contained large DIF. More importantly, we also found that the DIF adjusted model reduced the bias of group mean estimates across the simulation condition and 50% of the bias was reduced on average. However, the DIF adjusted model still produced bias; −3.61 points on average for positive DIF and 3.52 points for negative DIF on the PISA scale. These results indicate that the DIF items could yield biased group mean estimates, and the DIF adjustment can be implemented for the IRT scaling procedure to obtain the valid group mean estimates in the context of LSAs. In addition, the direction of DIF items should be carefully monitored to avoid the possible under or overestimation of the group mean estimates from the multigroup IRT model.

Interestingly, we also found that the group mean bias was mainly evident with the uniform DIF items, and the nonuniform DIF items had a minimum impact. This finding was consistent across the simulation conditions. This result implies that the group mean estimates are mainly affected by the uniform DIF items, and in operational settings, uniform DIF should be more explicitly investigated than nonuniform DIF. This result also highlights previous DIF studies in the context LSAs where uniform DIF is generally detected with high power than nonuniform DIF (e.g., Buchholz & Hartig 2019). Based on this finding, we recommend researchers and practitioners investigate DIF items more precisely by plotting ICCs, using common and group-specific item parameters.

In addition, we found that the group standard deviation estimates were also biased by the DIF items from the simulation study. Although the DIF adjustment somewhat addressed the bias, the group standard deviation estimates were underestimated overall, mainly with the negative direction DIF items, and the bias was as high as −8.19 points on the PISA scale. The corresponding RMSE value was 4.73. This finding has an important implication in the context of educational research. For example, if educational researchers and practitioners are interested in meta-analyzing student performance, it is common to obtain standardized effect sizes by using standard deviation estimates to make a valid cross-country performance comparison. Moreover, statistical inferences, such as interval estimates and hypothesis testing for country scores, also heavily rely on the valid standard deviation estimates. To obtain accurate standardized effect sizes and make a valid statistical inference, DIF items should be properly revised.

However, it is worthwhile to note that the simulation study results should be interpreted with caution. In operational assessments such as LSAs, the group score comparability is the main interest and estimating unique item parameters for DIF fundamentally decreases the comparability of the scale because the number of international item parameters reduces (Note that comparability is defined as the proportion of international item parameters in this study). From the psychometric perspective, it is critical to maintain the high comparability of the scales across groups and obtain the comparable group scores. Although we showed that the DIF items can cause the nonignorable bias of the group mean and standard deviation estimates from the simulation study, increasing the number of unique item parameters to address DIF items reduces the comparability of the scales and increases the model complexity. To obtain the comparable group scores in LSAs, it is important to primarily consider the high level of scale comparability and measurement invariance across groups (Rutkowski & Svetina, 2014). We also emphasize that the statistical decision on the DIF adjustment does not always relate to

construct-irrelevancy (Robitzsch & Lüdtke, 2020). As previous DIF studies discussed, DIF detection and adjustment should depend on statistical decisions and reviews from item experts and developers (Penfield & Camilli, 2007).

Finally, we acknowledged the limitations of the studies. Specifically, the simulation study we designed only investigated the limited data generation conditions. For example, the number of groups in the simulation was fixed at ten, and the number of items administered to students was fixed at 40. Although the numbers of fixed groups and items in this study are commonly observed in typical LSAs, to increase the generalizability of the results, more data generation conditions should be explored. Increasing the number of groups and items could affect the group mean and standard deviation estimates from the multigroup IRT model, and a future study is needed to examine the impact. In addition, in the simulation study, we only considered the dichotomous item response model (e.g., 2PLM) to generate the data. Given that the LSAs in general often include mixed-format tests, it is important to consider both dichotomous and polytomous item response data and investigate the impact of the DIF items on the group score estimates. Furthermore, the DIF detection method using RMSD assumes that the functional form of the fitted model adequately describes the data. In our empirical investigation, we used 2PL and GPCM for the dichotomous and polytomous responses in accordance with the PISA and PIAAC operational procedures. The future research should investigate the impact of DIF in LSAs using nonparametric DIF detection methods such as logistic regression. Finally, we did not include BIB design in the data generation procedure. The BIB design is commonly used in LSAs to cover a wide range of content and obtain reliable group score estimates. A future study should include the BIB design in the data generation conditions and investigate the impact of the BIB design along with the DIF items on group score estimates.

The results from the two studies provide important evidence that the DIF adjustment in IRT scaling is important and effective to address possible bias in group score reporting. We believe that the study contributes to the measurement literature in general and specifically to large-scale group-score assessments, providing information about DIF items and their consequences. Additionally, the study would provide guidelines for researchers and practitioners on how to properly address DIF item issues in the context of LSAs. The study results could also help lead to the development of new methods or modeling frameworks that consider the magnitude of misfit and consequently improves the current operational work in national and international LSAs.

Acknowledgements
The authors would like to thank Emily Kerzabi for her editorial help.

Authors' contributions
SJ conducted all analyses and prepared the initial draft of the manuscript. UA and FR provided the guidance concerning the analyses and read and approved the manuscript. HS commented on the manuscript.

Funding
The authors received no financial support for the research, authorship, and/or publication of this article.

Data Availability
The datasets analyzed during the current study are available in the OECD PISA-data repository https://www.oecd.org/pisa/data/2018database.

Declarations

Ethics approval and consent to participate
This research was based on a desk-based systematic literature review and no ethics approval was required and no human subjects were involved in the research.

Consent for publication
The authors provide consent for publication of this paper in the journal.

Competing interests
The authors have no known competing interests to disclose.

Received: 8 January 2022 / Accepted: 26 September 2022

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723

Birnbaum, A. (1968). *On the estimation of mental ability (Series Report No. 15)*. USAF School of Aviation Medicine

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). New York, NY: Springer

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

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456–466

Cosgrove, J., & Cartwright, F. (2014). Changes in achievement on PISA: the case of Ireland and implications for international assessment practice. *Large-scale Assessments in Education*, *2*, 1–17

De Jong, M. G., Steenkamp, J. B. E., & Fox, J. P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research*, *34*, 260–278

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hill

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing*, *2*, 199–215

Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing*, *5*, 23–35

Fox, J. P., & Verhagen, J. (2018). Random item effects modeling for cross-national survey data. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 529–550). London: Routledge

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement*, *38*, 164–187

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hilldale, NJ: Lawrence Erlbaum Associates

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*, 329–367

Joo, S., Khorramdel, L., Yamamoto, K., Shin, H. J., & Robin, F. (2021). Evaluating item fit statistic thresholds in PISA: Analysis of cross-country comparability of cognitive items. *Educational Measurement: Issues and Practice*, *40*, 37–48

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking*. New York, NY: Springer

Köhler, C., Robitzsch, A., & Hartig, J. (2020). A bias-corrected RMSD item fit statistic: An evaluation and comparison to alternatives. *Journal of Educational and Behavioral Statistics*, *45*, 251–273

König, C., Khorramdel, L., Yamamoto, K., & Frey, A. (2021). The benefits of fixed item parameter calibration for parameter accuracy in small sample situations in large-scale assessments. *Educational Measurement: Issues and Practice*, *40*, 17–27

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hilldale, NJ: Erlbaum

Mazzeo, J., & von Davier, M. (2014). Linking scales in international large-scale assessments. In L. Rutkowski, von M. Davier, & D. Rutkowski (Eds.), *Handbook of international large scale assessment* (pp. 229–257). Boca Raton, FL: CRC Press

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*, 359–381

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Chapter 3: Scaling procedures in NAEP. *Journal of Educational Statistics*, *17*, 131–154

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176

Neumann, K., Fischer, H. E., & Kauertz, A. (2010). From PISA to educational standards: The impact of large-scale assessments on science education in Germany. *International Journal of Science and Mathematics Education*, *8*, 545–563

Organization for Economic Co-Operation and Development (2016). *PISA 2015 Technical Report*. http://www.oecd.org/pisa/data/2015-technical-report

Organization for Economic Co-Operation and Development (2019). *PISA 2018 Technical Report*.http://www.oecd.org/pisa/data/2018-technical-report

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

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

Robitzsch, A. (2020). Lp loss functions in invariance alignment and Haberman linking with few or many groups. *Stats*, *3*, 246–283

Robitzsch, A. (2021). Robust and nonrobust linking of two groups for the Rasch model with balanced and unbalanced random DIF: A comparative simulation study and the simultaneous assessment of standard errors and linking errors with resampling techniques. *Symmetry*, *13*, 2198

Robitzsch, A. (2022). Statistical properties of estimators of the RMSD item fit statistic. *Foundations*, *2*, 488–503

Robitzsch, A., & Lüdtke, O. (2019). Linking errors in international large-scale assessments: Calculation of standard errors for trend estimation. *Assessment in Education: Principles Policy & Practice*, *26*, 444–465

Robitzsch, A., & Lüdtke, O. (2020). A review of different scaling approaches under full invariance, partial invariance, and noninvariance for cross-sectional country comparisons in large-scale assessments. *Psychological Test and Assessment Modeling*, *62*, 233–279

Robitzsch, A., & Lüdtke, O. (2022). Mean comparisons of many groups in the presence of DIF: An evaluation of linking and concurrent scaling approaches. *Journal of Educational and Behavioral Statistics*, *47*, 36–68

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, *39*, 142–151

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

Rutkowski, D., Rutkowski, L., & Liaw, Y. L. (2018). Measuring widening proficiency differences in international assessments: Are current approaches enough? *Educational Measurement: Issues and Practice*, *37*, 40–48

Rutkowski, L., Rutkowski, D., & Zhou, Y. (2016). Item calibration samples and the stability of achievement estimates and system rankings: Another look at the PISA model. *International Journal of Testing*, *16*, 1–20

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*, 31–57

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

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

Svetina, D., & Rutkowski, L. (2014). Detecting differential item functioning using generalized logistic regression in the context of large-scale assessments. *Large-scale Assessments in Education*, *2*, 1–17

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*, 1292–1306

von Davier, M. (2005). *mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models [Computer software]*. Princeton, NJ: ETS

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful?. In von M. Davier, & D. Hastedt (Eds.), *Issues and methodologies in large scale assessments* (2 vol.). Hamburg, Germany: IEA-ETS Research Institute.

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

Wu, M. (2010). Measurement, sampling, and equating errors in largescale assessments. *Educational Measurement: Issues and Practices*, *29*, 15–27

Yamamoto, K., Khorramdel, L., & Von Davier, M. (2013). Scaling PIAAC cognitive data. *Technical report of the survey of adult skills (PIAAC)*, Paris, France: OECD

Zwitser, R. J., Glaser, S. S. F., & Maris, G. (2017). Monitoring countries in a changing world: A new look at DIF in international surveys. *Psychometrika*, *82*, 210–232

Publisher's Note