OECD Education Working Papers No. 244

**PISA-Based Test
for Schools: International
Linking Study 2020**

**Tomoya Okubo,
Tanja Bastiniac,
Tiago Fragoso,
Chi Sum Tse**

<https://dx.doi.org/10.1787/ef1356ae-en>

DIRECTORATE FOR EDUCATION AND SKILLS

Cancels & replaces the same document of 12 January 2021

**PISA-BASED TEST FOR SCHOOLS: INTERNATIONAL LINKING STUDY
2020**

OECD Education Working Paper No. 244

Tomoya OKUBO, Tanja BASTIANIC, Tiago FRAGOSO, Chi Sum TSE

This cancel and replace has been issued to amend the working paper number.

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

Joanne CADDY, Joanne.CADDY@oecd.org
Fiorella CIANCHI, Fiorella.CIANCHI@oecd.org
Tomoya OKUBO, Tomoya.OKUBO@oecd.org

JT03470218

Acknowledgement

The authors would like to thank Margaret Wu, Francesco Avvisati, and Bonaventura Francesco Pacileo for their constructive comments and support on the international linking study project. Fiorella Ciani provided editorial support for the final draft.

Abstract

An international linking study was conducted in order to link parameters of PISA-based Test for Schools (PBTS) cognitive items to PISA international scales. New booklets for the linking study were designed in which the PISA trend items were inserted as anchor items in addition to the PBTS items. Data was collected from four countries with over 95 000 students via Computer-Based Testing, and analysed with the finite mixture modelling in order to estimate the parameters of PBTS items under the constraint of fixed PISA item parameters. The estimated item parameters were validated in terms of reliability and international comparability. The linking study enabled the PBTS test to provide valid and reliable scores on PISA international scale.

Résumé

Une étude de comparabilité internationale a été menée afin de relier les paramètres des items cognitifs de l'Évaluation PISA pour les établissements scolaires (PBTS, l'acronyme anglais) aux paramètres internationaux du PISA. De nouveaux livrets ont été conçus pour cette étude, incluant des items d'ancrage issu du PISA et des item du PBTS. Les données ont été collectées auprès de quatre pays et rassemblent plus de 95 000 étudiants qui ont participé aux épreuves informatisés. Elles sont analysées à l'aide d'un modèle de mélange fini afin d'estimer les paramètres des items cognitifs du PBTS sous la contrainte de paramètres d'items cognitifs du PISA donnés. Les paramètres estimés de ces items remplissent les critères de fiabilité et de comparabilité internationale. L'étude de comparabilité a permis au PBTS de fournir des scores valides et fiables sur l'échelle internationale du PISA.

Table of Contents

<i>Acknowledgement</i>	2
<i>Abstract</i>	3
<i>Résumé</i>	3
<i>Table of Contents</i>	4
Figures	5
1. Introduction	6
1.1. Background of PISA-based Test for Schools (PBTS)	6
1.2. Objective of international linking study 2020	6
1.3. Overview of the 2020 international linking study	7
2. Survey design	8
2.1. Data collection	8
2.2. Anchor item cluster	8
2.3. PBTS item clusters	9
2.4. Booklet design	11
3. Scaling model	13
3.1. Item response model and its parameterisation	13
3.2. The multinomial logistic regression model and its parametrisation	15
3.3. Finite mixture modelling with known classes	16
4. Linking procedures	19
4.1. Equating model	19
4.2. Item validation	20
5. Results	22
5.1. Item fit	22
5.2. Item comparability	23
5.3. Estimated item parameters	26
6. Discussion	31
References	32
Appendix	35

Tables

Table 1. The number of participants in international linking study	8
Table 2. The number of items and units of PISA anchor item clusters	8
Table 3. The number of PBTS items	10
Table 4. The number of items and units in each item cluster in the PBTS	10
Table 5. Item clusters and their presenting order	12
Table 6. The proportion of the anchor items to each item cluster	13
Table 7. Excluded anchor items from analysis	22
Table 8. Excluded PBTS items from analysis	22
Table 9. PBTS items which have national parameters	26
Table 10. International linking study booklets (1A, 1B, and 1C)	35
Table 11. International linking study booklets (2A, 2B, and 2C)	35
Table 12. International linking study booklets (3A, 3B, and 3C)	36
Table 13. International linking study booklets (4A, 4B, and 4C)	36
Table 14. International linking study booklets (5A, 5B, and 5C)	37
Table 15. International linking study booklets (6A, 6B, and 6C)	37

Figures

Figure 1. Test information functions of the PISA anchor item cluster	9
Figure 2. Test information functions of the PBTS booklets	11
Figure 3. Item characteristic curve of a MATH item without a national parameter	23
Figure 4. Item characteristic curve of a READ item without a national parameter	24
Figure 5. Item characteristic curve of a MATH item with a national parameter	25
Figure 6. Estimated MATH item parameters in scatterplot	28
Figure 7. Estimated READ item parameters in scatterplot	29
Figure 8. Estimated SCIE item parameters in scatterplot	30

1. Introduction

1.1. Background of PISA-based Test for Schools (PBTS)

The PISA-based Test for Schools (PBTS) is an international educational assessment used by schools to support international benchmarking and school improvement. The PBTS provides schools and policy-makers with descriptive information and analyses on the skills and creative application of knowledge of 15-year-old students in mathematics (MATH), reading (READ), and science (SCIE). The scores calculated based on the student responses are fully comparable to PISA scales. It also provides information on how different factors within and outside school relate with scores. A student questionnaire forms part of the assessment.

Cognitive items (MATH, READ and SCIE) were developed under the PISA assessment framework in 2012 (OECD, 2013^[1]) by the Australian Council for Educational Research (ACER) which was responsible for the item development of the first four PISA cycles. Over 400 cognitive items were developed by ACER, and they were checked through psychometric analysis with data obtained in field trials. At the end, 140 items out of 400 items were chosen after confirming their validity and reliability. Further, the student questionnaire mainly contains two components: questions on the student's family background and school experience, and subject-specific questions. The items of the questionnaire are the same to those of the PISA survey. See PBTS technical report (OECD, 2016^[2]) for details of the context questionnaire.

To ensure comparability between the PBTS scores with the PISA scores, linking of mathematics, reading and science to the PISA scales was undertaken in 2012 using statistical methods. In 2012, the mathematics and science items were scaled on the PISA 2003 and PISA 2006, respectively by anchoring the PISA-linked items at their 2003 and 2006 item parameters, where tests were delivered in paper-and-pencil mode at the time. The reading items were placed on the PISA 2000 scale by adjusting the PISA 2009 item parameters by a linear transformation (ACER, 2012^[3]).

In 2019, PBTS transited from paper-and-pencil delivery to computer-based testing (CBT). In accordance with the transition, minor adaptation for cognitive items such as formatting were conducted in order to adopt the items to CBT environment. Furthermore, PISA 2015 changed the types item response theory (IRT) models from the Rasch model (Rasch, 1960^[4]) and partial credit model (Masters, 1982^[5]) to the two-parameter logistic model (Lord and Novick, 1968^[6]) and generalised partial credit model (Muraki, 1992^[7]), since concerns over the insufficiencies of the models which do not satisfy the condition of slope parameter had been raised (Oliveri and von Davier, 2014^[8]; Kreiner and Christensen, 2014^[9]). Consequently, a linking study was expected to be undertaken in accordance with the transition of the test delivery mode and the calibration models of PISA and PBTS.

1.2. Objective of international linking study 2020

The purpose of the international linking study (the Study, hereafter) was to estimate the parameters of the PBTS items on the PISA international scale so that the PBTS' scores could be fully compatible with PISA scores. In the wave of PISA 2015, PISA's main mode of delivery was transitioned to CBT delivery from paper-based testing (PBT). The IRT models for scaling was also updated (OECD, 2017^[10]). As mentioned above, PISA for Schools project estimated its item parameters in 2012; hence, item parameters of the PBTS items also needed to be rescaled on the PISA international scale calibrated in PISA 2018.

Moreover, the digitalised PBTS items would need to be validated in countries that participated in the international linking study.

This technical report aims at providing technical details of the linking procedures and equating methodology. This report consists of three parts: the first part is on survey design, especially booklet design; the second part describes the scaling model; and the third part shows the linking procedures. This report concludes with the results of the linking study.

1.3. Overview of the 2020 international linking study

There are 140 items in the current PBTS item bank, distributed as 40 MATH items, 46 READ items and 54 SCIE items. Items are clustered within units, and there are 62 units in total. As mentioned above, the aim of the linking study was at calibrating the 140 items (62 units) on the PISA international scale. 36 PISA cognitive items were inserted into booklets so that the PBTS cognitive items could be presented together with the PISA items to students. All the linking study booklets included the same cluster of PISA items (i.e., anchor items) in order to link different booklets via the anchor items (i.e., common item design). The anchor items were selected from the PISA item bank scaled in PISA 2015 under the CBT delivery mode. In section 2, it will be explained how the linking study booklets were assembled.

In the Study, over 95 000 students from 4 countries (Brazil, Spain, Russian Federation, and United States) participated. Each student was assigned randomly one of the linking study booklets that contained both the PBTS items and the PISA items. The four countries had different proficiency levels of participants; further, there was a possibility that the items worked differently in each country since the languages of the items were different. To incorporate these issues and take such behaviour into account, a statistical modelling was employed (see section 3).

Parameters of the PISA items were already estimated and validated in the past PISA cycles. Therefore, during this Study only parameters of the PBTS items were estimated with the linking study data under the fixed parameters of the PISA items. By fixing the PISA item parameters at their international values, the estimated PBTS item parameters will be comparable to the fixed PISA parameters as they are estimated on the same scale of the fixed parameters. It was also analysed the linking study data obtained from four different countries in order to check measurement invariance of the items among the countries. The procedures utilised for checking for invariance are presented in section 4.

The results of the analyses are shown in section 5. It can be seen how the statistical modelling that was applied fits to the international linking study data. Further, it was confirmed comparability validity of the estimated parameters of the PBTS items.

2. Survey design

2.1. Data collection

In the Study, data was collected from Spain, United States, Brazil, and Russian Federation between October 2019 and February 2020. Table 1 shows the number of schools and the number of students that participated in the Study. In total, over 95 000 students participated in it. Data was collected by a representative sample in Russian Federation and convenient samples from the other three countries. Student responses of cognitive items and student questionnaires were collected strictly following the procedures from the PISA for Schools Technical Report (OECD, 2016_[2]).

Table 1. The number of participants in international linking study

	Schools	Students
Brazil	231	13 645
Spain	33	2 351
Russian Federation	1 899	80 811
United States	23	1 698

During the data analysis, student responses from each country were weighted so that the four countries contributed equally to the likelihood function when estimating parameters. This means that the results of statistical analysis are not affected by large sample countries.

2.2. Anchor item cluster

In order to estimate new item parameters of PBTS items on the PISA international scale, it was inserted PISA items with known item parameters. In addition, item clusters for each domain were constructed. Table 2 shows the number of anchor items and units of domains. In each domain, the number of anchor items was set to be more than ten. On the other hand, the number of PBTS items in a booklet was set not to exceed 10. Thus, the total number of items in a booklet that students took was 60 or below, and the percentage of the anchor items in a booklet reached 60%. The criteria were set in order to keep the content validity of the anchor items and to obtain reliable parameter estimates of PBTS items (Vale et al., 1981_[11]; McKinley and Reckase, 1981_[12]; Wingersky and Lord, 1984_[13]). Consequently, the number of anchor items were set 11, 12, and 13 for MATH, READ, and SCIE, respectively. Details of the selection process are shown below.

Table 2. The number of items and units of PISA anchor item clusters

PISA item	MATH	READ	SCIE	Total
N of items	11	12	13	36
N of units	6	4	5	15

The PISA anchor items were selected by the following criteria:

- The proportion of multiple-choice items and self-constructed items should be almost the same with that of PBTS.
- The test information function of the selected items should widely cover (Wingersky and Lord, 1984_[13]) scores around the average of the population scores in each domain, and form a symmetrical unimodal distribution around the population scores.
- The test information function of the selected items should be higher than any other combinations of the items around the target student proficiency levels.
- Anchor items should be selected from units that have been presented for over 20 countries in the PISA main survey. In addition to it, the anchor items should have been taken by the four countries.
- Percentage of inequivalent parameters of the selected items (i.e., “unique item parameters” and “excluded from scaling” in PISA item parameters table) over the PISA participating countries should be below 20%. The criterion was set to minimise the effect of unique item parameters based on a simulation study by Jurich, et al. (2012_[14]).

Therefore, the selected anchor items were expected to be the best combination of the items in terms of reliability, under the above-mentioned criteria. Definition of the information function is provided in the next section (Equation 2).

Figure 1 shows the test information functions of the selected anchor items. The test information functions of the anchor item clusters covered a wide range of the ability/trait. Scores are around 500, which is the centre of the scale representing the average on each PISA domain set in the first PISA cycle. The test information function of SCIE is the highest due to the higher number of items in the SCIE compared to the other two domains.

Figure 1. Test information functions of the PISA anchor item cluster

2.3. PBTS item clusters

In PISA-based Test for Schools (PBTS), the number of items for each domain is as shown in Table 3. There are 140 items (62 units) in the current PBTS item bank.

Table 3. The number of PBTS items

	Math	Reading	Science	Total
<i>Items</i>	40	46	54	140
<i>Units</i>	25	13	24	62

For PBTS item clusters, items were divided into six clusters in each domain so that it satisfied the conditions mentioned above. The number of items in a cluster in MATH was 6-7, and that in READ and SCIE were 7-9 and 8-10, respectively. Three item clusters from three domains were combined and assembled in one combined cluster. The combined clusters consisted of 23 or 24 items (Table 4).

Table 4. The number of items and units in each item cluster in the PBTS

	PBTS item cluster						
Math	Ma01	Ma02	Ma03	Ma04	Ma05	Ma06	Total (N)
<i>Items</i>	7	7	7	6	6	7	40
<i>Units</i>	4	5	4	3	4	5	25
Reading	Re01	Re02	Re03	Re04	Re05	Re06	Total (N)
<i>Items</i>	7	7	8	9	8	7	46
<i>Units</i>	2	2	2	3	2	2	13
Science	Sc01	Sc02	Sc03	Sc04	Sc05	Sc06	Total (N)
<i>Items</i>	10	9	8	8	10	9	54
<i>Units</i>	4	4	4	4	4	4	24
Total (Items)	24	23	23	23	24	23	140
Total (Units)	10	11	10	10	10	11	62

An important point to discuss is the method of making equivalent booklets (i.e. parallel booklets). A heuristic method is used to divide items into some levels according to their difficulty, and to assign these to each booklet so that the proportion of items at each level will be equivalent among the booklets. However, this method does not ensure equivalence between booklets as it does not take item discriminations into account. Furthermore, this method is not able to propose the best B equivalent booklets in terms of test reliability. Preferably, the booklets should be equivalent not only in reliability (information function), but also in proportion of the sub scales (area, domain) and average time limits of items, etc.

However, it is impossible to manually construct B booklets that have the highest information functions and are equivalent among booklets, under many constraints. The best method to make B best equivalent booklets is to use an automatic assembly algorithm for booklet generation (Luecht, 1998^[15]; van der Linden and Adema, 1998^[16]; van der Linden, 2005^[17]). This method allows to generate equivalent B booklets from an item bank under the conditions that test administrators require. In general, the booklets will be generated so that the information functions of the forms overlap each other. This method has to be employed when constructing booklets in the linking study. Concretely, the Maximum Clique Algorithm proposed by Ishii et.al. (2015^[18]) was used for this Study.

In order to make PBTS booklets as equivalent as possible among them, it was employed an automatic assembly algorithm in which shape of information function and the proportion

of multiple-choice items were taken into account in the optimisations. The test information functions of the booklets are presented in Figure 2. The information functions were calculated based on the previous version of the PBTS item parameters.

Most of the test information functions of the item clusters overlap each other within each domain. However, item information function of Re05 is higher than the rest. This is because there are only 13 units for READ in PBTS; therefore, it was not possible to constitute perfectly equivalent booklets among the booklets given the available item bank and other assembly constraints. On the other hand, the test information functions of SCIE booklets overlap each other.

Figure 2. Test information functions of the PBTS booklets

2.4. Booklet design

The 140 PBTS items have been divided into 6 item clusters as shown in Table 4. An item cluster is a set of items in a domain, while a test booklet is a set of item clusters which contains all items that a student takes. The number of items in a booklet could be different among the item clusters, because units contain one or more items that should be presented to the students consecutively.

For example, item clusters Ma01-Ma06 consisted of 6-7 math items, Re01-Re06 consisted of 7-9 items, and Sc01-Sc06 consisted of 8-10 items in each cluster. In addition, we set the number of anchor items for math, reading, and science as 11, 12, and 13, respectively. We also set the number of PBTS items not to exceed 10, thus the proportion of the anchor items exceeded 60%, which was considered sufficient to obtain valid and reliable estimates of the PBTS item parameters. Only one item cluster was prepared for the anchor items in each domain (i.e., MaAn, ReAn, and ScAn), making anchor items common to all booklets in each domain. A booklet consisted of the anchor item cluster plus one of the PBTS item clusters; consequently, the total number of items that students took was 59-60 for a booklet.

The most important thing here is that although the number of items are different among booklets, the information functions of the booklets overlap to each other as shown in Figure 2. This implies the booklets can be considered as equivalent in terms of reliability of the measurement.

In order to balance order effects by the domain location, we rotated the order of the domains when forming booklets. It was thus formulated three different presenting order regarding domains. Consequently, this assembly resulted in 18 (6 item clusters by 3 presenting orders) booklets. Table 5 shows the assignment for the item clusters for the 18 booklets. Furthermore, the proportions of the anchor items are shown in Table 6. Details of the booklets are presented in the Appendix.

Table 5. Item clusters and their presenting order

Booklet	1st session	2nd session	3rd session
1A	Ma01&MaAn	Re01&ReAn	Sc01&ScAn
1B	Sc01&ScAn	Ma01&MaAn	Re01&ReAn
1C	Re01&ReAn	Sc01&ScAn	Ma01&MaAn
2A	Ma02&MaAn	Re02&ReAn	Sc02&ScAn
2B	Sc02&ScAn	Ma02&MaAn	Re02&ReAn
2C	Re02&ReAn	Sc02&ScAn	Ma02&MaAn
3A	Ma03&MaAn	Re03&ReAn	Sc03&ScAn
3B	Sc03&ScAn	Ma03&MaAn	Re03&ReAn
3C	Re03&ReAn	Sc03&ScAn	Ma03&MaAn
4A	Ma04&MaAn	Re04&ReAn	Sc04&ScAn
4B	Sc04&ScAn	Ma04&MaAn	Re04&ReAn
4C	Re04&ReAn	Sc04&ScAn	Ma04&MaAn
5A	Ma05&MaAn	Re05&ReAn	Sc05&ScAn
5B	Sc05&ScAn	Ma05&MaAn	Re05&ReAn
5C	Re05&ReAn	Sc05&ScAn	Ma05&MaAn
6A	Ma06&MaAn	Re06&ReAn	Sc06&ScAn
6B	Sc06&ScAn	Ma06&MaAn	Re06&ReAn
6C	Re06&ReAn	Sc06&ScAn	Ma06&MaAn

Table 6. The proportion of the anchor items to each item cluster

	PBTS item cluster					
	Ma01	Ma02	Ma03	Ma04	Ma05	Ma06
Math						
Proportion of anchor items	61.1%	61.1%	61.1%	64.7%	64.7%	61.1%
Reading						
Proportion of anchor items	63.2%	63.2%	60.0%	57.1%	60.0%	63.2%
Science						
Proportion of anchor items	56.5%	59.1%	61.9%	61.9%	56.5%	59.1%
Total	60.0%	61.0%	61.0%	61.0%	60.0%	61.0%

3. Scaling model

3.1. Item response model and its parameterisation

In this Study, the generalised partial credit model (GPCM) and the 2-parameter logistic model (2PL) were employed for graded response items and binary response items, respectively. Note that some PISA anchor items are modelled with the partial credit model (PCM) and the Rasch model in which discrimination parameter of item j (a_j) is fixed to 1.0. The followings are definitions of the IRT models used in PISA.

In the GPCM, the response probability of category k ($1, 2, \dots, K_j$) of item j by student i is described as:

- Equation 1

$$p_{jk}(\theta_i) = \frac{\exp\left(\sum_{k'=0}^k Da_j(\theta_i - (b_j - d_{jk'}))\right)}{\sum_{l'=0}^{K_j-1} \exp\left(\sum_{l'=0}^l Da_j(\theta_i - (b_j - d_{jl'}))\right)}$$

where D is the scale constant, and is set to 1.7 in PISA. Here, we set $d_{j0} = 0$ and $\sum_{l'=1}^{K_j-1} d_{jl'} = 0$ as to avoid location indeterminacy. Note that it holds

$$\sum_{k=0}^{K_j-1} p_{jk}(\theta_i) = 1$$

for all the items.

If the number of categories of item j (K_j) is two, the 2PL is employed in which $p_{j1}(\theta_i)$ is defined as:

$$p_{j1}(\theta_i) = \frac{\exp(Da_j(\theta_i - b_j))}{\exp(0) + \exp(Da_j(\theta_i - b_j))}$$

while $p_{j0}(\theta_i)$ is expressed as:

$$p_{j0}(\theta_i) = \frac{\exp(0)}{\exp(0) + \exp(Da_j(\theta_i - b_j))}$$

As mentioned above, the PCM and the Rasch are special cases of GPCM and 2PL in which discrimination parameter is imposed to 1.0. IRT parameters of PBTS are all expressed as GPCM and 2PL with the scale constant $D = 1.7$.

As Equation 1 shows, the IRT models have indeterminacy between distributions of θ and the item parameters (i.e., factor indeterminacy), meaning the item parameters are dependent on the distribution of θ .

In PISA, the item parameters are scaled with $\theta \sim N(0, 1)$ of samples from OECD countries in past PISA cycles (See chapter 12 of the OECD (2017_[10]) for details.). Because of the factor indeterminacy problem, $\theta \sim N(0, 1)$ was set in general when estimating item parameters for the first time. However, in the Study, the item parameters of PISA items were fixed in order to avoid the factor indeterminacy problem, instead of setting $\theta \sim N(0, 1)$. Details of the model are shown in section 4.

The item information function ($I_j(\theta)$) of the GPCM is calculated as follows:

- Equation 2

$$I_j(\theta) = D^2 a_j^2 \sum_{k=0}^{K_j-1} k'^2 p_{jk} - D^2 a_j^2 \left(\sum_{k'=0}^{K_j} k p_{jk} \right)^2$$

Furthermore, an information function of a set of items, $I(\theta)$, can be calculated as $I(\theta) = \sum_{j=1}^J I_j(\theta)$ where J denotes the total number of items in the set of items. The reciprocal the root of the information function ($1/\sqrt{I(\theta)}$) represents standard error at θ .

3.2. The multinomial logistic regression model and its parametrisation

Here, a multinomial logistic regression (MLR) model parametrisation of the IRT models will be introduced. This type of parametrisation is used in some software and R packages (Mplus, TAM, mirt, etc.) as a default setting. Therefore, it is important to confirm the default model parametrisation adopted by the software of choice for the scaling.

The MLR parametrisation is:

- Equation 3

$$p_{jk}(\theta_i) = \frac{\exp(\sum_{k'=0}^K (\lambda_j \theta_i - \tau_{jk'}))}{\sum_{l=0}^{K-1} \exp(\sum_{l'=0}^l (\lambda_j \theta_i - \tau_{jl'}))}$$

where λ and τ are called slope and threshold, respectively. If a software such as mirt package is used, which employs the multinomial logistic regression parametrisation, the following MLR to GPCM transformations are needed:

- Equation 4

$$\begin{aligned} a_j &= \lambda_j \sqrt{\sigma^2} \\ b_j &= \frac{\sum_{k'=0}^{K_j-1} (\tau_{jk'} - \lambda_j \mu)}{\lambda_j \sqrt{\sigma^2} (K_j - 1)} \\ d_{jk} &= \frac{\sum_{k'=0}^{K_j-1} (\tau_{jk'} - \lambda_j \mu)}{\lambda_j \sqrt{\sigma^2} (K_j - 1)} - \frac{(\tau_{jk} - \lambda_j \mu)}{\lambda_j \sqrt{\sigma^2}} \end{aligned}$$

Note that $d_{j0} = 0$. Further, μ and σ^2 are parameters of normal distribution of the target population. On the contrary, transformation from GPCM parametrisation to MLR parametrisation is described as follows:

- Equation 5

$$\begin{aligned} \lambda_j &= \frac{a_j}{\sqrt{\sigma^2}} \\ \tau_{jk} &= a_j (b_j - d_{jk}) + \frac{a_j}{\sqrt{\sigma^2}} \mu \end{aligned}$$

Generally, they are set to 0 and 1, respectively. Consequently, the GPCM model (Equation 1) is equivalent to the MLR model¹ (Equation 3).

In the model parameterisation in PISA and PBTS, the IRT parameterisation is employed; however, in case the transformation is needed, the above equations need to be applied (Equation 4 or Equation 5) according to the software or the R package that is used.

¹ On the other hand, the graded response model (Samejima, 1969^[33]) is equivalent to the categorical factor analysis model. Those parameters can also be converted to each other.

3.3. Finite mixture modelling with known classes

In the Study, it was employed the finite mixture modelling (multi-group analysis) in order to estimate both parameters of the PBTS items and the parameters of the group distribution simultaneously, where the classes of the model are treated as known information.

Since the participating countries had different levels of proficiencies, it was necessary to take the proficiency levels of group into account so that the PBTS item parameters could be estimated adequately. Furthermore, the finite mixture modelling allows to estimate item parameters for each group (e.g., country, booklet, etc.) if needed. These points were imperative conditions for the choice of statistical modelling.

Hence, for the modelling performed exposed in this section, the log-likelihood function of the finite mixture IRT modelling with known classes is presented below. Assuming local (or conditional) independence, the probability of observing response vector $\mathbf{u}_i$ from student i with latent trait θ is described as:

$$p(\mathbf{u}_i|\theta, \boldsymbol{\zeta}) = \prod_{j=1}^J p(u_{ij}|\theta, \zeta_j)$$

where ζ_j is the parameter vector of item j .

Here, it is considered the case in which a student is drawn from one of G different groups ($g = 1, 2, \dots, G$) in which θ in group g follows a continuous distribution $h_g(\theta|\xi_g)$ where ξ_g is the parameter vector of the distribution h_g . The group trait distribution (h_g) can be any continuous distribution, such as a beta distribution or a normal distribution (Thissen and Steinberg, 1997^[19]; Baker and Kim, 2004^[20]). In the linking study analyses, it was employed a normal distribution for h_g . Therefore, $\xi_g = (\mu_g, \sigma_g)$, representing each groups' unknown mean and standard deviation parameters.

The marginal probability of $p_g(u|\zeta, \tau_g)$ in group g represents the expected probability of a response pattern in the group and is described as:

$$p_g(\mathbf{u}|\zeta_g, \xi_g) = \int_{-\infty}^{+\infty} p_g(\mathbf{u}|\theta, \zeta_g) h_g(\theta|\xi_g) d\theta$$

where ζ_g denotes the item parameters of group g . This implies parameters of item j can be different among groups thus describing differential behaviour between groups.

Furthermore, the probability of observing the joint response data $p(\mathbf{U}|\boldsymbol{\zeta}, \boldsymbol{\xi})$ throughout all groups can be expressed as the multinomial distribution:

$$p(\mathbf{U}|\boldsymbol{\zeta}, \boldsymbol{\xi}) = \prod_{g=1}^G \frac{N_g!}{\prod_{m=1}^M N_{gm}!} \prod_{m=1}^{M_g} p_g(\mathbf{u}_m|\zeta_g, \xi_g)^{N_{gm}}$$

where $m = 1, 2, \dots, M_g$ is an index for the response pattern, and N_{gm} denotes the number of respondents who took response pattern m in group g . Consequently, the log-likelihood function of the model is described as:

- Equation 6

$$\log L(\mathbf{U}|\boldsymbol{\zeta}, \boldsymbol{\xi}) = \sum_{g=1}^G \sum_{m=1}^{M_g} N_{gm} \log p_g(\mathbf{u}_m|\boldsymbol{\zeta}_g, \boldsymbol{\xi}_g)$$

In the Study, g was set for booklets in a country, not the country itself. Booklets that have same item combinations were treated as the same group; therefore, there were 24 (6 types of booklets $\times$ 4 countries) groups in the model.

Item parameter estimation of the model is performed via marginal maximum likelihood estimation. The first derivative function of the log-likelihood function with respect to the item parameters $\boldsymbol{\zeta}_{gj}$ is shown in a general form below:

- Equation 7

$$\frac{\partial \log L(\mathbf{U}|\boldsymbol{\zeta}, \boldsymbol{\xi})}{\partial \boldsymbol{\zeta}_{gj}} = \sum_{g=1}^G \sum_{m=1}^{M_g} \frac{N_{gm}}{p_g(\mathbf{u}_m|\boldsymbol{\zeta}_g, \boldsymbol{\xi}_g)} \int_{-\infty}^{+\infty} \left(\sum_{k'=0}^{K_j} \frac{v_{imk'} \frac{\partial p_{jk'}}{\partial p_{jk}}}{p_{jk'}} \right) \frac{\partial p_{jk}}{\partial \boldsymbol{\zeta}_{gj}} p_g(\mathbf{u}_m|\theta, \boldsymbol{\zeta}_g) h_g(\theta|\boldsymbol{\xi}_g) d\theta$$

where v_{imk} is an indicator variable with value 1 if $u_{im} = k$, and 0 otherwise.

The first derivative function of the log-likelihood function with respect to distribution parameters $\boldsymbol{\xi}$ is given by:

- Equation 8

$$\frac{\partial \log L(\mathbf{U}|\boldsymbol{\zeta}, \boldsymbol{\xi})}{\partial \boldsymbol{\xi}_g} = \sum_{m=1}^{M_g} \frac{N_{gm}}{p_g(\mathbf{u}_m|\boldsymbol{\zeta}_g, \boldsymbol{\xi}_g)} \int_{-\infty}^{+\infty} p(\mathbf{u}_m|\theta, \boldsymbol{\zeta}) \frac{\partial h_g(\theta|\boldsymbol{\xi}_g)}{\partial \boldsymbol{\xi}_g} d\theta$$

Under the normality assumption of the student proficiency levels in a group, the following first derivative function was obtained with respect to μ_g and σ_g^2 :

$$\begin{aligned} \frac{\partial \log L(\mathbf{U}|\boldsymbol{\zeta}, \boldsymbol{\xi})}{\partial \mu_g} &= \frac{1}{\sigma_g^2} \sum_{m=1}^{M_g} N_{gm} \int_{-\infty}^{+\infty} (\theta - \mu_g) \frac{p(\mathbf{u}_m|\theta, \mu_g, \sigma_g^2)}{p_g(\mathbf{u}_m|\boldsymbol{\zeta}_g, \mu_g, \sigma_g^2)} h_g(\theta|\mu_g, \sigma_g^2) d\theta \\ \frac{\partial \log L(\mathbf{U}|\boldsymbol{\zeta}, \boldsymbol{\xi})}{\partial \sigma_g^2} &= \frac{1}{2} \sum_{m=1}^{M_g} N_{gm} \int_{-\infty}^{+\infty} (\sigma_g^2 - (\theta - \mu_g)^2) \frac{p(\mathbf{u}_m|\theta, \mu_g, \sigma_g^2)}{p_g(\mathbf{u}_m|\boldsymbol{\zeta}_g, \mu_g, \sigma_g^2)} h_g(\theta|\mu_g, \sigma_g^2) d\theta \end{aligned}$$

When $\boldsymbol{\zeta}$ is known, estimates of μ_g and σ_g^2 are given as:

- Equation 9

$$\hat{\mu}_g = \frac{1}{N_g} \sum_{m=1}^{M_g} \frac{N_{gm}}{p_g(\mathbf{u}_m | \boldsymbol{\zeta}_g, \mu_g, \sigma_g^2)} \int_{-\infty}^{+\infty} \theta p(\mathbf{u}_m | \theta, \mu_g, \sigma_g^2) h_g(\theta | \mu_g, \sigma_g^2) d\theta$$

$$\hat{\sigma}_g^2 = \frac{1}{N_g} \sum_{m=1}^{M_g} N_{gm} \left((\bar{\theta}_{gm} - \hat{\mu}_g)^2 + \frac{1}{p_g(\mathbf{u}_m | \boldsymbol{\zeta}_g, \mu_g, \sigma_g^2)} \int_{-\infty}^{+\infty} (\theta - \bar{\theta}_{gm})^2 p(\mathbf{u}_m | \theta, \mu_g, \sigma_g^2) h_g(\theta | \mu_g, \sigma_g^2) d\theta \right)$$

which can be used for calculation of expected log-likelihood in the so-called E-step of the commonly utilised EM algorithm (Dempster, Laird and Rubin, 1977_[21]), which it was employed in this analysis following its wide use in the literature (Bock and Aitkin, 1981_[22]).

Further, by maximising the log-likelihood function with respect to the parameters (Equation 7 and Equation 8) based on the expected log-likelihood computed using Equation 9 one perform the Maximisation step (M-step) of the EM algorithm. Through sufficient iterations of the E and M steps, the estimates for $\boldsymbol{\zeta}$ and $\boldsymbol{\xi}$ were obtained.

In the Study, a normal distribution for $h_g(\theta | \boldsymbol{\xi}_g)$ was assumed for all groups. Note that the PISA anchor item parameters were fixed when estimating the rest of the item parameter vector $\boldsymbol{\zeta}$. Furthermore, an equality constraint was set for all $\boldsymbol{\zeta}_g$, which implies the items are invariant among groups. If model fit is not good enough, the equality constraint for $\boldsymbol{\zeta}_{gj}$ was not enforced. Also, do note that in this setting the group-specific parameters, $\boldsymbol{\xi}$ were treated as nuisance parameters.

4. Linking procedures

4.1. Equating model

The purpose of the Study was to establish a link between the PBTS and PISA scales by estimating item parameters of the PBTS items on the PISA 2018 international scale. This can be achieved by maximising likelihood function of PBTS item parameters (ζ_{PBTS}) under the constraint of fixed parameters of PISA items (ζ_{PISA}). This constraint enables to solve factor indeterminacy problem and locate ζ_{PBTS} on the PISA scale.

Neither the representativeness of each country sample nor the normality of the ability/traits regarding the proficiency levels are the necessary conditions to set the PBTS item parameters on the PISA scale. This is because the PISA item parameters are already calibrated with representative samples of the world's 15-years-old students, and they are fixed in the (log) likelihood function defined in Equation 6. Thus, maximising Equation 6 under the fixed parameters (ζ_{PISA}) gives the estimates of ζ_{PBTS} on the PISA scale. With regard to the distribution of the proficiency levels of the linking study samples, it can be any continuous distribution as shown in Equation 7 (Baker and Kim, 2004^[20]).

The important thing here is that it was confirmed the equivalence of the factor structure through equivalences of the factor loadings ($\lambda_j^{(g)} = \lambda_j^{(h)}$) and the intercepts ($\tau_j^{(g)} = \tau_j^{(h)}$) for the items that constitute the latent trait being assessed, the proficiency on each of the three PISA main cognitive domains in this case². Violation of the factor structure equivalence in the context of IRT modelling is called non-uniform differential item functioning (non-uniform DIF) and uniform differential item functioning (uniform DIF). In general, the topic is collectively treated as measurement invariance (Reise, Widaman and Pugh, 1993^[23]; Holland and Wainer, 1993^[24]; Steenkamp and Baumgartner, 1998^[25]; Chen, 2007^[26]; West, Taylor and Wu, 2012^[27]; Asparouhov and Muthén, 2014^[28]).

In the Study, a common (anchor) item design was employed for parameter equating. In the common item design, all groups are connected through the anchor items, so that the items and the student proficiency levels can be calibrated all on the common PISA scale. The simplest pattern of the common item design is the one in which all samples take the same set of anchor items. In this linking study, all the students took the set of PISA anchor items in addition to the PBTS items to be estimated.

The Fixed Item Parameters Calibration (FIPC) was employed for the parameter equating. The FIPC is a model that equates a group of items onto a group of items whose item parameters are already known. In this study, the PISA items were treated as anchor items, and were the target items that set the scale for the freely estimated PBTS items. To this effect, the parameters of PISA items are considered known and fixed at their international values.

Linking separate calibration using equating coefficients from the characteristic curve method (Haebara, 1980^[29]; Stocking and Lord, 1983^[30]) could be an alternative for the linking model (Kim and Cohen, 1998^[31]). However, as the parameters of the PISA items were treated as fixed parameters, the FIPC was selected as the equating model in the Study. Because the countries, even in booklets, have different distribution of the proficiency levels, it was necessary to incorporate student proficiency levels into account when estimating the PBTS item parameters. To achieve this, the finite mixture modelling was employed in which the parameters of PISA item (ζ_{PISA}) were fixed and the parameters of

² The factor analysis model with categorical data and the item response theory (the difference models (Thissen and Steinberg, 1986^[39])) are mathematically the same model. See Takane and de Leeuw (1987^[38]) for the proof.

the PBTS items (ζ_{PBTS}) were estimated. In the analysis, ξ_g for each g were estimated. Consequently, all the estimated parameters could be calibrated on the PISA international scale under the FIPC model.

4.2. Item validation

Although the PISA items have already been validated based on statistical analyses, PBTS items need to be confirmed to possess adequate properties in terms of model fit and differential item functioning (DIF).

DIF is a measure that shows inequality of item functioning towards groups. In this linking study, data were collected from four different countries, from which students took the test in four different languages meaning four translation processes. In an international educational assessment, there is a possibility that a country has a translation problem in an item, which can lead to inappropriate or biased item parameter estimates. Therefore, it is important to collect data from different countries to obtain valid item parameters in order to avoid risk of treating group-specific item parameters as international parameters.

If DIF is found in items, the item might be excluded or group-specific group parameters might be estimated under the assumption of partial invariance (PI) of the latent trait. The reason of keeping DIF items under the PI assumption is that the approach is favoured over dropping the group-specific item data from the analysis in order to retain the information from these responses (OECD, 2017_[10]). Obtaining data from many countries will reduce the risk of calibrating items on a biased scale, and at least dilute negative effects of the DIF items for the calibration.

Nonetheless, it is necessary to confirm that the estimated international item parameters are not biased according to country-specific item functioning. In practice, equivalence, not equality of factors, is the necessary condition to compare outputs between groups.

To check the equivalence of measurements (i.e., absence of DIF) among countries, the root mean square deviation (RMSD) index was employed, and is defined as:

- Equation 10

$$\text{RMSD}_j = \frac{1}{K_j} \sum_{k=0}^{K_j-1} \sqrt{\int \left(p_{jk}^{(obs)}(\theta) - p_{jk}^{(irt)} \right)^2 f(\theta) d\theta}$$

Where $f(\theta)$ is the density function of standardised normal distribution and the integration is calculated over the real line and approximated by standard numerical quadrature methods. Here, $p_{jk}^{(irt)}(\theta)$ is an item characteristic curve based on the estimated item parameters of IRT, while $p_{jk}^{(obs)}(\theta)$ is the pseudo observed item characteristic curve at point θ which is defined as:

- Equation 11

$$p_{jk}^{(obs)}(\theta) = \frac{\sum_{i=1}^N v_{ijk} p(\theta | \mathbf{u}_i, \zeta, \xi)}{\sum_{i=1}^N p(\theta | \mathbf{u}_i, \zeta, \xi)}$$

The RMSD is a measure of a country-specific discrepancy between the IRT and observed item characteristic curves for each country, which was employed in PISA 2015 (OECD, 2017_[10]). If there is no country-specific discrepancy, the RMSD index will be zero.

In the analysis of PISA data, a RMSD value of 0.12 was set as a cut-off criterion. If lack of measurement equivalence is detected among countries in the items, procedures for the DIF items have to be applied in order to keep comparability among countries and stability of the scaling procedure.

5. Results

5.1. Item fit

Table 7 lists anchor items that did not fit to the data obtained. Although PISA items have already been validated and confirmed in their reliability and validity, a few did not work in some countries as expected. The SCIE item had an issue in recording student responses; therefore, it was excluded. Parameters of the two MATH items were estimated with the PCM, which had been employed until PISA 2012. As OECD (2017_[10]) mentions, the PCM provides slightly worse data-model fit than GPCM, which has been employed since PISA 2015. In the linking study, the same exclusion criteria was applied to all items regardless the response model employed. The two MATH items estimated with the PCM showed poorer data-model fit than the other items; therefore, they were excluded.

However, said items were not eliminated from all country analyses but only from the group modelling step. This means a SCIE anchor item was excluded from only country C's model, and thus was not excluded from the other countries' models. All other items had adequate model fit, and used as anchor items when estimating item parameters of the PBTS items. Consequently, only 3 item/country out of 144 item/country (=36 anchor items * 4 countries) were excluded from the modelling.

Table 7. Excluded anchor items from analysis

Treatment	Domain	Cluster	Item	Country
Exclusion	MATH	Anchor	PISA item	Country C
Exclusion	MATH	Anchor	PISA item	Country D
Exclusion	SCIE	Anchor	PISA item	Country C

Based on the anchor items with the exception of those listed in Table 8 the statistical analysis presented in the previous section was performed, and model fit for the PBTS items was checked. Table 8 shows the items that did not fit to the linking study data. Items whose discrimination parameters were negative or $a_j/SE < 1.96$ were excluded. Looking into the contents of the excluded items, some of the items had translation problems or marking issues. Thus, 3 items/country out of 560 item/country (=140 items * 4 countries) were excluded for later analysis at this phase.

Table 8. Excluded PBTS items from analysis

Treatment	Domain	Subdomain	Cluster	Item	Country
Exclusion	MATH	Employing	5	PM5015Q02	Country C
Exclusion	READ	Access and retrieve	2	PR6008Q02	Country A
Exclusion	SCIE	Scientifically interpret	6	PS7224Q02	Country B

5.2. Item comparability

The PBTS items were checked for DIF by country based on Equation 10. The RMSD index was calculated for each item by each country based on the parameter estimates of the modelling introduced in the previous sections. Note that an equality constraint for all ζ_g in the initial model was set. An equality constraint of an item/country was set free if the RMSD index of an item/country exceeded 0.12 (OECD, 2017_[10]). The RMSD index was checked for all item/country in the initial model. This process was iterated until all index values fell below 0.12. An item parameter of a country, which was set free from the equality constraint, is called a national parameter.

Figure 3 and Figure 4 show item characteristic curves (ICC) of a MATH and a READ item, respectively. The figures show their RMSD index values as well. In the figures, solid lines represent the ICC of each country (mode-based ICC; $p_{jk}^{(irt)}(\theta)$), while dashed lines show pseudo observed response probability calculated in the EM cycles (Equation 11). Colours of the lines represent countries. The two figures show how we checked the item traits and item comparability among countries. The two items shown in the figures work equivalently among the four countries, since all the dashed lines overlap to the solid line, meaning the student responses of each country fit to the ICC of the international parameters.

Figure 3. Item characteristic curve of a MATH item without a national parameter

Figure 4. Item characteristic curve of a READ item without a national parameter

Figure 5 shows the ICC of a MATH item which a national parameter value is adopted in one country since it has shown a discrepancy between student responses in country and the ICC of the international parameter. In initial analysis, the RMSD index of the country coloured in red exceeded 0.12. Thus, we released an equality constraint of the item parameters only for the country so that we could estimate the country-specific item parameter (i.e., national parameter).

The RMSD indices shown in the figure are indices calculated with the international parameter used in the three countries and the national parameter used in the red coloured country. We checked all ICCs of the PBTS items, and confirmed that all PBTS items work equivalently to the concepts that are measured.

Figure 5. Item characteristic curve of a MATH item with a national parameter

Table 9 shows the list of item/country for which national parameters were estimated. 27 item/country out of 557 item/country (=140 items * 4 countries – 3 excluded items) were provided the national parameters. The rest of the items show no DIF and their parameters are thus treated as the international item parameters. Apart from country DIF, gender DIF was checked and no differential item functioning in gender was found in any the items.

Table 9. PBTS items which have national parameters

Treatment	Domain	Subdomain	Cluster	ItemID	Country
National parameter	MATH	Employing	1	PM5020Q02	Country A
National parameter	MATH	Employing	3	PM5021Q02	Country A
National parameter	MATH	Employing	6	PM5008Q01	Country B
National parameter	MATH	Interpreting	2	PM5013Q03	Country C
National parameter	READ	Integrate and interpret	2	PR6018Q03	Country A
National parameter	READ	Integrate and interpret	4	PR6010Q06	Country C
National parameter	READ	Access and retrieve	5	PR6017Q10	Country C
National parameter	READ	Reflect and evaluate	3	PR6023Q01	Country C
National parameter	READ	Reflect and evaluate	4	PR6004Q05	Country D
National parameter	READ	Integrate and interpret	2	PR6018Q03	Country D
National parameter	SCIE	Explain	3	PS7013Q02	Country A
National parameter	SCIE	Explain	5	PS7020Q04	Country A
National parameter	SCIE	Scientifically interpret	1	PS7201Q01	Country A
National parameter	SCIE	Explain	3	PS7216Q01	Country A
National parameter	SCIE	Explain	4	PS7000Q01	Country B
National parameter	SCIE	Scientifically interpret	4	PS7000Q02	Country B
National parameter	SCIE	Explain	4	PS7014Q03	Country B
National parameter	SCIE	Explain	5	PS7212Q02	Country B
National parameter	SCIE	Explain	3	PS7216Q01	Country B
National parameter	SCIE	Explain	6	PS7217Q02	Country B
National parameter	SCIE	Scientifically interpret	3	PS7002Q04	Country C
National parameter	SCIE	Explain	1	PS7012Q01	Country C
National parameter	SCIE	Scientifically interpret	2	PS7208Q01	Country C
National parameter	SCIE	Explain	4	PS7000Q01	Country D
National parameter	SCIE	Explain	1	PS7012Q01	Country D
National parameter	SCIE	Explain	3	PS7013Q02	Country D
National parameter	SCIE	Explain	3	PS7216Q01	Country D

5.3. Estimated item parameters

Parameters of the PBTS items were estimated on PISA scales using the FIPC with the international linking study data. Figure 6, Figure 7, and Figure 8 show scatterplots of the estimated item parameters of MATH, READ, and SCIE of the four countries.

In each scatterplot, X-axis represents difficulty of the item, while Y-axis denotes average raw score of the item scaled to $[0 - 1]$ in a country. Note that the range of the four X-axes are all the same. Colour of the plot represents discrimination parameter of the items where blue means high discrimination (above 1.0), while red means low discrimination (below 0.5). Item ID of the PISA anchor items are shown as “Anchor item” in the scatterplot. The figures are shown instead of showing all the estimated parameters explicitly on a table.

The four scatterplots have very similar structure in terms of the relationship between the X and Y-axis of the plots. However, they are put on different levels of Y-axis, which implies the four countries had different base lines of student proficiencies. The plots including both the PISA anchor items and the PBTS items form an S-curve in each country. This implies the PISA anchor items and the PBTS items are on the same scale, although there are some outliers in each country. For example, items PM5187Q01 and PM5102Q01 are slightly off the S-curve. This happens because the discrimination parameters of the items are low, and

the difficulty parameters are therefore shifted to an extreme. When estimating the IRT score of a student with the two low discrimination items, they will not contribute to the likelihood function of the student much, meaning the items will not affect the IRT score estimation much.

Looking into the differences of distribution of the plots, READ items are less discriminative than MATH and SCIE. Moreover, READ items are more diverted from the S-curve comparing to MATH and SCIE.

Figure 6. Estimated MATH item parameters in scatterplot

Figure 7. Estimated READ item parameters in scatterplot

Figure 8. Estimated SCIE item parameters in scatterplot

6. Discussion

In this section, the international linking study 2020 will be summarised. New booklets were assembled for the linking study, which were composed of both the PISA items and the PBTS items. Thus, sufficient number of the PISA items were inserted to the booklets. All the booklet contained the same set of the PISA items, meaning the set of the PISA items played a role of anchor in this Study. Over 95 000 students from 4 countries (Brazil, Spain, Russian Federation, and United States) participated, and each of the students took one of the booklets. The proportion of the PISA items to the PBTS items were 6:4 in the booklets, which enabled the PBTS items to establish a robust link to the PISA international scale.

The finite mixture modelling was employed as the statistical model for parameter estimation for the linking study data, considering the following three points: 1) the four countries have different distribution in terms of student proficiency levels; 2) the four countries have the equivalent factor structure regarding the concepts that it is measured (i.e., MATH, READ, and SCIE), but subject to be confirmed; and 3) the parameters of the PISA items are already estimated and validated. Since the purpose of the Study is to estimate the international parameters of PBTS items, student distribution of the countries should be modelled when estimating the item parameter. In addition, by fixing the parameters of anchor items, the estimated PBTS item parameters are fully aligned with the PISA international scale.

For the PISA anchor items, the results showed that only 3 out of 144 PISA items/country were excluded from the analysis due to poor model-fit. Therefore, approximately 60% of items were used as anchors in the booklets for the equating as planned in this linking study; thus, a strong connection between the estimated PBTS item parameters and the PISA scale could be built.

For the PBTS items, only 3 item/country out of 560 item/country were excluded because the estimated discrimination parameters were negative or too low, ensuring quality of the PBTS items. Based on the rest of the items and through the maximisation on Equation 6, the international parameters of the PBTS items were estimated under the equality constraint of item parameters among groups. Item characteristic response functions were checked of all the items including PISA items and confirmed they all work well as cognitive items.

By checking the RMSD index (Equation 10), 27 item/country out of 557 item/country were freed and thus a national parameter was estimated, which indicates that 5% of the item/country worked differently for a given country. The majority of PBTS items work equivalently in students from the four different countries. Hence, a sufficient level of international comparability can be expected for the scores calculated based on the item parameters using both the international parameters and the national parameters.

As observed from Figure 6, Figure 7, and Figure 8, the distribution of item difficulties in the PBTS is well balanced and widely distributed in all domains. Therefore, the test information function of the PBTS booklets in each domain shown in the PBTS technical report will cover a wide range of proficiency levels. Considering the purpose of the PISA for Schools project to support school-level analysis, this is important as sufficient reliability of the PBTS scores can be expected in a wide range of proficiency levels.

This validation analysis can be performed constantly as new data is collected in the future, further ensuring quality of the measurement. As international item parameters were estimated, the same validation analyses and procedures can be applied for newly participating countries. This is practical and useful from a test management view. In the future, continuous item validation study with collected data will reinforce the validity of the items and scores calculated based on the item parameters.

References

- ACER (2012), “PISA-based Test for Schools: Technical Report”, *Australian Council for Educational Research (ACER)*. [3]
- Asparouhov, T. and B. Muthén (2014), “Multiple-group factor analysis alignment”, *Structural Equation Modeling*, Vol. 21/4, pp. 495–508. [28]
- Baker, F. and S. Kim (2004), *Item Response Theory: Parameter Estimation Techniques. 2nd Edition*, CRC Press. [20]
- Bock, R. and M. Aitkin (1981), “Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm.”, *Psychometrika*, Vol. 46, pp. 443–459. [22]
- Bock, R. and M. Zimowski (1997), *Multiple group IRT*, Springer-Verlag. [37]
- Chen, F. (2007), “Sensitivity of goodness of fit indexes to lack of measurement invariance.”, *Structural Equation Modeling*, Vol. 14, pp. 464–504. [26]
- Dempster, A., N. Laird and D. Rubin (1977), “Maximum likelihood from incomplete data via the EM algorithm.”, *Journal of the Royal Statistical Society, Series B.*, Vol. 39, pp. 1–38. [21]
- Haebara, T. (1980), “Equating logistic ability scales by a weighted least squares method.”, *Japanese Psychological Research*, Vol. 22, pp. 144–149. [29]
- Holland, P. and H. Wainer (1993), *Differential item functioning*, Erlbaum. [24]
- Ishii, T., P. Songmuang and M. Ueno (2015), “Maximum clique algorithm and its approximation for uniform test form assembly.”, *IEEE Transactions on Learning Technologies*, Vol. 7/1, pp. 83–95. [18]
- Jurich, D., C. DeMars and J. Goodman (2012), “Investigating the impact of compromised anchor items on IRT equating under the nonequivalent anchor test design.”, *Applied Psychological Measurement*, Vol. 36/4, pp. 291–308. [14]
- Kim, S. and A. Cohen (1998), “A comparison of linking and concurrent calibration under item response theory.”, *Applied Psychological Measurement*, Vol. 22/2, pp. 131–143. [31]
- Kreiner, S. and K. Christensen (2014), “Analyses of model fit and robustness: A new look at the PISA scaling model underlying ranking of countries according to reading literacy”, *Psychometrika*, Vol. 79/2, pp. 210–231. [9]
- Lord, F. and M. Novick (1968), *Statistical theories of mental test scores.*, Addison-Wesley. [6]

- Luecht, R. (1998), "Computer-assisted test assembly using optimization heuristics", *Applied Psychological Measurement*, Vol. 22/3, pp. 224–236. [15]
- Masters, G. (1982), "A Rasch model for partial credit scoring", *Psychometrika*, Vol. 47, pp. 149–174. [5]
- McKinley, R. and M. Reckase (1981), "A comparison of procedures for constructing large item pools", *Columbia MO: University of Missouri, Department of Educational Psychology*. [12]
- Mislevy, R. (1984), "Estimating latent distributions", *Psychometrika*, Vol. 49, pp. 359–381. [32]
- Muraki, E. (1992), "A generalized partial credit model: Application of an EM algorithm", *Applied Psychological Measurement*, Vol. 16, pp. 159–176. [7]
- OECD (2020), *PISA 2018 Technical Report*. [34]
- OECD (2017), *PISA 2015 Technical Report*, OECD Publishing. [10]
- OECD (2016), *PISA-Based Test for Schools Technical Report 2016*. [2]
- OECD (2013), *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*, PISA, OECD Publishing. [1]
- Oliveri, M. and M. von Davier (2014), "Toward increasing fairness in score scale calibrations employed in international large-scale assessments", *International Journal of Testing*, Vol. 14/1, pp. 1–21. [8]
- Rasch, G. (1960), *Probabilistic models for some intelligence and attainment tests*, Nielsen and Lydiche. [4]
- Reise, S., K. Widaman and R. Pugh (1993), "Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance", *Psychological Bulletin*, Vol. 114/3, pp. 552–566. [23]
- Samejima, F. (1969), "Estimation of latent ability using a response pattern of graded scores", *Psychometric Monograph*, Vol. 17. [33]
- Steenkamp, J. and H. Baumgartner (1998), "Assessing measurement invariance in crossnational consumer research", *Journal of Consumer Research*, Vol. 25, pp. 78–90. [25]
- Stocking, M. and F. Lord (1983), "Developing a common metric in item response theory", *Applied Psychological Measurement*, Vol. 7, pp. 201–210. [30]
- Takane, Y. and J. de Leeuw (1987), "On the relationship between item response theory and factor analysis of discretized variables", *Psychometrika*, Vol. 52, pp. 393–408. [38]
- Thissen, D. and L. Steinberg (1986), "A taxonomy of item response models", *Psychometrika*, Vol. 51/4, pp. 567–577. [39]
- Vale, C. (1986), "Linking item parameters onto a common scale", *Applied Psychological Measurement*, Vol. 10, pp. 333–344. [35]
- Vale, C. et al. (1981), "Methods for linking item parameters", *Brooks Air Force Base TX*. [11]
- van de Vijver, F. and K. Leung (1997), "Methods and data analysis for cross-cultural research", *Cross-cultural psychology series*, Vol. 1. [36]

- van der Linden, W. (2005), *Linear methods for optimal test design*, Springer. [17]
- van der Linden, W. and J. Adema (1998), “Simultaneous assembly of multiple test forms”, *Journal of Educational Measurement*, Vol. 35/3, pp. 185–198. [16]
- van der Linden, W. and R. Hambleton (eds.) (1997), *A Response Model for Multiple-Choice Items*, Springer. [19]
- West, S., A. Taylor and W. Wu (2012), *Model fit and model selection in structural equation modeling*, Guilford Press. [27]
- Wingersky, M. and F. Lord (1984), “An investigation of methods for reducing sampling error in certain IRT procedures”, *Applied Psychological Measurement*, Vol. 8/3, pp. 347-364. [13]

Appendix

Table 10. International linking study booklets (1A, 1B, and 1C)

1A			1B			1C		
1st session	2nd session	3rd session	1st session	2nd session	3rd session	1st session	2nd session	3rd session
PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item
PM5020Q02	PISA item	PS7017Q01	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item
PISA item	PISA item	PS7017Q04	PISA item	PM5169Q01	PISA item	PR6002Q01	PISA item	PM5142Q01
PISA item	PR6027Q01	PISA item	PS7012Q01	PM5169Q02	PISA item	PR6002Q02	PS7012Q01	PM5142Q02
PM5109Q01	PR6027Q04	PISA item	PS7012Q03	PISA item	PR6002Q01	PR6002Q04	PS7012Q03	PISA item
PM5109Q02	PR6027Q10	PISA item	PS7012Q04	PISA item	PR6002Q02	PR6002Q05	PS7012Q04	PM5169Q01
PISA item	PISA item	PS7201Q03	PISA item	PISA item	PR6002Q04	PISA item	PISA item	PM5169Q02
PISA item	PISA item	PS7201Q01	PISA item	PM5142Q01	PR6002Q05	PISA item	PS7231Q01	PISA item
PISA item	PISA item	PS7201Q02	PISA item	PM5142Q02	PISA item	PISA item	PS7231Q04	PISA item
PISA item	PISA item	PISA item	PS7231Q01	PISA item	PISA item	PISA item	PISA item	PM5020Q02
PM5169Q01	PISA item	PISA item	PS7231Q04	PM5020Q02	PISA item	PISA item	PISA item	PISA item
PM5169Q02	PISA item	PISA item	PISA item	PISA item	PR6027Q01	PISA item	PISA item	PISA item
PISA item	PR6002Q01	PS7012Q01	PS7201Q03	PM5109Q01	PR6027Q04	PR6027Q01	PS7017Q01	PISA item
PM5142Q01	PR6002Q02	PS7012Q03	PS7201Q01	PM5109Q02	PR6027Q10	PR6027Q04	PS7017Q04	PISA item
PM5142Q02	PR6002Q04	PS7012Q04	PS7201Q02	PISA item	PISA item	PR6027Q10	PISA item	PISA item
PISA item	PR6002Q05	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PM5109Q01
PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PM5109Q02
PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PS7201Q03	PISA item
	PISA item	PS7231Q01	PS7017Q01		PISA item	PISA item	PS7201Q01	
		PS7231Q04	PS7017Q04				PS7201Q02	
		PISA item	PISA item				PISA item	
		PISA item	PISA item				PISA item	
		PISA item	PISA item				PISA item	

Table 11. International linking study booklets (2A, 2B, and 2C)

2A			2B			2C		
1st session	2nd session	3rd session	1st session	2nd session	3rd session	1st session	2nd session	3rd session
PISA item	PR6008Q02	PS7205Q02	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item
PISA item	PR6008Q04	PISA item	PS7205Q02	PISA item	PISA item	PISA item	PISA item	PM5013Q01
PISA item	PR6008Q05	PISA item	PISA item	PISA item	PISA item	PR6018Q01	PISA item	PM5013Q03
PISA item	PISA item	PISA item	PISA item	PISA item	PR6018Q01	PR6018Q03	PS7220Q01	PISA item
PM5188Q01	PISA item	PS7208Q01	PISA item	PISA item	PR6018Q03	PR6018Q08	PS7220Q02	PISA item
PISA item	PISA item	PS7208Q02	PS7208Q01	PM5013Q01	PR6018Q08	PR6018Q09	PS7220Q03	PISA item
PM5102Q01	PISA item	PISA item	PS7208Q02	PM5013Q03	PR6018Q09	PISA item	PISA item	PISA item
PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item
PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PM5107Q01
PM5163Q01	PISA item	PS7220Q01	PISA item	PM5188Q01	PISA item	PISA item	PS7228Q02	PM5107Q02
PISA item	PISA item	PS7220Q02	PS7220Q01	PISA item	PISA item	PISA item	PS7228Q03	PISA item
PM5013Q01	PISA item	PS7220Q03	PS7220Q02	PM5102Q01	PISA item	PISA item	PS7228Q04	PISA item
PM5013Q03	PR6018Q01	PISA item	PS7220Q03	PISA item	PR6008Q02	PISA item	PISA item	PM5188Q01
PISA item	PR6018Q03	PISA item	PISA item	PISA item	PR6008Q04	PR6008Q02	PISA item	PISA item
PM5107Q01	PR6018Q08	PISA item	PISA item	PM5163Q01	PR6008Q05	PR6008Q04	PISA item	PM5102Q01
PM5107Q02	PR6018Q09	PISA item	PISA item	PM5107Q01	PISA item	PR6008Q05	PS7205Q02	PISA item
PISA item	PISA item	PS7228Q02	PS7228Q02	PM5107Q02	PISA item	PISA item	PISA item	PISA item
PISA item	PISA item	PS7228Q03	PS7228Q03	PISA item	PISA item	PISA item	PISA item	PM5163Q01
	PISA item	PS7228Q04	PS7228Q04		PISA item	PISA item	PISA item	
		PISA item	PISA item				PS7208Q01	
		PISA item	PISA item				PS7208Q02	
		PISA item	PISA item				PISA item	

Table 12. International linking study booklets (3A, 3B, and 3C)

3A			3B			3C		
1st session	2nd session	3rd session	1st session	2nd session	3rd session	1st session	2nd session	3rd session
PISA item	PISA item	PISA item	PISA item	PISA item	PR6023Q01	PR6023Q01	PS7013Q01	PM5010Q01
PISA item	PISA item	PISA item	PISA item	PISA item	PR6023Q06	PR6023Q06	PS7013Q02	PM5010Q02
PM5021Q01	PISA item	PISA item	PISA item	PM5010Q01	PR6023Q07	PR6023Q07	PISA item	PISA item
PM5021Q02	PISA item	PS7216Q01	PS7002Q03	PM5010Q02	PR6023Q08	PR6023Q08	PISA item	PISA item
PISA item	PR6013Q03	PS7216Q03	PS7002Q04	PISA item	PISA item	PISA item	PISA item	PM5026Q01
PM5002Q01	PR6013Q04	PS7021Q02	PISA item	PM5021Q01	PISA item	PISA item	PISA item	PISA item
PM5002Q02	PR6013Q06	PS7021Q04	PISA item	PM5021Q02	PISA item	PISA item	PISA item	PM5002Q01
PISA item	PR6013Q08	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PM5002Q02
PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PR6013Q03	PS7002Q03	PISA item
PISA item	PISA item	PISA item	PS7216Q01	PM5002Q01	PR6013Q03	PR6013Q04	PS7002Q04	PISA item
PISA item	PISA item	PISA item	PS7216Q03	PM5002Q02	PR6013Q04	PR6013Q06	PISA item	PISA item
PM5010Q01	PISA item	PS7002Q03	PISA item	PISA item	PR6013Q06	PR6013Q08	PISA item	PISA item
PM5010Q02	PISA item	PS7002Q04	PISA item	PISA item	PR6013Q08	PISA item	PISA item	PM5021Q01
PISA item	PISA item	PISA item	PISA item	PM5026Q01	PISA item	PISA item	PISA item	PM5021Q02
PISA item	PISA item	PISA item	PS7021Q02	PISA item	PISA item	PISA item	PS7021Q02	PISA item
PISA item	PISA item	PISA item	PS7021Q04	PISA item	PISA item	PISA item	PS7021Q04	PISA item
PM5026Q01	PR6023Q01	PS7013Q01	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item
PISA item	PR6023Q06	PS7013Q02	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item
	PR6023Q07	PISA item	PISA item		PISA item	PISA item	PISA item	
	PR6023Q08	PISA item	PS7013Q01		PISA item	PISA item	PS7216Q01	
		PISA item	PS7013Q02				PS7216Q03	

Table 13. International linking study booklets (4A, 4B, and 4C)

4A			4B			4C		
1st session	2nd session	3rd session	1st session	2nd session	3rd session	1st session	2nd session	3rd session
PISA item	PISA item	PISA item	PISA item	PISA item	PR6010Q01	PISA item	PISA item	PISA item
PISA item	PISA item	PISA item	PISA item	PM5110Q01	PR6010Q03	PISA item	PISA item	PISA item
PISA item	PISA item	PISA item	PISA item	PM5110Q02	PR6010Q06	PISA item	PISA item	PM5110Q01
PISA item	PR6026Q02	PS7000Q01	PS7230Q02	PISA item	PR6010Q07	PISA item	PS7014Q01	PM5110Q02
PISA item	PR6026Q05	PS7000Q02	PS7230Q03	PISA item	PISA item	PR6004Q01	PS7014Q03	PISA item
PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PR6004Q05B	PISA item	PISA item
PISA item	PISA item	PS7018Q03	PISA item	PISA item	PR6026Q02	PR6004Q06	PISA item	PISA item
PM5104Q01	PISA item	PS7018Q04	PISA item	PM5104Q01	PR6026Q05	PISA item	PISA item	PISA item
PISA item	PR6004Q01	PISA item	PS7000Q01	PISA item	PISA item	PISA item	PS7018Q03	PM5124Q01
PM5124Q01	PR6004Q05B	PISA item	PS7000Q02	PISA item	PISA item	PISA item	PS7018Q04	PM5124Q02
PM5124Q02	PR6004Q06	PISA item	PISA item	PISA item	PISA item	PR6010Q01	PISA item	PM5124Q03
PM5124Q03	PISA item	PS7230Q02	PISA item	PISA item	PISA item	PR6010Q03	PS7230Q02	PISA item
PISA item	PISA item	PS7230Q03	PISA item	PM5124Q01	PISA item	PR6010Q06	PS7230Q03	PISA item
PISA item	PR6010Q01	PISA item	PS7018Q03	PM5124Q02	PISA item	PR6010Q07	PISA item	PM5104Q01
PM5110Q01	PR6010Q03	PISA item	PS7018Q04	PM5124Q03	PISA item	PISA item	PISA item	PISA item
PM5110Q02	PR6010Q06	PISA item	PISA item	PISA item	PR6004Q01	PISA item	PISA item	PISA item
PISA item	PR6010Q07	PS7014Q01	PS7014Q01	PISA item	PR6004Q05B	PISA item	PS7000Q01	PISA item
	PISA item	PS7014Q03	PS7014Q03		PR6004Q06	PR6026Q02	PS7000Q02	
	PISA item	PISA item	PISA item		PISA item	PR6026Q05	PISA item	
	PISA item	PISA item	PISA item		PISA item	PISA item	PISA item	
	PISA item	PISA item	PISA item		PISA item	PISA item	PISA item	

Table 14. International linking study booklets (5A, 5B, and 5C)

5A			5B			5C		
1st session	2nd session	3rd session	1st session	2nd session	3rd session	1st session	2nd session	3rd session
PISA item	PISA item	PISA item	PS7212Q01	PISA item	PISA item	PISA item	PISA item	PISA item
PISA item	PISA item	PISA item	PS7212Q02	PISA item	PISA item	PISA item	PS7211Q01	PISA item
PM5145Q01	PISA item	PISA item	PS7212Q04	PM5101Q01	PISA item	PR6017Q04	PS7211Q02	PISA item
PISA item	PR6011Q03	PISA item	PISA item	PM5101Q02	PR6017Q04	PR6017Q05	PISA item	PISA item
PISA item	PR6011Q05	PISA item	PS7222Q01	PISA item	PR6017Q05	PR6017Q08	PISA item	PISA item
PM5017Q02	PR6011Q07	PISA item	PS7222Q02	PISA item	PR6017Q08	PR6017Q10	PISA item	PM5015Q02
PM5017Q03	PISA item	PS7222Q01	PS7222Q03	PISA item	PR6017Q10	PR6017Q11	PS7212Q01	PISA item
PISA item	PISA item	PS7222Q02	PISA item	PISA item	PR6017Q11	PISA item	PS7212Q02	PM5017Q02
PM5101Q01	PISA item	PS7222Q03	PISA item	PM5145Q01	PISA item	PISA item	PS7212Q04	PM5017Q03
PM5101Q02	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item
PISA item	PISA item	PISA item	PISA item	PM5017Q02	PISA item	PISA item	PISA item	PISA item
PM5015Q02	PISA item	PISA item	PISA item	PM5017Q03	PISA item	PISA item	PISA item	PM5101Q01
PISA item	PR6017Q04	PS7212Q01	PISA item	PISA item	PISA item	PISA item	PS7020Q03	PM5101Q02
PISA item	PR6017Q05	PS7212Q02	PS7020Q03	PISA item	PISA item	PR6011Q03	PS7020Q04	PISA item
PISA item	PR6017Q08	PS7212Q04	PS7020Q04	PISA item	PISA item	PR6011Q05	PS7222Q01	PISA item
PISA item	PR6017Q10	PISA item	PISA item	PM5015Q02	PR6011Q03	PR6011Q07	PS7222Q02	PM5145Q01
PISA item	PR6017Q11	PISA item	PISA item	PISA item	PR6011Q05	PISA item	PS7222Q03	PISA item
	PISA item	PISA item	PISA item		PR6011Q07	PISA item	PISA item	
	PISA item	PS7020Q03	PS7211Q01		PISA item	PISA item	PISA item	
	PISA item	PS7020Q04	PS7211Q02		PISA item	PISA item	PISA item	
		PS7211Q01	PISA item				PISA item	
		PS7211Q02	PISA item				PISA item	
		PISA item	PISA item				PISA item	

Table 15. International linking study booklets (6A, 6B, and 6C)

6A			6B			6C		
1st session	2nd session	3rd session	1st session	2nd session	3rd session	1st session	2nd session	3rd session
PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item
PM5187Q01	PISA item	PISA item	PISA item	PM5183Q01	PISA item	PISA item	PISA item	PISA item
PISA item	PISA item	PISA item	PISA item	PM5183Q02	PISA item	PISA item	PISA item	PM5008Q01
PM5185Q01	PISA item	PS7217Q01	PS7224Q02	PISA item	PR6014Q03	PR6015Q01	PS7221Q01	PISA item
PM5185Q02	PISA item	PS7217Q02	PISA item	PM5187Q01	PR6014Q06	PR6015Q03	PS7221Q02	PM5181Q02
PISA item	PISA item	PISA item	PISA item	PISA item	PR6014Q07	PR6015Q04	PS7221Q03	PISA item
PM5183Q01	PR6015Q01	PISA item	PISA item	PISA item	PISA item	PR6015Q06	PS7221Q04	PISA item
PM5183Q02	PR6015Q03	PISA item	PS7217Q01	PM5185Q01	PISA item	PISA item	PISA item	PM5185Q01
PISA item	PR6015Q04	PS7215Q02	PS7217Q02	PM5185Q02	PISA item	PISA item	PISA item	PM5185Q02
PISA item	PR6015Q06	PS7215Q03	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item
PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PISA item	PS7215Q02	PM5183Q01
PISA item	PISA item	PS7224Q02	PISA item	PM5008Q01	PR6015Q01	PISA item	PS7215Q03	PM5183Q02
PM5008Q01	PISA item	PISA item	PS7215Q02	PISA item	PR6015Q03	PISA item	PISA item	PISA item
PISA item	PR6014Q03	PISA item	PS7215Q03	PM5181Q02	PR6015Q04	PR6014Q03	PISA item	PISA item
PISA item	PR6014Q06	PISA item	PISA item	PISA item	PR6015Q06	PR6014Q06	PISA item	PISA item
PM5181Q02	PR6014Q07	PS7221Q01	PISA item	PISA item	PISA item	PR6014Q07	PS7224Q02	PISA item
PISA item	PISA item	PS7221Q02	PISA item	PISA item	PISA item	PISA item	PISA item	PM5187Q01
PISA item	PISA item	PS7221Q03	PS7221Q01	PISA item	PISA item	PISA item	PS7217Q01	PISA item
	PISA item	PS7221Q04	PS7221Q02		PISA item	PISA item	PS7217Q02	
		PISA item	PS7221Q03				PISA item	
		PISA item	PS7221Q04				PISA item	
		PISA item	PISA item				PISA item	