


## Evaluating Item Fit Statistic Thresholds in PISA: Analysis of Cross-Country Comparability of Cognitive Items

Seang-Hwane Joo, Educational Testing Service, Lale Khorramdel, National Board of Medical Examiners, Kentaro Yamamoto, Hyo Jeong Shin , and Frederic Robin, Educational Testing Service

**Abstract:** In Programme for International Student Assessment (PISA), item response theory (IRT) scaling is used to examine the psychometric properties of items and scales and to provide comparable test scores across participating countries and over time. To balance the comparability of IRT item parameter estimations across countries with the best possible model fit, a partial invariance approach is used in PISA. In this approach, international or common item parameters are estimated for the majority of items, while unique or country-specific item parameters are allowed for item-country combinations where a misfit to the common parameters can be identified. The goal of the current study is to establish item fit statistic thresholds for identifying such misfits. We investigated the impact of various thresholds on scale and score estimation. To evaluate the impact of various item fit thresholds, we systematically examined the number of unique item parameters and country performance distributions and compared the overall model fit statistics using data from PISA 2015 and 2018. Results showed that  $RMSD = .10$  provides the best fitting model while still establishing stable parameter estimations and sufficient comparability across groups. The applications and implications of the results are discussed.

**Keywords:** item fit statistics, measurement invariance, Programme for International Student Assessment, scale comparability

### Evaluating Item Fit Statistic Thresholds in PISA: Analysis of Cross-Country Comparability of Cognitive Items

International large-scale assessments (ILSAs), such as the Programme for International Student Assessment (PISA), the Progress in International Reading Literacy Study (PIRLS), and Trends in International Mathematics and Science Study, play an important role in the field of education because their data provide country-level comparisons for cognitive and noncognitive domains. Educational policymakers, practitioners, and researchers pay a significant amount of attention to ILSA results for evaluating current educational systems and platforms. In addition, the country-level comparison results directly or indirectly influence the educational decision making and curriculum implementation for each country (Cosgrove & Cartwright, 2014; Neumann, Fischer, & Kauertz, 2010). One of the primary assumptions for these country-level comparisons is the comparability of scales and scores across the participating countries (Oliveri & von Davier, 2014). Consequently, establishing a high level of comparability also referred to as measurement invariance (MI) is essential in ILSAs. To evaluate the MI across countries, item response theory (IRT) models are incorporated for scaling the data and systematically examining the psychometric properties of items. In PISA operational settings, the two-parameter logistic model (2PLM; Birnbaum, 1968) and the generalized partial credit model (GPCM; Muraki,

1992) are currently employed for dichotomous and polytomous items, respectively. Because PISA takes place every 3 years and multiple countries participate (more than 80 countries), MI must be established across countries and languages within one assessment cycle and across assessment cycles overtime for the measurement of trends (Mazzeo & von Davier, 2014). Therefore, it is necessary to place the estimated item parameters and scores on a common scale.

One popular method to calibrate IRT parameters on a common scale across multiple groups is *multi-group IRT* (Bock & Zimowski, 1997) or *concurrent calibration* (Kolen & Brennan, 2014). These are similar methods in the sense that they estimate item parameters and scores simultaneously across groups. In the concurrent calibration, a set of common items is administered on at least two test forms to establish a linkage across multiple test forms. Because concurrent calibration estimates the IRT model parameters for multiple groups within a single run, it is more efficient and accurate than the procedures where group-specific IRT models are separately estimated first and then the parameters from multiple groups are linked to being on the same metric via IRT linking (Joo, Lee, & Stark, 2017; Kim & Cohen, 2002; Kolen & Brennan, 2014). Concurrent calibration has been used for the PISA IRT scaling since the 2015 main survey (Organization for Economic Co-Operation and Development [OECD], 2016).

## Comparability and MI in ILSAs

In the general context of comparative multi-group analysis, the comparability of the measured constructs and test scores across groups is an essential and necessary condition. Although there are a number of studies examining MI methodologies and their efficiencies (e.g., Cheung & Rensvold, 2002; Steenkamp & Baumgartner, 1998), only a few of them investigated in the context of ILSAs. MI in ILSAs is unique because (a) they involve a large number of groups, and groups are to be tested simultaneously, (b) they use a balanced incomplete block design, where examinees respond to only a subset of items, resulting in a large amount of missingness in the data, and (c) group-level distribution parameters are estimated using the latent regression model (Muraki, 1992). To establish MI in the context of ILSAs, IRT modeling approaches are utilized considering different levels of MI, such as scalar invariance, metric invariance, and configural invariance. For an overview of MI, readers are referred to Davidov et al. (2014), Meredith (1993), and Millsap (2010).

## Examining MI in PISA 2015

The multi-group IRT model used in PISA assumes all item parameters to be equal across country and language groups, referred to as *international* or *common* item parameters, in the initial step of the analysis. However, this assumption is very strong and inapplicable for cross-country comparison studies because of students' ethnic diversities, socioeconomic status, and cultural backgrounds (De Jong, Steenkamp, & Fox, 2007; Kreiner & Christensen, 2014; Sachse, Roppelt, & Haag, 2016). The heterogeneity exists not only between countries but also within countries for subpopulations, such as by gender, socioeconomic status, and educational background (Rutkowski, Rutkowski, & Liaw, 2018). Previous studies explicitly emphasized the cross-country differences and the impact of the differences in item functioning and MI using ILSA data (e.g., Ercikan, 2002; Ercikan & Koh, 2005; Gierl & Khaliq, 2001). An empirical-based simulation study also showed that a scaling method that uses common item parameter estimates only for the cross-country comparison introduces a bias in item parameter and group-level proficiency estimates (Rutkowski, Rutkowski, & Zhou, 2016). Given the individual- and country-level heterogeneities are prevalent in the context of educational studies, measurement noninvariance should be considered. Previously, several methods for identifying measurement noninvariance were developed and evaluated (e.g., Joo & Kim, 2019; Kim, Cao, Wang, & Nguyen, 2018; Rutkowski & Svetina, 2014, 2017; Svetina & Rutkowski, 2014).

As an approach to deal with violations of MI, a multiple-group IRT model with a group-specific item parameter estimation approach was proposed (Oliveri & von Davier, 2011, 2014) and implemented in the PISA 2015 analysis (OECD, 2016). This approach incorporates the examination of item-by-country interaction effects during the IRT scaling procedure, wherein either unique item parameters are estimated separately for a specific country or a group of countries. More precisely, if item fit statistics reveal a misfit of the common (or international) item parameters established in the first step of the analysis, the common parameter constraints are released in a second step and in consecutive steps to estimate unique item parameters. Unique item parameters are only estimated in the case of misfitting items. Items may be excluded

from the analysis for certain groups if an error can be identified (like translation or other technical errors) and cannot be fixed (e.g., in the case of scoring errors). The unique item parameters can be estimated for a single country or for a group of countries that show a similar pattern or direction of item misfit; in the latter case, a group-specific unique item parameter is estimated with an equal constraint for the group of countries. Note that the common and unique item parameters are estimated simultaneously in the same iterative analysis step. To achieve higher comparability across countries, it is desired to obtain more common item parameters and fewer unique item parameters specific to a country-language group.

## Advantages of Estimating Unique Item Parameters

Previous studies have shown that allowing unique item parameters in the IRT scaling significantly improved the overall model fit and removed country-specific measurement bias (Oliveri & von Davier, 2011, 2014; Rutkowski et al., 2016). This approach is comparable to the concept of partial invariance in confirmatory factor analysis (CFA; Byrne, Shavelson, & Muthén, 1989). Rather than assuming full MI across multiple groups, partial invariance allows some item parameter differences across groups per construct and still maintains the measurement comparability and validity across multiple groups.

The impact of such item-by-country interactions on scale calibrations has been systematically and empirically investigated. For example, Oliveri and von Davier (2011) compared various IRT models, including single-group Rasch and 2PL models, mixture Rasch and 2PL models, and multi-group Rasch and 2PL models, and multi-group 2PL with partially unique item parameters. They applied these models to the PISA 2006 Math, Science, and Reading domains and found that the multi-group 2PL model with partially unique item parameters showed the best fit, based on the Akaike Information Criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978). They also found that the best-fitting model substantially improved the item-level fit proportions, from 17% to 74% across countries, while the country mean agreement remained acceptable (e.g., the country mean correlations were above .98 for all three domains). In addition, a previous study conducted on PIRLS 2006 data also found model fit improvement (AIC and BIC) when unique and partially unique item parameters were estimated in addition to common item parameters (Oliveri & von Davier, 2014). More recently, Rutkowski and Rutkowski (2018) empirically examined the impact of partial invariance in the context of background questionnaires using PISA data. They fitted a series of CFA models to several Nordic countries and found that the model with common item parameters produced the worst model fit based on the chi-square test and the root means squared error of approximation and allowing unique item parameters significantly improved the model fit.

In sum, although it is desirable to obtain as many common item parameters as possible to increase the comparability of scales, estimating unique item parameters in the case of misfit is still advantageous in cross-country comparison studies because (a) commonality of item parameters is a very strong assumption and is difficult to achieve due to the group heterogeneity, (b) the model fit can be improved and the measurement bias for a specific country can be reduced, and (c) a stable and reliable scale, as well as sufficient score comparability across countries, can still be provided.

## Item Fit Statistics and Thresholds

In the IRT scaling in PISA, mean deviation (MD) and root mean squared deviation (RMSD) are generally computed for each item in each country-by-language group. The model with constrained common item parameters is fitted in the initial scaling step, and the item fit statistics are evaluated for each item and group. An item is then flagged as a misfit based on the MD and RMSD values for each country group. In principle, the direction of item misfit (e.g., positive or negative) is evaluated with MD and the magnitude of item misfit is evaluated with RMSD. A positive MD indicates that the item is easier for a certain group and a negative MD indicates that the item is more difficult. The misfitting item for a specific country is considered as a measurement noninvariant item, and unique item parameters are estimated separately. In the previous studies (e.g., Buchholz & Hartig, 2019; Oliveri & von Davier, 2011, 2014), the item fit statistic thresholds for determining a misfitting item were  $\text{RMSD} = .10$  for the cognitive domains and  $\text{RMSD} = .30$  for the noncognitive domains. Based on the PISA 2015 technical report, the thresholds of  $\text{RMSD} = .12$  for the cognitive domains and  $\text{RMSD} = .30$  for the noncognitive domains were used in the PISA main survey scaling (OECD, 2016).

However, to date, very few studies have evaluated these item fit thresholds. Buchholz and Hartig (2019) conducted a simulation study examining the performance of RMSD by manipulating item parameters of multiple groups and their variabilities. Although the study recommended using an RMSD threshold of .05 as a conservative approach, this finding is limited to the simulation conditions and cannot be generalized without considering additional conditions or empirical evidence. More specifically, the study did not provide an empirical example of the impact of RMSD thresholds on group score estimates or standard errors of the measurement in the context of PISA. Furthermore, Rutkowski and Svetina (2014) investigated the performance of various overall model fit indices based on simulated data. However, their study was specifically aimed at multiple-group analysis under the structural equation modeling (SEM) framework, and specific item-level fit sensitivity was not investigated.

Therefore, the current study aims to examine the validity of the item fit thresholds currently used in PISA operational settings and provide empirical evidence for establishing such thresholds. It is expected that higher RMSD values for thresholds (i.e., a more liberal approach) will result in better comparability of scales and scores for the cognitive domains. The reason is that higher RMSD values would result in the detection of fewer misfitting items, and therefore, fewer unique item parameters would be estimated. A higher number of common item parameters and fewer unique item parameters result in higher comparability. However, this could also result in a poor model fit and incremental measurement bias. Conversely, if lower RMSD values are used as a threshold (i.e., a more conservative approach), a higher number of unique parameters would be estimated, resulting in lower comparability but a better overall model fit. In addition, the estimated model could suffer from overparameterization, as the number of estimated unique item parameters increases. Overparameterization implies that the model produces a large number of different group-specific item parameter estimates, resulting in several unique but very similar item parameter estimates. Hence, RMSD thresholds should be established with a balance of model accuracy (model fit) and international

scale comparability. To the best of our knowledge, no research has examined this issue using a systematic approach based on empirical data.

## Purpose

The purpose of the current study is to investigate the impact of item fit statistic thresholds on the scale and score comparability in the context of PISA. More specifically, the focus of the study is on evaluating various RMSD thresholds and providing empirical evidence about the efficacy of the threshold currently used in operational settings. Several models with RMSD thresholds were estimated and the cross-country comparability was systematically investigated. In addition, the model fit statistics for the various models were compared to identify the best-fitting model.

## Method

### *PISA Main Survey Data*

We used the PISA 2015 and 2018 main survey data to investigate the impact of various item fit statistic thresholds on the item parameters and ability estimates. We considered the core cognitive domains of Reading, Math, and Science. Note that Science and Reading were major domains in PISA 2015 and 2018, respectively, in the sense that all students received major domain items in addition to one or two minor domain items. From the 2015 cycle, the PISA assessment was conducted in two different administration modes: computer-based assessment (CBA) and paper-based assessment (PBA). For the CBA countries, 88–92% of students received a test form that assembled from one major domain and one minor domain. The two-domain test form consisted of four 30-minute clusters, resulting in a total of 2 hours of testing time per student. Eight to twelve percent of students received a form that was assembled from three domains, consisting of four 30-minute clusters covering one major and two minor domains. This balanced incomplete block design is typical for ILSAs, in general, resulting in missing responses. Following the PISA operational settings, the missing responses were excluded from the likelihood function in the IRT scaling. For more detailed information about the PISA main survey design, data, and scaling method, readers are referred to the PISA 2015 and 2018 technical reports (OECD, 2016; OECD, 2019).

In the current study, we used data from CBA countries only because (a) most countries have participated in CBA (e.g., 57 CBA as opposed to 15 PBA countries in PISA 2015 and 70 CBA as opposed to 8 PBA in PISA 2018) and (b) only CBA countries received new items in major domains, resulting in more assessment items for CBA countries than PBA countries. In PISA 2015, CBA countries received a total 184 of items for Science (99 items were new) whereas PBA countries received a total of 85 items (trend items only). Similarly, in PISA 2018, CBA countries received a total of 244 items for Reading (172 items were new), whereas PBA countries received a total of 72 items. Given that more valid and precise measurement are generally found with a larger number of items, we analyzed CBA countries only in this study.

The total number of country-by-language groups for each cognitive domain was 83 and 103, respectively, for PISA 2015 and 2018 data. We considered the same country-by-language groups to be consistent with the operational PISA analyses to

take the effect of multiple languages within a country into account, and we used Senate weights in our analysis to, again, follow the procedure taken in the operational setting. The senate weight constrains the sample size for each country to be equal (e.g., 5,000) so that all countries contribute equally to IRT scaling.

### Item Fit Statistics Threshold Analysis

We scaled the PISA main survey data with the 2PLM and GPCM for the dichotomous and polytomous responses, respectively. Because the large-scale assessment data include multiple country-by-language groups and missing responses as a result of the assessment design (e.g., balanced incomplete block design), we calibrated scales and scores using the concurrent calibration strategy implemented in the *mdltm* program (von Davier, 2005).

The concurrent calibration was conducted over multiple steps. In the initial calibration, we estimated item parameters with equality constraints across the country-by-language groups using the common (international) item parameters. Once the model was estimated, MD and RMSD were computed and evaluated for each country-by-language group. MD and RMSD were computed for each country-by-language group as:

$$MD_g = \int [p_g^{obs}(\theta) - p_g^{exp}(\theta)] f_g(\theta) d\theta \quad (1)$$

$$RMSD_g = \sqrt{\int [p_g^{obs}(\theta) - p_g^{exp}(\theta)]^2 f_g(\theta) d\theta} \quad (2)$$

MD and RMSD are computed for a country-by-language group  $g = 1, \dots, G$  based on the deviation between the observed and expected item characteristic curves (ICCs). In Equations (1) and (2),  $p_g^{obs}(\theta)$  represents the group-specific observed ICC and  $p_g^{exp}(\theta)$  represents the group-specific expected ICC given the student's ability  $\theta$ . In addition,  $f(\theta)$  indicates the weight of the students' ability scales in the group. The observed ICC is obtained from the observed responses across students for each item and the expected ICCs are computed based on the IRT model using the estimated item parameters. The integrals in Equations (1) and (2) are approximated with Gaussian quadrature points ranging from  $-5$  to  $5$  (Bock & Aitkin, 1981; von Davier, 2005). As shown in Equations (1) and (2), the deviation between the observed and expected ICCs in item fit statistics is computed across the student's ability levels and two computational approaches are used to produce MD and RMSD. MD is computed from the ICC difference over the latent trait distribution, whereas RMSD is computed from the squared ICC difference over the latent trait distribution (the final RMSD is obtained by taking the square root of the mean squared difference estimates). Thus, MD generally ranges from negative to positive values, whereas RMSD ranges from 0 to 1. Larger MD and RMSD values indicate a larger item misfit. Because MD does not include the square term, the ICC deviation can be impacted by both direction and magnitude. A positive MD indicates that the item is easier for a certain group because the observed ICC has shifted, overall, toward the left and the observed proportion of correct is higher than the expected proportion. A negative MD indicates that the item is more difficult. However, MD has a limitation when observed and expected ICCs are not uniformly different across the latent trait distribution. For example, if the ICC differences have opposite directions with

the same magnitude, the overall ICC difference is counter-balanced across the latent trait distribution. As a result, the observed and expected ICC difference is negligible.

In our study, we considered the RMSD thresholds varied from .40 to .00 with an initial decrement of .10 (i.e., RMSD thresholds = .40, .30, .20, .10, and .00). To investigate the impact more precisely, we additionally considered the RMSD thresholds with finer grids, a decrement of .02 from .20 to .15 (i.e., RMSD thresholds = .20, .18, and .16) and a decrement of .01 from .15 to .10 (i.e., RMSD thresholds = .15, .14, .13, .12, .11, and .10). For the threshold interval from .10 to .00, we varied with a decrement of .02 (i.e., RMSD thresholds = .10, .08, .06, .04, .02, and .00). As a result, a total of 16 RMSD thresholds were applied for identifying misfitting items and estimating unique item parameters. Based on the preliminary investigation, we found that no item had an RMSD greater than .50 for any group.

In the second step of IRT scaling, we re-estimated the IRT model with unique item parameters based on the predetermined RMSD thresholds. Re-estimating the model was an iterative process, such that the model estimation continued until no item was identified as having misfit. When unique item parameters were estimated, we also evaluated MD along with RMSD. If an item showed misfit based on the RMSD threshold, the direction of MD was investigated. If two or more misfitting country-by-language groups had the same MD direction (positive or negative), the equality constraint was applied for those groups so that the same unique item parameters were estimated for those groups. This approach was chosen to increase the comparability of the scales across the country-by-language groups and reduce possible model over-parameterization.

### Outcome Variables

*Percentage of unique item parameters.* As a measure of comparability, we computed the number and percentage of unique item parameters for each domain. It is expected that the smaller RMSD threshold would produce more unique item parameters for the country-by-language groups and consequently, comparability of scales and scores would decrease. Note that an RMSD threshold of .00 would produce unique item parameters for all the country-by-language groups. To compute the percentage of unique item parameters, we considered the total number of item-by-group cells by multiplying the numbers of items and country-by-language groups. For example, in PISA 2015, the total number of item-by-group cells for the Science domain was  $184 \text{ (items)} \times 83 \text{ (country-by-language)} = 15,272$ . Similarly, the total number of item-by-group cells was  $103 \times 83 = 7,286$  for Math and  $81 \times 83 = 5,727$  for Reading. In addition, we separately computed the numbers of unique and partially unique item parameters. The percentages of unique and partially unique item parameters were then obtained by computing the numbers of unique and partially unique item parameters over the total number of item-by-group combinations. Note that in the PISA 2015 main survey, the percentages of the unique and partially unique item parameters were approximately 5% and 8% at most, respectively, for all cognitive domains using the RMSD threshold of .12 (OECD, 2016).

*Group mean and standard deviation.* To investigate the accuracy of scale and score calibrations, we computed the

country-by-language group means and standard deviation estimates across various RMSD thresholds. More specifically, the group mean and standard deviation estimates for all country-by-language groups were computed for each model based on the RMSD thresholds, and the estimates were averaged across the groups. Thus, the averaged mean and standard deviation estimates were computed for each model across RMSD thresholds. For group means and standard deviations, we used the direct estimates from the multi-group IRT models. The average group mean and standard deviation estimates show the accuracy and stability of the group-level scores across RMSD thresholds. Although in the PISA analysis, group mean and standard deviation estimates are not directly used to report country-level performances, they are useful to provide information about the within-country score stability across various RMSD thresholds. In addition, mean and standard deviation estimates also provide the distributions of the country-by-language groups across the RMSD threshold models.

*Group mean score agreement.* To investigate the agreement of group mean scores across the different threshold models, we computed the correlation between scores obtained from the model with all common item parameters versus the models with common, unique, and partially unique item parameters. The correlations of scores can provide information about the agreement of scores across the various RMSD thresholds. In a previous study, robust score agreement results were found with the unique item parameter IRT scaling approach (Oliveri & von Davier, 2014). In their study, the RMSD threshold of .10 was considered, and the score correlation across 40 countries was over .99. The correlation result implies that the country performance was minimally affected by unique item parameter estimation. However, we expect that with smaller RMSD thresholds, the group-level score correlation will decrease alongside the comparability of item parameter estimates across countries.

*Model fit statistics.* To examine the model fit for the various RMSD threshold models, we computed the overall model fit statistics, including AIC, BIC, and consistent AIC (CAIC). AIC, BIC, and CAIC include the penalty terms for the number of estimated parameters. Also, it is worthwhile to note that BIC and CAIC additionally penalize the number of examinees from the deviance, and AIC tends to choose models with more parameters. The model fit statistics recommended that under small sample size conditions, in which errors occur from the under-fitting model, AIC is more reasonable than BIC and CAIC (Dziak et al., 2012). On the other hand, under the large sample size conditions, in which errors occur from the over-fitting model, BIC and CAIC are more appropriate than AIC (see Dziak et al., 2012 p. 23). A consistent pattern was also found in previous studies that examined the model fit indices in the SEM framework (Kim et al., 2016). Based on the previous findings, BIC and CAIC seem to be the more appropriate indices for determining the better-fitting model in the context of large-scale assessment.

## Results

### *Unique and Partially Unique Item Parameters*

Tables 1 and 2 represent the number and percentage of the unique and partially unique item parameters across the

RMSD thresholds for the Math, Reading, and Science domains in PISA 2015 and 2018. Overall, the highest comparability was retained for Math, followed by Science and Reading. As shown in Table 1, when the RMSD threshold was set for .12, the percentages of unique item parameters were 1.13% for Math, 2.75% for Science, and 3.01% for Reading. For the same condition, the corresponding percentages of partially unique item parameters were 3.81%, 6.53%, and 7.32%, respectively. Similarly, in PISA 2018, the unique item percentages were .84% for Math, 1.00% for Science, and 2.97% for Reading, and the partially unique percentages were 3.26% for Math, 5.50% for Science, and 7.65% for Reading. The numbers and percentages of unique and partially unique item parameters were comparable across two cycles and similar values were found in the PISA operational results (OECD, 2016, 2019).

Regarding the RMSD thresholds, the scale comparability was reasonably accomplished with the RMSD thresholds from .40 to .10, but the RMSD thresholds below .10 produced a substantially large number of unique and partially unique item parameters for all cognitive domains. In PISA 2015, the percentages of unique item parameters for Math, Science, and Reading were 2.95%, 4.43%, and 8.66% when the RMSD threshold of .10 was applied, whereas the corresponding percentages jumped to 13.27%, 25.86%, and 27.33% when the RMSD threshold of .08 was applied. Similarly, in PISA 2018, the percentages of unique item parameters for Math, Science, and Reading were .81%, 1.11%, and 5.12% with the RMSD threshold of .10, whereas the corresponding values considerably increased to 9.02%, 17.54%, and 29.80% with the RMSD threshold of .08. The increase in the percentage of the unique item parameter estimates was also substantial as the RMSD threshold decreased from .08 to .00 in both PISA 2015 and 2018 data.

### *Group Mean and Standard Deviation Estimates*

Figure 1 illustrates the country-by-language group mean and standard deviation estimates for all cognitive domains. Figure 1 provides information about within-group consistency for the mean and standard deviation estimates across the RMSD thresholds. Note that, for the model specification, the group mean and standard deviation of the first group were constrained to be 0 and 1, respectively, and the rest of the groups' means and standard deviations were freely estimated. For each cognitive domain, the means and standard deviations were averaged across the country-by-language groups. As shown in Figure 1 panels (a) and (b), the mean estimates were stable as the RMSD thresholds decreased from .50 to .10. However, within the RMSD threshold interval between .10 and .00, the mean estimates were inconsistent and fluctuated for all cognitive domains. The group mean fluctuation across the RMSD thresholds was substantial for the PISA 2015 cognitive domains and relatively moderate for the PISA 2018 cognitive domains. In addition, as shown in Figure 1 panels (c) and (d), the average standard deviation estimates were consistent as the RMSD thresholds decreased from .50 to .10, whereas the values considerably increased as the RMSD threshold decreased from .10 to .00.

### *Group Mean Score Correlation*

Figure 2 presents the score correlation between the RMSD-based unique parameter model and the all common item parameter model. The score correlation provides information



**Table 1. Number and Percentage of Unique and Partially Unique Item Parameters Based on the RMSD Thresholds for PISA 2015 Data**

RMSD Thresholds	.40	.30	.20	.18	.16	.15	.14	.13	.12	.11	.10	.08	.06	.04	.02	.00
Total (Item*Group)	5,727	5,727	5,727	5,727	5,727	5,727	5,727	5,727	5,727	5,727	5,727	5,727	5,727	5,727	5,727	5,727
Common	5,726	5,722	5,688	5,672	5,628	5,604	5,568	5,520	5,444	5,372	5,187	4,433	2,750	564	360	0
Partially Unique	0	2	19	31	72	80	110	147	218	281	371	534	517	119	94	0
Unique	1	3	20	24	27	43	49	60	65	74	169	760	2,460	5,044	5,273	5,727
Partially Unique %	.00%	.03%	.33%	.54%	1.26%	1.40%	1.92%	2.57%	3.81%	4.91%	6.48%	9.32%	9.03%	2.08%	1.64%	.00%
Unique %	.02%	.05%	.35%	.42%	.47%	.75%	.86%	1.05%	1.13%	1.29%	2.95%	13.27%	42.95%	88.07%	92.07%	100.00%
RMSD	.40	.30	.20	.18	.16	.15	.14	.13	.12	.11	.10	.08	.06	.04	.02	.00
Thresholds																
Total (Item*Group)	7,286	7,286	7,286	7,286	7,286	7,286	7,286	7,286	7,286	7,286	7,286	7,286	7,286	7,286	7,286	7,286
Common	7,285	7,276	7,180	7,092	7,004	6,940	6,844	6,728	6,534	6,239	5,779	4,339	1,560	375	133	0
Partially Unique	0	0	55	115	203	244	336	431	533	701	876	956	438	85	33	0
Unique	1	10	51	79	79	102	106	127	219	346	631	1,991	5,288	6,826	7,120	7,286
Partially Unique %	.00%	.00%	.75%	1.58%	2.79%	3.35%	4.61%	5.92%	7.32%	9.62%	12.02%	13.12%	6.01%	1.17%	.45%	.00%
Unique %	.01%	.14%	.70%	1.08%	1.08%	1.40%	1.45%	1.74%	3.01%	4.75%	8.66%	27.33%	72.58%	93.69%	97.72%	100.00%
RMSD	.40	.30	.20	.18	.16	.15	.14	.13	.12	.11	.10	.08	.06	.04	.02	.00
Thresholds																
Total (Item*Group)	15,272	15,272	15,272	15,272	15,272	15,272	15,272	15,272	15,272	15,272	15,272	15,272	15,272	15,272	15,272	15,272
Common	15,272	15,256	15,086	14,975	14,776	14,642	14,440	14,175	13,854	13,473	12,949	9,465	4,107	3,599	1,485	0
Partially Unique	0	4	99	188	367	474	597	775	998	1,334	1,648	1,857	999	489	359	0
Unique	0	12	87	109	129	156	235	322	420	465	675	3,950	10,166	11,184	13,428	15,272
Partially Unique %	.00%	.03%	.65%	1.23%	2.40%	3.10%	3.91%	5.07%	6.53%	8.73%	10.79%	12.16%	6.54%	3.20%	2.35%	.00%
Unique %	.00%	.08%	.57%	.71%	.84%	1.02%	1.54%	2.11%	2.75%	3.04%	4.42%	25.86%	66.57%	73.23%	87.93%	100.00%

Note. RMSD = root mean square deviation. Total = total number of item-by-group parameters, Common = number of common item parameters, Partially Unique = number of partially unique item parameters, Unique = number of unique item parameters.

**Table 2. Number and Percentage of Unique and Partially Unique Item Parameters Based on the RMSD Thresholds for PISA 2018 Data**

<b>RMSD Thresholds</b>	<b>.40</b>	<b>.30</b>	<b>.20</b>	<b>.18</b>	<b>.16</b>	<b>.15</b>	<b>.14</b>	<b>.13</b>	<b>.12</b>	<b>.11</b>	<b>.10</b>	<b>.08</b>	<b>.06</b>	<b>.04</b>	<b>.02</b>	<b>.00</b>
<b>Total</b>	8,446	8,446	8,446	8,446	8,446	8,446	8,446	8,446	8,446	8,446	8,446	8,446	8,446	8,446	8,446	8,446
<b>(Item*Group)</b>																
Common	8,443	8,440	8,406	8,377	8,321	8,289	8,239	8,174	8,100	7,986	7,867	6,922	4,774	2,297	0	0
Partially Unique	0	4	26	40	68	103	147	203	275	390	511	762	659	115	0	0
Unique	3	2	14	29	57	54	60	69	71	70	68	762	3,013	6,034	8,446	8,446
Partially Unique %	.00%	.05%	.31%	.47%	.81%	1.22%	1.74%	2.40%	3.26%	4.62%	6.05%	9.02%	7.80%	1.36%	.00%	.00%
Unique %	.04%	.02%	.17%	.34%	.67%	.64%	.71%	.82%	.84%	.83%	.81%	9.02%	35.67%	71.44%	100.00%	100.00%
RMSD	.40	.30	.20	.18	.16	.15	.14	.13	.12	.11	.10	.08	.06	.04	.02	.00
<b>Thresholds</b>																
<b>Total</b>	25,235	25,235	25,235	25,235	25,235	25,235	25,235	25,235	25,235	25,235	25,235	25,235	25,235	25,235	25,235	25,235
<b>(Item*Group)</b>																
Common	25,228	25,196	24,899	24,721	24,358	24,121	23,793	23,354	22,556	21,977	21,199	14,826	4,988	267	0	0
Partially Unique	2	9	213	360	652	886	1,164	1,520	1,930	2,334	2,744	2,888	1,295	44	0	0
Unique	5	30	123	154	225	228	278	361	749	924	1,292	7,521	18,952	24,924	25,235	25,235
Partially Unique %	.01%	.04%	.84%	1.43%	2.58%	3.51%	4.61%	6.02%	7.65%	9.25%	10.87%	11.44%	5.13%	.17%	.00%	.00%
Unique %	.02%	.12%	.49%	.61%	.89%	.90%	1.10%	1.43%	2.97%	3.66%	5.12%	29.80%	75.10%	98.77%	100.00%	100.00%
RMSD	.40	.30	.20	.18	.16	.15	.14	.13	.12	.11	.10	.08	.06	.04	.02	.00
<b>Thresholds</b>																
<b>Total</b>	11,845	11,845	11,845	11,845	11,845	11,845	11,845	11,845	11,845	11,845	11,845	11,845	11,845	11,845	11,845	11,845
<b>(Item*Group)</b>																
Common	11,842	11,829	11,750	11,678	11,548	11,477	11,410	11,257	11,075	10,864	10,650	8,290	4,089	306	0	0
Partially Unique	0	6	56	98	212	288	362	485	651	863	1063	1477	833	106	0	0
Unique	3	10	39	69	85	80	73	103	119	118	132	2,078	6,923	11,433	11,845	11,845
Partially Unique %	.00%	.05%	.47%	.83%	1.79%	2.43%	3.06%	4.09%	5.50%	7.29%	8.97%	12.47%	7.03%	.89%	.00%	.00%
Unique %	.03%	.08%	.33%	.58%	.72%	.68%	.62%	.87%	1.00%	1.00%	1.11%	17.54%	58.45%	96.52%	100.00%	100.00%

Note: RMSD = root mean square deviation. Total = total number of item-by-group parameters, Common = number of common item parameters, Partially Unique = number of partially unique item parameters, Unique = number of unique item parameters.

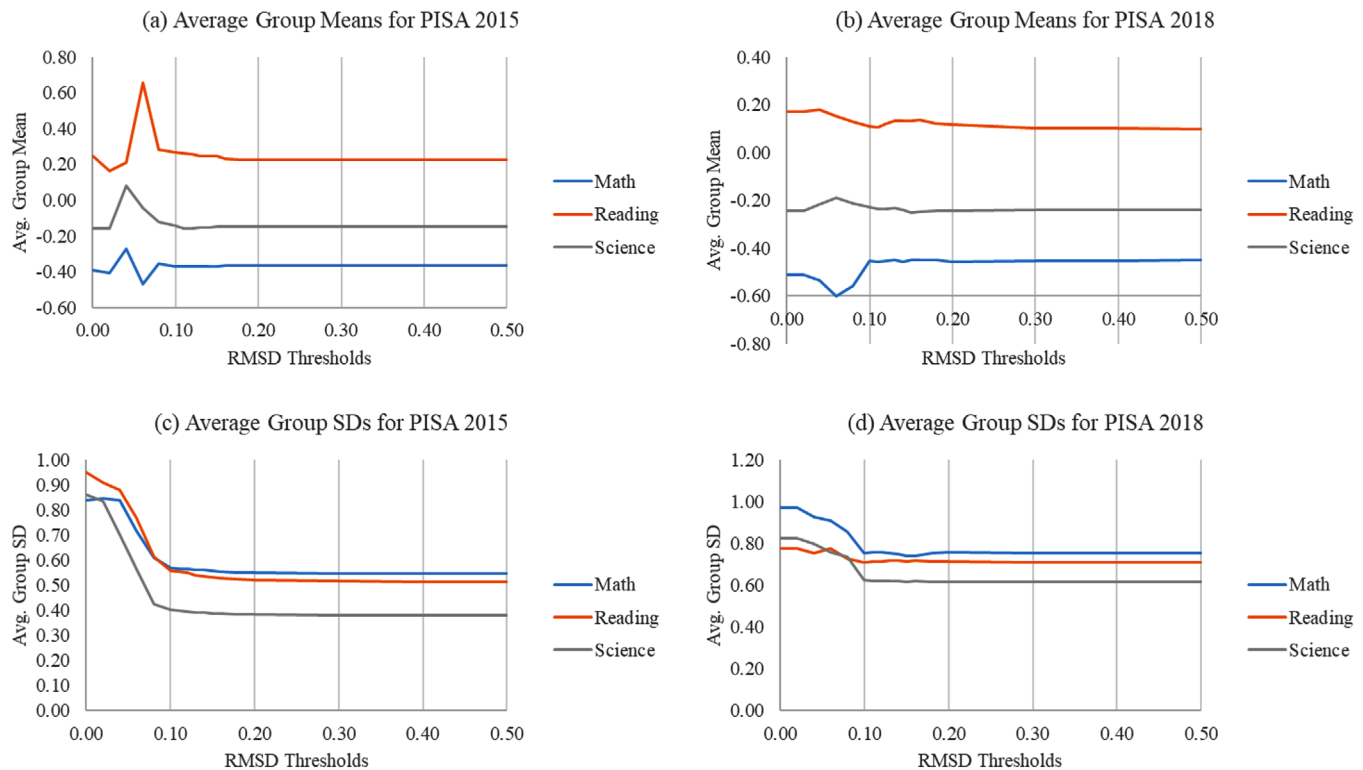


FIGURE 1. Average country-by-language group means and standard deviations across the RMSD thresholds. [Color figure can be viewed at [wileyonlinelibrary.com](#)]

about the agreement of group scores across the RMSD thresholds. We found that when the RMSD thresholds were relatively large (e.g., the RMSD thresholds were .10 and above), the score correlation was close to 1. This finding was expected because when a high level of scale comparability is obtained, consistent score agreement across countries tends to be observed, which is supported by previous studies (Jerrim et al., 2018; Oliveri, & von Davier, 2011, 2014).

However, when the RMSD thresholds were relatively low (e.g., the RMSD thresholds were .10 and below), the score correlation decreased. As shown in Figure 2, a notable decrement occurred when the RMSD threshold below .10 was used for all cognitive domains. For PISA 2015 data, the score correlations between all common versus all unique models (i.e., RMSD threshold was .00) were approximately .92, .93, and .92 for Math, Science, and Reading, respectively. For PISA 2018 data, the corresponding score correlations were .92, .94, and .96.

#### *Model Fit Statistics and the Number of Estimated Item Parameters*

Figure 3 shows the model fit statistics (AIC, BIC, and CAIC) results, and Figure 4 presents the number of estimated unique and partially unique item parameters across the various RMSD threshold models. As shown in Figure 3, BIC and CAIC consistently decreased (note that smaller values indicate better model fit) as the RMSD threshold value decreased from .50 to .10, and then increased substantially as the RMSD threshold decreased from .10 to .00. However, AIC consistently decreased as the RMSD threshold decreased. This pattern was consistent across all cognitive domains in both PISA 2015 and PISA 2018 data. The incremental BIC and CAIC

imply model overparameterization because BIC and CAIC tend to penalize the overparameterized models; conversely, AIC tends to penalize the underparameterized models (Dziak et al., 2012). Similarly, we found a substantial increase in the number of estimated parameters in the unique item parameter model when the RMSD decreased from .10 to .00, as shown in Figure 4. The results from Figures 3 and 4 indicate that the unique item parameter model with RMSD thresholds less than .10 are problematic, for which the model is estimated with too many parameters.

To investigate the cause of poor fit, we also computed the average item parameter differences for Math, Science, and Reading using the model with RMSD = .10. The item parameter differences were computed by taking squared differences between the common item parameters and the unique item parameters. For items that have more than one unique set of item parameters, an average of the multiple differences was computed. The item parameter differences were computed for discrimination and difficulty parameters separately and the results were summarized by taking the squared root of the average item parameter differences across items. In PISA 2015, the root mean squared differences of the difficulty parameters across items were .82, .69, and .76 for Math, Science, and Reading, respectively, whereas the differences of the discrimination parameters were .17, .16, and .17. Similarly, in PISA 2018, the root mean squared differences of the difficulty parameters were .91, .67, and .73 for Math, Science, and Reading, respectively, whereas the differences of the discrimination parameters were .21, .15, and .16. Note that this finding is highlighting the previous study, where the difficulty parameter was the main source on causing misfitting items as opposed to the discrimination parameter (Buchholz & Hartig, 2019).



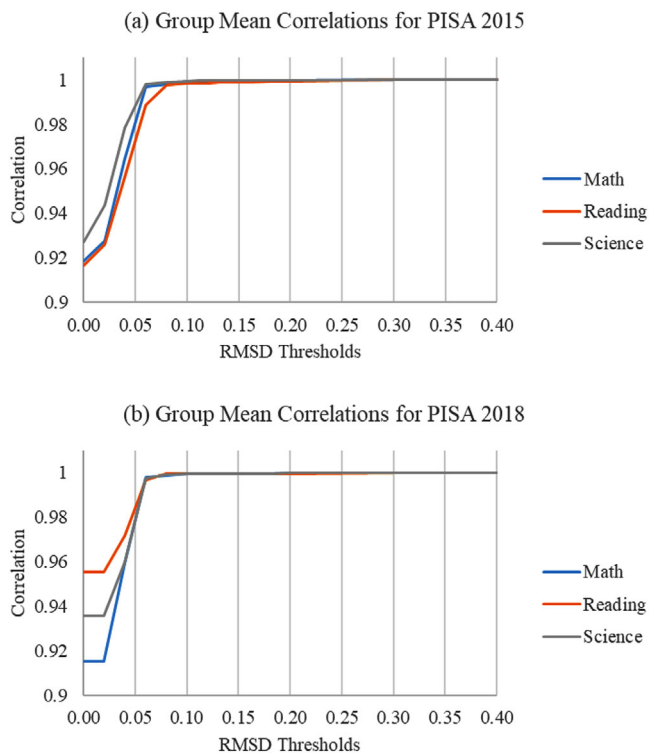


FIGURE 2. Correlations of country-by-language group means for the common item parameter model versus group-specific unique item parameter models across the RMSD thresholds. [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

## Discussion and Conclusion

In this study, we aimed to examine the adequacy of the item fit statistics thresholds used in IRT scaling for evaluating item misfit and the estimation of common (international) and unique (country-specific) item parameters in PISA. The current study empirically examined the impact of RMSD thresholds on score and scale calibrations using the PISA 2015 main survey data for the cognitive domains. We systematically varied RMSD thresholds by allowing the estimation of unique item parameters in cases of misfit to the common item parameters and evaluated the cross-country comparability of item parameters and scores. In addition, we compared the overall model fit indices and the number of estimated unique item parameters in the fitted models.

The results showed the use of the RMSD threshold of .12, which yields 3% or fewer unique item parameters for all cognitive domains, or the RMSD threshold of .10, which yields 10% or fewer unique parameters. A high level of cross-country comparability of item parameters was retained with both RMSD thresholds, and the findings were comparable to the results published in the PISA 2015 technical report (OECD, 2016). Moreover, the country-by-language group mean estimates remained consistent across the common item parameter model and the unique item parameter model. However, models with the RMSD threshold below .10 produced unstable and inconsistent group mean estimates across country-by-language groups. A similar pattern was observed for the country-by-language group standard deviation estimates: the RMSD threshold below .10 increased the group standard deviation considerably. Furthermore, the group mean correlation across the country-by-language groups was close to 1 and remained consistent as the RMSD threshold decreased

from .40 to .10, indicating that the country-by-language group mean estimates were stable across the fitted models. However, the group mean correlation decreased notably as the RMSD threshold decreased from .10 to .00, indicating that the country-by-language group mean estimates were unstable across the fitted models. These results imply that RMSD greater or equal to .10 is an agreeable threshold for unique item parameter estimation while maintaining a high score agreement across countries.

More importantly, the study empirically supports that the item fit threshold in the PISA 2015 main survey scaling for cognitive domains was adequately established. The simultaneous estimation of common and unique item parameters reduces the number of misfitting items and maintains the comparability of scales and scores across country-by-language groups. Although estimating unique item parameters decreases the group mean estimation error (Rutkowski et al., 2016) and improves overall model fit indices in the IRT scaling (Oliveri & von Davier, 2011), it is also important to maintain the cross-country scale comparability with common item parameters. Note that in some cases where a small number of items are administered for many groups, it might be more desirable to maintain comparable scales with common item parameters than estimate unique item parameters for a slight increment of model fit, particularly when the fit statistics are based on a smaller sample size. Ensuring a high number of common items across groups is necessary in large-scale assessments for the valid score comparison. For this reason, it is encouraging for practitioners that the current study confirmed the RMSD threshold of .10, which reduces the number of misfitting items, resulting in more accurate group mean estimates, while providing comparable assessment scores across the country-by-language groups.

We also compared model fit statistics and examined the number of estimated parameters for models with various RMSD thresholds. It is important to identify the RMSD threshold that can keep the balance between reducing item misfit and maintaining model parsimony because more conservative RMSD thresholds could yield overestimated parameters. Based on the study results, we found that the model with an RMSD threshold equal to .10 yielded the smallest model fit indices, indicating that this is the best-fitting model. As the RMSD threshold decreased from .10 to .00, BIC and CAIC indices, and the number of estimated parameters, increased substantially, although AIC consistently decreased. Given that PISA and general large-scale assessments tend to have large sample size, BIC and CAIC are more appropriate because BIC and CAIC penalize the deviance for large sample size and AIC tends to choose more complex (i.e., overparameterized) models (Dziak et al., 2012; Kim et al., 2016). The study results imply that RMSD thresholds below .10 would yield overparameterized models. This finding is important because it empirically shows that the RMSD threshold used in operational settings retains the balance of the number of estimated parameters for the unique item parameter approach and still gains the advantages of improving model fit.

In addition, it is worthwhile to note that the RMSD threshold from the current study result is a more conservative approach than the previous study's RMSD suggestion (Buchholz & Hartig, 2019). We suspect the difference comes from the data characteristics as the current study used empirical data, whereas the previous study used simulated data. Given that large-scale assessment data generally include more errors induced by measurement, sampling, and equating procedures

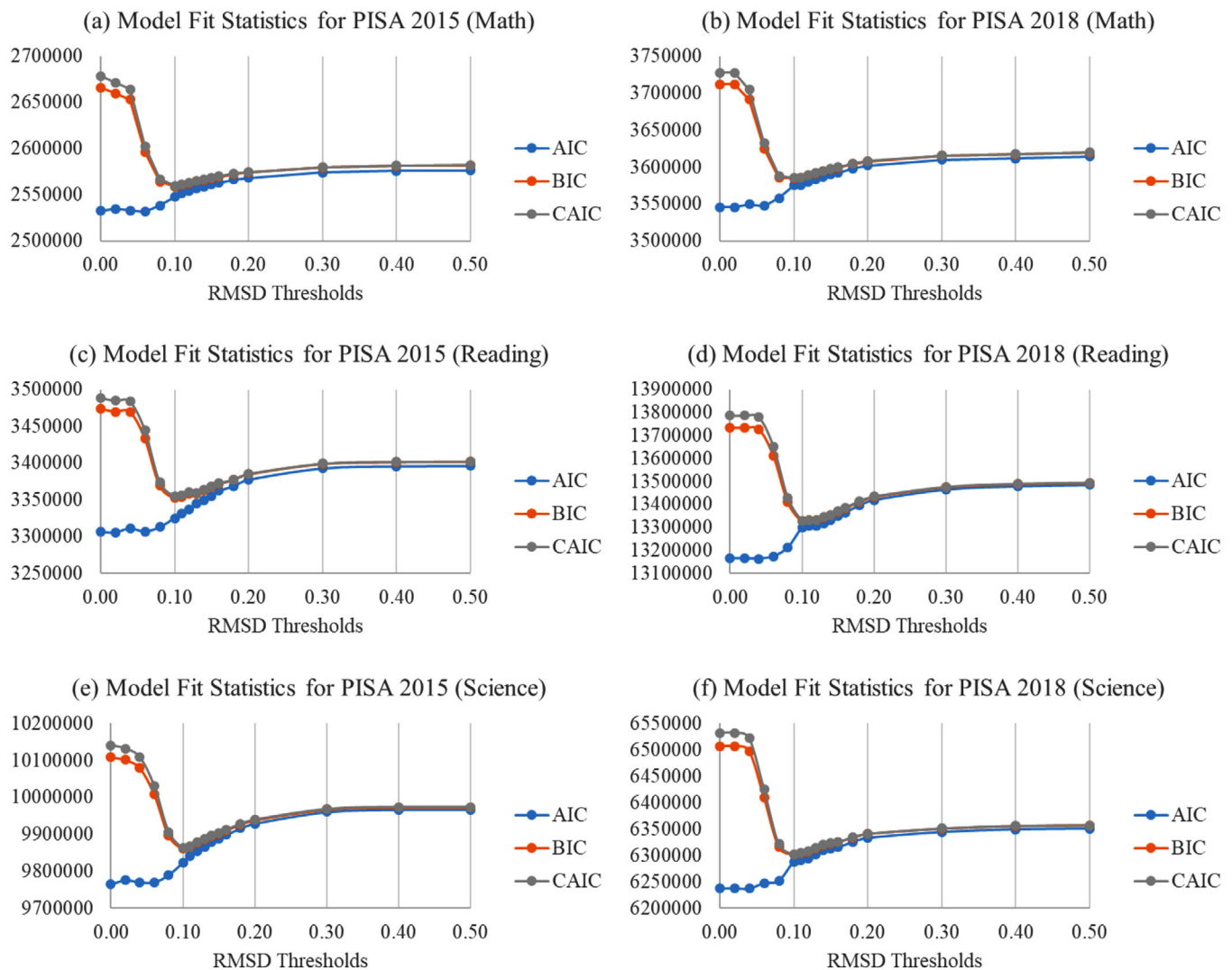


FIGURE 3. Model fit statistics for the RMSD-based unique item parameter models. [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

(Wu, 2010), the conservative RMSD threshold might be more reasonable for operational settings.

Although the results provide informative evidence regarding IRT scaling in PISA, we also recognize the limitations of the study. First, this study only investigated the cognitive domains in the PISA 2015 main survey because the main scope of the study was to evaluate and validate the item fit threshold for cognitive domains currently used in operational settings. Note that PISA includes more than 50 noncognitive scales and each noncognitive scale tends to show distinct psychometric properties, including reliability, validity, and underlying response process. It is more challenging to obtain the comparability of the scales in noncognitive assessment because students from various countries have a different conceptual understanding and cultural bias for questionnaires (Pokropek, Borgonovi, & McCormick, 2017). In addition, noncognitive scales tend to show response styles, including random, extreme, and central tendency in low-stake assessments (Buckley, 2009; Khorramdel & von Davier, 2014; Khorramdel et al., 2017), consequently, less reliable scales are often observed. In operational settings, for this reason, the RMSD threshold for the noncognitive assessment is more liberal (e.g.,  $\text{RMSD} = .30$ ) compared to the cognitive

assessment (OECD, 2016). However, the empirically derived item fit threshold for noncognitive assessment has not been established yet and should be investigated. Given that distinctive and challenging scales are present in noncognitive assessments, a systematic and explicit investigation for the item fit threshold is needed in the future. Second, in the current analysis, we only considered data from one cycle, so it would be interesting to compare the consistency of item fit statistic thresholds from different cycles. Given that PISA is conducted every 3 years and a number of trend items are retained from cycle to cycle, it is worthwhile to investigate the behavior of RMSD thresholds from different cycles. For example, although the current study estimated item parameters as new items, it is possible to fix item parameters using values obtained from previous cycles for trend items and investigate the scale and score comparability with various RMSD thresholds. Third, the current study considered all administered items for the item fit threshold analysis from the cognitive domains in PISA. Given that PISA includes the relatively large number of items for cognitive and noncognitive assessments, the study may not be ideal for other ILSAs. Future studies should investigate the behavior of item fit thresholds by reducing the number of items.

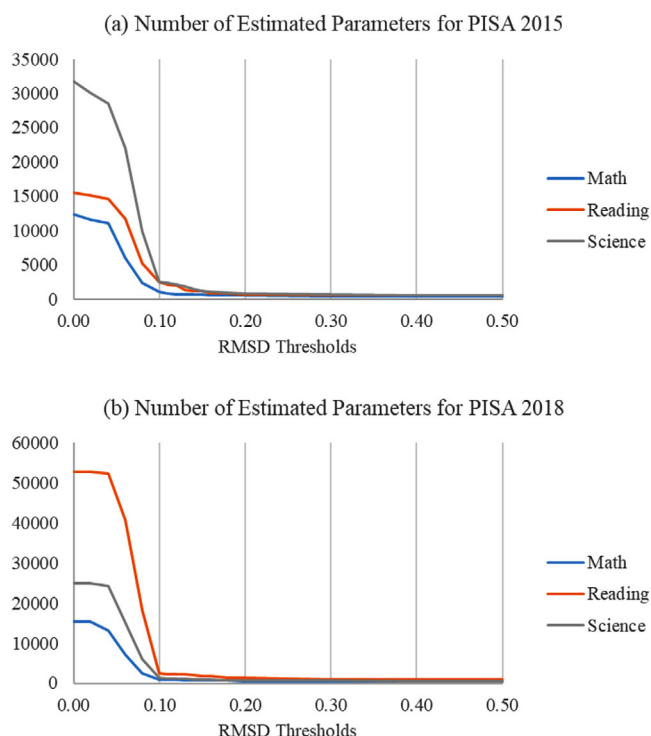


FIGURE 4. The number of estimated parameters for the RMSD-based unique item parameter models. [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

Finally, although it is important to investigate various RMSD thresholds based on empirical data, it is also critical to investigate simulation-based RMSD thresholds to generalize the current study findings for the wide range of ILSAs. We acknowledge the limitation of the real data analysis where we do not know the truth. Real data-based RMSD thresholds may not be generalizable for the other ILSAs because the number of participating country and language systems, scaling methodologies, and measurement errors associated with the construct are different from each other. To evaluate more generalizable RMSD thresholds, one needs to conduct a simulation study where a variety of conditions can be manipulated as the empirical data, and the truth parameter is known.

## References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.
- Birnbaum, A. (1968). *On the estimation of mental ability (Series Report No. 15)*. Randolph Air Force Base, TX: USAF School of Aviation Medicine.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443–459.
- Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In *Handbook of modern item response theory* (pp. 433–448). New York: Springer.
- Buchholz, J., & Hartig, J. (2019). Comparing attitudes across groups: An IRT-based item-fit statistic for the analysis of measurement invariance. *Applied Psychological Measurement*, 43, 241–250.
- Buckley, J. (2009). Cross-national response styles in international educational assessments: Evidence from PISA 2006. *NCES Conference on the Program for International Student Assessment: What we can learn from PISA*. Washington, DC.
- Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105, 456.
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233–255.
- Cosgrove, J., & Cartwright, F. (2014). Changes in achievement on PISA: The case of Ireland and implications for international assessment practice. *Large-Scale Assessments in Education*, 2, 1–17.
- Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, 40, 55–75.
- De Jong, M. G., Steenkamp, J. B. E., & Fox, J. P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research*, 34, 260–278.
- Dziak, J. J., Coffman, D. L., Lanza, S. T., & Li, R. (2012). *Sensitivity and specificity of information criteria*. Pennsylvania, PA: The Methodology Center and Department of Statistics, The Pennsylvania State University.
- Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing*, 2, 199–215.
- Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing*, 5, 23–35.
- Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement*, 38, 164–187.
- Jerrim, J., Parker, P., Choi, A., Chmielewski, A. K., Sälzer, C., & Shure, N. (2018). How robust are cross-country comparisons of PISA scores to the scaling model used? *Educational Measurement: Issues and Practice*, 37, 28–39.
- Joo, S. H., & Kim, E. S. (2019). Impact of error structure misspecification when testing measurement invariance and latent-factor mean difference using MIMIC and multiple-group confirmatory factor analysis. *Behavior Research Methods*, 51, 2688–2699.
- Joo, S. H., Lee, P., & Stark, S. (2017). Evaluating anchor-item designs for concurrent calibration with the GGUM. *Applied Psychological Measurement*, 41, 83–96.
- Khorramdel, L., & von Davier, M. (2014). Measuring response styles across the Big Five: A multi-scale extension of an approach using multinomial processing trees. *Multivariate Behavioral Research*, 49, 161–177.
- Khorramdel, L., von Davier, M., Bertling, J. P., Roberts, R. D., & Kyllonen, P. C. (2017). Recent IRT approaches to test and correct for response styles in PISA background questionnaire data: A feasibility study. *Psychological Test and Assessment Modeling*, 59, 71–92.
- Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling*, 24, 524–544.
- Kim, E. S., Joo, S. H., Lee, P., Wang, Y., & Stark, S. (2016). Measurement invariance testing across between-level latent classes using multi-level factor mixture modeling. *Structural Equation Modeling*, 23, 870–887.
- Kim, S. H., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement*, 26, 25–41.
- Kolen, M. J., & Brennan, R. L. (2014). Nonequivalent groups: Linear methods. In *Test equating, scaling, and linking* (pp. 103–142). New York: Springer.
- Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness: A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika*, 79, 210–231.
- Mazzeo, J., & von Davier, M. (2014). Linking scales in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 229–258). CRC Publishing.

- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525-543.
- Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives*, 4, 5-9.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176.
- Neumann, K., Fischer, H. E., & Kauertz, A. (2010). From PISA to educational standards: The impact of large-scale assessments on science education in Germany. *International Journal of Science and Mathematics Education*, 8, 545-563.
- Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, 53, 315.
- Oliveri, M. E., & von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing*, 14, 1-21.
- Organization for Economic Co-Operation and Development. (2016). PISA 2015 technical report. Retrieved from <http://www.oecd.org/pisa/data/2015-technical-report>
- Organization for Economic Co-Operation and Development. (2019). PISA 2018 technical report. Retrieved from <http://www.oecd.org/pisa/data/2018-technical-report>
- Pokropek, A., Borgonovi, F., & McCormick, C. (2017). On the cross-country comparability of indicators of socioeconomic resources in PISA. *Applied Measurement in Education*, 30, 243-258.
- Rutkowski, D., Rutkowski, L., & Liaw, Y. L. (2018). Measuring widening proficiency differences in international assessments: Are current approaches enough? *Educational Measurement: Issues and Practice*, 37, 40-48.
- Rutkowski, L., Rutkowski, D., & Zhou, Y. (2016). Item calibration samples and the stability of achievement estimates and system rankings: Another look at the PISA model. *International Journal of Testing*, 16, 1-20.
- Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, 74, 31-57.
- Rutkowski, L., & Svetina, D. (2017). Measurement invariance in international surveys: Categorical indicators and fit measure performance. *Applied Measurement in Education*, 30, 39-51.
- Sachse, K. A., Roppelt, A., & Haag, N. (2016). A comparison of linking methods for estimating national trends in international comparative large-scale assessments in the presence of cross-national DIF. *Journal of Educational Measurement*, 53, 152-171.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.
- Steenkamp, J. B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, 25, 78-90.
- 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.
- von Davier, M. (2005). mdlm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models [Computer software]. Princeton, NJ:ETS.
- Wu, M. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practices*, 29, 15-27.