Psychological Test and Assessment Modeling, Volume 62, 2020 (1), 107-135

Comparability of response time scales in PISA

Hyo Jeong Shin[1], *Emily Kerzabi*[2], *Seang-Hwane Joo*[2], *Frederic Robin*[2] & *Kentaro Yamamoto*[2]

Abstract

The primary goal of this study was to explore the possibility of establishing common response time (RT) scales in the Programme for International Student Assessment (PISA) across participating countries and economies. We use categorized item-level RTs, which affords improved handling of non-lognormal RT distributions with outliers (observed in PISA RT data) and of missing data stemming from PISA's complex rotated booklet design. Categorized RT data were first analyzed using unidimensional multiple-group item response theory (IRT) models assuming a single latent trait in the RT data. Due to systematic patterns of misfit, the RT data were then analyzed using multidimensional multiple-group IRT models, in which RT scales were assumed to vary by item properties, specifically by item type or cognitive demand. Results indicate that PISA RT scales appear to be multidimensional by item type (multiple-choice and constructed-response). The present study provides implications for the analytical procedures involving RT in international large-scale assessments.

Keywords: Response time, process data, large-scale assessment, Programme for International Student Assessment (PISA), comparability, dimensionality, measurement invariance

[1] *Correspondence concerning this article should be addressed to:* Hyo Jeong Shin, PhD, Educational Testing Service, 660 Rosedale Road, 13-E, Princeton, NJ 08541, USA; email: hshin@ets.org

[2] Educational Testing Service, Princeton, USA

Introduction

One of the main goals of international large-scale assessments (ILSAs) is to produce group-level score results that are comparable across countries and cycles (Kirsch, Lennon, von Davier, Gonzalez, & Yamamoto, 2013). To that end, ILSAs fit a statistical model that integrates a variety of information about respondents, including responses to the cognitive assessment and contextual information, such as gender and socio-economic status (von Davier & Sinharay, 2014). This analytic procedure is called population modeling, which combines the item response theory (IRT) model and a latent regression model, and it has been used in the Programme for International Student Assessment (PISA) and in many other large-scale assessments. This methodology allows proficiency scales to be established for each major cognitive domain (e.g., mathematics, reading, and science) that are comparable across groups and cycles (Organisation for Economic Co-operation and Development [OECD], 2017; National Center for Education Statistics, 2018a, 2018b). With the introduction of computer-based assessments (CBA) for ILSAs and the extended research possibilities therein, there is a need to more closely investigate response time (RT) and other types of process data in ILSAs for the purpose of ensuring continued comparability and accurate interpretations across heterogeneous groups (Ercikan, Guo, & He, 2019). In this paper, we focus on the comparability of the RT scales across countries in PISA for two main reasons: to address possible data validity issues and to provide practical guidance for improving the population modeling of PISA.

In ILSAs, evaluating data quality is one of the first critical analysis steps because data validity issues may jeopardize the reliability of large-sale population surveys and potentially reduce the comparability across countries and cycles. Validity issues can range from errors in the data collection or data handling (e.g., incorrect coding of variables in the data files) to data fabrication issues (Yamamoto & Lennon, 2018) and must be fixed or accounted for before applying more complex data analysis or modeling approaches. RT and other process data, such as the number of actions, can provide very useful information for investigating suspicious or unexpected response patterns at the individual- or school- or country-level. Using RT data to identify potentially problematic individuals or (sub)groups (e.g., countries, schools) allows researchers and data analysts to account for possible data validity issues, which leads to more accurate and comparable results across groups (Yamamoto & Lennon, 2018). In addition to evaluating the validity of cognitive responses, RT and process data also can be used to provide insights into the test-taking strategies, motivation, and engagement of both individuals and groups (Goldhammer, Martens, Christoph, & Lüdtke, 2016; Lee & Haberman, 2016; Lee & Jia, 2014; Lee & Chen, 2011; Meyer, 2010; Shin & von Davier, 2018; Wise & Kong, 2005). Using raw RT variables can sometimes be sufficient for this purpose within a country. However, establishing common RT scales across countries, if possible, would be even more beneficial for evaluating data quality at the international level and for understanding and comparing test takers' behaviors across countries, with different cultures and languages.

Furthermore, an exploration of the comparability of RT scales could improve PISA population modeling by providing implications about how RT information can be incorporated into the model. In large-scale assessments, only a limited number of cognitive

items can be administered to each respondent within the testing time limit. However, incorporating contextual information (collected through a background questionnaire) in the latent regression model further reduces the measurement error associated with utilizing a limited number of cognitive item responses. Using estimated model parameters for skill distributions from a latent regression model, a set of multiple imputations, called plausible values, are generated and reported in the PISA database (OECD, 2017). Plausible values are not intended for making inferences about individual proficiency; however, they do provide less biased estimates of group-level proficiency distributions (Mislevy, 1991; von Davier, Gonzalez & Mislevy, 2009). Therefore, this methodology is known as "state-of-the-art" for secondary analyses in ILSAs (Braun & von Davier, 2017).

Given this methodology, a concerning issue arises when RT data are available in the public database – as in PISA – but were not incorporated in the generation of plausible values. In such cases, utilizing RT variables in secondary analyses may result in biased estimates of the relationship between proficiency measures (plausible values) and RT data (Meng, 1994; Mislevy, 1991). Therefore, there is a need to incorporate RT data in the population modeling when generating plausible values (von Davier, Khorramdel, He, Shin, & Chen, 2019). Recent studies (Shin, Yamamoto, Khorramdel, Robin, von Davier, Gamble, & Zhao, 2019; Shin, Jewsbury, & van Rijn, 2019) have shown the potential benefits of incorporating RT into the PISA population modeling: Inclusion of RT resulted in an improved predictive power of covariates and a substantial increase in measurement precision, for example, of about 16% in one country of the PISA 2015 data. However, there is a room for improvement, and little is known about how to best incorporate RT data in population modeling. Moreover, the nature of RT in low-stake cross-country skill surveys, like PISA, is vastly unexplored, specifically, how different types of items should be handled, and if the same pre-processing of RT variables can be applied across diverse groups. Note that if RT information were to be included in population modeling, it would need to be applicable for all countries, not just some (because RT information is available in the online public data file for all countries that administer PISA as a CBA). In this context, establishing common RT scales that are comparable across participating countries can be one way to understand the nature of RT data, as well as determine if there are any significant factors to be considered in processing RT data for secondary analyses. For example, using the predicted RT scale scores for individual respondents, such as a proxy for working speed, could serve as comparable and stable covariates in the population model. In an effort to establish comparable RT scales across countries, feasible data handling procedures can also be sought for how to process and incorporate RT information into the PISA population model that are applicable to all participating countries.

Therefore, this study aims to explore the possibility of establishing comparable RT scales across participating countries in PISA. To address the comparability of RT scales, we examine the measurement invariance of the item-level RT parameters across all countries. If measurement invariance of RT parameters holds for all items across all countries (the same slope and the same intercept parameters fit the items independent of the groups), we consider the RT scale comparable. To that end, we start with a unidimensional multiple-group IRT model with equality constraints across countries (i.e. item

parameters are constrained to be equal across countries), assuming a single latent trait underlying the RT data. Because RT scales could vary by item characteristics or properties (such as item type or cognitive demand), but remain comparable across countries for the given item characteristic, we further evaluate if multidimensional multiple-group IRT models, with invariance constraints across countries but different dimensions based on item characteristics, fit the data relatively better. Comparing unidimensional and multidimensional models allows the dimensionality of RT scales to be examined with regard to the item characteristics while simultaneously evaluating the measurement invariance of RT parameters across countries. When fitting these IRT models, the estimation of the item-level RT parameters was based on the categorized item-level RT data, which affords improved handling of non-lognormal RT distributions with outliers (observed in PISA RT data) and of missing data stemming from PISA's complex rotated booklet design (OECD, 2017).

In the following sections, we first introduce the challenges of studying RT in PISA, focusing on the distribution of RT data and the factors related to the dimensionality of RT scales. Then, we present our unidimensional and multidimensional modeling approaches and analysis procedures for examining the measurement invariance of RT parameters across countries, explaining the rationale behind them. Findings are followed by the implications of the study in the concluding remarks. Throughout the paper, we use the term *RT data* or *RT variables* as manifest variables that are observed at the item-level and *RT scales* as the latent person constructs that are estimated based on the item-level RT data.

Understanding response times in PISA

Although common proficiency scales for each domain have been established, to date, common RT scales have not been investigated, nor has the comparability of RT across countries been evaluated. It is a complicated issue to investigate the comparability of RT scales in PISA because 1) the assessment is not intended to measure RT scales and 2) RT data may not have identical distributions across countries. In particular, PISA RT data has the potential to be affected by item properties (such as item type: multiple-choice [MC] items and constructed-response [CR] items), assessment languages (administered in over 100 multiple languages), and country variations (more than 72 countries and economies participate).

One approach for solving issues surrounding the comparability of RT scales is to estimate RT parameters that are common across participating countries in the measurement model and to conduct a measurement invariance analysis across countries. In the next section, we explain two important challenges that we would like to address and investigate in this paper.

Distribution of RT data

The first fundamental challenge of modeling RT data is that the data are quite messy and the underlying distribution of RT data in PISA is not fully known. Most previous studies on modeling RT assumed a lognormal distribution (e.g., van der Linden, 2007). The lognormal distribution of RT data has been popular due to its simplicity in parameterization and interpretation; however, empirical RT distributions in PISA do not follow lognormal distributions and show varying patterns of outliers across items (e.g., Shin & von Davier, 2018). In this case, there have been studies that suggest different transformations including the Weibull distribution (e.g., Rouder, Sun, Speckman, Lu, & Zhou, 2003), the gamma distribution (e.g., Maris, 1993), and the Box-Cox transformation for RT data to assume a normal distribution (Klein Entink, van der Linden, & Fox, 2009).

Instead of these parametric approaches with specified distributions for RT data, we consider an alternative semiparametric approach, which categorizes the continuous RT variables into ordinal RT variables. This approach is referred to as semiparametric because the assumption on the RT distribution is less stringent than in the parametric lognormal distributions. The major reason for categorizing RT data is because an incorrectly specified distribution for RT data may result in biased parameter estimates (e.g., Molenaar, Bolsinova, & Vermunt, 2017). These studies argued that the semiparametric approach was proven to be effective in handling the typically non-lognormal RT distributions with outliers. Although, converting continuous RT variables to categorical RT variables does result in the loss of some information, we selected the semiparametric approach to model non-lognormal RT distributions with outliers observed in PISA RT data. This approach is particularly helpful when the best practices of preprocessing RT data (e.g., detection and treatment of outliers, transformation, and standardization of the RTs) are still under investigation. For example, outliers may have an impact on estimation, while removing outliers case-wise or replacing outliers with missing values (or expected values, or more sophisticated methods of imputations) will present different types of issues. Additionally, this approach also affords the use of the *mdltm* software (Khorramdel, Shin, & von Davier, 2019; von Davier, 2005) which is readily able to incorporate the complex features of PISA designs, including sampling weights and considerable missingness due to the rotated matrix sampling design.

Therefore, the present study utilizes categorized RT data at the item-level, with and without outliers. Among the various methods to identify and treat outliers, we selected the median absolute deviation (MAD)-based method, which is known to be one of the most robust methods (Rousseeuw & Croux, 1993) and has been used in PISA since the 2015 cycle. When categorizing RT data at the item-level, we apply the same thresholds across countries and language groups. A consistent threshold provides a more straightforward interpretation of categorized item-level RT data when the comparability of RT scales is examined across groups; however, this application is still under investigation. Recent studies on RT data from PIAAC and PISA revealed little evidence to support identifying item-by-country specific thresholds (Weeks, von Davier, & Yamamoto, 2016; Robin, Shin, Khorramdel, Yamamoto, & Pohl, 2018). This is mainly because country interactions do not appear to be necessary in the identification of thresholds

when the variability in RT data is accounted for by the examinees and items. We further examine the feasibility of this approach by analyzing two different categorizations of item-level RT (binary and five-category equal percentile) according to previous studies: equal-percentile, multiple percentiles (e.g., Molenaar et al., 2017) and binary categories based on the median values (e.g., Partchev & de Boeck, 2012).

Dimensionality of RT scales

When modeling observed RT data with the latent trait measurement model, the next challenge is determining the dimensionality of RT scales. Most literature has, so far, treated RT data as a unidimensional entity for a single latent construct (e.g., mental speed or processing power) that is measured through the test (e.g., van der Linden, 2007; Goldhammer & Klein Entink, 2011); however, while item difficulty may vary, it seems that only single item types or homogeneous sets of items were investigated in these studies.

In a mixed-format test, such as PISA, which uses a variety of item types (i.e., mixture of MC and CR) and elicits a wide range of cognitive demands to measure broad constructs, RT data can be affected by such item characteristics. For example, while different items naturally elicit shorter or longer raw RT, different item types are expected to produce different RT distributions. Assuming respondents are not guessing but solving the items, respondents are more able to rapidly respond to an MC item than to a CR item, where they must type multiple characters rather than select an appropriate response. Another interesting factor could be an item's cognitive demand. In the design framework, PISA's content developers defined three possible levels (high/medium/low) of cognitive demand required for solving Science items in the PISA 2015 cycle and refer to this variable as *depth of knowledge* (OECD, 2016, p. 41). It is expected that students spend more time on high-demanding items and less time on low-demanding items given the same level of proficiency.

We view these two factors, item type and cognitive demand, as the major sources that can violate the measurement invariance of RT parameters in PISA. One can consider modeling the effects of item type and cognitive demand on the RT data by treating them as fixed-effects (e.g., through time intensity parameters) in the generalized linear modeling framework. This approach may provide more straightforward parameterization and interpretation of main effects and interaction effects of those two factors. However, in this study, we are more concerned with the possibility that the effect of item type or cognitive demand could differ across respondents, sampled from a widely heterogeneous population (from different countries where different languages are spoken). If properly modeled and handled, predicted RT scale scores for individual respondents would be comparable across participating countries; thus, they can serve as more comparable and stable characteristics of students in the population modeling, for example, as proxies of working speed. Therefore, we have attempted to fit the multidimensional RT models to examine the possibility that RT scales measure different constructs by different item types and levels of cognitive demand.

Research questions

Two main research questions are investigated in this study using categorized RT data in PISA.

1. Is it possible to establish a unidimensional RT scale in PISA that is comparable across participating countries and economies? In other words, does the measurement invariance of RT hold (i.e., RT parameters fit well across countries) across participating groups under the unidimensional model?
2. If not, are there alternative multidimensional RT scales that provide meaningful interpretations and implications for PISA analytical procedures?

The first research question is addressed through a unidimensional model assuming a single latent trait underlying the RT data. We first fit the unidimensional multiple-group IRT model using the categorized RT data and estimate the item-level RT parameters. Then, we evaluate the item fit statistics for each group and for each item to see if measurement invariance of RT parameters holds across participating groups. If the same RT slope and intercept parameters fit well across countries and language groups, scalar or strong invariance can be assumed, and the unidimensional RT scale would be supported (Millsap, 2010). However, if there is any systematic misfit pattern observed in the RT data under the unidimensional model (the second research question), it is reasonable to investigate the possibility of alternative multidimensional RT scales to provide a more meaningful and comparable interpretation of RT data. This will also provide implications for the PISA analytical procedures with regards to RT data treatment.

Methods

Data

We use RT data from the PISA 2015 Science domain, which was the major domain in that cycle, and included over 50 CR items along with over 100 MC items.[3] Based on the balanced incomplete block design (OECD, 2017), the major domain of Science was administered to all participating students and yielded extensive RT data. There were 184 cognitive CBA Science items in the PISA 2015 cycle in total. However, there was one pair of items presented together on the same screen, so only one RT variable was available. This resulted in 183 RT variables in total, from 98 newly developed items (65 MC items and 33 CR items) and 85 trend items (54 MC items and 31 CR items) that had been administered in previous cycles. Tables 1 and 2 present the classifications of the RT variables by the item type (MC vs. CR) and by a depth of knowledge (cognitive demand; classified into high/medium/low).

[3] The major domains for PISA included a larger number of items that were both new and trend. The minor domains included a small number of items and were only trend.

Table 1*:*
Classification of CBA science items from the PISA 2015 main survey by item type

	Multiple Choice	Constructed Response
New	65	33
Trend	54	31
Total	**119**	**64**

Table 2:
Classification of CBA science items from the PISA 2015 main survey by cognitive demand

	Low	Medium	High
New	26	64	8
Trend	30	48	7
Total	**56**	**112**	**15**

When categorizing the item-level RT data, two different approaches were attempted: 1) use of five categories of equal percentiles, and 2) use of two categories (e.g., slow vs. fast) split along the item-specific median RT values. Following Molenaar et al. (2017), we attempted these two different methods of categorization to examine the robustness of the results. Furthermore, we also looked at the effects of outliers by 1) including outliers (also labelled as 'not censored') and 2) excluding the outliers based on the MAD-based method, which is operational in PISA (also labelled as 'censored as missing'). MAD was determined based on the international data, which pooled countries' RTs and converted outliers to missing values. Thus, excluding the outliers relies on a strong assumption that RT outliers occur randomly (missing at random). Taken together, this resulted in four different datasets, dependent on the number of categories of RT data and the inclusion of the outliers: five-category RT data excluding outliers, five-category RT data including outliers, binary RT data excluding outliers, and binary RT data including outliers.

Measurement model: Multiple-group IRT model

For each of the four datasets, we first fit the unidimensional multiple-group IRT model (Bock & Zimowski, 1997; von Davier & Yamamoto, 2004) based on the two-parameter logistic model (2PL) for the binary RT data and the generalized partial credit model (GPCM; Muraki, 1992) for the five-category RT data. This is the same approach that was operationally performed at the IRT-based scaling stage in the PISA 2015 cycle (OECD, 2017). The unidimensional multiple-group IRT model enables the estimation of RT parameters (i.e., slopes and intercepts) that are common across different populations, as well as unique group means and standard deviations. Let j denote a person in group k

responding in category h of item i, and suppose there are K groups and a test composed of n items. Assuming conditional independence of responses, the probability of observing the pattern of response ($X_j = [X_{1j}, X_{2j}, \ldots, X_{nj}]$) can be written as:

$$P\left(X_j \mid \theta\right) = \prod_{i}^{n} P_i(X_{ij} = h \mid \theta), \tag{1}$$

which applies to all groups and persons, given the person attribute θ. For analyses in this study, the *mdltm* software was used (Khorramdel et al., 2019; von Davier, 2005), which provides marginal maximum likelihood estimation via the expectation-maximization algorithm (Sundberg, 1974, 1976; also Dempster, Laird & Rubin, 1977).

When RT parameters are estimated for individual items, either the RT parameters can be constrained to be the same across different groups or allowed to be unique for each group. A latent person ability, or attribute θ, follows a normal distribution with a finite mean and variance in the population of persons corresponding to group k. With the probability density function denoted as $g_k\left(\theta\right)$, the marginal probability of response pattern X_j in group k can be expressed as

$$\overline{P_k}\left(X_j\right) = \int_{-\infty}^{\infty} P\left(X_j \mid \theta\right) g_k\left(\theta\right) d\theta \,. \tag{2}$$

In this study, we consider three different sub-grouping strategies to appropriately account for the effects of languages and countries on RT data. Overall, there were 55 CBA countries with RT data available (PBA participants did not collect RT) with 39 unique languages in PISA 2015; this resulted in a maximum of 84 unique country-by-language groups. The first approach was to model the full 84 possible country-by-language groupings; this separated all unique country-by-language groups, even those with small samples (less than 250 participants). Next, the PISA 2015 groupings were modeled, which contained 78 country-by-language groups; this grouping combined small samples (typically the minor language in a country) with the larger sample of the same country. To investigate the impact of language, the final assessed grouping was solely based on language, resulting in 39 different language groups. That is, the same language groups across different countries were treated as the same group, expecting that the dominant significant grouping factor in RT variables was language, but the differences across countries would be negligible. We also used the final sample senate weights so that each group contributed equally in estimating the RT parameters.

Testing the measurement invariance of RT parameters

To test the measurement invariance of RT parameters across groups, we used the fit statistics obtained from the unidimensional multiple-group IRT model to see if scalar invariance (i.e., same slopes and intercepts across groups) holds when the RT scale is assumed to be unidimensional. We use the root mean square deviation (RMSD) that is calculated for each group and for each item against the common RT parameters. This

quantity has proved to be useful in testing the measurement invariance assumptions across country-by-language groups for responses on cognitive assessments and on background questionnaires (von Davier et al., 2019; Buchholz & Hartig, 2017).

RMSD is computed for each group based on the deviation between the observed and expected item characteristic curves (ICCs) as below.

$$\text{RMSD} = \sqrt{\int \left[P_{obs}(\theta) - P_{exp}(\theta) \right]^2 f(\theta) d\theta} \tag{3}$$

In Equation 3, $P_{obs}(\theta)$ represents the observed ICC, and $P_{exp}(\theta)$ represents the expected ICC given ability θ. In addition, $f(\theta)$ indicates the group-specific density distribution on the ability scale. The observed ICCs are obtained from the pseudo observed response counts across students computed from the MML-EM algorithm, and the expected ICCs are computed from the IRT model with the estimated item parameters. The integrals in Equations 2 and 3 are generally approximated with Gaussian quadrature points (Bock & Aitkin, 1981; von Davier, 2005). RMSD ranges from 0 to 1, and a large RMSD value indicates that the item does not fit the model.

In several studies with ILSAs (e.g., Buchholz & Hartig, 2017; Oliveri & von Davier, 2011, 2014), the threshold for determining a misfit was an RMSD > 0.10 for cognitive domains and an RMSD > 0.30 for noncognitive domains. Similar thresholds were used in operational IRT-scaling for the PISA 2015 cycle (OECD, 2017): an RMSD > 0.12 for the cognitive domains and an RMSD > 0.30 for the noncognitive domains. We chose a threshold of 0.20, between the thresholds set for the cognitive and noncognitive domains, but this selection is somewhat arbitrary since there is not yet a rule of thumb to evaluate the fit statistics of RT data.

If the same RT slope and intercept parameters fit well across groups, measurement invariance holds under the unidimensional RT scale, supporting a comparable RT scale across groups. Conversely, if any systematic pattern is observed in the item fit statistics, multidimensional modeling of RT data may provide more meaningful and comparable RT scales. For this, we would consider item type and cognitive demand as two potential factors relating to RT scale dimensionality.

Results

Following the order of our research questions, we first present the results from the unidimensional measurement model for the RT scale in PISA. In particular, we focus on testing measurement invariance using item fit statistics under the unidimensional RT scale. We then present two alternative types of multidimensional measurement models for RT scales that may provide comparable interpretations and meaningful implications for the PISA analytical procedures.

Measurement invariance of RT parameters under the unidimensional RT scale

First, we fit the unidimensional multiple-group IRT model to four different datasets with three types of different grouping methods, which resulted in 12 fitted models (3 country/language grouping methods [39, 78, 84] * 2 categories [binary fast/slow, 5-equal RT percentiles] * 2 methods of outlier treatment [MAD-based excluding outliers, including outliers]). Tables 3, 4, 5, and 6 with the first column (# of Dim) as "Unidim" present the model fit statistics resulting from the unidimensional models using the Akaike information criterion (AIC; Akaike, 1974), "Consistent" AIC (CAIC; Bozdogan, 1987), and the Bayesian information criterion (BIC; Schwarz, 1978). Due to the rotated matrix sampling design administered in PISA 2015, which inherently bears considerable missingness by design, BIC_sp and CAIC indices, which adjust for sample size, were deemed most appropriate for the sparse data (Khorramdel et al., 2019). Model fits are presented by the number of RT data categories: five-category data in Tables 3 and 4 and two-category (binary) data in Tables 5 and 6, each depending on the treatment of outliers. Only the fitted models presented in the same table are comparable in terms of model fits.

Among the results of the unidimensional model, the most notable pattern is that the group of 39 (combining different countries with the same language) always showed the worst fit. This implies that RT data are not solely affected by language factors, and we need to take into account the differences from countries as well. This observation is consistent with that of Lee and Haberman (2016), who found that test-taking strategies vary between respondents from different countries. In their study, Chinese and Korean participants tended to speed up or slow down depending on the item, whereas French and German participants progressed at a relatively steady rate. Hence, more variability in RT data was observed in the pacing of test-takers from China and Korea than in the pacing of test-takers from France and Germany. Considering such language and country differences is in line with the IRT-based scaling procedures of the PISA cognitive domains. Note that the groups of 78 and 84 both consider differences across countries, and that the only difference is the consideration of sample size.[4] Although the grouping of 84 by country/language showed better model fit in the unidimensional models with five-category data, fit statistics were quite close to the group of 78 in many conditions. Therefore, in the following sections, we focus on the results from the 78 groups used in PISA 2015.

Next, the measurement invariance of RT parameters was tested when the unidimensional measurement model was fit to the RT data with 78 groups. Among the 78 groups, fit statistics were evaluated for 66 groups with a sufficient number of responses per item (N>100). Note that fit statistics from the unidimensional model can be most informative to see if there are any meaningful patterns in the fit statistics that can provide implications for processing the RT data in the PISA analytical procedures.

[4] 84 groups differentiated multiple language groups for the given country regardless of the sample size, while 78 groups combined those minor language groups with major language groups if sufficient sample size was not reached (N=250). Minor language groups were combined with major language groups for six countries (Hong Kong [China], Ireland, Italy, Luxembourg, Macao [China], Sweden).

Table 3:
Five-category RT data excluding outliers: Comparison of model fit statistics

# of Dim	# of Groups	# of Parameters	AIC	BIC	BIC_sp	CAIC_sp	Log Penalty	Log Penalty (ind)	% of Improvement
Unidim	39	991	23556696	23567120	23565270	23566261	1.522	1.609	90.011
	78	1069	23546251	23557496	23555500	23556569	1.521	1.609	90.718
	84	1081	23546125	23557496	23555478	23556559	1.521	1.609	90.728
2 (Item Types)	78	1459	23407930	23423277	**23420553**	**23422012**	**1.512**	1.609	100
	84	1501	23407688	**23423477**	23420675	23422176	1.512	1.609	100
3 (Cognitive Demand)	78	1849	23698306	23717755	23714303	23716152	1.531	1.609	
	84	1921	23698228	23718435	23714849	23716770	1.531	1.609	

Table 4:
Five-category RT data including outliers: Comparison of model fit statistics

# of Dim	# of Groups	# of Parameters	AIC	BIC	BIC_sp	CAIC_sp	Log Penalty	Log Penalty (ind)	% of Improvement
Unidim	39	991	25304926	25315350	25313568	25314559	1.527	1.609	88.015
	78	1069	25293961	25305206	25303283	25304352	1.527	1.609	88.736
	84	1081	25293 850	25293859	25303277	25304358	1.527	1.609	88.744
2 (Item Types)	78	1459	25120801	25136148	**25133524**	**25134983**	**1.516**	1.609	100
	84	1501	25120491	**25136280**	25133580	25135081	1.516	1.609	100
3 (Cognitive Demand)	78	1849	25466494	25485944	25482618	25484467	1.537	1.609	
	84	1921	25466404	25486611	25483156	25485077	1.537	1.609	

Table 5:
Binary RT data excluding outliers: Comparison of model fit statistics

# of Dim	# of Groups	# of Parameters	AIC	BIC	BIC_sp	CAIC_sp	Log Penalty	Log Penalty (ind)	% of Improvement
Unidim	39	991	442	9830212	9834861	9834036	9834478	0.693	84.31
	78	1069	520	9821813	9827282	9826312	9826832	0.693	85.116
	84	1081	532	9821721	9827317	9826323	9826855	0.693	85.126
2 (Item Types)	78	1459	910	9664465	**9674037**	**9672339**	**9673249**	0.693	100
	84	1501	952	**9664204**	9674218	9672441	9673393	0.693	100
3 (Cognitive Demand)	78	1849	1300	9791611	9805286	9802859	9804159	0.693	
	84	1921	1372	9791469	9805901	9803340	9804712	0.693	

Table 6:
Binary RT data including outliers: Comparison of model fit statistics

# of Dim	# of Groups	# of Param.	AIC	BIC	BIC_sp	CAIC_sp	Log Penalty	Log Penalty (ind)	% of Improvement
Unidim	39	991	442	10545619	10550269	10549474	10549916	0.693	82.367
	78	1069	520	10535956	10541425	10540490	10541010	0.693	83.229
	84	1081	532	10535875	10541471	10540514	10541046	0.693	83.239
2 (Item Types)	78	1459	910	10345811	**10355383**	**10353747**	**10354657**	0.693	100.000
	84	1501	952	**10345515**	10355529	10353817	10354769	0.693	100.000
3 (Cognitive Demand)	78	1849	1300	10506876	10520550	10518212	10519512	0.693	
	84	1921	1372	10506637	10521069	10518601	10519973	0.693	

Figure 1 shows the histogram of countries based on the counts of misfitting items (RMSD > 0.20) when the binary RT data were used with and without outliers. This pattern is quite consistent when the five-category RT data were used; thus, results from the five-category RT data are not presented in this paper. Overall, nearly 60 groups out of 66 groups exhibited fewer than 10 misfitting items out of 183 Science items given the threshold of RMSD > 0.20. Nearly half of those 60 groups showed no misfits at all: 34 groups when outliers were excluded and 28 groups when outliers were included. Most groups did not threaten the measurement invariance of RT parameters resulting from the unidimensional model; however, some groups had a relatively larger number of misfits for more than 5.5% of items (10 out of 183 items). Five out of eight groups using Chinese or Chinese-based writing (Japanese) showed 10-26% misfitting RT items, Singapore (English) had about 12% of misfitting RT items, and the other two groups, Brazil (Portuguese) and the Dominic Republic (Spanish), each showed about 6% of misfitting RT items.

Proportions of misfitting RT items and proportions by item types (CR or MC) for these eight groups are presented in Table 7. One notable finding is that a systematic pattern appears by item type, whether the item is a CR or MC, for those groups with a larger misfit. When we looked at the histograms of RMSD values by item type (Figures 2 and 3), Brazil and the Dominic Republic showed relatively higher proportions of misfits for MC items, while Chinese language groups (Hong Kong, Macao, China, and Taiwan) and Japan, as well as Singapore (English), demonstrated much higher proportions of misfits

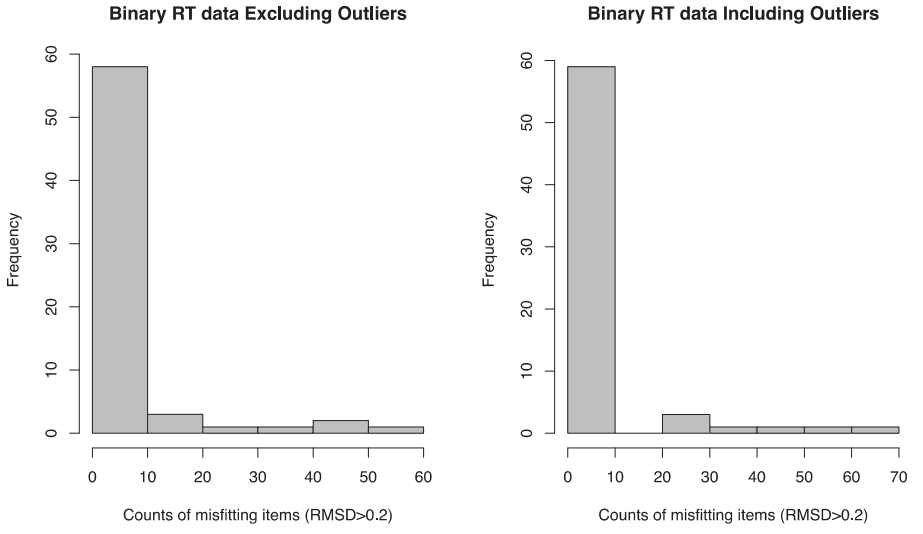
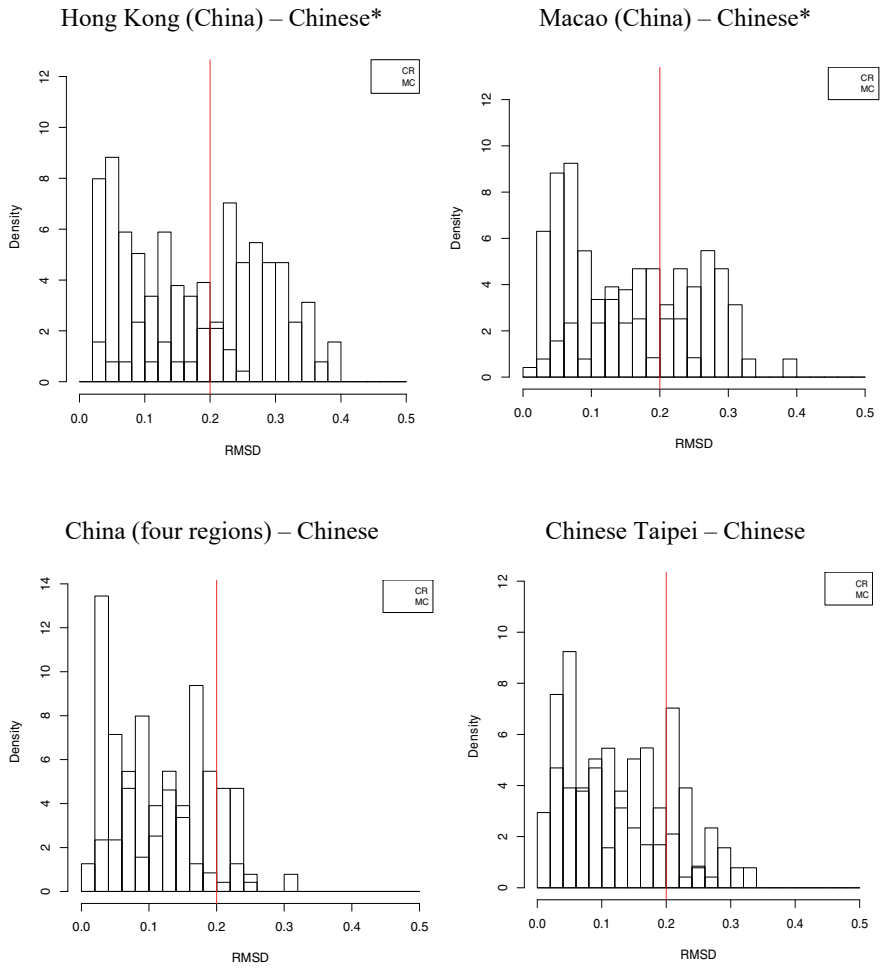


Figure 1:
Histogram of misfit counts across countries from the unidimensional model for the Binary RT data

Table 7:
Percentage (%) of misfitting RT items by Item Types in eight groups with larger number of misfits

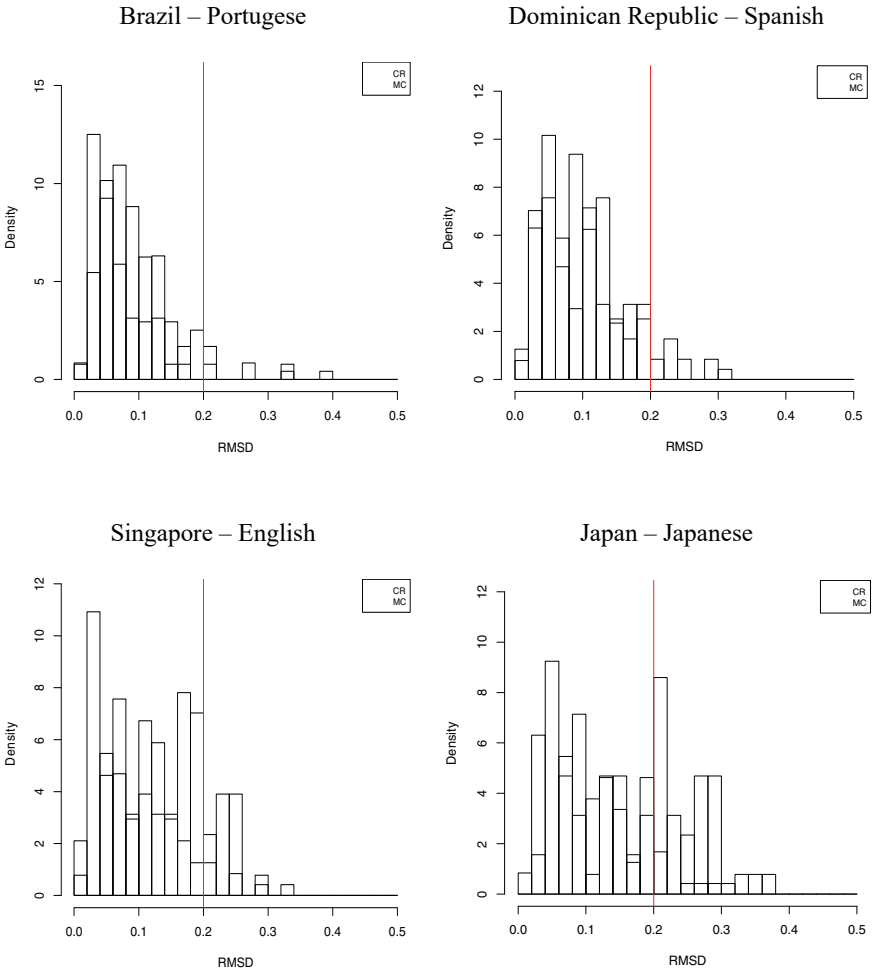
Country	Language	Excluding Outliers			Including Outliers		
		Total (183 items)	CR (64 items)	MC (119 items)	Total (183 items)	CR (64 items)	MC (119 items)
Brazil	Portuguese	6.01	3.13	6.72	3.28	1.56	4.20
Dominic Republic	Spanish	6.56	0.00	9.24	14.21	9.38	16.81
Hong Kong	Chinese*	31.15	73.44	7.56	34.97	81.25	10.08
Japan	Japanese	22.95	51.56	6.72	24.04	53.13	8.40
Macao	Chinese*	26.78	53.13	11.76	31.15	62.50	14.29
China	Chinese	10.93	21.88	4.20	11.48	25.00	4.20
Singapore	English	12.02	21.88	5.88	12.02	25.00	5.04
Taiwan	Chinese	17.49	34.38	7.56	16.94	39.06	5.04

Note. Hong Kong and Macao had two languages of assessment; Chinese was the major language of assessment in both groups but RT data from 121 English-speaking respondents (2.3%) in Hong Kong and from 41 Portuguese-speaking respondents (1.2%) in Macao are included in this table. The Japanese language also uses Chinese-based characters in writing.



Note: Hong Kong and Macao had two languages of assessment; Chinese was the major language of assessment in both groups but RT data from 121 English-speaking respondents (2.3%) in Hong Kong and from 41 Portuguese-speaking respondents (1.2%) in Macao are included in these figures.

Figure 2:
Histogram of RMSD in *eight groups with larger number of misfits*
where Chinese was used as language of assessment.



Note: The Japanese language also uses Chinese-based characters in writing.

Figure 3:
Histogram of RMSD in *eight groups with larger number of misfits* where other languages were used as language of assessment

Table 8:
Percentage (%) of misfitting RT items by Cognitive Demand in eight groups with the greatest number of misfits

Country	Language	Excluding Outliers				Including Outliers			
		Total (183 items)	Low (56 items)	Medium (112 items)	High (15 items)	Total (183 items)	Low (56 items)	Medium (112 items)	High (15 items)
Brazil	Portuguese	6.01	16.07	0.89	0.00	3.28	10.71	0.00	0.00
Dominic Republic	Spanish	6.56	14.29	2.68	0.00	14.21	19.64	12.50	6.67
Hong Kong	Chinese*	31.15	17.86	33.93	53.33	34.97	23.21	38.39	53.33
Japan	Japanese	22.95	10.71	25.89	40.00	24.04	12.50	27.68	40.00
Macao	Chinese*	26.78	19.64	26.79	46.67	31.15	19.64	33.93	53.33
China	Chinese	10.93	7.14	8.93	33.33	11.48	7.14	9.82	40.00
Singapore	English	12.02	10.71	10.71	20.00	12.02	8.93	10.71	33.33
Taiwan	Chinese	17.49	14.29	16.96	26.67	16.94	8.93	18.75	33.33

Note. Hong Kong and Macao had two languages of assessment; Chinese was the major language of assessment in both groups but RT data from 121 English-speaking respondents (2.3%) in Hong Kong and from 41 Portuguese-speaking respondents (1.2%) in Macao are included in this table. The Japanese language also uses Chinese-based characters in writing.

for CR items[5]. For example, when the outliers were excluded, the Dominic Republic showed misfits exclusively for MC items, while Japan showed misfits for 51.5% of CR items and 6.7% for MC items.

Misfit patterns in RT data by cognitive demand are also interesting (Table 8). In particular, Brazil and the Dominic Republic showed relatively higher proportions of misfits for low-level cognitive demand items. Chinese language groups, as well as Japan and Singapore, showed rather evenly distributed misfits across the three levels of cognitive demand, although slightly more misfits were observed among high-level cognitive demand items. One thing to note here is that high-level cognitive demand items were mostly CR items (10 out of 15), and low-level cognitive demand items were mostly MC items (46 out of 56). This may suggest potential interactions between cognitive demand and item type, but the main effects from each factor are not clear yet. In general, it was not enough to conclude that the measurement invariance of RT parameters holds across different countries and language groups, given the higher proportion of misfits in those countries or certain language groups. Although we have focused on eight groups with prominent misfit patterns, other countries may also benefit from having more comparable and interpretable RT scales when the item type and cognitive demand is considered. Therefore, we evaluate if the multidimensional multiple-group IRT model fits better when such item properties are taken into account.

Multidimensionality of RT scales by item types in PISA

The model fit statistics are presented in the same series of tables (Tables 3 to 6) with the first column *(# of Dim*) as *2 (Item Types)* for item types, and *3 (Cognitive Demand)* for levels of cognitive demand. The results indicate that, regardless of the treatment of outliers (excluding or including), the number of data categories (binary or five-category), or the number of country/language groups (39, 78, 84), BIC_sp and CAIC indices always favored the two-dimensional model by item types. The difference in model-fit improvement based on the Gilula and Haberman (1994) log-penalty measure was considerable and supported the two-dimensional model by item types as well. The unidimensional model ranged from 82.3% to 90.7% model-fit improvement over the baseline model (independence) compared to the more general two-dimensional model. The three-dimensional model by cognitive demand always showed worse fit, suggesting that multiple RT scales by cognitive demand are not necessary to describe the RT data. Together with the evaluation of fit statistics to test for measurement invariance, it seems reasonable to conclude that the RT scales appear to be differentiated mainly between MC items and CR items in the PISA mixed format test and misfits observed by cognitive demand have no systematic pattern when item type is taken into account.

[5] Further interesting comparison can be made for Japanese and Korean against Chinese. Although these two languages (Japanese and Korean) are somewhat related with Chinese characters, the dependence on Chinese characters in reading and writing on the computer are very different between the two countries: Korean is purely phonetic and does not use any Chinese characters in item questions or answers, while Japanese is quite character-dependent and uses many Chinese characters.

When the two-dimensional model by item type was fitted, most significant misfits observed from the unidimensional model disappeared, except in the Dominic Republic. In particular, counts of misfitting RT items among Chinese-speaking countries were considerably reduced: from 57 to 8 in Hong Kong, from 49 to 8 in Macao, from 22 to 2 in China, and from 32 to 6 in Taiwan. Although less obvious than in Chinese-speaking countries, the same pattern was observed in the other countries with high misfit under the unidimensional model: from 11 to 8 in Brazil, from 42 to 10 in Japan, and from 22 to 9 in Singapore. Moreover, there were only four groups whose counts of misfitting RT items increased slightly: Slovakia, Luxemburg, Switzerland, and Canada (French); in the unidimensional model, these groups did not show any misfits, but one misfitting RT item appeared in the two-dimensional model in these groups. When the remaining misfitting RT items were monitored, no systematic pattern could be found; thus, items could be more sensitive in terms of RT data behavior in certain countries or language groups, just as item-by-country interactions (unique parameters) are allowed for items in the PISA cognitive domains (OECD, 2017). Taken together, it seems reasonable to say that measurement invariance holds when RT scales are separated by item type, suggesting that the comparability of RT scales can be established by item type.

Further, robustness of the results was examined by inclusion or exclusion of outliers and the number of RT data categories. Figure 4 presents a comparison of the group-level means of RT scales depending on the treatment of outliers and the number of categories

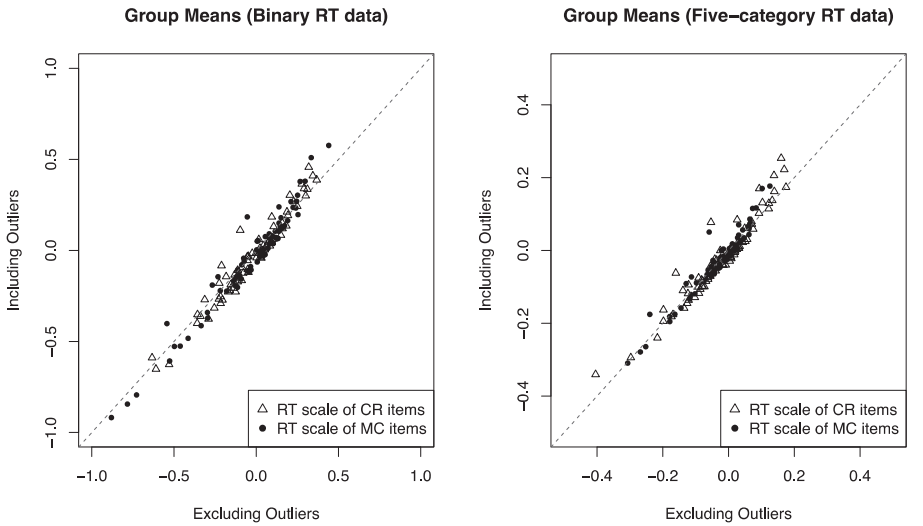


Figure 4:
Comparison of group means by censoring and number of categories.

in the data.[6] More negative group means indicate faster responding behavior of that group compared to the other groups. Because the two outlier treatment methods and two types of categorizations resulted in different datasets, direct comparison would not be applicable; thus, only the qualitative tendency of the results could be examined. As two panels in Figure 4 show, there is a strong linear relationship, and the rank order of countries is generally consistent regardless of the censoring or categorization method.

Interestingly, as implied by the best fitting two-dimensional model by item type, the three fastest responding countries were different by item types: Korea, the Netherlands, and Qatar were consistently the fastest responding countries for CR items (empty triangles located on the left bottom of each panel), while Hong Kong, Korea, and Taiwan were consistently the fastest responding for MC items (solid dots located on the left bottom of each panel). As for the slowest responding countries, Brazil and Peru were consistently identified as such on MC items (solid dots located on the right top of each panel), but no countries were identified as being consistently the slowest on CR items in terms of group means.

Furthermore, in order to investigate if fast responding was solely related to the language administered in the test or country, group means of the same language and multiple language groups per country were compared. For example, a comparison of Dutch-speaking groups in the Netherlands and Belgium indicated that the Netherlands was one of the fastest responding, but Belgium was not. In another example, a comparison of the Qatar English-speaking group and the Qatar Arabic-speaking group indicated that the Arabic-speaking group was consistently fast, but the English-speaking group was not. Therefore, the fast responding pattern on items seems to be an interaction of language and other various aspects of the country, which was also suggested by several remaining misfits in the RT data.

The latent correlations between the RT scales of CR items and MC items were also estimated. The distribution of correlations by categorization methods and censoring methods are summarized in Table 9 and Figure 5. Except for one case with five-category RT data without outliers (third row in Table 7), the range of distributions appears consistent: The lowest latent correlation among 78 groups was about 0.2, the mean and median were about 0.5, and the highest correlation was about 0.7. Overall, all groups showed moderate to strong linear associations in the RT scales of CR items and MC items. That is, RT scales measured by different item types are quite distinct, and each RT scale provides somewhat unique information that is not captured by the RT scale of another item format. One interesting finding here is that Chinese-speaking groups (China, Hong Kong, and Macao) consistently showed the lowest correlations, less than 0.3 between item types. This suggests that Chinese-speaking groups are unique in yielding different RT patterns when responding to MC items and CR items. In a similar vein, Chinese character-related language groups (China, Hong Kong, Japan, Macao, and Taiwan) showed the largest differences in their group means between MC items and CR items.

[6] In order to make two latent group means comparable – RT scale based on MC items and RT scale based on CR items – group means of RT scale based on the MC items were aligned using linear transformations.

Table 9:
Distribution of Latent Correlations Between RT scales of CR Items and MC Items

Outliers	Categorization	Min	Mean	Median	Max
Excluded	Binary	0.243	0.525	0.532	0.697
Included	Binary	0.187	0.506	0.516	0.666
Excluded	Five-category	0.367	0.607	0.615	0.769
Included	Five-category	0.246	0.570	0.576	0.686

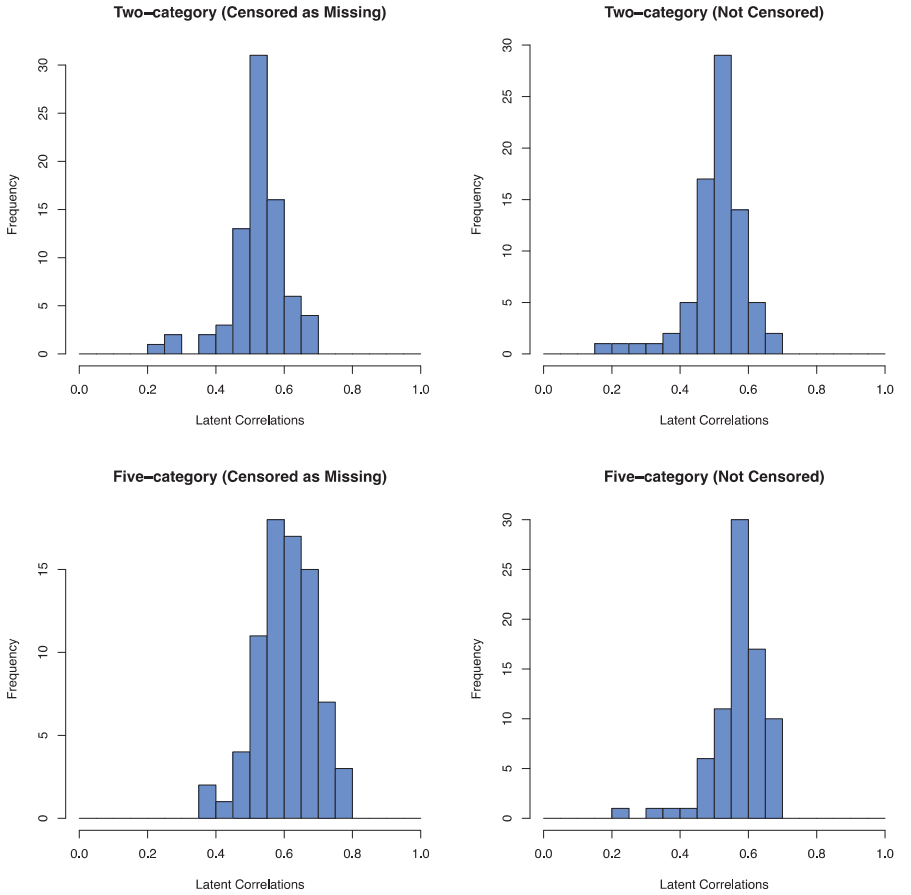


Figure 5:
Distribution of latent correlations between item types by censoring methods and categorization methods.

In summary, given that the two-dimensional model is the best fitting model and that the highest latent correlation of RT scales between item types is lower than 0.7 for most cases, a multidimensional measurement model for the RT data by item type would contribute to the comparability of RT scales for all participating countries. Furthermore, the uniqueness in groups using Chinese-based characters seems to be better treated by fitting a two-dimensional model by item type.

Conclusion and discussion

The current study aims to examine the possibility of generating RT scales in PISA that are comparable across countries for their utilization in data quality assurance and for obtaining practical guidance for possibly improving the population model. RT data have been available in the PISA public data files for secondary analysis since CBA became the mode of test administration for the majority of participating countries. We examined the comparability of RT scales across countries by evaluating the measurement invariance of RT parameters and by accounting for item properties as different dimensional structures. Results indicate that a unidimensional RT scale, assuming a single latent trait underlying the RT data, cannot be supported across all items and participating country-by-language groups in PISA. Measurement invariance of the RT parameters under the unidimensional multiple-group IRT model was threatened by huge misfits observed in some country-language groups, which appeared systematic by item type, particularly in Chinese-based character language groups. Alternatively, the two-dimensional model by item type (different working speed given the type of item) was the better fitting model. Both model fit statistics and latent correlations showed that RT scales measured by different item types are quite distinct. Therefore, the present study suggests that analytical procedures in PISA involving RT should consider item types to increase the comparability of the RT scales: separate working speed by MC items and CR items.

Although the present study drew some implications for the comparability of RT scales in PISA, we also recognize the limitations of the study. One of the major limitations is related with the categorization of RT data. Creating categories based on equal percentiles may increase the differences where abundant data points are available: RT variables with minor differences, only 1 or 2 seconds, could be classified into different categories. In addition, intercept parameters of RT remained uninterpretable due to equal-percentiles applied at the item-level. In fact, the current approach was not appropriate to estimate time intensity parameters (i.e., expected processing time of the item) on which many previous studies have focused. The categorization of RT in this study is amenable to the present undertaking, but other alternative methods can be sought. For example, fractioning RT variables by equidistant intervals (e.g., 10 seconds) may preserve the level information, and time intensity parameters (corresponding to the item difficulty parameters) could remain meaningful.

Another limitation is related with the arbitrary selection of RMSD thresholds. As Pokropek, Borgonovi, & McCormick (2017) revealed, it is more difficult to obtain the comparability of the scales in noncognitive assessment because students from different coun-

tries have a different understanding of questionnaires based on their country, language, and cultural backgrounds. The same argument can be made for the RT data, where RT distributions are subject to item types, language, and cultural backgrounds, as shown in the present study. That is the main reason why we chose an RMSD threshold of 0.20, the relative mid-point threshold between cognitive and noncognitive assessments.

Finally, alternative methodologies and data from different PISA cycles would be useful to see if the current findings are generalizable and would enrich the interpretation of findings in the present study. In particular, the present study modeled the effects of item type and cognitive demand in the dimensional structure of RT scales with a focus on estimating latent correlations between them and evaluating the fit statistics for testing measurement invariance. On the other hand, specifying them as fixed-effects in the generalized linear modeling framework would provide a more straightforward interpretation of the main effects and interaction effects of those two factors. Furthermore, with the multiple cycles of CBA available (PISA 2015 and PISA 2018), the comparability of RT scales across cycles within a country would be interesting to assure the stability of processing RT data by item types.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. doi:10.1109/TAC.1974.1100705

Braun H. & von Davier, M. (2017). The use of test scores from large-scale assessment surveys: psychometric and statistical considerations. *Large-scale Assessments in Education, 5*(17) doi:10.1186/s40536-017-0050-x

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), 241–250. doi:10.1177/0146621617748323

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 46–443. doi:10.1007/BF02293801

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). doi:10.1007/978-1-4757-2691-6_25

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*(3), 345–370. doi:10.1007/BF02294361

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B, 39*(1), 1–38. doi:10.1.1.133.4884

Ercikan, K, Guo. H, & He, Q. (2019). Use of response process data in large-scale assessments for cross cultural comparisons. Presented at the panel session at the 2019 Comparative and International Education Society. San Francisco, CA.

Gilula, Z., & Haberman, S. J. (1994). Conditional log-linear models for analyzing categorical panel data. *Journal of the American Statistical Association, 89*(426), 645–656. doi:10.1080/01621459.1994.10476789

Goldhammer, F., & Klein Entink, R. H. (2011). Speed of reasoning and its relation to reasoning ability. *Intelligence, 39,* 108–119. doi:10.1016/j.intell.2011.02.001

Goldhammer, F. Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC*, OECD Education Working Papers 133. Paris, France: OECD Publishing. doi:10.1787/5jlzfl6fhxs2-en

Khorramdel L., Shin H., & von Davier M. (2019) GDM Software *mdltm* Including Parallel EM Algorithm. In M. von Davier & Y.S. Lee. (Eds.), *Handbook of Diagnostic Classification Models: Methodology of Educational Measurement and Assessment*. Springer.

Kirsch I., Lennon M., von Davier M., Gonzalez E., & Yamamoto K. (2013). *On the Growing Importance of International Large-Scale Assessments.* In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), *The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research*. Springer, Dordrecht

Klein Entink, R. H., van der Linden, W. J., & Fox, J.-P. (2009). A Box-Cox normal model for response times. *British Journal of Mathematical and Statistical Psychology, 62*, 621–640. doi:10.1348/000711008X374126

Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling, 53*(3), 359–379. Retrieved from: https://psycnet.apa.org/record/2011-28090-006

Lee, Y.-H., & Haberman, S. (2016). Investigating Test-Taking Behaviors Using Timing and Process Data. *International Journal of Testing*, *16*(3), 240-267. doi:10.1080/15305058.2015.1085385

Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-Scale Assessments in Education, 2*(8). doi:10.1186/s40536-014-0008-1

Maris, E. (1993). Additive and multiplicative models for gamma distributed variables, and their application as psychometric models for response times. *Psychometrika, 58,* 445–469. doi:10.1007/BF02294651

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science, 9*, 538–558. Retrieved from https://www.jstor.org/stable/2246252

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement, 34*, 521–538. doi:10.1177/0146621609355451

Millsap, R. E. (2010), Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives, 4*, 5–9. doi:10.1111/j.1750-8606.2009.00109

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177–196. doi:10.1007/BF02294457

Molenaar, D., Bolsinova, M., & Vermunt, J. (2017). A semi-parametric within-subject mixture approach to the analyses of responses and response times. *British Journal of Mathematical and Statistical Psychology, 72*(2), 205–228. doi:10.1111/bmsp.12117

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2). doi:10.1177/014662169201600206

National Center for Education Statistics (2018a). *2017 NAEP reading report card*. Washington, DC: Institute of Education Sciences, U.S. Department of Education. Retrieved from https://www.nationsreportcard.gov/reading_2017

National Center for Education Statistics. (2018b). *2017 NAEP mathematics report card.* Washington, DC: Institute of Education Sciences, U.S. Department of Education. Retrieved from https://www.nationsreportcard.gov/math_2017/

Oliveri, M., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling, 53*(3), 315. Retrieved from https://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011_20110927/04_Oliveri.pdf

Oliveri, M., & von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing 14*(1), 1–21. doi:10.1080/15305058.2013.825265

Organisation for Economic Co-operation and Development (2016). *PISA 2015 assessment and analytical framework: Science, reading, mathematical and financial literacy.* Paris, France: OECD Publishing. doi:10.1787/9789264255425-en

Organisation for Economic Co-operation and Development (2017). *PISA 2015 technical report*. Paris, France: OECD Publishing. Retrieved from http://www.oecd.org/pisa/data/2015-technical-report/

Partchev, I., & De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence, 40*(1), 23–32. doi:10.1016/j.intell.2011.11.002

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. doi:10.1080/08957347.2017.1353985

Robin, F. Shin, H., Khorramdel, L., Yamamoto, K., & Pohl, S. (2018). Examining item specific response time thresholds for the coding of omitted responses in PISA. Paper presented at the 11th International Test Commission Conference (ITC), Montreal, Canada.

Rouder, J., Sun, D., Speckman, P., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika, 68*(4): 589–606. doi:10.1007/BF02295614

Rousseeuw, P. J., & Croux C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association, 88*(424), 1273–1283. doi:10.2307/2291267

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–464. Retrieved from https://projecteuclid.org/euclid.aos/1176344136

Shin, H., Jewsbury, P., & van Rijn, P. (2019). *Conditional Dependencies between Cognitive Responses and Response Times in Large-scale Assessments.* Poster presented at the Foundational and Applied Statistics and Psychometrics (FASP) Initiative, Educational Testing Service, Princeton, NJ.

Shin, H., & von Davier, M. (2018). Mixture models for response accuracy and categorized response times. Paper presented at the International Meeting of Psychometrics Society (IMPS), New York, NY.

Shin, H., Yamamoto, K., Khorramdel, L., Robin, F., von Davier, M., Gamble, H., & Zhao, W. (2019, April). Incorporating response time into the PISA population model. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Toronto, Canada.

Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics, 1*(2), 49–58. Retrieved from https://www.jstor.org/stable/4615553

Sundberg, R. (1976). An iterative method for solution of the likelihood equations for incomplete data from exponential families. *Communications in Statistics—Simulation and Computation, 5*(1), 55–64. doi:10.1080/03610917608812007

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*. doi:10.1007/s11336-006-1478-z

von Davier, M. (2005). mdltm [computer software]. Princeton, NJ: Educational Testing Service.

von Davier, M., Gonzalez, E. & Mislevy, R. (2009) What are Plausible Values and why are they useful? In: M. von Davier & D. Hastedt (Eds.): *IERI Monograph Series: Issues and Methodologies in Large Scale Assessments, Vol. 2*. Obtained from: http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02_Chapter_01.pdf

von Davier, M. & Sinharay, S. (2014). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 155-175). Boca Raton, FL: Chapman Hall/CRC.

von Davier, M., Khorramdel, L., He, Q., Shin, H., & Chen, H. (2019). Developments in psychometric population models for data from innovative large-scale assessments. *Journal of Educational and Behavioral Statistics*. doi:10.3102/1076998619881789

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement, 28*(6), 389–406. doi:10.1177/0146621604268734

von Davier, M., Yamamoto, K., Shin, H., Chen, H., Khorramdel, L., Weeks, J., Davis, S., Kong, N., & Kandathil, M. (2019). Evaluating item response theory linking and model fit for data from PISA 2000-2012. *Assessment in Education: Principles, Policy & Practice.* doi:10.1080/0969594X.2019.1586642

Weeks, J., von Davier, M., & Yamamoto, K. (2016). Using response time data to inform the coding of omitted responses. *Psychological Test and Assessment Modeling, 58*(4), 671–701.

Wise, S., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2). doi:10.1207/s15324818ame1802_2

Yamamoto, K., & Lennon, M. L. (2018). Understanding and detecting data fabrication in large-scale assessments. *Emerald Insight, 26*(2), 196–212. doi:10.1108/QAE-07-2017-0038