JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education
DOI: 10.1111/jedm.12263 *Journal of Educational Measurement*

Sensitivity of the RMSD for Detecting Item-Level Misfit in Low-Performing Countries

Jesper Tijmstra
Tilburg University
Maria Bolsinova
ACTNext
Yuan-Ling Liaw
IEA Hamburg
Leslie Rutkowski and David Rutkowski
Indiana University

Although the root-mean squared deviation (RMSD) is a popular statistical measure for evaluating country-specific item-level misfit (i.e., differential item functioning [DIF]) in international large-scale assessment, this paper shows that its sensitivity to detect misfit may depend strongly on the proficiency distribution of the considered countries. Specifically, items for which most respondents in a country have a very low (or high) probability of providing a correct answer will rarely be flagged by the RMSD as showing misfit, even if very strong DIF is present. With many international large-scale assessment initiatives moving toward covering a more heterogeneous group of countries, this raises issues for the ability of the RMSD to detect item-level misfit, especially in low-performing countries that are not well-aligned with the overall difficulty level of the test. This may put one at risk of incorrectly assuming measurement invariance to hold, and may also inflate estimated between-country difference in proficiency. The degree to which the RMSD is able to detect DIF in low-performing countries is studied using both an empirical example from PISA 2015 and a simulation study.

Introduction

In the context of international large-scale assessment, the issue of establishing measurement invariance across countries is of great importance, as valid between-country comparisons can only be made if the latent variable is measured equivalently in each considered country. If measurement invariance is violated, the existence of *bias* risks advantaging or disadvantaging one or more countries. In order for measurement invariance to hold, the items should not display any form of differential item functioning (DIF), meaning that the relationship between the trait or ability to be measured and the probability of getting a particular score on the item should be the same for all countries. Various DIF detection methods exist in the classical measurement and item response theory (IRT) literature. Examples include Lord's chi-square (Lord, 1980), a likelihood ratio method (Thissen, Steinberg, & Wainer, 1993), the Mantel-Haenszel chi-square statistic (Holland & Thayer, 2013), and logistic mixed models (Van den Noortgate & De Boeck, 2015), among others. In an IRT framework, items are said to exhibit DIF if the item characteristic curve (ICC) varies across groups (Embretson & Reise, 2000), typically, either in location,

discrimination, or both. Uniform bias occurs where location-based DIF exists, while nonuniform bias involves group differences in the discrimination parameter (Mellenbergh, 1982). In either case, for a given level of the latent variable, the probability of a particular (e.g., behavior or attitude items) or correct (e.g., achievement item) answer will differ across groups. Such item parameter differences raise questions about test fairness and the degree to which test or other instrument results can be compared across groups of interest. With measurement invariance playing such a crucial role for obtaining valid inferences about individuals and groups, the importance of critically checking the performance of each item in each group is difficult to overstate. An insightful example of how one could proceed to perform these checks and assess the practical impact of possible violations is provided by van Rijn, Sinharay, Haberman, and Johnson (2016).

Although most DIF methods are developed for and demonstrated with a single reference and focal group, their application to a larger numbers of groups is increasingly common. For example, the OECD's Programme for International Student Achievement (PISA) includes dozens of highly heterogeneous countries that differ in language, culture, geography, and economic development. This degree of heterogeneity across groups poses inherent challenges for comparable measurement. Efficient evaluation of possible DIF for each item and for each country becomes especially important in this context, and in this sense large-scale educational assessment can be said to pose important challenges for item and model fit assessment methods.

One relatively new method for DIF detection is the root-mean squared deviation (RMSD; von Davier, 2017), which has also been considered under the label RMSEA (root-mean squared error of approximation) (Kunina-Habenicht, Rupp, & Wilhelm, 2009). Briefly, this measure, based on classic methods of model evaluation that rely on average squared errors, quantifies the weighted distance between the model-based and empirical ICCs and is sensitive to both location and discrimination differences. Although the RMSD has a small, applied literature (e.g., Kunina-Habenicht et al., 2009; Oliveri & von Davier, 2011) and it is the dominant method for DIF detection in PISA (OECD, 2017b) and is also used in the Programme for the International Assessment of Adult Competencies (Yamamoto, Khorramdel, & von Davier, 2013), research that evaluates the performance of this measure is limited. One recent exception involves research that examines the general behavior of the RMSD under a generalized partial credit model (GPCM; Buchholz & Hartig, 2017). The authors offer general cutoff guidelines; however, as we describe subsequently, important questions remain. In response, we take both an analytical and simulation-based perspective to query the degree to which the RMSD is well-suited for detecting group-specific departures from common item parameters. With an emphasis on international assessments like PISA, we especially focus on the RMSD performance for low-performing countries (i.e., countries whose proficiency distribution is located on the low end of the considered scale). Our emphasis on the low end of the proficiency scale reflects the fact that international assessments in general, and PISA in particular, have experienced meaningful growth in participation, covering a diverse set of countries. For example, PISA participation has extended well beyond the group of economically developed countries for which it was originally designed, with the latest completed round in 2015 comprising 72 educational systems.[1] Furthermore, newcomers

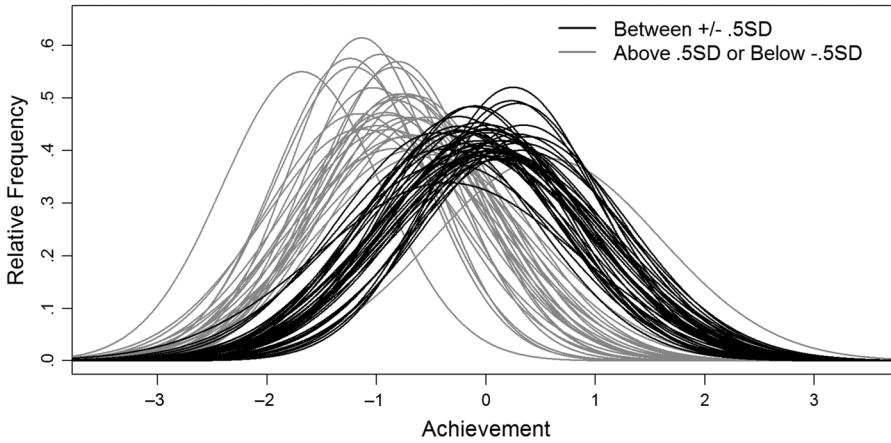

Figure 1. Estimated proficiency distributions of each participating country for the PISA 2015 Science scale.

to PISA tend to come from less economically developed countries, which, historically, tend to be low performers (OECD, 2017a). Although we situate our research in the context of international assessments and will use PISA as an illustrative example, the measure and our investigation of it are general enough to apply broadly to DIF detection in international large-scale assessment in general.

To provide some context for the problem at hand, Figure 1 illustrates the PISA Science proficiency distribution of the different participating countries for the 2015 cycle. Here, the zero point on the ordinate axis matches the historic Science mean and the unit of the scale is the historic standard deviation. The dark curves represent countries whose proficiency mean is within $\pm.5$ standard deviations from the historic mean. Lighter curves represent countries that are below .5 standard deviations. At the extreme, the mean of the lowest performer (the Dominican Republic) is more than 1.5 standard deviations from the historic math mean. With a spread of more than two logits between the mean of the highest and lowest performers, large differences in average proficiency levels are apparent (OECD, 2017b).

Country heterogeneity manifests itself in ways that extend beyond differences in proficiency level. Economic, language, cultural, and geographic differences are striking. In 2015, each populated continent was represented. Furthermore, many countries tested in multiple languages (e.g., Canada in French and English; Switzerland in Germany, French, and Italian) and cross-country dialect differences necessitated country-specific adaptations (e.g., Israel's Arabic version differed from the Qatar's Arabic version). Although most countries adopted a computer-based platform, 16 countries opted for a paper-based assessment. These different forms of heterogeneity may increase the risk of items showing DIF across the participating countries, further increasing the importance of being able to adequately detect such violations of measurement invariance.

Starting with the 2015 PISA cycle, DIF measures were based on the difference between the empirical and expected ICCs via the mean deviation (MD) and the

RMSD. Specifically, items were flagged as differentially functioning if $|MD| > .12$ or $RMSD > .12$ (OECD, 2017c, p. 151). As noted elsewhere (von Davier, 2017), the MD is sensitive to differences in item location between the international sample and the considered country while the RMSD is sensitive to both location and discrimination. Given that the two-parameter logistic model and the GPCM are used in PISA and other international assessments, we limit our investigation to the RMSD. When considering dichotomously scored items, the RMSD is designed to capture differences between the expected probability of giving a correct or positive answer (i.e., based on the item response function as estimated in the international sample that consist of all countries) and the observed probability of giving a correct or positive answer. Formally, it is defined as

$$RMSD = \sqrt{\int (P_o(\theta) - P_e(\theta))^2 f(\theta) d\theta}, \tag{1}$$

where θ denotes the latent variable measured by the scale, $f(\theta)$ denotes the density function of θ in the focal group (i.e., the country that is considered), and $P_e(\theta)$ and $P_o(\theta)$ denote the model-based (i.e., expected) and empirical (i.e., observed) probability of obtaining a correct response given θ, respectively.[2] If the model is correct, with increasing sample size the empirical probability will converge on the model-based probability for each θ and the RMSD will tend to 0, with an RMSD of 0 indicating a perfect overlap between the two probabilities and hence the absence of DIF.

As will be further explicated in the next section, the fact that the RMSD considers differences in item response probabilities means that its value will not only depend on the degree to which the curves capturing the expected and empirical response probabilities differ in location, but also on the proficiency distribution of the country or group that is considered. That is, if an item shows the same degree of uniform DIF in two countries while these countries differ in their proficiency distribution, their RMSD values for that item can be expected to differ. Consequently, it is important to address the question to what extent the RMSD is influenced by differences in the proficiency distribution of the countries or groups that are considered.

This paper systematically assesses the performance of the RMSD and its usefulness for detecting uniform DIF for dichotomous items when faced with groups that have heterogeneous proficiency distributions. Specifically, we examine whether and in which situations it can constitute a useful measure for DIF detection in international large-scale assessment. This paper focuses specifically on the application of the RMSD to dichotomously scored items, as heterogeneity in countries' proficiency distributions can be expected to be most impactful for dichotomous items, since these are characterized by a single ICC whose location may differ notably from the mean of the ability distribution in the considered country. Since polytomous items are characterized by multiple response curves (matching each of the categories), it can be expected that a strong mismatch between the location of the curves of the item and the mean of the focal group's proficiency distribution will be much more rare.

The remainder of the paper is organized as follows. The next section presents a technical discussion of the RMSD and its properties. In the subsequent section, an empirical example using the PISA 2015 Science scale is considered, where the

RMSD is used to evaluate DIF for specific countries. The following section presents a simulation study, which systematically considers the impact that a mismatch between the location of the item and of the proficiency distribution has on the ability of the RMSD to detect different degrees of uniform DIF.

Properties of the RMSD

The RMSD is an item-fit measure designed to capture discrepancies between the model-based and the empirical item-response probabilities, similar in spirit to other RMSEA measures for determining model fit. As Equation 1 shows, for dichotomously scored items the RMSD provides a quantification of the discrepancy between the model-based ICC ($P_e(\theta)$) and an approximation of the empirically observed ICC ($P_o(\theta)$), weighted using the density function of θ in the (sub)population that is considered (OECD, 2017c). $P_e(\theta)$ is obtained by using the estimates of the item parameters (i.e., it is the estimated item response function in, for example, the international sample), and $f(\theta)$ is estimated as part of the analysis. Since $P_o(\theta)$ denotes the empirically observed ICC, if one makes use of EM-estimation procedures it can be specified as the obtained observed ICC that is based on the pseudo-counts in the E-step of the algorithm (OECD, 2017c; von Davier, 2017). The integral in Equation 1 can be approximated using quadrature, such that if Q quadrature points $(\theta_1, \ldots, \theta_Q)$ are used the RMSD is approximated through

$$\hat{RMSD} = \sqrt{\sum_{q=1}^{Q} (P_o(\theta_q) - P_e(\theta_q))^2 \Pr(\theta = \theta_q \mid \theta \in \{\theta_1, \ldots, \theta_Q\})}, \tag{2}$$

where

$$\Pr(\theta = \theta_q \mid \theta \in \{\theta_1, \ldots, \theta_Q\}) = \frac{\mathcal{N}(\theta_q \mid \hat{\mu}, \hat{\sigma}^2)}{\sum_{s=1}^{Q} \mathcal{N}(\theta_s \mid \hat{\mu}, \hat{\sigma}^2)}, \tag{3}$$

and where

$$P_o(\theta_q) = \sum_p x_{pi} \Pr(\theta = \theta_q \mid \theta \in \{\theta_1, \ldots, \theta_Q\}, \mathbf{X} = \mathbf{x}_p), \tag{4}$$

where $\mathbf{X}$ is a response vector with realization $\mathbf{x}_p$ for person p, where x_{pi} is the item score of person p on item i, and where

$$\begin{aligned} &\Pr(\theta = \theta_q \mid \theta \in \{\theta_1, \ldots, \theta_Q\}, \mathbf{X} = \mathbf{x}_p) \\ &\quad = \frac{\Pr(\theta = \theta_q \mid \theta \in \{\theta_1, \ldots, \theta_Q\}) \prod_i \Pr(X_i = x_{pi} \mid \theta = \theta_q, \hat{\boldsymbol{\gamma}}_i)}{\sum_{s=1}^{Q} \Pr(\theta = \theta_s \mid \theta \in \{\theta_1, \ldots, \theta_Q\}) \prod_i \Pr(X_i = x_{pi} \mid \theta = \theta_s, \hat{\boldsymbol{\gamma}}_i)}. \end{aligned} \tag{5}$$

Here, $\hat{\boldsymbol{\gamma}}_i$ denotes a vector containing the item parameter values that are used to specify the model-based ICC, where in the current context these values are fixed to the maximum likelihood estimates obtained for the international sample (i.e., assuming the model-based ICC to be the same for all considered countries).

As a formal quantification of the degree of misfit, it has several advantages that are worth mentioning. First of all, it allows one to obtain an objective measure of the degree of observed misfit, which may hold an arguable advantage over the more subjective process of "eyeballing" the differences between the two curves. Second of all, the measure is sensitive to both a mismatch in the location of the two curves (i.e., a discrepancy in the item difficulty matching the two curves) and a mismatch in the slope of the two curves (i.e., differences in the item discrimination matching the two curves). A third advantage of the measure is that by taking the square of the discrepancy, negative and positive discrepancies observed over the range of the latent variable do not cancel each other out: A value of 0 is only obtained if the two curves overlap perfectly, and conversely, if the curves overlap perfectly then the estimated value of the RMSD will tend to 0 as sample size increases. Fourth, the value of the measure does not depend on sample size. As such, substantively irrelevant misfit due to very large samples will not be flagged. Finally, the formulation of the RMSD is general, and is therefore not tied to the choice of a particular IRT model. This means that one can apply it regardless of which specific IRT model one wants to employ.

Given these advantages of the RMSD, considering its behavior in further detail may be worthwhile to assess whether it is an adequate tool for determining whether an item shows misfit in a particular group or country. When using the RMSD, the quantification of misfit depends on the average absolute difference between persons' observed and expected probability of passing the item. This means that for an item to be flagged as misfitting, the expected and empirical probability of passing the item need to show notable differences on average. For example, if one uses the suggested cutoff value of .12 (OECD, 2017b, p. 151), only sufficiently large absolute differences in probabilities would cause an item to be flagged, since an item where for most persons the absolute difference in probabilities does not exceed .12 would likely not get flagged.

It is important to note that Equation 1 shows us that the RMSD of an item (and hence whether an item gets flagged) does not solely depend on the two curves that are compared, but that it is also population-dependent: For a given $P_e(\theta)$ and $P_o(\theta)$, the RMSD will change depending on the proficiency distribution that is considered. This can be illustrated by examining a Rasch item with a difficulty parameter of 0 that shows a location shift of 1 point in the focal group (i.e., the item actually has a difficulty of 1 for that group, while the model assumes the difficulty to be 0). While the shift in location impacts all persons in that group in the same way (i.e., their expected performance on the item matches that of a person in the international population whose score on the latent variable is one point lower), the difference between the two probabilities depends heavily on their value on the latent variable: A person with a value of 0 for θ will have a difference of $-.27$ (i.e., $\frac{e^{-1}}{1+e^{-1}} - \frac{e^{0}}{1+e^{0}}$), while a person with a value of -2 for θ will have a difference in probabilities of only $-.07$ (i.e., $\frac{e^{-3}}{1+e^{-3}} - \frac{e^{-2}}{1+e^{-2}}$). Since the RMSD depends on the average absolute difference between these two probabilities, the RMSD for this item will change if we consider a different population distribution.

Concretely, this has two important implications: (1) The RMSD does not quantify item misfit in isolation from the focal group's population distribution of the latent

variable, and (2) using fixed RMSD values implies that assigning a misfit flag to an item depends on the extent to which that item's location is aligned with the focal group's proficiency distribution. A consequence of (1) is that for a given empirical and observed ICC there is no unique value for the RMSD, and hence the RMSD does not provide an unconditional quantification of the amount of misfit between two curves. Thus, there is no one-to-one mapping between DIF magnitude and RMSD value. Furthermore, as we show subsequently, RMSD values can depend on whether DIF is positive or negative. Consequently one cannot interpret the RMSD directly as a measure of item misfit.

Implication (2) follows from the fact that if the proficiency distribution is located far away from the international item location, most persons will have an expected probability close to 0 or 1. Then, even for very large differences between the international and country-specific difficulty parameters, only small differences in probabilities will exist, leading to small RMSD values. Consequently, using fixed thresholds for the RMSD for detecting misfit in different groups can be expected to result in a lower sensitivity for detecting misfit in those groups for which the considered item is very difficult (or very easy) than in those groups whose overall proficiency level matches the difficulty of the item. This also means that using a fixed RMSD cutoff value could lead to notably different conclusions about whether DIF is present, depending on whether the focal group's proficiency distribution is located near to or far from the estimated international item location.

A corollary of (2) is that shifts in item location of a certain magnitude may be easier to detect if the shift is in one direction (i.e., moving toward the mean of the focal group's proficiency distribution) and more difficult to detect in the other direction (i.e., moving away from that mean). This can be observed by considering a person with a θ value of -1 who is presented with an item that the model assumes has a difficulty parameter of 1. If the actual difficulty is one point lower (i.e., 0), we see a change in probability of .15, while if it is one point higher than the model indicates (i.e., 2), the absolute change is only .07. In general, this means that for low-performing countries, if there is an equal amount of positive and negative location shifts, one can expect to detect more of the negative shifts (i.e., the item having a lower-than-expected difficulty) than of the positive shifts. If by and large only negative shifts are picked up and corrected for while positive shifts are not detected, one can expect this to result in negative bias for the estimated proficiency mean of that group (i.e., resulting in a mean estimate for that country that is too low). This would lead to an inflation of the estimated differences in proficiency between high- and low-performing countries, since this negative bias for the estimated mean would not be expected for countries whose proficiency distribution is aligned with the difficulty of the test.

Empirical Example: PISA 2015 Science

Starting with the 2015 PISA cycle, the RMSD has been used as a tool for detecting country-specific item-level misfit (OECD, 2017c). We will consider its outcomes in the context of the 2015 Science scale. In PISA 2015, the GPCM was used for item

calibration. For complete details on item parameter estimation, please see the PISA technical report (OECD, 2017c).

To evaluate the behavior of the RMSD in practice, it may be helpful to consider the RMSD values obtained for countries that differ notably in their proficiency distribution. The population distribution with the lowest average performance belongs to the Dominican Republic, with an estimated proficiency mean of 332 (OECD, 2017a), where the scale is defined by the historic mean of 500 and standard deviation of 100. Consequently, this is the country for which the average probability of providing correct answers is close to zero for many items, since the item locations in the Science domain had a mean of .045 and a standard deviation of .562 (OECD, 2017a), where 0 and 1 are the historic, untransformed mean and standard deviation, respectively. Therefore, the test can be considered relatively misaligned for the Dominican Republic compared to countries that have a proficiency distribution closer to the "baseline" point of the scale. In contrast, with a mean of 496 the proficiency distribution of the United States seems well-aligned with the average location of the test items.

Based on the suspected property of the RMSD to be less sensitive to detecting item-level misfit when the item difficulty and proficiency distribution are far apart, one could expect that detecting misfit for items on this test should generally be easier for the United States compared to the Dominican Republic, as the Dominican Republic is a relative outlier in terms of proficiency. It is also an outlier in other demographic areas. For example, the Dominican Republic reported: (1) a 2016 per capita gross domestic product of 6,793 USD (World Bank, 2018a) compared to the OECD average of 42,183 USD (OECD, 2018a); (2) a 2016 life expectancy of 73.86 years (World Bank, 2018b), compared to an OECD average of 80.84 years (OECD, 2018b); and (3) a secondary school enrollment rate of 77%, compared to 97% in the United States (World Bank, 2018c). In light of these meaningful differences, it might also be realistic to expect a larger proportion of the items on the test to behave differently (i.e., have DIF compared to the behavior of the item in the international sample) in the Dominican Republic than in the United States. That is, while one might expect the number of false negatives to be larger in the Dominican Republic than in the United States, it may also be expected that there are a larger number of items with DIF for the former country than for the latter. Thus, simply comparing the number of flagged items in each country does not directly inform us whether there are issues with using the RMSD for DIF detection in heterogeneous populations, since there is no guarantee that in each country a similar number of items show DIF. Hence, having a similar number of flagged items does not provide evidence of the RMSD functioning as intended. It is however still relevant to consider the distribution of RMSD values in these two countries to gain some insight into its behavior in different countries, and to consider concrete items to investigate which degree of observed misfit can match a particular RMSD value in countries with proficiency distributions similar to these two.

To investigate the behavior of RMSD for detecting misfit in the Dominican Republic compared to the United States, the following analyses were performed: The GPCM was applied to the Science items data from the Dominican Republic and the United States, with item parameters fixed at the estimated values that were published in the PISA technical report (see OECD, 2017a). The model was fitted with

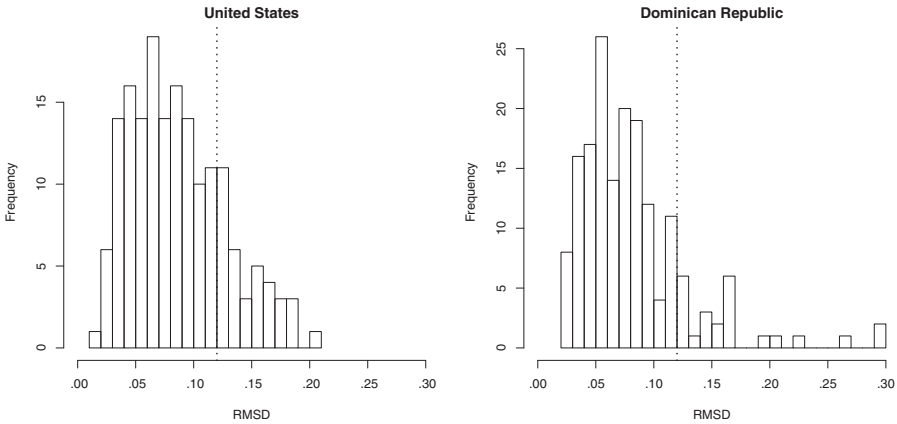

Figure 2. Empirical distribution of the RMSD for the 171 binary items on the PISA 2015 Science scale, for the Dominican Republic and for the United States. The dotted line indicates the cutoff value of .12 used to flag DIF-items.

marginal maximum likelihood assuming a normal distribution of ability in each country. The analysis was performed using the R-package TAM (version 2.8-21; Robitzsch, Kiefer, & Wu, 2018). After fitting the model, the RMSD for each item was computed for the two countries separately. The means of ability were estimated to be .005 and -1.267 in the United States and the Dominican Republic, respectively. The standard deviations were estimated to be .848 and .584, respectively.

Figure 2 displays the distribution of the RMSD values obtained for the dichotomous items on the Science scale for both the Dominican Republic and the United States. The commonly considered cutoff value of .12 is indicated using a vertical dotted line. Using this cutoff value, 24 items are classified as showing misfit for the Dominican Republic while 36 items are flagged for the United States. Taken at face value, these results suggest that DIF may be slightly more of an issue for the United States, at least in terms of the number of compromised items. However, as indicated in the sections above, it may be that the RMSD has higher power to detect DIF for countries whose ability distribution is more aligned with the difficulty level of the items on the test. Hence, it could be that this difference between the two countries in terms of the number of flagged items is partly due to differences in the proficiency distribution rather than there actually being fewer problematic items for the Dominican Republic than for the United States. That is, it could be that a particular item has DIF of the same magnitude in both countries (e.g., the item-response function is the same in both countries but differs from that in the international sample), but that in using the RMSD this DIF is more easily detected for the United States. This could explain why despite the fact that PISA was originally designed for OECD countries (including the United States), the Dominican Republic counted fewer flagged items on this test than the United States.

The fact that the RMSD depends not only on the discrepancy between the two curves but also on the proficiency distribution can be illustrated by considering the expected and observed ICCs of item DS649Q02C for these two countries, which

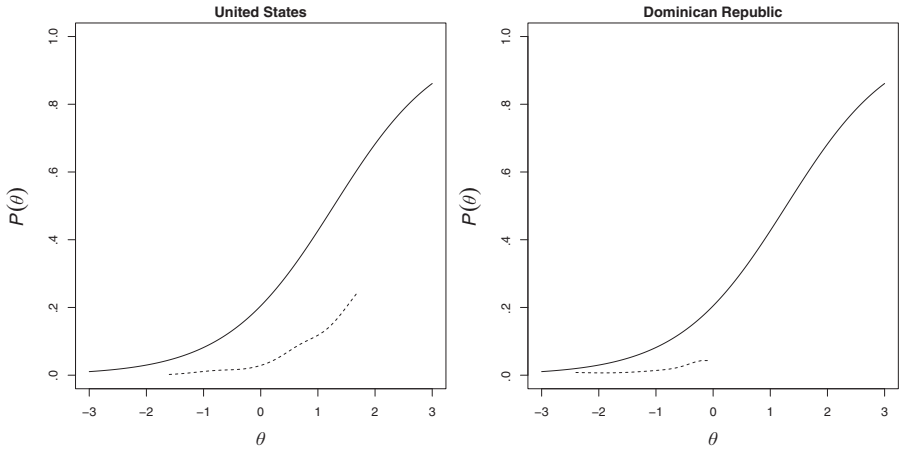

Figure 3. Expected (solid line) and observed (dashed line) item characteristic curves for item DS649Q02C, for the United States and for the Dominican Republic. The observed curve is plotted in the range of $\hat{\mu} \pm 2\hat{\sigma}$ of the country's proficiency distribution.

are displayed in Figure 3. The expected ICC is the estimated international item response function based on the two-parameter logistic model, where the difficulty is estimated to be 1.279 and the item discrimination 1.061. The observed ICC for each country is computed based on pseudo-counts of the E-step of the EM algorithm after convergence (see Equations 2– 5) and plotted in the range of $\hat{\mu} \pm 2\hat{\sigma}$ of the country's proficiency distribution. First of all, it can be observed that this is a challenging item for both countries, but especially for the Dominican Republic, where (based on the international ICC) even respondents located two standard deviations above the country mean have only a small chance (18.9%) to answer the item correctly. For this item, an RMSD of .205 is obtained for the United States, while a value of .072 is obtained for the Dominican Republic. Although the RMSD may seem to suggest stronger misfit for the United States than for the Dominican Republic, the observed ICCs seem to be comparable for the two countries. Thus, this exemplifies the issue that whether the RMSD flags an item as showing misfit does not only depend on the extent to which the model-based and empirical ICC of an item differ (i.e., the degree to which DIF can be said to be present for the item), but also on the proficiency distribution of the country for which those curves are contrasted. This suggests that comparing for each country the number of items flagged by the RMSD as showing DIF will not provide a good indication of the extent to which measurement invariance holds in the different countries, since for some countries one can expect a larger proportion of false negatives (i.e., nonflagged DIF items).

Simulation Study

Method

To more systematically study the impact that the item location and degree of item misfit have on the sensitivity of the RMSD to detect item misfit in low-proficiency

countries, a simulation study was conducted. To ensure that the results of the simulation study have practical relevance and empirical validity, for this study we used the empirical results obtained for the PISA 2015 Science scale to specify the parameter values for the generating model. Both the sample size (4740) of the Dominican Republic, the distribution of these respondents over the different booklets of the Science test, and the missing value patterns were kept identical to that of the PISA 2015 cycle.

For the ability distribution, the estimated proficiency distribution of the Dominican Republic was used. Ability parameters for 4,740 respondents were generated from a normal distribution with a mean of -1.267 and a standard deviation of .584. For the items, the reported GPCM estimates (OECD, 2017b) of the item parameters of the 184 items on the Science scale were used, with the exception of one item that served as the focal item to which DIF is introduced and for which the RMSD is studied.

As the focal item for which the sensitivity of the RMSD is evaluated, we took the item presented earlier (i.e., DS649Q02C) as our starting point (estimated difficulty of 1.279 and discrimination of 1.061). This item was picked as the starting point because it represents the type of items for which one would like to critically assess the performance of the RMSD in low-performing countries for: it is of a difficulty level that is both realistic in the context of standard large-scale assessment (difficult, but not unreasonably difficult for groups of average proficiency level), and at the same time it is an item that few respondents in a low-performing country would be able to answer correctly. Hence, it captures a realistic scenario where one could expect the RMSD to run into issues with DIF detection.

We introduced uniform DIF to the focal item for the considered population by shifting the generating ICC away from the expected ICC (i.e., the curve obtained based on the international sample) in increments of .25. An item location shift (i.e., uniform DIF) ranging from -2 to 2 was considered, such that 17 degrees of DIF were studied. This means that uniform DIF of up to 2 units on the scale was considered. Since absolute DIF of magnitudes above .5 standard deviations on the proficiency scale (roughly matching .5 points on the considered scale) is generally considered to be an issue, most of the DIF conditions considered in the simulation represent scenarios in which it would be highly problematic to fail to detect this DIF. While uniform DIF of absolute magnitude above 1.0 would not be common in practice, we included these extreme conditions to provide a more complete illustration of the behavior of the RMSD in all possible scenarios.

In addition to varying the degrees of DIF for the item, the location of the item in the international population was also varied to study the behavior of the RMSD for different expected item difficulties. This was done to study what would have happened if the discrepancy between the population mean of the Dominican Republic and the location of the item in the international sample would have been larger or smaller than it was for the actual item, and hence to allow for more general conclusions about the behavior of the RMSD. To capture a broad range of item difficulties, we considered 17 equidistantly spaced item difficulties ranging from -1.976 (i.e., 3.25 points lower than the actual item difficulty, representing a very easy item) to 2.024 (i.e., .75 points higher than the actual item difficulty, representing a very

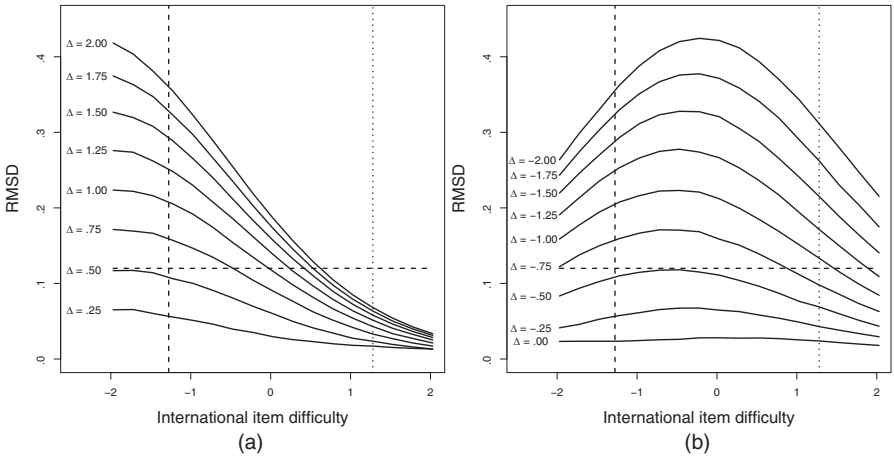

Figure 4. Results for the conditions with (a) positive DIF and (b) negative DIF. The dashed vertical line indicates the location of the mean of the Dominican Republic, the dotted vertical line indicates the international item difficulty of the item DS649Q02C, and the dashed horizontal line matches the commonly considered cutoff value of the RMSD of .12. Each curve represents the average RMSD value for a given DIF magnitude and direction (based on 500 replications).

difficult item), to ensure that item difficulties were considered that roughly matches the range of item difficulties of the PISA Science items.

With 17 degrees of DIF and 17 different international item difficulties, a total of 289 conditions were considered. In each condition, 500 data sets were generated. For each data set, the value of the RMSD was estimated, using the specified item response function of the international population as the expected ICC and estimating the observed ICC based on the generated data. To simplify the analyses, all persons received equal weight, and hence—in contrast to empirical analyses in large-scale assessment with a substantive focus rather than our current methodological focus—no person weights were used in the analyses. The analysis was performed using the R-package TAM (version 2.8-21; Robitzsch et al., 2018).

Results

Figure 4 depicts the simulation results for both positive (Panel a) and negative (Panel b) DIF, with the no-DIF condition included in Panel (b). In both plots, the estimated mean of the country's proficiency distribution (-1.267) is indicated by a dashed vertical line to illustrate how differences between that proficiency mean and the difficulty of the item in the international sample (the latter being a factor that was varied) affect the value of the RMSD obtained for particular degrees of uniform DIF. Each curve in the two plots corresponds to a magnitude of simulated uniform DIF (denoted by Δ) and should be read relative to a chosen point on the horizontal axis that represents a value for the international difficulty parameter. Thus, when considering the situation where $\Delta = 2$, we are considering a situation where the country-specific item difficulty is two points higher than the difficulty in the

international sample. If we consider the value of the difficulty in the international sample that was obtained for the original item (1.279, indicated by the dotted vertical line), this shows us that the expected value of the RMSD is .06 if positive DIF of magnitude 2 would be present (top curve). The curve shows us that if the difficulty of the item is more aligned with the population proficiency mean, higher values of the RMSD will be obtained. Consequently, whether positive DIF of a certain magnitude can be expected to result in a flag depends quite strongly on how far removed the item location is from the population's proficiency mean. The unfortunate conclusion is that for items that are at least somewhat difficult (international item location above .75), the RMSD cannot be expected to flag the item as misfitting for this population, even if the item difficulty in the country is two points higher than in the international population. In contrast, if the item is expected to be easy (international item location below $-.75$), positive DIF of magnitude .75 or above can realistically be detected for this country by the RMSD. For items that according to the international analyses are of average difficulty (i.e., 0), picking up positive DIF of magnitude 1 is already uncertain, since the expected RMSD is approximately .12 in that case. Thus, the sensitivity of the RMSD to detect positive DIF depends strongly on the location of the item in the international sample. It is also clear that regardless of item difficulty, DIF of magnitude .50 or less is unlikely to result in a flag for the item if the .12 cutoff value is used for the RMSD. It may also be noted that in the case of positive DIF the highest values for the RMSD are not actually obtained exactly at the point where the item difficulty in the international population matches the mean proficiency of the considered country, but rather when the international difficulty is somewhat lower than that mean. This can be explained by considering that the RMSD contrasts the international ICC with that of the country, and that the discrepancy between these two curves is maximized when the international and country-specific item difficulty are equally far removed from the country proficiency mean.

Panel (b) in Figure 4 displays the results that were obtained for the negative DIF conditions, and can be read similar to Panel (a).[3] Since this panel concerns negative DIF, situations are considered where the item is easier in the Dominican Republic than in the international population. If we again consider the original value of the international difficulty parameter (1.279), we can observe that much higher values are obtained for the RMSD than was the case when positive DIF was present. For DIF of magnitude -2, an RMSD of .30 can be expected, and even DIF of magnitude -1 can be expected to result in a flag, since the average RMSD exceeds .12 in that condition. If the item had a lower difficulty in the international population, one would have obtained even higher values for the RMSD for DIF of these magnitudes. For international item difficulties ranging from roughly -2 to 1, DIF of magnitude $-.75$ can already be expected to result in a flag, since the average RMSD exceeds .12 in those conditions. Thus, since most items on the test will have a difficulty falling somewhere in that range, one can expect the RMSD to be able to detect DIF of magnitude $-.75$ for most items on the test. This is in contrast to what was observed for DIF of magnitude .75, for which the average RMSD only exceeded .12 for items with a difficulty lower than $-.5$. These findings strongly suggest that for countries whose proficiency mean is not aligned with the difficulty of an item, the RMSD will not be equally sensitive to positive and negative DIF, but rather is much more likely

to pick up on one of the two: negative DIF if the item is difficult and the country has a low proficiency level, and positive DIF if the item is easy while the country has a high proficiency level. Alarmingly, this may mean that depending on the proficiency level of the country, by and large only negative (or only positive) DIF is likely to be detected for most of the considered items.

Discussion

The RMSD is a measure of country-specific item-level misfit that focuses on contrasting observed and expected item-response probabilities. As such, it is most suitable for detecting misfit in cases where the presence of DIF can be expected to result in a notable shift in probabilities for most persons in the population that is considered. This situation is realized when the location of the item is close to the mean of the proficiency distribution of the population, that is, if the average person in that country has about a 50% chance to provide the correct answer. In that case, item location shifts slightly above .5 points (in either direction) can be detected if a cutoff value of .12 is used for the RMSD, and in this sense the RMSD can be said to be helpful in detecting misfit. However, even in these optimal conditions, a shift of .5 would often not be detected, suggesting that when using a cutoff value of .12 for the RMSD one can expect DIF of moderate size to often remain undetected even under optimal conditions.

The RMSD becomes much less sensitive to uniform DIF if the location of the item is not close to the mean of the population that is considered. If the item is relatively difficult (or easy) for the considered population, we can expect many respondents to have a probability of correctly answering the item that is close to zero (or to one), in which case even a notable amount of uniform DIF does not affect the item probability heavily. That is, due to the logistic shape of the ICC, uniform DIF results in a larger shift in probability of passing the item for respondents located close to the item location than for those far removed from it. Consequently, for a given degree of uniform DIF the value of the RMSD depends on the proficiency distribution of the considered population, and the RMSD will be less sensitive to detect misfit for items that are not "aligned" in difficulty with the proficiency distribution of the considered population. Thus, even if an item has the same amount of DIF in two countries, the item is less likely to be flagged as having misfit in the country whose proficiency distribution is further removed from the item location. Consequently, the risk of not detecting misfit for items on the test is notably larger for countries whose population distribution is far removed from the location of most of the items on the test. The RMSD therefore does not seem ideal for detecting misfit when a notable discrepancy between the proficiency level of a country and the overall difficulty of the considered test is present, which is a condition that may be realized when including heterogeneous populations in international large-scale assessments.

An additional problem that arises when using the RMSD to detect misfit is that the RMSD may not be equally sensitive to positive and negative shifts in item difficulty, as was illustrated in the simulation study. If a country is notably less proficient than the item is difficult, uniform positive DIF is much more difficult to detect than uniform negative DIF, because in the latter case the appropriate ICC for that country is

located closer to the mean proficiency of that country. As a consequence, if positive and negative DIF occurs to an equal extent across items in the test, the RMSD is likely to flag more of the items that show negative DIF than those that show positive DIF. This unequal flagging of positive and negative DIF items can be a source of confounding: If one deals with this signalled misfit by allowing the item response functions of the flagged items to be estimated freely while keeping the function of the nonflagged items fixed to that in the international sample, this unbalanced correction can be expected to result in bias for the estimated mean proficiency of that country.[4] That is, by correcting for negative DIF (i.e., an item being easier than expected) while often ignoring positive DIF (due to it not being detected), the mean of low-performing countries may be underestimated. For high-performing countries, the opposite effect may occur: Positive DIF would frequently be flagged and corrected for (resulting in an upward adjustment of the estimated mean) while negative DIF would often not be corrected for (resulting in the absence of a downward adjustment of the estimated mean). Thus, the fact that the sensitivity of the RMSD to detect DIF depends on the location of the population distribution may mean that when the RMSD is used as the main tool for detecting and consequently correcting for DIF, differences in proficiency between low- and high-performing countries may be overestimated to some degree. Concretely, this means that there may be a serious risk that the bias that occurs due to missing DIF items—and hence ignoring the item misfit for those items—does not "cancel out" at the test level (which one might hope for if both positive and negative DIF are equally likely to be missed). Hence, it is not just an issue of formal model misfit, but rather this issue can be expected to have relevant practical implications for the model inferences and should not be ignored. A discussion on how to assess the practical implications of ignoring model and item misfit is for example provided by Sinharay and Haberman (2014), by van Rijn et al. (2016), and by Köhler and Hartig (2017).

It should be noted that in this manuscript the only type of misfit that has been studied extensively is uniform DIF. Thus, it has always been assumed that each item's ICC is of the same shape in all countries, but that the item location in a given country can differ from the item location estimated for the international sample. This choice was motivated by the idea that it is specifically uniform DIF that can be expected to threaten between-country comparisons, as unmodeled uniform DIF entails that for every person in a particular country the actual item probability is lower (or higher) than expected, and can therefore arguably be considered to be the most problematic kind of item misfit in international large-scale assessment if left undetected. For nonuniform DIF, the conclusion that DIF becomes more difficult to detect the further population mean is removed from the item difficulty may not hold in general, due to the two curves having different shapes. However, also in the case of nonuniform DIF will the differences between the expected and observed item probability depend on the considered proficiency level, and hence the sensitivity of the RMSD to detecting nonuniform DIF is likely to also depend on the proficiency distribution of the considered population.

It should also be noted that—like all DIF detection methods in the literature—usage of the RMSD for DIF detection requires one to make assumptions about the identification of the model in each of the considered groups. That is, in each

considered country the proficiency scale needs to be fixed, and the choices one makes to achieve this may have an impact on which items are flagged as showing DIF. Since this paper shows that—even when there are no issues with correctly fixing the scale in each group—the RMSD often fails to detect DIF, and since the issue of model identification essentially applies to all existing DIF detection methods, this identification issue has not been the focus of the current paper.

Given the discussed limitations of relying on the RMSD for determining which (and how many) items show DIF in each considered country, a sensible solution would be to always supplement its use with the use of other methods that are less sensitive to possible heterogeneity between countries. Since the RMSD intends to quantify the discrepancy between the expected and observed ICC, supplementing its use with a graphical check of the two curves (e.g., as presented in Figure 1) can be considered highly advisable, since it will give the user more information about the degree to which these two curves appear to deviate from each other. This should allow users to get a more concrete and complete picture of the extent to which relevant discrepancies appear to be present. This should be helpful for avoiding the high proportion of false negatives that our simulation results indicated should be expected when relying only on the RMSD, a proportion that—as our simulation results illustrated—will be especially high for countries whose average proficiency level deviates notably from the overall difficulty level of the test. Thus, we recommend to always use the RMSD in tandem with an inspection of the corresponding item curves, and to place additional importance on these graphs when considering low-performing countries to reduce the risk of not detecting problematic items.

Notes

[1]Purely for ease of reading, we will use the term "country" to refer to particular educational systems participating in PISA.

[2]The technical details of the RMSD and its estimation in practice are provided in the section "Properties of the RMSD."

[3]It may be noted that the curves in Panel (a) and those in Panel (b) are roughly mirror images of each other, reflected vertically around the proficiency mean of the country. This result is in line with expectations, as the formulation of the RMSD (Equation 1) shows that one can switch $P_o(\theta)$ and $P_e(\theta)$ without it affecting the value of the RMSD. Consequently, in the case of negative DIF, the RMSD is highest for items whose international location is somewhat higher than the proficiency mean of the country.

[4]This issue can be expected to also be present if one opts to exclude flagged items from the analysis of that country, since the imbalance in terms of the proportion of deleted items with positive versus negative DIF remains.

References

Buchholz, J., & Hartig, J. (2017). Comparing attitudes across groups: An IRT-based item-fit statistic for the analysis of measurement invariance. *Applied Psychological Measurement*, *43*(3), 1–10.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahway, NJ: Lawrence Erlbaum Associates, Inc.

Holland, P. W., & Thayer, D. T. (2013). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–144). New York; London: Routledge.

Köhler, C., & Hartig, J. (2017). Practical significance of item misfit in educational assessments. *Applied Psychological Measurement*, *41*(5), 388–400.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation*, *35*(2–3), 64–70.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Lawrence Erlbaum Associates, Inc.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, *7*(2), 105–118.

OECD (2017a). *PISA 2015 Assessment and Analytical Framework*. Paris: OECD Publishing.

OECD (2017b). *PISA 2015 Technical Report*. Paris: OECD Publishing.

OECD (2017c). *Scaling PISA data*. PISA 2015 Technical Report. Paris: OECD Publishing.

OECD (2018a). Gross domestic product. Retrieved from http://data.oecd.org/gdp/gross-domestic-product-gdp.htm

OECD (2018b). Life expectancy at birth. Retrieved from https://data.oecd.org/healthstat/life-expectancy-at-birth.htm

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, *53*(3), 315–333.

Robitzsch, A., Kiefer, T., & Wu, M. (2018). *TAM: Test analysis modules*.

Sinharay, S., & Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice*, *33*(1), 23–35.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*, (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Van den Noortgate, W., & De Boeck, P. (2015). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics*, *30*(4), 443–464.

van Rijn, P. W., Sinharay, S., Haberman, S. J., & Johnson, M. S. (2016). Assessment of fit of item response theory models used in large-scale educational survey assessments. *Large-Scale Assessments in Education*, *4*(1), 10.

von Davier, M. (2017). *Software for multidimensional discrete latent trait models*. ETS.

World Bank. (2018a). Dominican Republic GDP per capita (current US$). Retrieved from https://data.worldbank.org/country/dominican-republic

World Bank. (2018b). Dominican Republic life expectancy. Retrieved from https://data.worldbank.org/country/dominican-republic

World Bank. (2018c). School enrollment, secondary (% gross). Retrieved from https://data.worldbank.org/indicator/se.sec.enrr

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

Authors

JESPER TIJMSTRA is an Assistant Professor at the Department of Methodology and Statistics, Tilburg University, PO Box 90153, Tilburg, The Netherlands; j.tijmstra@uvt.nl. His

primary research interests include psychometric methods, with greatest emphasis on item response theory.

MARIA BOLSINOVA is a Research Scientist at ACTNext, 500 ACT dr., Iowa City, IA 52243; maria.bolsinova@act.org. Her primary research interests include psychometric methods.

YUAN-LING LIAW is a Research Analyst in the Research and Analysis Unit of the IEA Hamburg, Überseering 27, 22297 Hamburg, Germany; yuan-ling.liaw@iea-hamburg.de. Her primary research interests include psychometric methods, with greatest emphasis on international large-scale assessment.

LESLIE RUTKOWSKI is an Associate Professor of Inquiry Methodology, Indiana University, Bloomington, IN, 47405; lrutkows@iu.edu. Her primary interests include latent variable models, especially for cross-cultural comparisons.

DAVID RUTKOWSKI is an Associate Professor of Educational Policy and Leadership Studies, Indiana University, Bloomington, IN, 47405; drutkows@iu.edu. His primary interests include educational policy and assessment, especially with a focus on international assessment.