Article

Journal of Educational and Behavioral Statistics
Vol. XX, No. X, pp. 1–29
DOI: 10.3102/10769986231183326
Article reuse guidelines: sagepub.com/journals-permissions

Alternatives to Weighted Item Fit Statistics for Establishing Measurement Invariance in Many Groups

Sean Joo
University of Kansas

Montserrat Valdivia
Dubravka Svetina Valdivia
Leslie Rutkowski
Indiana University Bloomington

Evaluating scale comparability in international large-scale assessments depends on measurement invariance (MI). The root mean square deviation (RMSD) is a standard method for establishing MI in several programs, such as the Programme for International Student Assessment and the Programme for the International Assessment of Adult Competencies. Previous research showed that the RMSD was unable to detect departures from MI when the latent trait distribution was far from item difficulty. In this study, we developed three alternative approaches to the original RMSD: equal, item information, and b*-norm weighted RMSDs. Specifically, we considered the item-centered normalized weight distributions to compute the item characteristic curve difference in the RMSD procedure more efficiently. We further compared all methods' performance via a simulation study and the item information and* b*-norm weighted RMSDs showed the most promising results. An empirical example is demonstrated, and implications for researchers are discussed.*

Keywords: *measurement invariance; differential item functioning; root mean square deviation; international large-scale assessments; PISA; PIAAC*

In large-scale educational testing, valid inferences rely on proving that the specified measurement model fits the data. In an international assessment context, this effort is complicated by the fact that model fit must hold within and across educational systems (given the complexity and ambiguity in conceptions of the "nation-state," particularly for city-states, nonnational systems, or territories with disputed or ambiguous political status, we refer to participating jurisdictions as educational systems). This effort is a particular challenge, given that, for instance, the 2018 cycle of the Organization for Economic Cooperation and Development's (OECD) Programme for International Student Assessment

(PISA) included about 80 highly heterogeneous educational systems that differed in language, culture, geography, and economic development. In addition, the OECD's Programme for International Assessment of Adult Competencies (PIAAC) Round 3 assessment included over 40 heterogeneous countries/economies to measure the major cognitive and workplace skills. The sheer number of systems and their heterogeneity substantially raise the complexity associated with validating model-data consistency.

Implicit in model fit evaluation is an assumption of measurement equivalence or measurement invariance (MI). That is, if a common model specification fits the data in each educational system according to agreed-upon criteria, the measure should function in the same way across these systems. Deviations from model-data consistency demonstrate that, in the studied systems, measurement differences exist, raising concerns about scale comparability. A relatively new method for determining measurement equivalence is the root mean square deviation (RMSD). This measure, which serves as the standard for detecting differential item functioning (DIF) in PISA and PIAAC, quantifies the weighted distance between the model-based and empirical item characteristic curves (ICCs). In the current context, the weight refers to the latent trait distribution of the group under consideration. We describe this measure in detail subsequently.

Given PISA's and PIAAC's prominence in international assessments and the dominant role that the RMSD plays in determining MI, an emergent body of literature offers several insights into the measure's performance. For example, an early study on the RMSD suggested that the measure was sensitive to group differences in item parameters; however, a cutoff much more stringent than used in operational settings was suggested (Buchholz & Hartig, 2019). Although not as extreme, recent research also recommended a more conservative RMSD threshold than currently used for cross-cultural assessments (Joo et al., 2021). This new cutoff is only slightly stricter than what is currently used in PISA and PIAAC operations (OECD, 2019; Yamamoto et al., 2013). Despite these studies, related research showed that—regardless of the chosen threshold value—the RMSD is unable to detect departures from measurement equivalence when the group for which DIF is considered has a latent trait distribution much lower than the item's location (Tijmstra et al., 2020). Importantly, this finding held no matter how severe the violation of MI. In particular, using a simulation method, Tijmstra et al. (2020) showed that the RMSD values were substantially below the cutoff value when DIF occurred for a low-performing country across varied measurement differences (e.g., difficulty parameters differed between .25 and 2.00 on the measurement scale). Empirically, this result bears out, for example, in the 2015 PISA technical report (OECD, 2016), which shows that, based on the operational RMSD threshold, there were far more DIF items in a medium-performing country like the United States than in the Dominican Republic, the lowest performing country. This report and associated research point to

substantial shortcomings for this measure in the context of international assessments (Köhler et al., 2021; Robitzsch & Lüdtke, 2022; von Davier & Bezirhan, 2021).

Considering recent research on the RMSD and the importance of the measure in international assessments, we pursue this topic in the current article. In particular, we develop three alternatives to the "original" RMSD, and we use a Monte Carlo method to evaluate their performance in a setting that mimics a typical international context. We also demonstrate these alternative measures using an empirical example. In what follows, we describe the properties of the original RMSD and then describe three candidate variations of the RMSD. Next, we compare these three measures in various experimental conditions. We summarize the results and recommend the best-performing measures for establishing MI.

RMSD and Its Properties

To orient our discussion, we begin with a description of the item response theory (IRT) model used operationally in the current article. In particular, PISA and PIAAC use a two-parameter logistic model (2PLM; Birnbaum, 1968) for dichotomously scored items:

$$P(X_i = 1|\theta) = \frac{\exp\{a_i(\theta - b_i)\}}{1 + \exp\{a_i(\theta - b_i)\}}, \tag{1}$$

where $P(X_i = 1|\theta)$ is the probability of responding correctly ($X_i = 1$), given their proficiency level θ, and b_i and a_i are item difficulty (location) and discrimination (scale) parameters for an item i, respectively. For polytomously scored items, the generalized partial credit model (GPCM; Muraki, 1992) is used:

$$P(X_i = x|\theta) = \frac{\exp\left\{\sum_{v=0}^{x} a_i(\theta - b_i + d_{iv})\right\}}{\sum_{c=0}^{m} \exp\left\{\sum_{v=0}^{c} a_i(\theta - b_i + d_{iv})\right\}}, \tag{2}$$

where $P(X_i = x|\theta)$ is the probability of selecting response category x ($x = 0, 1, \ldots, m$), given their proficiency level θ, and d_{iv} is an item threshold parameter ($d_{i0} = 0$ and $\sum_{v=1}^{m} d_{iv} = 0$ by definition). The RMSD is an item fit measure that describes the discrepancy between the model-based ICC (expected ICC) and an approximation of the empirical ICC (observed ICC), given as $P_{exp}(X_i|\theta)$ and $P_{obs}(X_i|\theta)$, respectively. The squared difference between these two curves is further weighted by the normalized density function of the latent variable θ, given as $f(\theta)$. The RMSD is defined as

$$RMSD_i = \sqrt{\int \left\{P_{obs}(X_i|\theta) - P_{exp}(X_i|\theta)\right\}^2 f(\theta) d\theta}. \tag{3}$$

For the dichotomous case, the expected ICC is directly computed from the item response probability $P(X_i = 1|\theta)$ described in Equation 1 using the

estimated item parameters. For the polytomous case, the expected ICC is computed from the expected score probability described as

$$P_{exp}(X_i|\theta) = \sum_{x=0}^{m} xP(X_i = x|\theta), \tag{4}$$

where $P(X_i = x|\theta)$ is the GPCM probability described in Equation 2. The observed ICC is computed based on the observed pseudo counts from marginal maximum likelihood (MML) estimation and can be calculated as

$$P_{obs}(X_i = x|\theta) = \sum_{p=1}^{N} \frac{x_{pi} L(\theta|\boldsymbol{X}_p) A(\theta)}{\sum_{q=1}^{Q} L(\theta_q|\boldsymbol{X}_p) A(\theta_q)}, \tag{5}$$

where x_{pi} is the observed item response from an examinee p for an item i, N is the total number of examinees ($p = 1, \ldots, N$), Q is the total number quadrature of points ($q = 1, \ldots, Q$) in MML estimation, and $A(\theta)$ is the normalized prior distribution in MML estimation. Moreover, $L(\theta|\,\mathbf{X}_p)$ is the likelihood function for an examinee p computed from

$$L(\theta|\,\mathbf{X}_p) = \prod_{i=1}^{J} P(X_{pi} = x_{pi}|\theta), \tag{6}$$

where J is the total number of items ($i = 1, \ldots, J$). For the dichotomous case, the observed ICC is computed directly from $P_{obs}(X_i = 1|\theta)$ using Equation 5. For the polytomous case, the observed ICC is computed by aggregating the category response probabilities

$$P_{obs}(X_i|\theta) = \sum_{x=0}^{m} xP_{obs}(X_i = x|\theta). \tag{7}$$

An important feature of the RMSD is that the location of a specific group proficiency distribution, $f(\theta)$ in Equation 3, functions as a weight and partly determines the value of the measure. This implies that if the center of $f(\theta)$ is far away from the location of an item, it can be challenging for the RMSD to detect DIF because the weight distribution disproportionally weighs the ICC differences regardless of severity. This issue has frequently occurred particularly in the case of positive DIF for a low-performing group.

To explicitly demonstrate the limitations of RMSD, Figure 1 illustrates two plots for low- and medium-performing groups. For each plot, two ICCs (e.g., a dashed line indicates the international ICC, and a dotted line indicates the group-specific ICC) and the group proficiency distribution (solid line) are included. For DIF analysis in the context of large-scale assessments, the international ICC can be viewed as belonging to the reference group and the group-specific ICC can be viewed as belonging to the focal group. In this example, an international item difficulty was fixed at 0 and a group-specific item difficulty was fixed at 2 (i.e., positive DIF). In Figure 1 panel a, the low-performing group proficiency distribution $f(\theta)$ was centered at -1.28, which corresponds to the

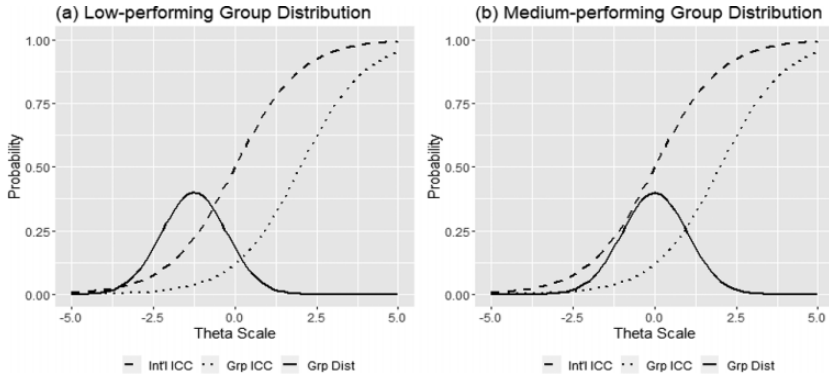

FIGURE 1. *International and group-specific item characteristic curves for the low- and medium-performing groups. (a) Low-performing group distribution. (b) Medium-performing group distribution.*

mean of the lowest performing country in the PISA Science domain in 2015 (OECD, 2016). On the other hand, in Figure 1 panel b, the medium-performing group proficiency distribution $f(\theta)$ was centered at 0, which matches the international item difficulty parameter. For the low-performing group, the RMSD was computed as .06, which is well below the operational threshold of .12. By contrast, for the medium-performing group, the RMSD was computed as .13 and the item would be detected as DIF with the operational threshold.

The asymmetry inherent in the RMSD performance is of particular concern, given that most newcomers to PISA and similar studies tend to be from less economically developed countries (e.g., PISA for development), which also usually perform substantially below the OECD (2019) average. Further, empirical results show that even recent adjustments to PISA to better accommodate these lower performing countries still result in a meaningful mismatch between item difficulty and country-level proficiency estimates (Rutkowski & Rutkowski, 2021). When considering all the evidence, it suggests that DIF could be present in lower performing groups but not easily detectable, despite being common and significant. Therefore, enhancing the RMSD becomes even more important in the context of international large-scale assessments.

Modified RMSD Measures

To address the problems associated with a mismatch between ICC differences and the disproportional RMSD weight distribution for the low-performing groups as shown in Figure 1, in this study, we propose three alternative RMSD statistics. More specifically, we applied three alternative weights in the RMSD statistic to increase DIF detection rates: equal weight, item information weight, and *b*-norm weight. We describe each of these subsequently.

The first alternative weight we considered is the equal weight distribution. The equal weight RMSD uses a uniform distribution $f_{eq}(\theta)$ rather than the group proficiency distribution $f(\theta)$ in the RMSD statistic in Equation 3. The uniform distribution $f_{eq}(\theta)$ can be obtained from

$$f_{eq}(\theta) = \frac{1}{U - L}, \text{ if } L < \theta < U \text{ and } f_{eq}(\theta) = 0, \text{ otherwise.} \tag{8}$$

Note that U and L are the upper and lower limits of the group-specific distribution. Given that the group-specific distribution is asymptotic to zero for both extremes of the latent trait continuum, we fixed the lower and upper limits that cover 99% of the distribution. The range between the lower and upper limits is approximately equivalent to the range of 3 standard deviations away from the mean of the group-specific distribution. The equal weight RMSD statistic can be considered the "average" difference between the international and group-specific ICCs across the group-specific latent continuum. It is important to note that because the range varies with the group-specific distribution, the equal weight RMSD statistic is influenced by the population distribution. We consider the group-specific range rather than the fixed range for the equal weight RMSD statistic because the observed ICC is formed in the range of the group-specific distribution and the observations between two ICCs are most frequent within the group-specific distribution. It is valuable to explore the equal weight approach because it has the potential to minimize the impact of disproportionally weighting the ICC difference in the original RMSD.

The second alternative for the weight distribution in the RMSD statistic we considered is the item information function $f_{info}(\theta)$. The item information weight RMSD uses the 2PLM or GPCM item information function (Donoghue, 1994) based on international item parameter estimates. We considered the international parameters to compute the item information function because the international parameters serve as the reference group in the context of large-scale assessments. In addition, we used the *normalized* item information distribution in the RMSD statistic. The marginalized item information function for the 2PLM can be obtained from

$$f_{info}(\theta) = \frac{a_i^2 P(X_i|\theta)\{1 - P(X_i|\theta)\}}{\sum_{q=1}^{Q} a_i^2 P(X_i|\theta_q)\{1 - P(X_i|\theta_q)\}}. \tag{9}$$

Similarly, the marginalized item information function for the GPCM can be obtained from

$$f_{info}(\theta) = \frac{a_i^2 \left[\sum_{x=0}^{m} x^2 P(X_i = x|\theta) - \left\{\sum_{x=0}^{m} x P(X_i = x|\theta)\right\}^2\right]}{\sum_{q=1}^{Q} a_i^2 \left[\sum_{x=0}^{m} x^2 P(X_i = x|\theta_q) - \left\{\sum_{x=0}^{m} x P(X_i = k|\theta_q)\right\}^2\right]}. \tag{10}$$

As shown in Figure 1, the ICC difference may not always be efficiently computed, because the original RMSD uses the group-specific distribution as a weight distribution, which centers on the mean of the group-specific distribution. Thus, it is reasonable to consider an item-centered weight distribution instead of the group-centered weight distribution. In principle, the item information function provides the highest value at the location of the item difficulty parameter (b) and the distribution forms around the center value across the latent trait continuum. The discrimination parameter (a) also controls the steepness and level of the item information function. In this study, we consider the normalized item information function as the alternative weight distribution to maintain consistency in the area under the item information curve and to guarantee that the function integrates to one for all items. Furthermore, because the item information distribution is normalized, the scale of the item information weight RMSD statistic is on the same metric as the original RMSD statistic (between 0 and 1). We expect that using the item-centered weight distribution rather than the group-centered weight distribution would increase the chance to detect the difference in ICC between the two groups. This is because the item-centered distribution focuses on the range of the ICC, where the international ICC has the steepest slope, which is also the range where uniform DIF is likely to result in the most significant increase or decrease in item probability.

To illustrate how the information function can serve as an RMSD weight in contrast to group achievement, Figure 2 continues our previous example. The left plot illustrates international and group-specific ICCs with the group proficiency distribution, and the right plot shows the two ICCs with the item information curve. Note that the group-specific ICC was plotted in the range, where the group proficiency probability is greater than .01. As shown in Figure 2, the ICC difference can be better captured with the item information curve, and as a result, a greater RMSD value can be obtained. Based on the item parameters in the example, the RMSD increased from .06 to .24 by weighting with the item information curve.

One potential problem associated with the item information distribution is that the distribution shape and weight density would depend on the item discrimination parameter. Consequently, it is difficult to obtain consistent results across all items because the item discrimination parameters are allowed to vary for the 2PLM and GPCM (i.e., confounding effect). To address this issue, we consider the b-norm weight distribution $f_b(\theta)$ in the RMSD statistic, which we will describe next. The b-norm weight that computes the standard normal probability density centered on international item difficulty (b) parameters. The b-norm weight distribution can be computed from

$$f_b(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{1}{2}(\theta - b_i)^2\right\}. \tag{11}$$

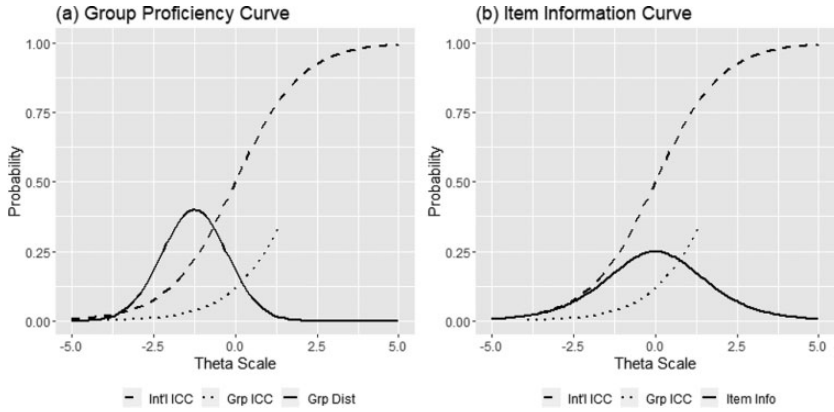

FIGURE 2. *International and group-specific item characteristic curves and (a) the group proficiency and (b) the item information curves.*

As is typical in the IRT framework, *b* parameters represent the value where the probability of correct response is 0.5 (i.e., center). Therefore, the *b*-norm weight is distributed symmetrically and centered on the international ICC. The *b*-norm weight is similar to the item information curve in the sense that both curves are centered at the location of the international item parameters. However, the *b*-norm weight fixes the distribution shape and density at a constant value, and only the center of the distribution is affected by the item difficulty parameter. As a result, it is possible to obtain consistent weight density across all items. In addition, because the *b*-norm distribution is adapted from the standard normal distribution, the sum of the area under the curve is always one.

To illustrate the difference between the item information and the *b*-norm weight distributions specifically, Figure 3 shows the *b*-norm and item information curves with various *a* parameter values using the 2PLM. As shown in Figure 3, the shape of item information curves (i.e., kurtosis) changes based on the value of *a* parameters. The smaller *a* parameter values tend to produce platykurtic item information curves, whereas the larger *a* parameter values tend to produce leptokurtic item information curves. In preliminary investigations, we found that item information with an *a* parameter of 1.60 produces a similar shape as the *b*-norm distribution. We found the *a* parameter value of 1.60 by minimizing the area of two curves using the quasi-Newton minimization function (Fletcher & Reeves, 1964) implemented in the statistical program R (*optim* function). To that end and based on Figure 3, we expect that the *b*-norm weight and the item information weight RMSD statistics perform similarly for the item with discrimination equal to 1.60.

We believe that both item information and *b*-norm are practical alternative weight distributions for computing RMSDs because those densities are centered

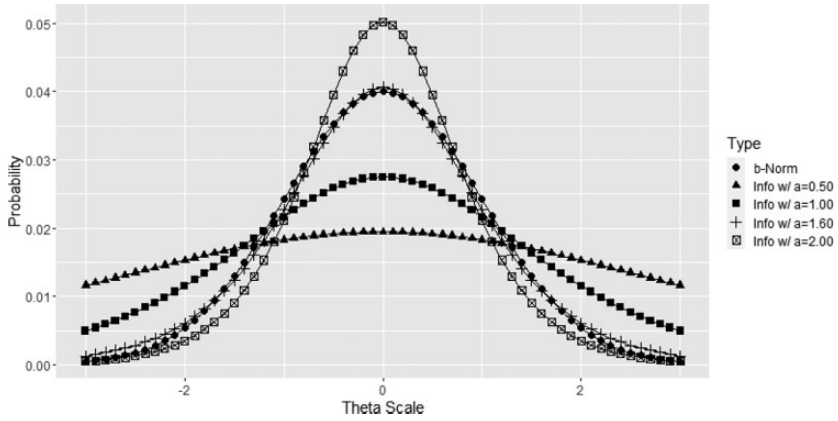

FIGURE 3. *Normalized item information densities with various item discrimination (*a*) parameters and the standard normal distribution density (*b*-norm weight) centered at zero.*

at the international item difficulty parameters rather than the group means. If the group-specific ICCs were to deviate from the international ICC, the difference would be most evident in the range where the ICC slope is the steepest. Hence, calculating the RMSD statistics based on the weight distribution centered at the international ICC would be a reasonable alternative approach for detecting differences in ICCs. By doing so, we expect that the RMSD with item information or *b*-norm weights would capture the ICC difference more clearly and efficiently across the latent trait continuum, especially for the lower performing countries.

Purpose of the Study

The purpose of this study is to develop alternative weighting approaches for computing RMSD statistics and to compare their performance with the original RMSD statistic. Given that the RMSD statistic in general does not follow an asymptotic distribution (Köhler et al., 2020) and the proposed RMSD statistics are newly developed, it is important to empirically investigate the extent to which the proposed RMSD statistics control Type I error under various conditions. More importantly, we need to explore the extent to which three alternative RMSD statistics improve the DIF detection rate (or power) under various large-scale assessment designs (e.g., the number of participating groups and their proficiency distributions), especially for those which does not match the location of the item and the proficiency distribution. To answer these research questions, we conducted a simulation study and compared the performance of the proposed RMSD methods under various conditions. The simulation study results will serve as empirical evidence for providing practical guidelines and recommendations to

those who analyze large-scale assessment data in applied settings. We also applied these alternative measures to an empirical data set.

Method

To explore the relative performance of three alternative RMSD measures (equal, item information, and *b*-norm), we conducted a Monte Carlo simulation study, where we manipulated several factors. Our design choices were informed by the PISA 2018 cognitive domain data. Below, we describe and provide a rationale for the study design decisions, including fixed and manipulated factors, the process of item and person parameters selection, data generation, and an outline of the plan of analysis to guide the interpretation of results.

Simulation Design

In the current study, we fixed the number of groups, per-group sample size, and the number of items. Specifically, we fixed the number of groups at 80 and the respective sample size at 500 per group. In PISA, for example, the number of student responses per item is approximately 500 because of the balanced incomplete block design, in which students respond to only a subset of items. In addition, PISA uses normalized sampling weights, such that each country contributes equally for item parameter estimation. Based on this design, we fixed the sample size of 500 per group and administered all items to students in the simulation study design, which is equivalent to the number of item responses in operational settings. Similarly, we fixed the number of items at 40 to approximate PISA design, where all students receive two assessment blocks with each block consisting of about 20 items in a single domain (OECD, 2019). Furthermore, we limited the generated data to dichotomous responses and uniform DIF to keep the study manageable and to avoid confounding factors. We investigated uniform DIF because previous studies reported that uniform DIF is more prevalent in operational settings (Joo et al., 2022; Joo & Lee, 2022) and the RMSD is more sensitive to uniform DIF than nonuniform DIF (Buchholz & Hartig, 2019).

We manipulated several factors in this study, including the proportion of DIF items (5%, 15%, and 30%), the magnitude of DIF items (small, medium, and large), the direction of DIF items (unidirectional and bidirectional), the difficulty of DIF items (easy, moderate, and hard), and the group proficiency distribution (low- and medium-performing groups). We provide the details and rationale for each manipulated factor as follows.

Percentage of DIF items. We manipulated the percentage of DIF items per group based on the PISA 2018 analysis. In accordance with the PISA 2018 Reading results, the percentage of DIF items per group ranged from 3% to 34% across a total of 211 items. Thus, we examined 5% (small), 15% (medium), and 30% (large) of DIF items, which resulted in 2, 6, or 12 items

being chosen as DIF items. We expect that as the percentage of DIF items increases, the power to detect DIF using the original RMSD statistic would decrease due to the equality constraint across all items in the multigroup IRT model estimation (Buchholz & Hartig, 2019; Joo et al., 2021; Joo & Lee, 2022; von Davier et al., 2019). Consequently, as more items are contaminated with DIF, more item parameter estimates could be biased and DIF detection would decrease.

Magnitude of DIF items. To investigate the effect of DIF magnitudes, we analyzed the PISA 2018 Reading data. In particular, we conducted a two-step analysis procedure. We first fitted the IRT model by fixing the international item parameters obtained from Annex A in the PISA 2018 technical report (OECD, 2019). Subsequently, we obtained the group proficiency distribution estimates and used them to estimate the item parameters for each item-by-country combination. The estimated item parameters were considered group-specific item parameter estimates and they were compared with the international item parameters (obtained from the technical report) to find magnitudes and percentages of DIF items. Of the 211 PISA 2018 Reading items, 58 had DIF more extreme than the latent trait continuum ± 3 for one or more countries. Using the quantiles of the DIF magnitude distribution, we examined the DIF size of 0.48 (small), 0.77 (medium), and 1.28 (large) in this study.

Direction of DIF items. We also included as factors two DIF directions in this study: unidirectional DIF and bidirectional DIF. In the unidirectional DIF condition, all DIF items were generated to be in the positive direction. That is, DIF items were generated by adding positive values to the international item parameters. In the bidirectional DIF condition, half of the DIF items were generated in the positive direction and the other half of the DIF items were generated in the negative direction. We considered these DIF direction factors for two reasons. First, it is possible to observe both unidirectional and bidirectional DIF items in operational settings. For example, bidirectional DIF items are often observed in medium-performing groups, whereas unidirectional DIF items are often found in low- or high-performing groups (i.e., positive DIF for low-performing groups and negative DIF for high-performing groups). Second, in the context of international large-scale assessments such as PISA and PIAAC, the direction of DIF items would directly affect the location of international item parameters when item parameters are concurrently estimated across all countries. Specifically, the more positive (or negative) DIF items are included, the more international item parameter estimates would be impacted, and as a result, items from non-DIF countries could also be flagged as containing DIF (i.e., Type I error). Thus, it is worthwhile to explore the performance of the alternative RMSD measures for two different directional DIF conditions.

Item difficulty of DIF items. Three average levels of DIF item difficulties were examined: easy, moderate, or hard. The main purpose of the study is to examine the improvement of the alternative RMSD statistics when the relative position of item and group distributions were not aligned well. In addition, it is important to explore the performance of alternative RMSDs for a broad range of item difficulties. In the simulation study, we generated easy items from difficulty parameters ranging from -1.91 to -0.42, moderate items from difficulty parameters ranging from -0.41 to 0.54, and hard items from difficulty parameters ranging from 0.55 to 2.65. Note that we chose the item difficulty parameter values commonly observed in operational settings, specifically, from the analysis of PISA 2018 Reading data.

Group proficiency distributions. We investigated two levels of group proficiency distributions for the DIF group: low-performing and medium-performing. We were particularly interested in the performance of alternative RMSD statistics for the low-performing groups, given that the original RMSD statistic is problematic for detecting DIF items, especially for low-performing groups. We included the medium-performing groups as well to explore the performance of alternative RMSD statistics for the majority of participating groups. Based on the analysis of PISA 2018 Reading data, we considered -1.24 for the mean of the low-performing group distribution, which was the minimum proficiency average on the logit scale. Similarly, we considered 0.24 for the mean of the medium-performing group distribution, which aligned with the average of group proficiencies from the PISA 2018 Reading analysis. Note that we generated DIF items for one group only in this study. For the low-performing group, we expect the alternative RMSDs would better capture the ICC difference, resulting in a higher power to detect DIF than the original RMSD.

The fully crossed study design yielded 108 conditions ($= 3$ [percentage of DIF items] $\times$ 3 [magnitude of DIF items] $\times$ 2 [direction of DIF items] $\times$ 3 [item difficulty of DIF items]) $\times$ 2 [group proficiency distribution]). As a baseline condition, we also generated data with no DIF on any items for any group. Each condition was replicated 200 times as a means of ensuring stable results.

Data Generation

To simulate the data, item and person parameters were first randomly drawn depending on the simulation conditions. For the discrimination parameters of the data generation model, we randomly generated values from a U(0.75, 2.25) for each replication. The range was chosen based on the PISA 2018 Reading item parameters, and we randomly generated the discrimination parameter values for each replication to minimize possible confounding effects due to parameter selection. Across 40 items, we chose 13 easy items, 14 moderate items, and 13 hard items. As described earlier, the easy items were randomly generated from

a U(−1.92, −0.42). Similarly, the moderate and hard items were randomly generated from U(−0.42, 0.54) and U(0.54, 2.65), respectively. To generate uniform DIF items, we shifted *b* parameters from the international item parameter values. For example, for the 5% small, easy item DIF condition, we randomly selected two of 13 easy items (of 40 items total) and added a value of .48 (small DIF) to the selected items. Similarly, for the 30% large, hard item DIF condition, we randomly selected 12 of 13 hard items (of 40 items total) and added a value of 1.28 (large DIF) to the selected items. For the group proficiency distributions, we first randomly generated the group mean parameters from U(−0.75, 1). Once the group mean parameters were generated, the person parameters were randomly generated from a normal distribution with the generated group mean parameters and the standard deviation of one. Once item and person parameters were obtained, we generated the response data using Equation 1.

Analysis and Outcome Variables

To estimate item parameters across many groups simultaneously, we applied a multigroup IRT model with an equality constraint on item parameters across groups. The metric was set by fixing the mean and variance of the first group to be zero and one, respectively. We used the *mdltm* program (von Davier, 2005) for fitting the multigroup IRT model. Once the parameters of the multigroup IRT model were estimated, we evaluated the item fit statistics generated from the original RMSD and our three proposed RMSD alternatives. The item fit statistics were computed for each item-by-group combination. Finally, we used the value of .12 as the cutoff point to determine DIF, following the PISA operational scaling procedures (OECD, 2019). We used the same cutoff point to detect DIF because the alternative RMSD statistics were computed not from the ICC (international and group-specific item parameters) difference but from different weight distributions, implying that the original and alternative RMSD statistics are on the same scale. In addition, previous studies have shown the effectiveness of the fixed cutoff point of .12 using simulated (Buchholz & Hartig, 2019) and empirical data (Joo et al., 2021). The data generation, RMSD statistics computation, and analysis were simultaneously conducted using the statistical program R (R Core Team, 2020).

We reported Type I error and power rates to summarize the simulation study results. Type I error was computed from the proportion of non-DIF items incorrectly flagged as DIF (false positives) and power was computed from the proportion of DIF items correctly detected (true positives) across replications. We evaluated Type I error for both no DIF (baseline) and DIF conditions. Finally, we reported the average Type I error rates across all non-DIF items and the average power rates across all DIF items to summarize the results.

TABLE 1.
Type I Error Rates of the Alternative RMSD Methods for the No DIF Conditions

		RMSD Weights			
Group Level	Item Diff.	Original	Equal	Item Information	*b*-Norm
Low	Easy	.00	.00	.00	.00
	Moderate	.00	.00	.01	.02
	Hard	.00	.00	.03	**.08**
Medium	Easy	.00	.00	.04	**.06**
	Moderate	.00	.00	.00	.00
	Hard	.00	.00	.03	.05

Note. Values in bold indicate the Type I error rates greater than the nominal level (.05). Item Diff. = item difficulty level; RMSD = root mean square deviation; DIF = differential item functioning.

Results

Baseline Condition

To investigate the performance of the alternative RMSD methods under the correctly specified models, we first investigated the DIF detection rates under the no DIF item (baseline) conditions. The DIF detection rates were first computed for each item across replications and then averaged across items. In addition, we only considered the conditions not related to the DIF items.

Table 1 shows the DIF detection rates results under the no DIF conditions (i.e., Type I error). We found that the original, equal, and item information weight RMSD statistics controlled the Type I error rates well across simulation conditions. Type I error rates were less than the nominal level (.05) for all conditions. For the *b*-norm weight approach, Type I errors were also controlled reasonably well. The maximum Type I error was .08 when data were generated with the low-performing groups and hard items were considered.

Type I Error and Power

Tables 2 and 3 show the Type I error results of the bidirectional DIF conditions for low- and medium-performing groups, respectively. Overall, the original and alternative RMSDs showed well-controlled Type I error across the simulation conditions. The majority of the Type I error values were below the nominal level (.05) for both low-performing group and medium-performing group conditions. For the low-performing group conditions, the largest Type I error rates were .03, .03, .06, and .08 for the original, equal, item information, and *b*-norm RMSDs, respectively. Similarly, for the medium-performing group conditions, the largest Type I error rates of the original, equal, item information, and *b*-norm RMSDs were .02, .05, .06, and .07, respectively.

TABLE 2.
Type I Error Results of the Bidirectional DIF Conditions for the Low-Performing Groups

			RMSD Weights			
Item Difficulty	DIF Size	DIF %	Original	Equal	Item Information	*b*-Norm
Easy	Small	5	.00	.00	.01	.03
		15	.00	.00	.01	.03
		30	.00	.00	.01	.03
	Medium	5	.00	.00	.01	.03
		15	.00	.00	.01	.03
		30	.00	.00	.03	.05
	Large	5	.00	.00	.01	.02
		15	.00	.00	.03	.04
		30	.01	.03	.02	**.08**
Moderate	Small	5	.00	.00	.01	.03
		15	.00	.00	.01	.03
		30	.00	.00	.02	.04
	Medium	5	.00	.00	.01	.03
		15	.00	.00	.02	.04
		30	.00	.00	.03	**.07**
	Large	5	.00	.00	.01	.03
		15	.00	.00	.03	**.06**
		30	.03	.03	**.06**	**.08**
Hard	Small	5	.00	.00	.01	.02
		15	.00	.00	.01	.01
		30	.00	.00	.01	.01
	Medium	5	.00	.00	.01	.02
		15	.00	.00	.01	.02
		30	.00	.00	.01	.01
	Large	5	.00	.00	.01	.02
		15	.00	.00	.02	.03
		30	.00	.00	.05	.05

Note. Values in bold indicate the Type I error rates greater than the nominal level (.05). DIF % = percentage of DIF items; RMSD = root mean square deviation; DIF = differential item functioning.

Tables 4 and 5 show the Type I error results of the unidirectional DIF conditions for the low- and medium-performing groups, respectively. Overall, the Type I error results were considerably higher compared to the bidirectional DIF conditions for all RMSDs. Specifically, the Type I error rates were notably increased as the size of DIF and the percentage of DIF increased. In addition, the item information and *b*-norm RMSDs overall showed more frequent Type I error rates than the original and equal RMSDs across simulation conditions. For example, for the low-performing groups, the Type I error rates were .21, .42, .38,

TABLE 3.
Type I Error Results of the Bidirectional DIF Conditions for the Medium-Performing Groups

			RMSD Weights			
Item Difficulty	DIF Size	DIF %	Original	Equal	Item Information	*b*-Norm
Easy	Small	5	.00	.00	.01	.01
		15	.00	.00	.00	.01
		30	.00	.00	.01	.01
	Medium	5	.00	.00	.01	.01
		15	.00	.00	.01	.01
		30	.00	.00	.01	.01
	Large	5	.00	.00	.01	.01
		15	.00	.00	.01	.01
		30	.01	.00	.05	.02
Moderate	Small	5	.00	.00	.01	.01
		15	.00	.00	.01	.01
		30	.00	.00	.01	.02
	Medium	5	.00	.00	.01	.01
		15	.00	.00	.01	.02
		30	.00	.00	.03	.04
	Large	5	.00	.00	.01	.02
		15	.00	.00	.04	.05
		30	.02	.05	**.06**	**.07**
Hard	Small	5	.00	.00	.01	.01
		15	.00	.00	.01	.01
		30	.00	.00	.01	.01
	Medium	5	.00	.00	.01	.01
		15	.00	.00	.01	.01
		30	.00	.00	.01	.01
	Large	5	.00	.00	.01	.01
		15	.00	.00	.02	.02
		30	.01	.00	.04	.02

Note. Values in bold indicate the Type I error rates greater than the nominal level (.05). DIF % = percentage of DIF items; RMSD = root mean square deviation; DIF = differential item functioning.

and .56 for the original, equal, item information, and *b*-norm RMSDs, respectively, when the item difficulty was easy, the size of DIF was large, and the percentage of DIF was 30%. Similarly, for the medium-performing groups, the Type I error rates were .39, .38, .51, and .56 for the original, equal, item information, and *b*-norm RMSDs, respectively, when the item difficulty was moderate, the size of DIF was large, and the percentage of DIF was 30%. Interestingly, the increased Type I error was more evident for the conditions when DIF item

TABLE 4.
Type I Error Results of the Unidirectional DIF Conditions for the Low-Performing Groups

			RMSD Weights			
Item Difficulty	DIF Size	DIF %	Original	Equal	Item Information	*b*-Norm
Easy	Small	5	.00	.00	.01	.03
		15	.00	.00	.02	.03
		30	.01	.00	**.08**	**.09**
	Medium	5	.00	.00	.01	.03
		15	.01	.00	**.07**	**.07**
		30	**.07**	**.06**	**.24**	**.32**
	Large	5	.00	.00	.02	.03
		15	**.08**	.04	**.27**	**.30**
		30	**.21**	**.42**	**.38**	**.56**
Moderate	Small	5	.00	.00	.01	.03
		15	.00	.00	.03	**.06**
		30	.00	.01	**.07**	**.13**
	Medium	5	.00	.00	.02	.04
		15	.00	.01	**.07**	**.11**
		30	.00	**.06**	**.15**	**.24**
	Large	5	.00	.00	.03	.05
		15	.00	.04	**.14**	**.21**
		30	.03	**.18**	**.25**	**.38**
Hard	Small	5	.00	.00	.01	.03
		15	.00	.00	.01	.03
		30	.00	.00	.01	.01
	Medium	5	.00	.00	.01	.02
		15	.00	.00	.01	.03
		30	.00	.00	.02	.03
	Large	5	.00	.00	.01	.03
		15	.00	.00	.01	.04
		30	.00	.00	.03	.05

Note. Values in bold indicate the Type I error rates greater than the nominal level (.05). DIF % = percentage of DIF items; RMSD = root mean square deviation; DIF = differential item functioning.

difficulty was easy and moderate. When hard items were generated as DIF, all RMSDs showed acceptable Type I error rates across the simulation conditions.

Figures 4 and 5 show the power rates of the bidirectional DIF conditions for the low- and medium-performing groups. In addition, Figures 6 and 7 show the power rates of the unidirectional DIF conditions for the low- and medium-performing groups. In these figures, each of the columns shows the RMSD methods, the rows indicate the percentage of DIF, and the marker shapes

TABLE 5.
Type I Error Results of the Unidirectional DIF Conditions for the Medium-Performing Groups

			RMSD Weights			
Item Difficulty	DIF Size	DIF %	Original	Equal	Item Information	*b*-Norm
Easy	Small	5	.00	.00	.01	.01
		15	.00	.00	.02	.03
		30	.00	.00	.04	.03
	Medium	5	.00	.00	.01	.01
		15	.00	.00	**.06**	.05
		30	**.07**	.03	**.18**	**.14**
	Large	5	.00	.00	.02	.02
		15	.04	.03	**.23**	**.17**
		30	**.46**	**.29**	**.52**	**.51**
Moderate	Small	5	.00	.00	.01	.01
		15	.00	.00	.01	.02
		30	.01	.00	**.07**	.05
	Medium	5	.00	.00	.01	.02
		15	.00	.00	.05	.04
		30	**.13**	**.06**	**.27**	**.28**
	Large	5	.00	.00	.02	.02
		15	**.10**	.05	**.25**	**.24**
		30	**.39**	**.38**	**.51**	**.56**
Hard	Small	5	.00	.00	.01	.02
		15	.00	.00	.03	.04
		30	.00	.01	.02	.03
	Medium	5	.00	.00	.01	.02
		15	.00	.01	**.06**	**.07**
		30	.00	.01	.03	.03
	Large	5	.00	.00	.02	.03
		15	.00	.05	**.11**	**.12**
		30	.02	.03	.05	.05

Note. Values in bold indicate the Type I error rates greater than the nominal level (.05). DIF % = percentage of DIF items; RMSD = root mean square deviation; DIF = differential item functioning.

(i.e., circle, triangle, and square) represent the size of DIF. The *x*-axis conveys the levels of item difficulty, and the *y*-axis represents the power rates. The light gray lines illustrate the thresholds that serve as a reference criterion (.80).

Based on the power results, the item information and *b*-norm weight RMSDs overall outperformed the original and equal weight RMSDs. The average power rates across the simulation conditions were .61, .66, .78, and .82 for the original, equal, item information, and *b*-norm RMSDs, respectively, indicating that the

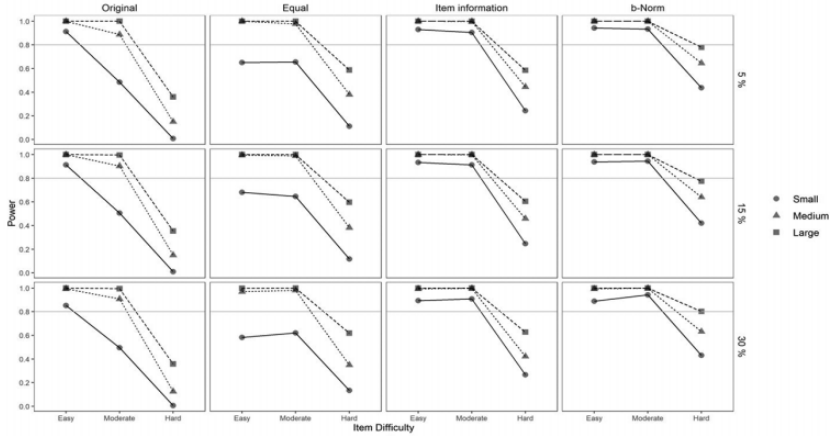

FIGURE 4. *Power results of the bidirectional differential item functioning conditions for the low-performing group.*

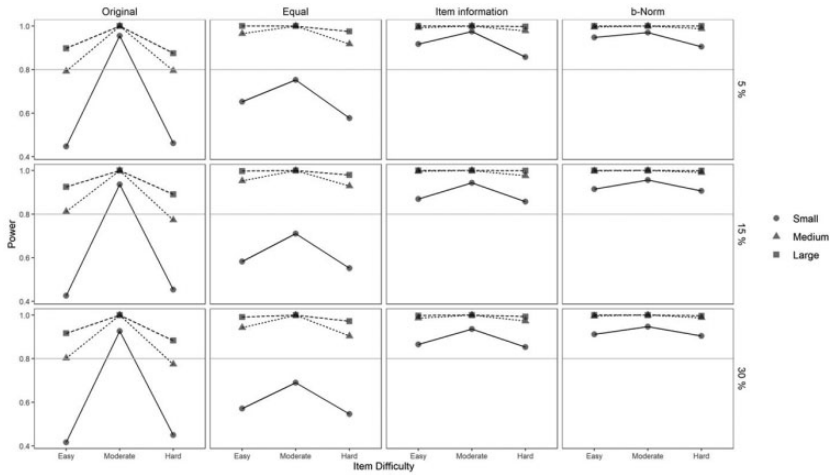

FIGURE 5. *Power results of the bidirectional differential item functioning conditions for the medium-performing group.*

b-norm weight RMSD yielded the highest power to detect DIF. The minimum power rates for the original, equal, item information, and *b*-norm RMSD statistics were .00, .00, .11, and .10 and the maximum power rates were 1.00 for all methods across the simulation conditions.

Notably, as the size of DIF increased, the power rates increased for all RMSD statistics, but the *b*-norm RMSD consistently showed the highest power rates

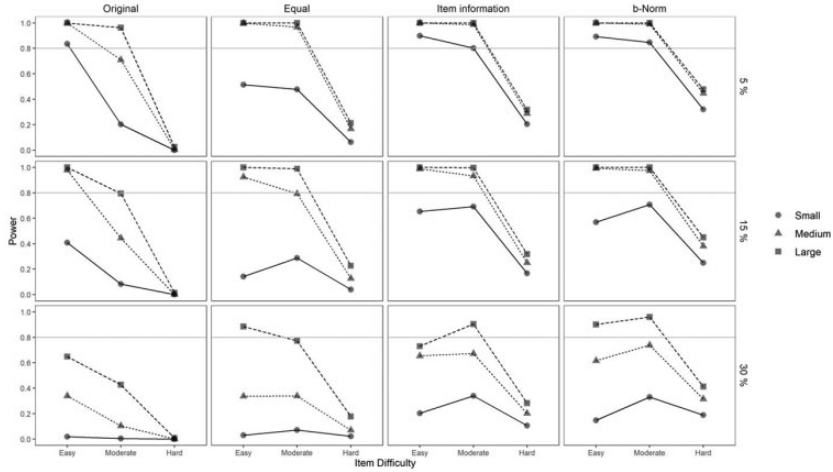

FIGURE 6. *Power results of the unidirectional differential item functioning conditions for the low-performing group.*

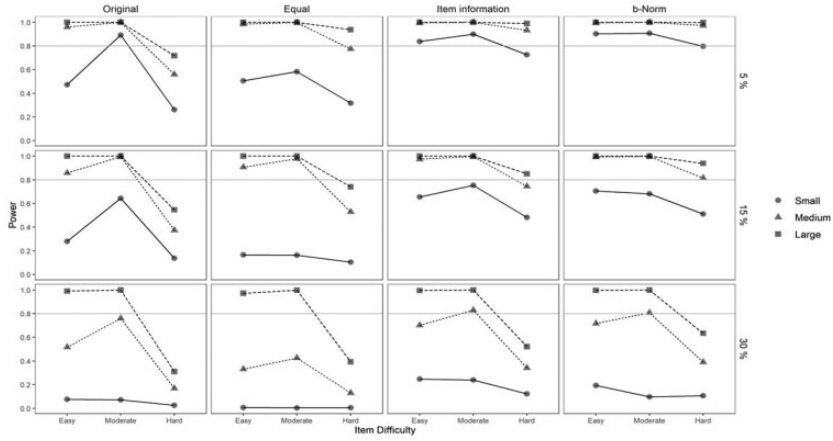

FIGURE 7. *Power results of the unidirectional differential item functioning conditions for the medium-performing group.*

across the DIF size conditions. For example, the average power rates of the small, medium, and large DIF conditions were .39, .66, and .77, respectively, for the original RMSD, whereas the corresponding values were .65, .83, and .88 for the item information RMSD and .68, .86, and .92 for the *b*-norm RMSD. In addition, as the percentage of DIF increased, the power rates decreased, but again the *b*-norm RMSD consistently produced the highest power rates. For example, the

average power rates of the 5%, 15%, and 30% DIF conditions were .68, .63, and .51 for the original RMSD, whereas the corresponding values were .89, .85, and .72 for the *b*-norm RMSD.

More importantly, the power rates differ substantially depending on the item difficulty and the group proficiency levels. For the low-performing group, the power rates reached the maximum when the item difficulty was easy, implying the RMSDs produced the best power rates when the item location and group proficiency matched. As expected, as the item difficulty increased, power rates decreased substantially. This pattern was consistent across all RMSD methods. However, the item information and *b*-norm RMSDs showed less decrement than the original and equal RMSDs across the conditions. For example, average power rates of easy, moderate, and hard DIF item conditions were .69, .62, and .18 for the original RMSD, whereas the corresponding values were .81, .84, and .24 for the item information RMSD and .81, .84, and .36 for the *b*-norm RMSD. Similarly, for the medium-performing group, power rates reached the maximum when item difficulty was moderate (i.e., when the item location and group proficiency matched) and power rates decreased when item difficulty was easy or hard. Interestingly, although the pattern was similar across all methods, item information, and *b*-norm RMSDs showed less decrement, and especially, power rates were above the acceptable criterion (.80) when bidirectional DIF was considered. For example, average power rates of easy, moderate, and hard DIF item conditions were .84, .89, and .44 for the original RMSD, whereas the corresponding values were .97, .98, and .69 for the item information RMSD and .97, .98, and .79 for the *b*-norm RMSD.

Finally, all RMSD methods showed higher power rates when bidirectional DIF was generated compared to unidirectional DIF. The average power rates of the original, equal, item information, and *b*-norm RMSDs were .49, .53, .69, and .72 for the unidirectional DIF condition as opposed to .72, .78, .88, and .92 for the bidirectional DIF condition.

Empirical Example

To illustrate the feasibility and performance of the alternative RMSD statistics in real data, we analyzed the PISA 2018 Reading data. Because the alternative RMSD statistics were proposed based on evidence that DIF is difficult to detect in lower performing countries, we selected the lowest performer (Dominican Republic) and a medium performer (United States) for our demonstration. Both countries participated in the computer-based assessment and a total of 245 items were administered. There were 224 dichotomous items and 21 polytomous (3-point scale) items. The dichotomous items were fitted with the 2PLM and the polytomous items were fitted with the GPCM using the *mdltm* program (von Davier, 2005). The total number of students was 5,674 for the Dominican Republic and 4,838 for the United States. For IRT scaling, we fixed the item parameters

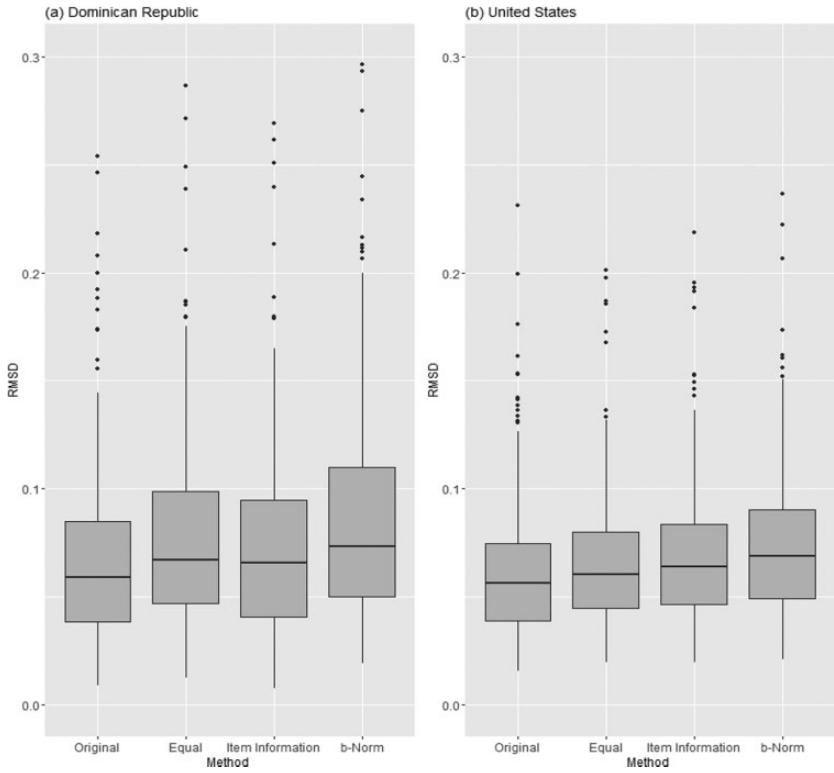

FIGURE 8. *The distributions of four root mean square deviations for the Dominican Republic (left) and the United States (right) using the Programme for International Student Assessment 2018 Reading data.*

to their internationally estimated values (OECD, 2019) and computed the four RMSD statistics (original, equal, item information, and *b*-norm weights).

Figure 8 shows the distributions of four RMSD statistics for the Dominican Republic and the United States. Overall, the RMSDs were mainly distributed between 0.0 and 0.1, which indicates that equal item parameters fit the data adequately. In addition, alternative RMSDs showed more values that exceeded the cutoff point than the original RMSD distribution. Using the RMSD cutoff point of .12, we detected 29 DIF items (12% DIF) with the original RMSD, 41 DIF items (17% DIF) with the equal weight RMSD, 32 DIF items (13% DIF) with the item information weight RMSD, and 55 DIF items (22% DIF) with the *b*-norm weight RMSD for the Dominican Republic. The average values were .07, .08, .08, and .09 for the original, equal, item information, and *b*-norm weight RMSDs, respectively, indicating that the *b*-norm weight RMSD yielded the highest RMSD values. On the other hand, for the United States, we detected

17 DIF items (7% DIF) with the original RMSD, 14 DIF items (6% DIF) with the equal weight RMSD, 19 DIF items (8% DIF) with the item information weight RMSD, and 29 DIF items (12% DIF) with the *b*-norm weight RMSD. The average values were .06 for the original, .07 for the equal, .07 for the item information, and .08 for the *b*-norm RMSDs, respectively.

Note that these findings are consistent with the simulation study results, highlighting that the alternative RMSD statistics (e.g., *b*-norm weight) have higher detection rates than the original RMSD statistic and they are more sensitive for the lower performing country than the medium-performing country. In addition, it is important to note that the percentages of DIF items were relatively higher for the Dominican Republic (e.g., ranged from 12% to 22% across methods) than the United States (e.g., ranged from 6% to 12% across methods). This is consistent with the RMSD distributions where the difference between the original and *b*-norm RMSDs was higher for the Dominican Republic than the United States. However, given that true DIF items are unknown in the empirical data set, readers should carefully interpret the results from the Dominican Republic and the United States. Note that the main purpose of the empirical example is to demonstrate the RMSD distributions across original and alternative RMSD approaches with real data.

Discussion

In the context of operational international large-scale assessments such as PISA and PIAAC, MI analysis across participating countries and assessment cycles solely depends on the conventional RMSD statistic (OECD, 2019; Yamamoto et al., 2013). However, previous studies have reported that the conventional RMSD statistic underdetects DIF items, especially for low-performing countries. Given that it is critical to maintaining the MI property across countries and assessment cycles (Rutkowski & Svetina, 2014); in this study, we explored alternative approaches to compute the RMSD statistic in the international large-scale assessment context. We proposed three alternative approaches for computing the RMSD statistic: equal weight, item information weight, and *b*-norm weight distributions. We conducted a simulation study to evaluate the performance of the proposed methods, including the original RMSD statistic under various conditions that mimic operational designs. We manipulated several simulation factors including DIF size, DIF percentage, DIF item difficulty, DIF direction, and group proficiency level. In addition, we used the PISA 2018 Reading data to empirically illustrate the feasibility of the proposed methods.

The simulation results empirically showed that the item information and *b*-norm weight RMSD statistics have satisfactory performance. When the model is correctly specified (i.e., when no DIF items are present), the Type I error rates for falsely detecting non-DIF items are well-controlled for both item information and *b*-norm RMSDs. Additionally, the power to detect DIF items is consistently

higher for the item information and *b*-norm RMSDs compared to the original and equal RMSDs. When the model is incorrectly specified (i.e., when DIF items were included), we found mixed results in terms of Type I error. Specifically, when bidirectional DIF items (positive and negative) were included, Type I error rates were well-controlled, consistent with the no DIF item condition. However, when the unidirectional DIF items (i.e., all positive) were included, all methods yielded an increased Type I error rate. We discuss the implications of the findings next.

First, the power to detect true DIF was consistently higher for item information and *b*-norm RMSD methods across simulation conditions. As DIF size increased, item information and *b*-norm RMSDs showed higher power than the original and equal RMSDs, regardless of group proficiency and item difficulty levels. In contrast, the original RMSD produced poor results, consistent with previous studies (e.g., power less than .80), especially when group proficiency did not match item difficulty. This finding highlights that item information or *b*-norm weights are practical alternatives for establishing measurement equivalence in international assessments, such as PISA and PIAAC. We believe that because the item information and *b*-norm are item-centered distributions, they are more effective at capturing the ICC differences between the international and group-specific item parameters. Regardless of the group proficiency level, the item information and *b*-norm distribution capture the ICC difference centered on the international item parameters. The underdetection issue associated with the original RMSD can be improved by using item information or *b*-norm weights in the RMSD computation. This development should prove useful to practitioners and applied researchers involved in operational work, given that accurate DIF identification is essential for reporting comparable group scores (Joo et al., 2022). Our simulation study results suggest that the use of item information or *b*-norm weights is recommended for calculating RMSD in operational settings. The R code for the alternative RMSD can be found in the publicly available GitHub repository (https://github.com/seunghwan-joo/rmsd.git).

Second, we also found that the *b*-norm weight distribution showed the highest power across all RMSD methods, indicating that the b-norm weight is a more consistent and effective alternative than the item information weight. One problem associated with the item information weight is that the density depends on both item discrimination and difficulty parameters. In PISA and PIAAC operations, discrimination parameters are allowed to vary across items (e.g., 2PLM and GPCM), and if an item has a low international discrimination parameter, the item information weight would have lower density, where the most severe ICC discrepancy occurs. Consequently, the RMSD value for the low discrimination item would be lower than the item that has a higher discrimination parameter. This item discrimination dependency can be solved by using the *b*-norm weight distribution, where the density is simply generated from a standard normal distribution centered at the *b* parameter. The *b*-norm distribution fixes the steepness

of the density at a constant value and only the *b* parameter affects the location of the distribution. Our simulation empirically showed that the *b*-norm distribution yielded consistently higher power across simulation conditions than the item information distribution. More importantly, the Type I error rates were comparably well-controlled as the other methods when the model is correctly specified and bidirectional DIF items were included.

It is important to note that all RMSD methods showed high Type I error rates for some conditions, especially when unidirectional DIF items were included. The Type I error increased as the percentage of DIF items increased. One of the possible explanations for the Type I error is due to the equality constraints imposed in the item parameters estimation procedure across groups. In general, when DIF analysis is conducted with contaminated anchor items, Type I error tends to be increased to some degree (Lopez-Rivas et al., 2008; Stark et al., 2006; Wang & Yeh, 2003; Wang & Woods, 2017). This is more problematic for the international assessment setting, because IRT scaling assumes that all items—even those with DIF—can (and do) serve as anchor items across all participating groups. Consequently, the estimated international item parameters would be shifted positively (or negatively) if all DIF items were generated in the positive (or negative) direction. This international item parameter shift would increase the Type I error rates because the non-DIF items would also be detected as DIF as the international item parameters depart from the generated (true) values. Our simulation study results empirically showed this evidence as the Type I error only increased when unidirectional (positive) DIF items were included. However, more importantly, unidirectional DIF is uncommon in operational settings, especially for medium-performing countries because DIF directions tend to be both positive and negative (e.g., some items are harder, and some items are easier than the international item difficulty as shown in the PISA technical report; OECD, 2019). In addition, in PISA and PIAAC operational scaling procedures, items from previous assessment cycles (i.e., trend items) are carried over to link the multiple assessment scales onto the same metric. Consequently, the trend items are kept constant at their pre-estimated parameters during the IRT modeling procedure. This approach ensures that any potential shifts in the international item parameters would not impact the DIF analysis for the trend items (König et al., 2021; OECD, 2019).

Finally, we evaluated the performance of our proposed alternative weight distributions using PISA 2018 Reading data. We considered the Dominican Republic and the United States as low- and medium-performing countries, respectively, in the empirical example. We found empirical evidence that the RMSD methods using the item information and *b*-norm distributions increased DIF detection rates and the rate was larger for the Dominican Republic. For example, for the Dominican Republic, the percentage of DIF-detected items increased by approximately 10% using the *b*-norm RMSD, whereas for the United States, the percentage of DIF-detected items increased by approximately

5% using the *b*-norm RMSD. The distributions of RMSD statistics also illustrated that the RMSD values were larger with the alternative RMSD methods. These results highlight that the alternative RMSDs are feasible to apply in real data in the context of large-scale assessments and showed that the proposed methods can improve DIF detection rates for low-performing countries.

At a reviewer's suggestion, we also investigated the performance of the proposed methods for nonuniform DIF conditions. We generated nonuniform DIF items by shifting the group-specific discrimination parameters of .10 (small DIF), .25 (medium DIF), and .60 (large DIF) from the international parameters. The size of DIF was chosen based on the empirical evidence from PISA 2018 Reading domain results (OECD, 2019) and previous DIF studies (Rutkowski & Svetina, 2014; Svetina & Rutkowski, 2014). Although all RMSD methods controlled Type I errors reasonably well (below the nominal level of .05), we did not find a substantial improvement in terms of power across all methods. Given that previous studies have reported the RMSD statistic is more sensitive to uniform DIF than nonuniform DIF (Buchholz & Hartig, 2019; Joo et al., 2021, 2022) and it is more common to observe uniform DIF or a combination of uniform and nonuniform DIF in PISA and PIAAC operations, we focused our investigation on uniform DIF in this study. However, the full simulation results for the nonuniform DIF are available in the GitHub repository.

The current study also includes several limitations. First, simulation studies are limited to the simulation conditions. For example, the number of groups in the simulation was fixed at 80 and the number of administered items was 40. In addition, DIF items were only generated for one group. Although the numbers of groups and items in this study are commonly observed in PISA and PIAAC, to increase generalizability, more simulation conditions could be explored. For example, it is common to have more than one group to have DIF in operational settings, given that the group-specific item parameters usually vary around the international item parameters. Various numbers of participating groups, DIF groups, and administered items could be explored to mimic various large-scale assessments, such as the Progress in International Reading Literacy Study or the Trends in International Mathematics and Science Study. In addition, in the simulation study, we investigated the performance of alternative RMSD statistics using dichotomous item responses only, although we computed the alternative RMSD statistics for the polytomous item responses in the empirical example. We did not explore the polytomous response condition because the evaluation of the alternative RMSD statistics could be confounded by the complexity of the simulation design, given that the proposed methods are newly developed. However, because it is common to include mixed-format tests, a future study should explore both dichotomous and polytomous item response data in the simulation conditions and compare the performance of the alternative RMSD methods. Finally, we used the cutoff value of .12 to identify DIF items for all RMSD statistics in the simulation study. This cutoff value is practically accepted in

operational settings, and research has shown its effectiveness (Buchholz & Hartig, 2019; Joo et al., 2021). However, this cutoff value was established based on the empirical evaluation and no theoretical support has been provided. In addition, previous studies have suggested that a fixed RMSD cutoff might be unrealistic (Köhler et al., 2020; Robitzsch, 2022). Although we used the cutoff of .12 to be practical and consistent with the operational scaling procedure, it is worthwhile to evaluate the effectiveness of various cutoff values for the alternative RMSD statistics. Furthermore, alternative DIF detection methods other than the RMSD statistic can also be investigated in the context of international large-scale assessments (e.g., Joo & Lee, 2022; Joo, Lee, & Stark, 2022).

Nonetheless, the current study provides important evidence that relatively simple adjustments to the RMSD statistic can improve the ability to detect departures from MI in international large-scale assessment settings such as PISA and PIAAC. Importantly, as newcomers to PISA tend to be lower performers, developing effective methods for detecting measurement differences, especially among low performers, is especially relevant.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Addison-Wesley.

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.

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement*, *31*, 295–311.

Fletcher, R., & Reeves, C. M. (1964). Function minimization by conjugate gradients. *The Computer Journal*, *7*, 149–154.

Joo, S., Ali, U., Robin, F., & Shin, H. J. (2022). Impact of differential item functioning on group score reporting in the context of large-scale assessments. *Large-Scale Assessments in Education*, *10*, 1–21.

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.

Joo, S., & Lee, P. (2022). Detecting differential item functioning using posterior predictive model checking: A comparison of discrepancy statistics. *Journal of Educational Measurement*, *59*, 442–469.

Joo, S., Lee, P., & Stark, S. (2022). Bayesian approaches for detecting differential item functioning using the generalized graded unfolding model. *Applied Psychological Measurement*, *46*, 98–115.

Köhler, C., Robitzsch, A., Fährmann, K., von Davier, M., & Hartig, J. (2021). A semiparametric approach for item response function estimation to detect item misfit. *British Journal of Mathematical and Statistical Psychology*, *74*, 157–175.

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.

Lopez-Rivas, G., Stark, S., & Chernyshenko, O. (2008). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement*, *33*, 251–265.

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

Organization for Economic Cooperation and Development. (2016). PISA 2015 technical report. http://www.oecd.org/pisa/data/2015-technical-report

Organization for Economic Cooperation and Development. (2019). PISA 2018 technical report. http://www.oecd.org/pisa/data/2018-technical-report

R Core Team. (2020). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing.

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

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, D., & Rutkowski, L. (2021). Running the wrong race? The case of PISA for development. *Comparative Education Review*, *65*, 147–165.

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.

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.

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.

Tijmstra, J., Bolsinova, M., Liaw, Y. L., Rutkowski, L., & Rutkowski, D. (2020). Sensitivity of the RMSD for detecting item-level misfit in low-performing countries. *Journal of Educational Measurement*, *57*, 566–583.

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

von Davier, M., & Bezirhan, U. (2021). A robust method for detecting item misfit in large scale assessments. https://doi.org/10.31234/osf.io/mnsdg

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.

Wang, M., & Woods, C. M. (2017). Anchor selection using the Wald test anchor-all-test-all procedure. *Applied Psychological Measurement*, *41*, 17–29.

Wang, W. C., & Yeh, Y. L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*, 479–498.

Yamamoto, K., Khorramdel, L., & von Davier, M. (2013). Scaling PIAAC cognitive data. *Technical report of the survey of adult skills (PIAAC)*, OECD. http://www.oecd.org/skills/piaac/Technical_Report_2nd_Edition_Chapters_17-23.pdf

Authors

SEAN JOO is an assistant professor in the Department of Educational Psychology at the University of Kansas, 1122 West Campus Rd., Lawrence, KS, 66045, USA; e-mail: sjoo@ku.edu. His research interests are item response theory, multilevel modeling, and international large-scale assessments.

MONTSERRAT VALDIVIA is a doctoral student in the Department of Counseling and Educational Psychology at the Indiana University Bloomington, 201 N. Rose Ave., Bloomington, IN, 47405, USA; e-mail: mbvaldiv@iu.edu. Her research interests are multistage testing, differential item functioning, and international large-scale assessments.

DUBRAVKA SVETINA VALDIVIA is an associate professor in the Department of Counseling and Educational Psychology at the Indiana University Bloomington, 201 N. Rose Ave., Bloomington, IN, 47405, USA; e-mail: dsvetina@iu.edu. Her research interests are educational and psychological measurement, item response theory, and multistage testing.

LESLIE RUTKOWSKI is a professor in the Department of Counseling and Educational Psychology at the Indiana University Bloomington, 201 N. Rose Ave., Bloomington, IN, 47405, USA; e-mail: lrutkows@iu.edu. Her research interests are quantitative methods, large-scale assessments, and latent variable modeling.

Manuscript received October 5, 2021
Revision received May 8, 2023
Accepted May 30, 2023