Chapter 9
Scaling PISA data

OVERVIEW

The test designs for the PISA 2018 and prior cycles were based on matrix sampling (and multistage adaptive testing design for the major domain of Reading in the PISA 2018) where each student is administered a subset of items from the total item pool. As a result, different groups of students answered different sets of items. Therefore, any statistic based on the number of correct responses cannot be used to report survey results. Differences in total scores, or statistics based on them, among students who took different sets of items may be due to variations in the difficulty of the test forms. Unless one makes very strong assumptions – for example, that different test forms and adaptive paths are perfectly parallel – the performance of groups measured through different set of items cannot be directly compared using total-score statistics. Moreover, item-by-item reporting ignores the dissimilarities of proficiencies of subgroups to which the set of items was administered. Finally, using the average percentage of items answered correctly (often called as P+) to estimate the mean proficiency of students in a given subpopulation does not provide any other information about the distribution of skills within that subpopulation (e.g. variances).

The limitations of number or percent correct scoring methods can be overcome by using item response theory (IRT) scaling. When responding to a set of items requires a given skill, the response patterns should show regularities that can be modelled using the underlying commonalities among the items. This regularity can be used to characterize students and items on the common scale, even when students take different sets of items. IRT makes it possible to describe the distributions of student performance in populations or subpopulations and to estimate the relationships between proficiency and background variables, as well as to build and select test forms that are matched in difficulty with the ability of students.

To further increase measurement accuracy and to facilitate the estimation of the relationships between proficiency and contextual variables from the background questionnaire (BQ), the IRT scaling of the test items can be combined with a latent regression model that uses information from the BQ in a population model. Once the population model is estimated, for each student, multiple plausible values can be drawn from a posterior distribution that account for the sources of uncertainty in the data.

This chapter first describes the quantity and quality of the data submitted by the participating countries and economies. Analyses were conducted to evaluate how well the assessment design was reflected in the data and to verify that the data were collected according to the test design and that data quality is appropriate for modelling to be applicable. In the following sections, the models and methods employed for IRT scaling, population modelling, and the generation of plausible values are described. How these models and methods were applied to the PISA 2018 data to produce the national and international item parameters and the plausible values are then detailed. Finally, the approach and methods used for estimating the linking errors between the 2018 main survey and the previous PISA cycles are explained. The specific analyses and results confirming the quality of the data collection for the new multi-stage adaptive test (MSAT) design, as well as the approaches used to scale the MSAT data and confirm the comparability of the MSAT results with results from prior non-adaptive PISA cycles, are also described throughout this chapter.

DATA YIELD AND DATA QUALITY

Before data were used for scaling and population modelling, analyses were carried out to examine the quality of data and to ensure that the test design requirements were met. The following subsections give an overview of these analyses and their results. Overall, the quality of the data and the quality of the cognitive instruments were confirmed. Item analysis results were communicated to countries for their review and feedback to the test development and psychometric teams. Taken together data yield and item analyses confirmed that the PISA 2018 computer platform successfully delivered, captured, and exported information for more than 600 CBA and 250 PBA items.

Targeted sample size, routing and data yield

Targeted sample size

The assessment design for the PISA 2018 main survey covered the core domains of mathematics, reading, and science to be delivered in both computer-based assessment (CBA) and the paper-based assessment (PBA), as well as two optional CBA domains of financial literacy and global competence (the innovative domain designed solely for the 2018 cycle). Participating countries were required to sample a minimum of 150 schools representing their national population of 15-year-old students. Countries taking the CBA with global competence needed to sample 42 students from each of 150 schools for a total sample of 6,300 students, while countries taking the CBA without global competence or the PBA needed to sample 35 students from each of 150 schools for a total sample of 5,250 students. CBA countries taking the financial literacy domain were also required to sample larger number of schools and/or students per schools to obtain an additional sample of 1,650 students. This group of 1,650 students who took financial literacy are called "Financial Literacy sample" and they are different from the "Main Sample" who did not take financial literacy. Note that this was different from the approach used in the 2015 cycle, when FL was administered to a portion of the main sample (see Chapter 2 for more details).

Regarding the assignment of forms in the main CBA designs, it is important to note that 88% or 92% of students (participating country assessing GC or not) received a form that consists of one hour of reading (major domain), and two 30-minute clusters for another domain. This resulted in one hour of assessment time per domain and a total of two hours of testing time per student (often called as "bivariate forms"). The rest of 12% or 8% of students received forms consisting of one hour of reading and two 30-minute clusters covering two of the other three domains (often called as "trivariate forms"; see Chapter 2 for more details).

Data yield

Table 9.1 shows the assessment languages and the sample sizes for the participating countries. In order for a student to be considered a "respondent" for PISA, the student needed to meet one of the following two criteria: 1) answered more than half of the cognitive items on their form/booklet, or 2) answered at least one cognitive item and completed a minimum amount of the student BQ (specifically, that the student answered at least one question regarding home possessions – ST012 or ST013).

Table 9.1 (1/2) Test mode, sample size per country and language for the main survey

Country	Language(s)	Test Mode	Main Sample	Financial Lit.	Total Sample	Schools
Albania	Albanian	CBA	6,359		6,359	327
Australia	English	CBA	14,273	9,411	23,684	763
Austria	German	CBA	6,802		6,802	291
Baku (Azerbaijan)	Russian and Azeri	CBA	6,827		6,827	197
Belarus	Russian and Belarusian	CBA	5,803		5,803	234
Belgium	Dutch, French, and German	CBA	8,475		8,475	288
Bosnia and Herzegovina	Serbian, Croatian, and Bosnian	CBA	6,480		6,480	213
Brazil	Portuguese	CBA	10,691	8,311	19,002	597
Brunei Darussalam	English	CBA	6,828		6,828	55
B-S-J-Z (China)*	Chinese	CBA	12,058		12,058	361
Bulgaria	Bulgarian	CBA	5,294	4,110	9,404	197
Canada	French and English	CBA	22,653	7,762	30,415	821
Chile	Spanish	CBA	7,621	4,485	12,106	256
Chinese Taipei	Chinese	CBA	7,243		7,243	192
Colombia	Spanish	CBA	7,522		7,522	247
Costa Rica	Spanish	CBA	7,221		7,221	205
Croatia	Croatian	CBA	6,609		6,609	183
Cyprus	English and Greek	CBA	5,503		5,503	90
Czech Republic	Czech	CBA	7,019		7,019	333
Denmark	Faroese and Danish	CBA	7,657		7,657	348
Dominican Republic	Spanish	CBA	5,674		5,674	235
Estonia	Russian and Estonian	CBA	5,316	4,167	9,483	230
Finland	Swedish and Finnish	CBA	5,649	4,328	9,977	214
France	French	CBA	6,308		6,308	252
Georgia	Russian, Georgian, and Azerbaijani	CBA	5,572	4,321	9,893	321
Germany	German	CBA	5,451		5,451	223
Greece	Greek	CBA	6,403		6,403	242
Hong Kong (China)	Chinese and English	CBA	6,037		6,037	152
Hungary	Hungarian	CBA	5,132		5,132	238
Iceland	Icelandic	CBA	3,296		3,296	142
Indonesia	Indonesian	CBA	12,098	7,133	19,231	397
Ireland	Irish and English	CBA	5,577		5,577	157
Israel	Hebrew, Arabic, and Hebrew	CBA	6,623		6,623	174
Italy	Italian and German	CBA	11,785	9,182	20,967	542
Japan	Japanese	CBA	6,109		6,109	183
Kazakhstan	Russian and Kazakh	CBA	19,507		19,507	616
Korea	Korean	CBA	6,650		6,650	188
Kosovo	Serbian and Albanian	CBA	5,058		5,058	211
Latvia	Russian and Latvian	CBA	5,303	3,151	8,454	308
Lithuania	Russian, Polish, and Lithuanian	CBA	6,885	4,076	10,961	362
Luxembourg	French, English, and German	CBA	5,230		5,230	44
Macao (China)	Chinese, Portuguese, and English	CBA	3,775		3,775	45
Malaysia	Malay and English	CBA	6,111		6,111	191
Malta	Maltese and English	CBA	3,363		3,363	50
Mexico	Spanish	CBA	7,299		7,299	286
Montenegro	Albanian and Serb (Yekavian)	CBA	6,666		6,666	61
Morocco	Arabic	CBA	6,814		6,814	179
Netherlands	Dutch	CBA	4,765	3,042	7,807	156
New Zealand	English	CBA	6,173		6,173	192
Norway	Bokmål and Nynorsk	CBA	5,813		5,813	251
Panama	Spanish and English	CBA	6,270		6,270	253
Peru	Spanish	CBA	6,086	4,734	10,820	340
Philippines	English	CBA	7,233		7,233	187
Poland	Polish	CBA	5,625	4,295	9,920	240
Portugal	Portuguese	CBA	5,932	4,568	10,500	276
Qatar	English and Arabic	CBA	13,828		13,828	188
Russian Federation	Russian	CBA	7,608	4,520	12,128	263
Serbia	Serbian and Hungarian	CBA	6,609	3,874	10,483	187
Singapore	English	CBA	6,676		6,676	166
Slovak Republic	Slovak and Hungarian	CBA	5,965	3,411	9,376	376
Slovenia	Slovenian	CBA	6,401		6,401	345
Spain	Valencian, Galician, Basque, Spanish, and Catalan	CBA	35,943	9,361	45,304	1,089
Sweden	Swedish and English	CBA	5,504		5,504	223
Switzerland	Italian, French, and German	CBA	5,822		5,822	228

Table 9.1 (2/2) Test mode, sample size per country and language for the main survey

Country	Language(s)	Test Mode	Main Sample	Total Sample	Schools
Argentina	Spanish	PBA	11,975	11,975	455
Jordan	Arabic	PBA	8,963	8,963	313
Lebanon	French and English	PBA	5,614	5,614	313
North Macedonia	Albanian and Macedonian	PBA	5,569	5,569	117
Republic of Moldova	Russian and Romanian	PBA	5,367	5,367	236
Romania	Romanian and Hungarian	PBA	5,075	5,075	170
Saudi Arabia	English and Arabic	PBA	6,136	6,136	234
Ukraine	Ukranian and Russian	PBA	5,998	5,998	250
Viet Nam	Vietnamese	PBA	5,377	5,377	151

* B-S-J-Z (China) data represent the regions of Beijing, Shanghai, Jiangsu, and Zhejiang.

1. Note by Turkey: The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document

Figures 9.1, 9.2 and 9.3 show the extent to which each participating country opting for the core CBA or the PBA assessments and the financial literacy assessment met or exceeded their sample requirements. In each figure, red horizontal line indicates the sample requirements for each design option. Some countries exceeded their requirements because of oversampling regions and/or minority languages. A few countries did not reach their sample requirement because of their small population size.

Figure 9.1 Main sample yield for countries and economies participating in the CBA

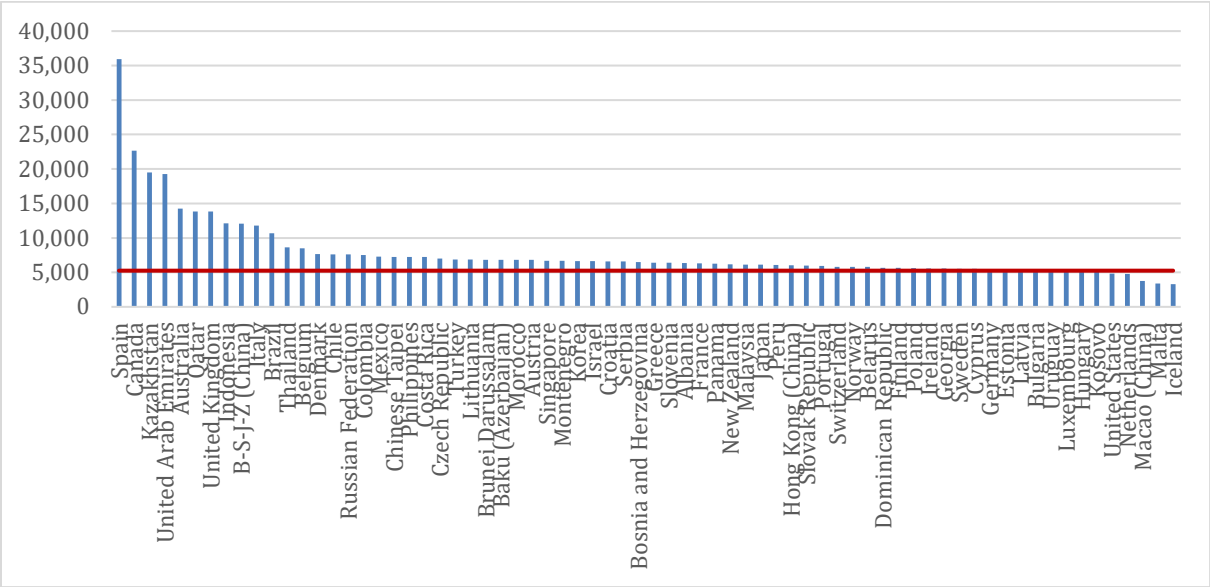


Figure 9.2 Financial literacy sample yield for countries and economies participating in the CBA

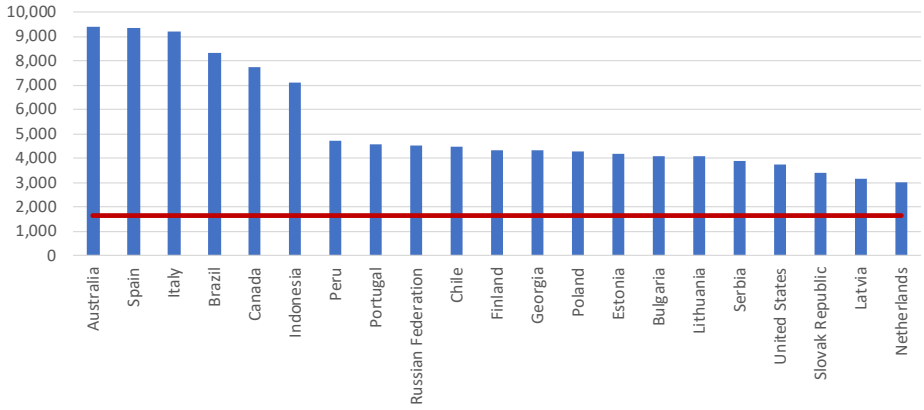
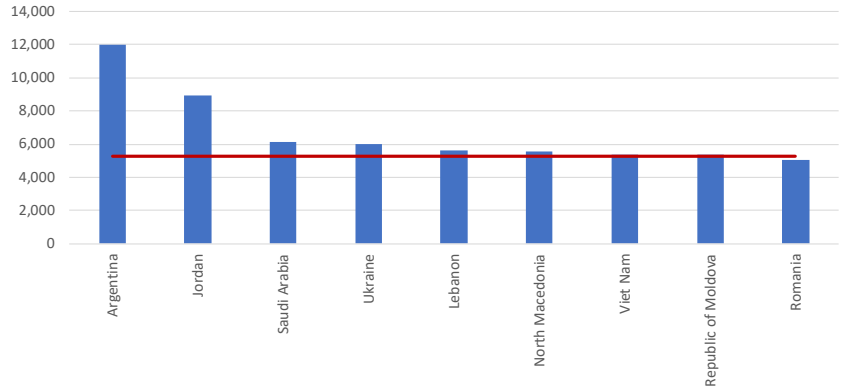


Figure 9.3 Sample yield for countries and economies participating in the PBA



Since the sample sizes varied greatly from country to country, the numbers of schools and the sample sizes from each school varied as well. As indicated in Table 9.1, the number of schools runs from 44 (Luxembourg) to 1,089 (Spain). But most countries met the requirement for the number of schools (a minimum of 150 schools).

The PISA assessment design also requires that students be randomly assigned to forms in the prescribed proportions. Results showed that this standard was met for all participating countries and confirmed that the sampling of students' responses to items was appropriate for item analyses and IRT scaling.

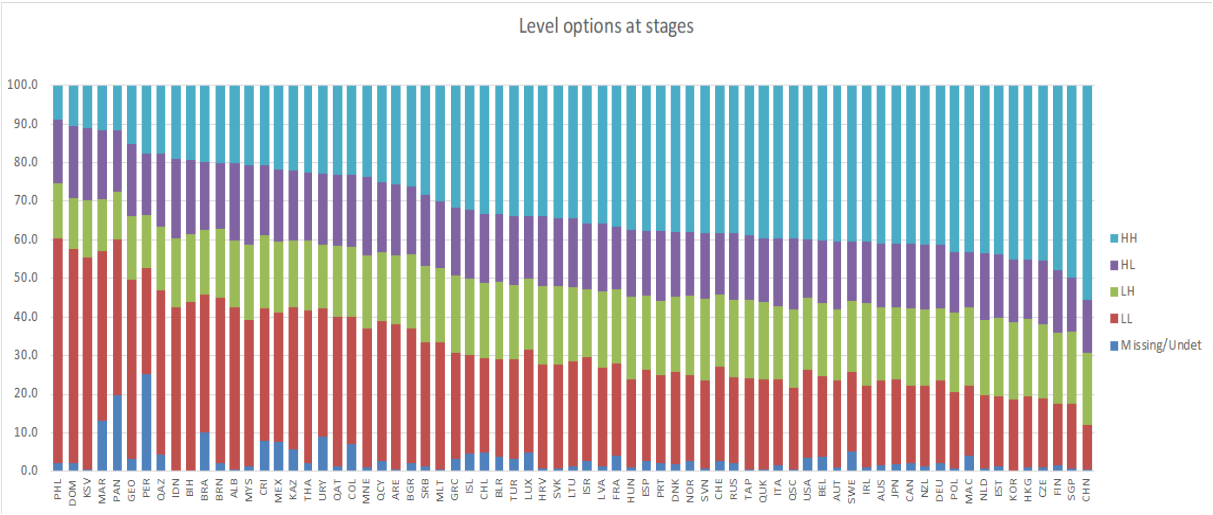
Reading MSAT data yield

The reading MSAT was designed to improve the measurement precision for countries and students across a wider range of proficiencies, and at the same time, collect the data needed for item analyses and IRT scaling (Chapter 2). Therefore, it was important to verify that the MSAT design was implemented to students as designed and intended. In particular, three aspects for MSAT design were closely monitored: 1) whether students were randomly assigned to each MSAT core testlet, 2) whether students were randomly assigned to either Design A (75%) or Design B (25%), and 3) whether students were routed to the stage 1 and stage 2 testlets according to the MSAT design (design A or design B).

First, results confirmed the random assignment at the Core stage worked well, with a uniform distribution of about 12.5% across eight different testlets. Second, Designs A and B were assigned in the desired proportions (75% and 25%, respectively) for all participating countries.

Third, the more complex routing of students from core to stage 1 and from stage 1 to stage 2 testlets were evaluated in detail. Figure 9.4 shows the proportions of students routed to difficult stage 1 and stage 2 testlets (labelled as "HH"; high options in both Stage1 and Stage2), the difficult stage1 and easy stage 2 testlets (labelled as "HL"; high option in Stage1 and low option in Stage2), the easy testlet in stage1 and difficult testlet in stage 2 testlets (LH), or easy testlets for both stage1 and stage 2 (LL). Some students were categorized as missing/undetermined when they did not complete certain stages, thus, were not assigned a stage 1 or stage 1 and stage 2 testlets. In the figure, the lowest to highest performing countries are shown from left to right. As intended, in lower performing countries, smaller proportions of students were assigned to more difficult testlets (skyblue and purple) while in higher performing countries, smaller proportions of students were assigned to easier testlets (red and green). However, as intended, every testlet was assigned to at least 25% of the total sample. As shown in Figure 9.4, adding up skyblue and purple (HH and HL) bars for the lowest performing countries, and adding up red and green (LL and LH) bars for the highest performing countries exceeded 25%. This confirmed that regardless of countries' proficiency distributions being lower or higher, the adaptive design always provided the minimum number of responses per item needed for IRT scaling and an appropriate item coverage across all range of item difficulties.

Figure 9.4 Reading MSAT proportions of students routed to higher and lower difficulty stage 1 and stage 2 testlets



Classical test theory statistics: item analysis

Classical item analyses were conducted on all computer and paper-based testing items at the national and international levels to verify that items functioned appropriately. Unexpected results were identified and explored for any indication of possible issues related to data collection, human- or machine-scoring, or other issues. Descriptive statistics for the observed responses and various missing response codes were provided to countries and the OECD for their review and feedback. Classical item analysis also provided nonmodel-based information useful for the review of IRT modelling outcomes.

The following statistics were computed:

- item scores frequencies (number of attempts, correct and incorrect responses, and non-responses—item omitted and not-reached)

- item difficulty
- item discrimination
- item score category statistics

Trend item statistics were compared with results from prior PISA cycles. Statistics were compiled separately for the paper-based and computer-based assessments and were examined at the aggregate level across countries. The analyses were also performed separately for each country to identify outlier items that worked poorly or differently across assessment cycles and countries/economies to detect flaws or obvious scoring rule deviations.

The PBA results included only paper-based student responses for the core domains of mathematics, reading, and science (trend items only). The CBA results included computer-based student responses for the core domains of mathematics, reading, and science, as well as financial literacy and global competence (trend items and new items in reading, financial literacy and global competence). Results were also produced by language within a country. *Une-heure* (UH) booklet results were provided for countries where applicable.

Tables 9.2 and 9.3 show examples of item analysis outputs. Table 9.2 shows the first three items in block/cluster M01. The first item, DM033Q01C, is the scored version of the paper item PM033Q01 (corresponding CBA output would be for item CM033Q01) – a multiple-choice item. Each table represents one item and the columns represent the different categories of response codes. The *total* column includes all categories except *NOT RCH* (not reached), which is determined by both non-response and position of the item; specifically, the students did not answer this item nor subsequent items in the cluster. For the CBA, timing data and process information could also be used to determine whether a student had sufficient time to respond to an item, but in this case, not reached for CBA items was determined solely on the non-response to an individual item. Items that did not perform properly in the field or were missing a human-coded response code were designated as NOT RCH to exclude them from item statistics. *OFF TSK* (off task) was used to differentiate the invalid missing category (when a student did not answer the question in the expected way, e.g., by giving a response not associated with the item or responded with more than one answer in an exclusive choice question). The mean score, standard deviation, biserial/polyserial, and point biserial/polyserial were based on the total block/cluster score.

The delta statistic, shown in Table 9.2, is an index of item difficulty based on P+ (proportion correct, or percent correct when expressed as a percentage) transformed so that it places items on a scale with a mean of 13.0 and a standard deviation of 4.0. Deltas ordinarily range from 6.0 for a very easy item (approximately 95% correct) to 20.0 for a very hard item (approximately 5% correct), with 13.0 corresponding to 50% correct.

The biserial or *R BIS* statistic is used to describe the relationship between performance on a single test item and a criterion (usually the total score on the test). It is an estimate of the correlation between the criterion and an unobservable variable—assumed to be normally distributed—that determines the performance on the item. The polyserial or *R POLY* statistic is the generalization of the biserial correlation for use with either dichotomous or polytomous items. It is a generalized form of the correlation between the criterion and the item score, where the item score is either (0, 1) or (0, 1, 2, 3….n), and the criterion is a continuous variable.

Table 9.2 Example output for examining response distributions.

BLOCK M01 (UNWEIGHTED)
Response Analysis

A View Room:Which plan best represents t

		1	NOT RCH	OFF TSK	OMIT	0	1				TOTAL	R BIS = 0.6064
ITEM 1	N	1	14	74	2054	5466					7608	PT BIS = 0.4551
	PERCENT	0.01	0.18	0.97	27.00	71.85					100.00	P+ = 0.7185
DM033Q01C	MEAN SCORE	7.00	5.00	1.22	3.59	7.31					6.25	DELTA = 10.69
	STD. DEV.	0.00	3.09	1.87	2.92	3.49					3.75	
TRN_MATH	RESP WT	0.00	0.00	0.00	0.00	1.00						ITEM WT = 1.00

Running Time:Which is the third fastest

		2	NOT RCH	OFF TSK	OMIT	0	1				TOTAL	R BIS = 0.6213
ITEM 2	N	8	0	98	2204	5299					7601	PT BIS = 0.4722
	PERCENT	0.11	0.00	1.29	29.00	69.71					100.00	P+ = 0.6971
DM474Q01C	MEAN SCORE	1.25	0.00	1.38	3.66	7.42					6.25	DELTA = 10.94
	STD. DEV.	1.48	0.00	1.66	3.06	3.41					3.75	
TRN_MATH	RESP WT	0.00	0.00	0.00	0.00	1.00						ITEM WT = 1.00

Population Pyramids:How many people (boy

		3	NOT RCH	OFF TSK	OMIT	00	11	12	13	21	TOTAL	R POLY = 0.8431
ITEM 3	N	20	1	1139	1639	335	530	201	3744		7589	PT POLY = 0.7118
	PERCENT	0.26	0.01	15.01	21.60	4.41	6.98	2.65	49.33		100.00	MEAN = 0.5636
DM155Q02C	MEAN SCORE	1.00	3.00	2.58	3.31	5.40	5.81	6.33	8.81		6.26	DELTA = 12.36
	STD. DEV.	0.55	0.00	2.02	2.45	2.48	2.69	2.71	2.84		3.74	
TRN_MATH	RESP WT	0.00	0.00	0.00	0.00	0.50	0.50	0.50	1.00			ITEM WT = 2.00

Table 9.3 has two parts. The first part shows a breakdown of the score categories and biserial correlations by category. The second part contains summary data for each item on a single row and reveals items that were flagged for surpassing certain thresholds. The thresholds are provided in Table 9.4. In this example, the third item is flagged for having an omit rate of greater than 10%.

Table 9.3 Example table of item score category analysis and item flags summary

BLOCK M01 (UNWEIGHTED)
Item Score Category Analysis (Partial credit model)

	Category	N	Pct. At	Pct. Below	Mean	Std. Dev.	Biserial	B *
ITEM 1	0	2142	28.15	0.00	3.52	2.93		
DM033Q01C	1	5466	71.85	28.15	7.31	3.49	0.6064	-0.9529
ITEM 2	0	2302	30.29	0.00	3.57	3.05		
DM474Q01C	1	5299	69.71	30.29	7.42	3.41	0.6213	-0.8303
ITEM 3	0	2779	36.62	0.00	3.01	2.31		
DM155Q02C	1	1066	14.05	36.62	5.78	2.65	0.6114	0.3033
	2	3744	49.33	50.67	8.81	2.84	0.5728	-0.8367

BLOCK M01 (UNWEIGHTED)
Item Analysis Flag Summary

Item ID	Num Resp	Type	R-BIS	P-PLUS	% NOTRCH	% OFFTSK	% OMIT	% MISS	Flags
DM033Q01	2	SCR	0.6064	0.7185	0.01	0.18	0.97	1.17	.....
DM474Q01	2	SCR	0.6213	0.6971	0.11	0.00	1.29	1.39	.....
DM155Q02	5	ECR	0.8431	0.5636	0.26	0.01	15.01	15.25	...O..

Table 9.4 Flagging criteria for items in the item analyses

	Criteria for flagging items
min rbis/rpoly	0.3
min P+	0.2
max P+	0.9
max Omit%	10
max Offtask%	10
max Not-Reached%	10

Reading MSAT Equated P+

In 2018, an *equated P+* statistic was developed and added to the statistics provided to CBA countries for reading. Because of the MSAT design for reading domain, the samples of students responding to different items are no longer randomly equivalent. The subsamples of students routed to the more difficult MSAT testlets are expected to be more proficient than the total sample of students, resulting in higher P+ values than would be expected from a sample randomly equivalent to the total sample. Conversely, the subsamples of students routed to the easier MSAT testlets are expected to be less proficient, resulting in lower P+ values than

expected from a randomly equivalent sample. Therefore, while classical observed P+ and other statistics are still helpful for identifying items with potential scoring or other issues, classical observed P+ values are no longer comparable across adaptive and non-adaptive designs (i.e., across cycles within countries/economies) or across countries/economies.

The equated P+ is equivalent to the P+ that would have been obtained from the non-adaptive designs used in previous PISA cycles. It accounts for the differences in proficiencies between the sample who responded to the item and the total sample. Therefore, the 2018 reading MSAT equated P+ can be compared with the 2015 (or earlier cycles) classical observed P+ as well as with each other across PISA 2018 countries. Computation details are provided later in this chapter.

Response time analyses

The computer-based platform captured response time data for all computer-based items delivered in the CBA countries in both the field trial and main survey. The timing data can be informative in evaluating the level of student engagement and effort over two-hour testing period for cognitive assessments. Very little time spent on the assessment was interpreted as low effort; too much time spend on the assessment could be an indication of technical problems or low ability. Response time information was aggregated as the amount of time spent by students on the full assessment, on each domain, and by cluster or testlet. Item response times by position and proficiency level were also computed. Overall, results indicate that the CBA data provided valid information that can be used to model items and estimate student performance within and across countries.

Outliers

Students were generally expected to complete the cognitive assessment within two one-hour periods separated by a break. Within each hour they had to follow the prescribed order of clusters or MSAT stages and units, at their own pace. Except for reading MSAT, students were expected to complete two clusters within one-hour, regardless of the positions within the assessment (cluster 1 and 2 in the first hour, cluster 3 and 4 in the second hour). Before or after the mid-test break – it was possible for some students to take additional time on the first cluster and less time on the second cluster. When assigned to reading, students had to complete the reading fluency part within the enforced limit of 3 minutes and were expected to complete their reading MSAT within an hour (taking 3 testlets; core stage 1 and stage 2), regardless of the position (first hour or second hour).

Focusing on larger than expected cluster response time, outliers were identified using the median absolute deviation approach (MAD; Rousseeuw & Croux, 1993; Leys et al., 2013). That is, when response time was greater than median$\{x_i\}$+4.4478*median$\{|x_i - \text{median}(x_j)|\}$, where $\{x_i\}$ is the collection of all sample values). In calculating the outliers, median values were computed across international data, not for each country-level data. In this way, the same criterion was used across countries and the identification of outliers was more stable.

Table 9.5 shows the percentages of response time outliers by domain. The proportions of outliers were small—less than 2 percent across all domains. Since, reading fluency was short and strictly time limited, reading fluency outlier analysis was not needed.

Table 9.5 Percentage of response time outliers by domain

Domain	Mathematics	Science	Reading	Global Competency	Financial Literacy
Number of Clusters	7	6	40 MSAT testlets*	4	2
Percent of Outliers	1.3%	1.5%	1.0%	1.7%	1.1%

* Instead of two clusters, reading MSAT forms include three testlets (core, stage 1 and stage 2) selected from a 40 testlet pool.

Cluster (or testlet) level response time

Table 9.6.a presents descriptive statistics for the cluster response times for the math, science, global competence and financial literacy domains, with the outliers excluded. These values are the sum of time each student spent on each item in a cluster, aggregated across students, countries and positions. The mean and standard deviation of cluster response times were similar across domains for all countries taking the CBA. On average, students spent about 20 minutes to respond items in each cluster, with 75% of the students completing the cluster in around 25 minutes. With the outliers removed, no student in any country took longer than 60 minutes to finish a given 30-minute cluster.

Some variability in assessment time was expected as test administrators had to log off the CBA during the break one by one. Still, students who took close to one hour to complete a given 30-minute cluster would be unlikely to have had enough time to finish the subsequent cluster with which it was paired. That is, for Math and Science, when the pair of clusters was administered before or after the mid-test break, the use of up to 60 minutes for the first of the two clusters would leave little to no time to finish the second cluster. Such long response times pointed to potential administration issues. Very short cluster response times of less than one minute also pointed to potential administration issues or technical problems with the data collection or pointed to a student very rapidly advancing through the items.

Table 9.6.b presents descriptive statistics for reading fluency and reading MSAT testlet response time. Most students responded to reading fluency items in less than 1 minute. With outliers removed, no students took more than 2 minutes. Regarding reading MSAT, on average, students took between 10.9 to 15.2 minutes to respond to the first, second and third testlets in the order testlets were presented[1]. Overall second testlet (Stage 1 items) included more items and took more time. Nevertheless, most students spent less than 20 minutes on any testlet. Considering reading fluency and reading MSAT altogether: on average students spent less than 40 minutes on the reading domain; most students spent less than 50 minutes; and very few reached the hour (all in Albania as shown in Figure 9.5 below). Therefore, ample time was provided to students to complete the reading fluency and reading MSAT as well as to complete any of the other domains assessed.

Table 9.6.a. Cluster response time (in minutes) descriptive statistics for non-adaptive domains

DOMAIN	MIN	Q1	MEDIAN	MEAN	Q3	MAX	SD	N
Mathematics	0.22	13.47	17.82	18.07	22.39	49.20	6.91	263,116
Science	0.17	15.42	20.41	20.68	25.55	56.49	8.18	263,207
Global Competency	0.16	15.39	20.20	20.43	25.05	55.32	8.05	71,172
Financial Literacy	0.14	15.59	20.61	21.49	27.27	56.12	8.87	46,181

[1] Note that "Reading Testlet1" indicates Core, "Reading Testlet2" indicates Stage 1 for Design A or Stage 2 for Design B, and "Reading Testlet 3" indicates Stage 2 for Design A or Stage 1 for Design A. Details about MSAT reading design is presented in Chapter 2

Table 9.6.b Stage response time (in minutes) descriptive statistics for reading fluency and MSAT reading domain

DOMAIN	MIN	Q1	MEDIAN	MEAN	Q3	MAX	SD
Reading Fluency	0.02	0.57	0.74	0.75	0.95	1.72	0.30
Reading Testlet 1	0.23	7.55	10.52	10.89	13.95	26.19	4.78
Reading Testlet 2	0.04	11.32	15.16	15.23	19.18	33.52	5.94
Reading Testlet 3	0.06	8.09	11.81	11.84	15.39	28.96	5.38
Reading MSAT	0.23	32.06	39.19	37.61	44.85	82.29	10.14
Reading Total (Reading Fluency + MSAT)	0.17	33.64	40.80	39.19	46.48	84.57	10.19

Response time and student performance

The relationship between response time and student performance was examined considering median of cluster-level response time by proficiency levels. The proficiency levels were computed based on the first plausible value (proficiency levels are explained in detail in Chapter 15). Table 9.7a and 9.7b show that, across all domains and up to level 4, more able students generally spent more time on each cluster or each MSAT testlet. The increase in time spent was most noticeable between students below level 1 and at level 2. Then, beyond level 4, students started to spend slightly less time.

Table 9.7a Cluster response time (in minutes) by proficiency level for non-adaptive domains

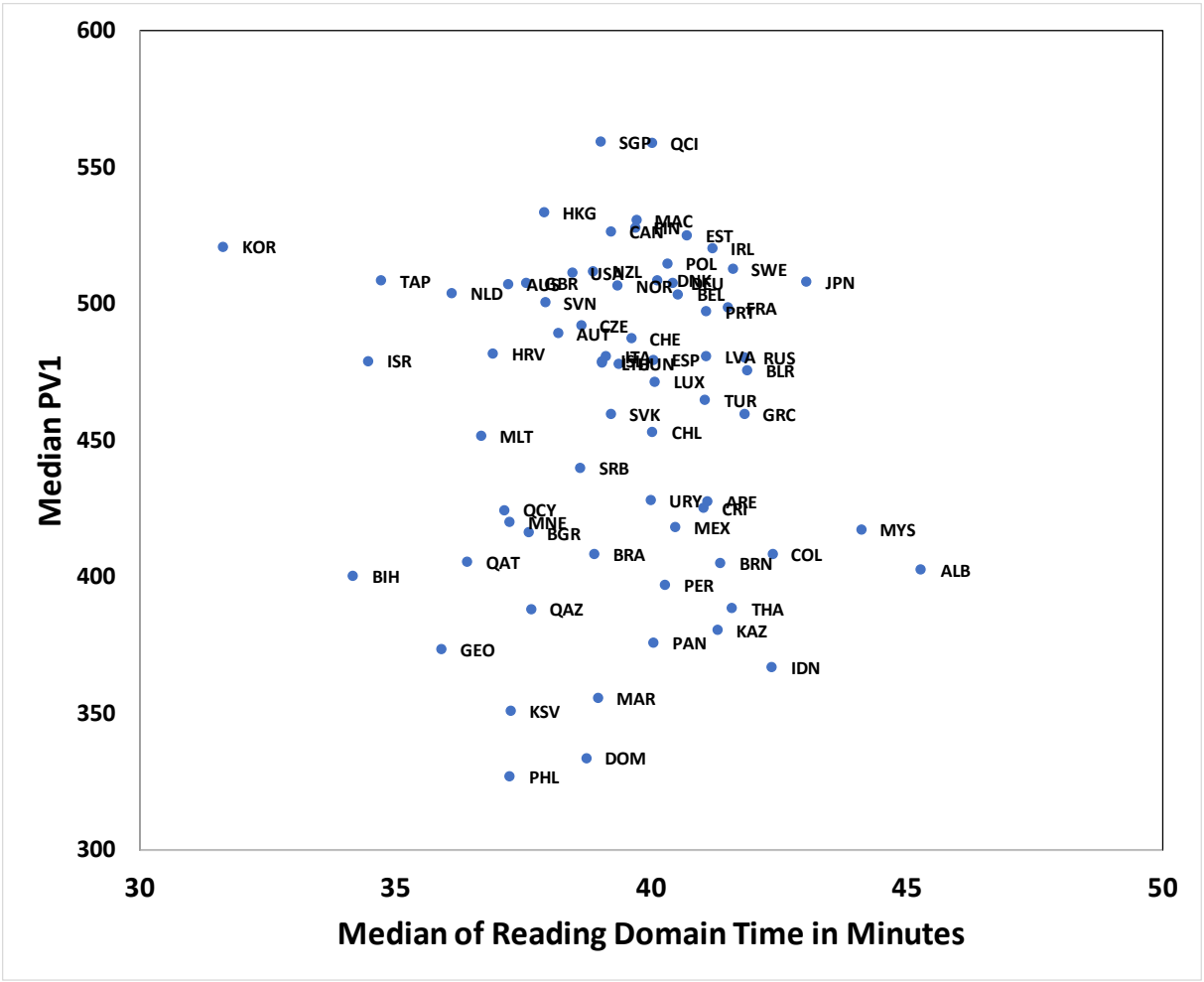
DOMAIN	*Below Level 1*	*Level 1*			*Level 2*	*Level 3*	*Level 4*	*Level 5*	*Level 6*	*N*
Mathematics	13.79	16.13			17.76	18.92	19.67	19.58	18.92	263,116
Science	11.58	**N/A**	14.43	17.67	20.31	21.78	22.16	22.13	21.72	263,207
Global Competency	17.23	19.66			20.97	21.34	21.02	20.55	**N/A**	71,172
Financial Literacy	14.63	17.46			20.00	21.40	21.97	21.74	**N/A**	46,181
		1c	**1b**	**1a**						

Table 9.7b Stage response time (in minutes) by proficiency level for MSAT reading

DOMAIN	*Below Level 1*	*Level 1*			*Level 2*	*Level 3*	*Level 4*	*Level 5*	*Level 6*	*N*
Reading Fluency	0.27	0.48	0.81	0.85	0.80	0.73	0.66	0.59	0.54	544,367
Reading Testlet 1	3.40	5.92	8.46	10.63	11.38	11.11	10.38	9.57	8.52	534,679
Reading Testlet 2	5.92	8.88	11.80	14.18	15.51	16.13	16.32	16.13	15.85	536,730
Reading Testlet 3	3.79	6.10	7.99	10.00	11.62	12.96	13.97	14.74	15.20	525,630
Reading MSAT	14.75	23.16	30.60	36.43	39.76	41.32	41.86	41.83	41.16	529,103
Reading Total (Reading Fluency + MSAT)	13.84	22.87	31.36	37.87	41.29	42.80	43.24	43.08	42.31	502,526
		1c	**1b**	**1a**						

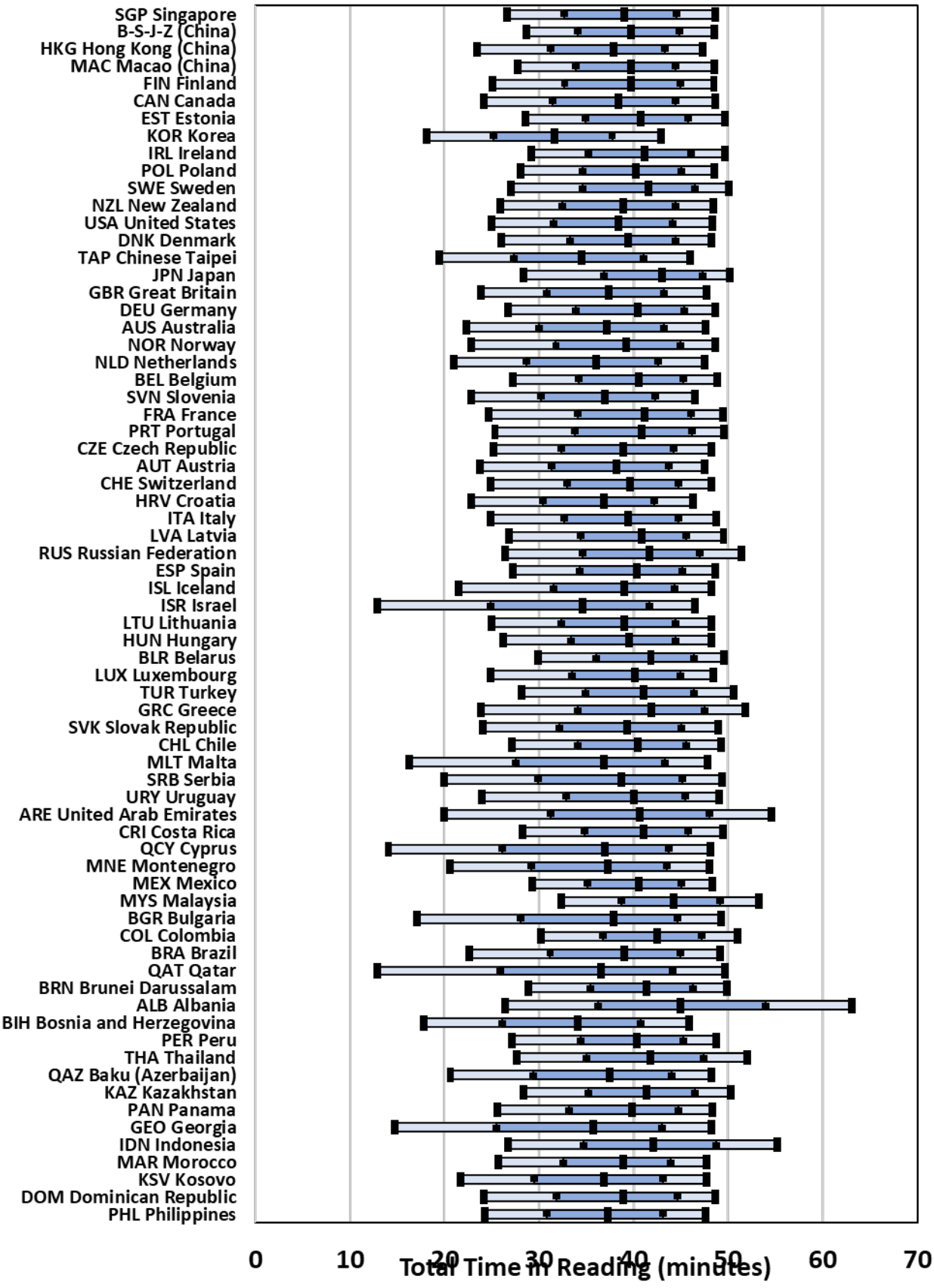
While the more able students generally need more time to complete the test, this is not the case when response time was aggregated at the country level. Figure 9.5 presents the relationship between median of total time spent for Reading domain and median of the first plausible value. Overall, Figure 9.5 shows that while countries do vary noticeably in their average proficiency, there is no clear relationship between average country proficiency and median total response time. For example, in the cases of Singapore and Korea, both have high average reading score, but Singapore's median response time is close to the overall median time while Korea's is unusually short.

Figure 9.5 Median response time versus country median proficiency score for all reading items



Because of differences in proficiency and other factors including motivation, the time it takes students to complete the assessment is expected to vary within each country. This is shown in Figure 9.6 indicating the distribution of total time spent on the reading MSAT for all countries sorted by performance. For each country, the middle black solid rectangle shows the median of total response time; the darker blue horizontal bars range from the 25th and 75th percentiles; and the lighter shader blue horizontal bars range from the 10th to the 90th percentile. The figure also suggests that the within-country response time variability is similar across countries regardless of differences in languages. Since reading is the major domain for PISA 2018 and all students are taking a one-hour MSAT, results are presented for reading only. Nevertheless, it can be noted that comparison with the science, the major domain in 2015 (OECD, PISA 2015 Technical report, chapter 9) show a similar pattern.

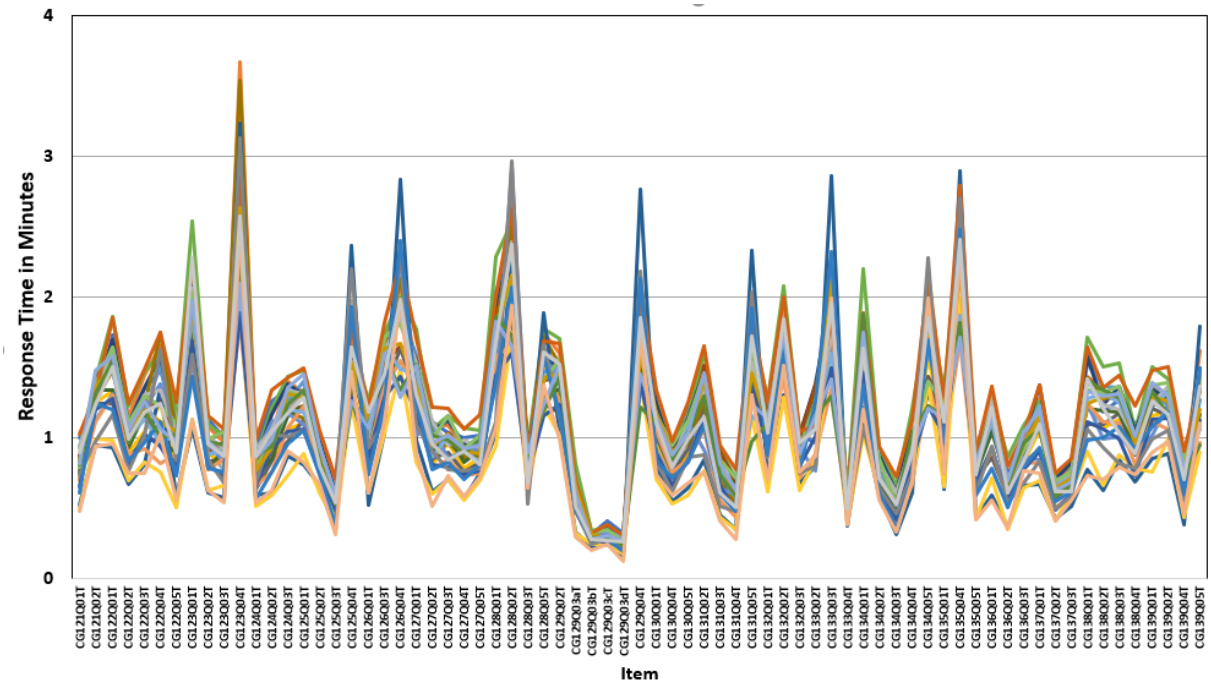
Figure 9.6 Variability of reading MSAT response time across countries


SGP Singapore
B-S-J-Z (China)
HKG Hong Kong (China)
MAC Macao (China)
FIN Finland
CAN Canada
EST Estonia
KOR Korea
IRL Ireland
POL Poland
SWE Sweden
NZL New Zealand
USA United States
DNK Denmark
TAP Chinese Taipei
JPN Japan
GBR Great Britain
DEU Germany
AUS Australia
NOR Norway
NLD Netherlands
BEL Belgium
SVN Slovenia
FRA France
PRT Portugal
CZE Czech Republic
AUT Austria
CHE Switzerland
HRV Croatia
ITA Italy
LVA Latvia
RUS Russian Federation
ESP Spain
ISL Iceland
ISR Israel
LTU Lithuania
HUN Hungary
BLR Belarus
LUX Luxembourg
TUR Turkey
GRC Greece
SVK Slovak Republic
CHL Chile
MLT Malta
SRB Serbia
URY Uruguay
ARE United Arab Emirates
CRI Costa Rica
QCY Cyprus
MNE Montenegro
MEX Mexico
MYS Malaysia
BGR Bulgaria
COL Colombia
BRA Brazil
QAT Qatar
BRN Brunei Darussalam
ALB Albania
BIH Bosnia and Herzegovina
PER Peru
THA Thailand
QAZ Baku (Azerbaijan)
KAZ Kazakhstan
PAN Panama
GEO Georgia
IDN Indonesia
MAR Morocco
KSV Kosovo
DOM Dominican Republic
PHL Philippines
0
10
20
30
40
50
60
70
Total Time in Reading (minutes)

Item-level response time

Response time and the relationships between response time and performance were also explored at the item-level. Typically, an item's median response time was highly consistent across countries. This is illustrated in Figure 9.7 showing the median item response time for all global competency items across countries, as an example. Each line with different color indicates different countries. Consistent pattern across countries suggests that students spent more time on the items that require longer time to solve, although there are relatively smaller differences across different languages and countries.

Figure 9.7 Median item response time by item in the global competency domain



Figures 9.8 and 9.9 show the median item-level response time in minutes of trend reading items and new reading items by students' proficiency levels (based on first plausible value; PV1), across all countries. The charts are sorted by the median of item-level response time. It is clear that low performing students (blue and orange lines) spent similar response times across all items but high performing students (grey and brown lines) showed larger variability in response times across items. In other words, the interaction between response time and performance by items was greater for high performing students than for low performing students. This pattern was consistently observed for both reading trend and new items.

Figure 9.8 Median item response time by proficiency level for reading trend items

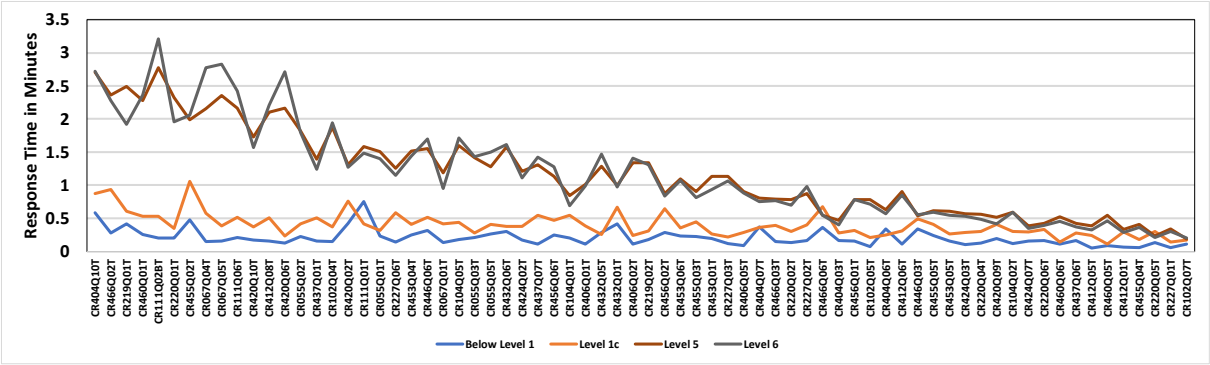
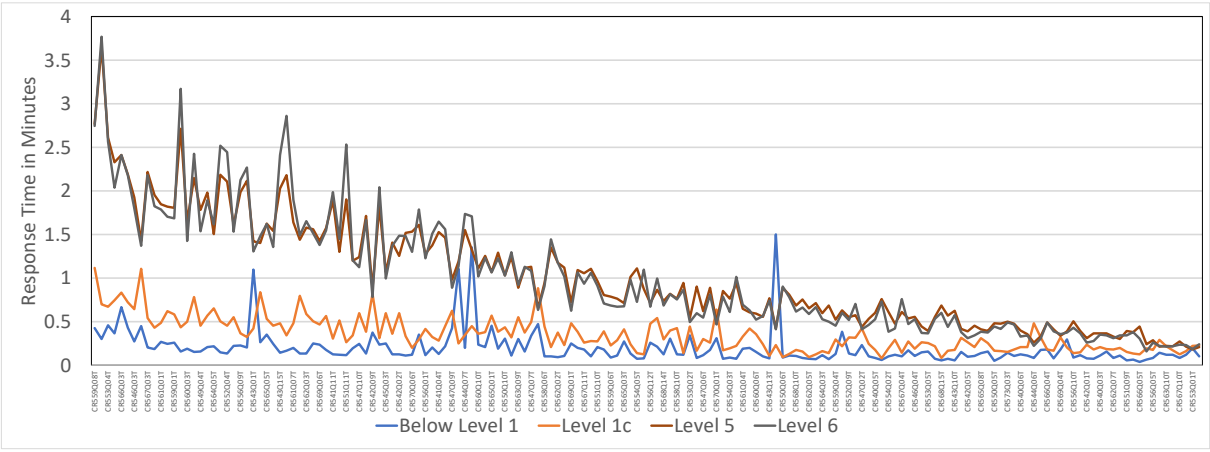


Figure 9.9 Median response time by proficiency level for reading new items



Response time reflecting possible motivation and administration issues

On average, students completed the entire test in 77.94 minutes, with SD of 17.68 and median response time of 80.47 minutes. Overall, about 1.9% of the students completed the test in less than 30 minutes. This type of students who spent extremely short time were found in nearly all countries. About 0.4% of the students across countries took longer than 120 minutes to complete the test. Some variability in assessment time was expected as test administrators had to log off the computer-based testing one-by-one. At the country-level, students in Albania, Colombia, Indonesia, and Malaysia took the longest median time to complete the test with 94.2, 89.0, 87.2, and 90.4 minutes, respectively. Students in Korea took the shortest median time, 63.5 minutes, to complete the test.

There were two countries where 3% or more of the students exceeded the time limit: Albania (11.5%) and Indonesia (3.3%); however, the unexpectedly high percentage of participants exceeding the time limit in Albania points to possible network issues rather than genuinely lengthy response times. Apart from this case, only a small proportion of respondents with very long or short total response times suggest that there were no systematic administration and/or motivation issues in specific schools. These students appear to be randomly distributed across schools and countries.

Position effects

Item position effects are a concern in large-scale assessment programmes because substantial position effects can increase measurement error and introduce bias in parameter estimation. For

example, a student may take the reading assessment in the first hour and then take two mathematics clusters in the second hour, while another student may take the same domains, reading and math, but in the reverse order. As in previous cycles, the PISA 2018 main survey design balanced item positions, particularly for Stage 2 items by introducing Design B in addition to Design A, in order to control and to monitor its impact on various statistics (Chapter 2, Figure 2.5). To evaluate and verify that the impact of item positions was minimal, item position effects were examined in terms of: 1) proportion of correct responses, 2) median response time, and 3) rate of omitted responses.

For all non-adaptive domains (all apart from reading), items were in fixed position within item clusters. A cluster may be first or second position within the first hour of testing, or in third or fourth position within the second hour of testing. The first and second hour of testing are separated by a short break. For the adaptive domain (reading), instead of assigning two non-adaptive clusters per hour period, three MSAT testlets (core, stage 1 and stage 2) were assigned to students in the first or second hour of testing. Furthermore, instead of item units being used in fixed position within a cluster, reading units were used in different positions depending on the testlets for a given MSAT stage (see Chapter 2). Therefore, position effects by unit (unit order effects), as well as by hour, were closely examined.

In particular, the PISA 2018 field trial was specifically designed to investigate unit order effects as a way to prepare the MSAT for the main survey. More specifically, the unit order was manipulated as either fixed or variable for randomly selected group of students. The field trial results confirmed the feasibility of introducing multistage adaptive testing in the main survey as unit order effects were found to be negligible. Nevertheless, two versions of the reading MSAT design were implemented in the main survey to help counterbalance the order of units and further ensure that unit order effects would be ignorable. These versions of the reading MSAT, referred to as Designs A and B, utilized the same testlets; however, stage 1 testlets in Design A became stage 2 testlets in design B and vice versa (stage 2 testlets became stage 1). See Chapter 2 for more detailed descriptions of the reading MSAT.

Tables 9.8a and 9.8b show the average proportion correct (P+) by cluster position for non-adaptive domains and by assessment hour for all domains in the CBA. As observed in PISA 2015 with mathematics, science, and financial literacy, the cluster position effects were computed as the difference between position 4 and 1 (Table 9.8a). The decreases in P+ between position 4 and 1 ranged from 0.02 in mathematics to 0.07 in financial literacy. The observed P+ position effects for the innovative global competency domain were comparable to the position effects observed in the other domains. Overall, cluster position effects were very similar to the values observed in 2015. Because of the MSAT design with three stages, reading was delivered for one hour instead of two independent 30-minute clusters. Thus, position effects were further evaluated by the assessment hour for reading as well as the other domains for comparisons (Table 9.8b). Generally, we see a smaller decrease in P+ between the 2nd and the 1st assessment hour than observed between 1st and 4th cluster position in the domains other than reading. For reading trend and new items, the decrease in average P+ between the 2nd and the 1st hour was relatively small and very similar to the decreases observed in the other domains.

Table 9.8a Average proportion correct (P+) by cluster position in the CBA for non-adaptive domains

DOMAIN	Position 1	Position 2	Position 3	Position 4	Position 4 - Position 1
Mathematics	0.42	0.41	0.40	0.39	-0.02
Science	0.47	0.42	0.44	0.41	-0.06
Financial Literacy	0.51	0.45	0.49	0.44	-0.07
Global Competence	0.41	0.38	0.38	0.36	-0.06

Table 9.8b Average proportion correct (P+) by assessment hour in the CBA for all domains

DOMAIN	1st Hour	2nd Hour	2nd Hour - 1st Hour
Mathematics	0.42	0.40	-0.02
Reading - trend	0.49	0.47	-0.02
Reading - new	0.55	0.52	-0.02
Science	0.45	0.43	-0.02
Financial Literacy	0.48	0.47	-0.01
Global Competence	0.40	0.37	-0.03

Tables 9.9a and 9.9b present the position effects in terms of median response time averaged by cluster position and by assessment hour. Math cluster results were similar to the 2015 results with about 4 minutes less spent in cluster position 4 than position 1. Science and financial literacy results showed some increase in median time difference between cluster position 4 and 1 when compared to 2015. Financial literacy in cluster position 1 took noticeably more time, resulting in larger differences between cluster positions 4 and 1. Global competency results were similar to the math and science results. As in 2015, there were indications that some students spent considerably more time on clusters 1 and 3, leaving them with less time for clusters 2 and 4. Table 9.9b shows the position effects by hour. As with P+, results by hour showed smaller differences than by cluster. For reading, decreases in median time between the 2nd and 1st hour were relatively small and somewhat lower than the decreases observed with the other domains.

Table 9.9a Median response time (in minutes) by cluster position for non-adaptive domains

DOMAIN	Position 1	Position 2	Position 3	Position 4	Position 4 - Position 1
Mathematics	20.42	17.17	17.06	16.09	*-4.33*
Science	26.52	16.89	21.98	17.47	*-9.05*
Financial Literacy	28.73	16.48	25.11	16.75	*-11.98*
Global Competence	25.40	18.20	21.29	17.77	*-7.62*

Table 9.9b Median response time (in minutes) by assessment hour for all domains

DOMAIN	1st Hour	2nd Hour	2nd Hour - 1st Hour
Mathematics	39.17	34.30	*-4.86*
Reading	40.75	37.98	*-2.78*
Science	44.73	40.98	*-3.75*
Financial Literacy	45.83	42.99	*-2.84*
Global Competence	44.62	40.01	*-4.61*

The omission rates at different positions for all CBA countries were analysed to further examine the quality of data affected by position. The omission rates are shown by cluster position and hour in Table 9.10a and 9.10b. These rates do not include 'not-reached' items, but omitted responses were treated as wrong responses in the scaling procedure, which affected item parameter estimation. Note that the omission rates for reading fluency are 0 since students had to respond to each item presented. Overall, omission rates by cluster and by hour and by

domain were very similar. As in PISA 2015, no omission rate for any domain in any position exceeded 0.10, and the omission rates in Positions 2 and 4 were higher than those in Positions 1 and 3, respectively.

Table 9.10a Omission rate by cluster position in the CBA for non-adaptive domains

DOMAIN	Position 1	Position 2	Position 3	Position 4	Position 4 - Position 1
Mathematics	0.06	0.07	0.07	0.08	0.02
Science	0.03	0.05	0.04	0.06	0.03
Financial Literacy	0.04	0.07	0.05	0.08	0.04
Global Competence	0.03	0.04	0.04	0.05	0.02

Table 9.10b Omission rate by assessment hour in the CBA for all domains

DOMAIN	1st Hour	2nd Hour	2nd Hour - 1st Hour
Mathematics	0.06	0.07	0.01
Reading - trend	0.07	0.08	0.01
Reading - new	0.05	0.06	0.01
Science	0.04	0.05	0.01
Financial Literacy	0.05	0.06	0.01
Global Competence	0.03	0.04	0.01

Position effects were also reviewed for the PBA. Tables 9.11a and 9.11b report the average proportion correct and average omission rates by cluster position. With so few countries participating in the paper-based version of the assessment, some discrepancies between previous cycles and between the PBA and CBA can be expected, however no major changes from past cycles or unusual differences between assessment modes were observed.

Table 9.11a Average proportion correct (P+) by cluster position in the PBA

DOMAIN	Position 1	Position 2	Position 3	Position 4	Position 4-Position 1
Mathematics	0.354	0.345	0.327	0.301	-0.053
Reading	0.517	0.497	0.472	0.430	-0.087
Science	0.425	0.417	0.398	0.364	-0.061

Table 9.11b Average proportion of omitted responses by cluster position in the PBA

DOMAIN	Position 1	Position 2	Position 3	Position 4	Position 4-Position 1
Mathematics	0.109	0.119	0.141	0.160	0.051
Reading	0.075	0.088	0.102	0.125	0.050
Science	0.066	0.079	0.094	0.114	0.048

Note: For reasons relating to data quality, Viet Nam was excluded from the calculations presented in these tables.

Position effects and motivation issues in reading MSAT

Further investigations for the reading MSAT were conducted regarding potential position effects and motivations issues.

Table 9.12 shows the average core, stage 1 and stage 2 testlet percent correct, not-reached and omission rates across all countries by stage position (Designs A and B). Note that position of stage 1 items and stage 2 items was swapped between Designs A and B, and exactly same set of items was used. When comparing the same testlets across Designs A and B, the results are very similar in terms of percent correct, not-reached ratio, and omission rates—for example,

not-reached ratio was 1.84 for Stage 1 in design A, and the corresponding value was 1.73 for Stage 2 in design B. Comparable cells between designs A and B are greyed.

Table 9.13 presents the cumulative median response time and completion rates at each stage by stage position. Similar to the above table, results by stage position were very close. For both Designs A and B, students spent about 11 minutes for solving Core stage items, about 36-37 minutes for solving Core and Stage1 items, and about 39 minutes for solving all three stages of items. Altogether, this confirms that Designs A and B provided comparable data although stage positions were switched for Stage 1 and Stage 2. Also, the completion rates were high in both Design A and Design B averaged across countries; about 96% of students could complete the MSAT reading regardless of stage position.

Table 9.12 Average proportion correct, not-reached and omitted rates for MSAT reading by stage position

	Proportion Correct		Not-Reached Rate		Omit Rate	
	Design A	Design B	Design A	Design B	Design A	Design B
Core Items	0.58	0.58	0.00	0.00	0.04	0.04
Stage 1 Items	0.55	0.48	0.02	0.13	0.05	0.07
Stage 2 Items	0.48	0.55	0.12	0.02	0.06	0.05

Table 9.13 Cumulative median response time and the completion rates by stage position

	Response Time (Median of Total min.)		Completion Rate (Proportion of valid cases)	
	Design A	Design B	Design A	Design B
Core	11.46	11.27	0.30%	0.30%
Stage1	37.02	36.33	3.80%	3.70%
Stage2	39.53	39.3	96.00%	96.10%

IRT MODELLING AND SCALING

The modelling and scaling of the PISA 2018 main survey data followed the general approach developed for PISA 2015 (PISA 2015 Technical report, Chapter 9). The following sections describe the IRT models and their assumptions, as well as the IRT scaling approach used in PISA 2018 for all domains. The scaling issues associated with the new reading MSAT design and how they were resolved are addressed. The model-based computations developed to produce the equated proportion correct (equated P+) described earlier are provided.

IRT Models and assumptions

As in the PISA 2015, the unidimensional multiple-group IRT model (Bock & Zimowski, 1997; von Davier & Yamamoto, 2004) based on the two-parameter logistic model (2PL) for the binary item responses and the generalized partial credit model (GPCM; Muraki, 1992) for the polytomous item responses was fit for each of the domains. The 2PLM is a generalization of the Rasch model (Rasch, 1960; Masters, 1982), which assumes that the probability of response x to item i depends on the difference between the respondent v's trait level θ_v and the difficulty

of the item β_i. In addition, the 2PLM postulates that for every item, the association between this difference and the response probability depends on an additional item discrimination parameter α_i:

9.1

$$P(x_i = 1|\theta, \beta_i, \alpha_i) = \frac{\exp(D\alpha_i(\theta - \beta_i))}{1 + \exp(\mathrm{D}\alpha_i(\theta - \beta_i))}$$

The probability of a positive response (e.g., solving an item correctly) is strictly monotonic, increasing with θ_v. The item discrimination parameter α_i, sometimes scaled by a constant D=1.7, characterizes how quickly the probability of solving the item approaches 1.00 with increasing trait level θ_v when compared to other items. In other words, the model accounts for the possibility that responses to different items do not have the same weight with relation to the latent trait. The discrimination parameter α_i describes how well a certain item discriminates between examinees with different trait levels compared to other items on the test.

The GPCM (Muraki, 1992), like the 2PLM, is a mathematical model for the probability that an individual will respond in a certain response category on a particular item. While the 2PLM is suitable for items with only two response categories (dichotomous items), the GPCM can be used with items with more than two response categories (polytomous items). The GPCM reduces to the 2PLM when applied to dichotomous responses. For an item i with m_i+1 ordered categories, the model equation of the GPCM can be written as:

9.2

$$P(x_i = k|\theta, \beta_i, \alpha_i, d_i) = \frac{\exp\{\sum_{r=1}^{k} \mathrm{D}\alpha_i\,(\theta - \beta_i + d_{ir})\}}{\sum_{u=0}^{m_i} \exp\{\sum_{r=1}^{u} \mathrm{D}\alpha_i\,(\theta - \beta_i + d_{ir})\}}$$

where d_i is the category threshold parameter.

Critical assumptions of most IRT models and the models used in the PISA are conditional independence (sometimes referred to as local independence) and unidimensionality. Under conditional independence, item response probabilities depend only on the latent trait θ and the specified item parameters—there is no dependence on any demographic characteristics of the students, responses to any other items presented in a test, or the survey administration conditions. Under the unidimensional assumption, a single latent variable, θ, accounts for performance on the full set of items. With past PISA data, these assumptions have been verified and item parameters have been estimated for each cognitive domain separately. These assumptions need to be confirmed for each domain in which any new items are used.

With this assumption, the formulation of the following joint probability of a particular response pattern $x = (x_1, \ldots, x_n)$ across a set of n items:

9.3

$$P(\boldsymbol{x}|\theta, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \prod_{i=1}^{n} P_i(\theta)^{x_i}\,(1 - P_i(\theta))^{1-x_i}$$

When replacing the hypothetical response pattern with the scored observed data, the above function can be viewed as a likelihood function that is to be maximised with respect to the item parameters. To do this, it is assumed that students provide their answers independently of one another and that the student's proficiencies are sampled from a distribution *f(□)*. The likelihood function is, therefore, characterised as:

9.4

$$P(\boldsymbol{X}|\boldsymbol{\beta},\boldsymbol{\alpha}) = \prod_{j=1}^{J} \int \left(\prod_{i=1}^{n} P_i(\theta)^{x_{ij}} \left(1 - P_i(\theta)\right)^{1-x_{ij}} \right) f(\theta) d\theta$$

Typically, the item parameters that provide the best possible fit to a given data set are estimated by maximising this function through a process called *item calibration*. The item parameters can then be used in the subsequent analyses such as the estimation of individual plausible values and population characteristics. However, it should be noted that IRT modelling does not provide an absolute scale, since any linear transformation of the item and latent trait parameters in the above formula lead to the exact same accounting of the data, often called as the scale indeterminancy. Therefore, as part of the calibration process, a choice must be made for the IRT scale to be determined.

For further information regarding the models discussed, see Fischer and Molenaar (1995) and van der Linden and Hambleton (1997, 2016), or von Davier and Sinharay (2014) for the use of these models in the context of international comparative assessments.

Item calibration and scaling

The PISA data collection designs are complex, and the assessments are adapted and translated for each participating country in one or more languages. To better account for cultural and language differences and to optimally scale the item parameters and proficiency estimates across countries, new calibration and scaling approaches were developed in 2015 (OECD, 2017, Chapter 9). For each domain, a series of multi-group concurrent calibrations of the historical data (2015 and prior PISA cycles) were conducted. As a result, all the items used in all the PISA cycles up to 2015 were estimated and scaled onto the same PISA IRT scale.

More specifically, during the first run of multi-group concurrent calibrations, the item parameters were constrained so that only one set of *common* or *international parameters* was estimated per item to model the data for all the country-by-language-by-cycle groups. As part of the calibration process, the fit (or misfit) of the common item parameters to the data for each pre-defined group (in PISA 2015, country-by-language-by-cycle group using the historic data) was evaluated. Then, item-by-group interactions were identified when the fit to the data was found to be poor (the fit statistic value was higher than a chosen threshold value). From the second calibration run, new *unique or group-specific* item parameters were estimated in the group or groups in which misfit was found. In the subsequent calibrations, the item fit threshold was gradually lowered until the target threshold is reached, thus allowing additional group-specific item parameters to be estimated. The fundamental of using this stepwise procedure is to optimize both the model data fit and the communality of item parameters across all groups—keeping common item parameters for as many groups as possible, or in other words minimizing the use of unique parameters. By allowing unique item parameters for items that show item-by-group interactions – in contrast to excluding such items or accepting poor common item parameter fit – the measurement error is reduced without introducing bias. The research base

for this approach can be found in Meredith (1993), Glas and Verhelst (1995), Glas and Jehangir (2014), Meredith and Teresi (2006), as well as Oliveri and von Davier (2014, 2011).

In PISA 2015 when the historic data was used, one set of common (international) parameters were estimated to model most country-by-language-by-cycle group data. Some items were allowed to have additional group-specific or unique item parameters used to model specific country-by-language-by-cycle groups (OECD, 2017, Chapter 12; von Davier, Yamamoto, Shin, Chen, Khorramdel, Weeks, Davis, Kong, & Kandathil, 2019).

The calibration and scaling for the PISA 2018 main survey followed the approach developed in 2015 and used the same IRT models. However, the historical data did not need to be included in the 2018 scaling since all trend items (reused from 2015 and/or prior PISA cycles) had already been calibrated and scaled in 2015. Therefore, in PISA 2018, a fixed item parameter linking approach was utilized with the trend item parameters fixed to their values established in the 2015 scaling. Then, along with the new items, the item fit analyses of the trend items were conducted to verify whether the fixed trend item parameters are applicable to the 2018 data and whether there is any necessity to re-estimate when it is not the case.

Item model data fit analyses (i.e., item fit or misfit) are a critical part of the scaling analyses described above. Different types of differential item functioning (DIF) statistics can be used to evaluate the extent to which the item model applied to a particular group fits the response data collected from that group. In the context of the IRT models used in PISA since 2015, the extent to which the model-based item characteristic curve (ICC, computed using formula 9.2 or 9.4 with the 2PL or the GPCM) and the empirical ICC differ can has been evaluated based on the mean deviation (MD) and the root mean square deviation (RMSD) statistics.

9.5

$$MD = \int \left[p_g^{obs}(X|\theta) - p_g^{exp}(X|\theta)\right] f_g(\theta) d\theta$$

9.6

$$RMSD = \sqrt{\int \left[p_g^{obs}(X|\theta) - p_g^{exp}(X|\theta)\right]^2 f_g(\theta) d\theta},$$

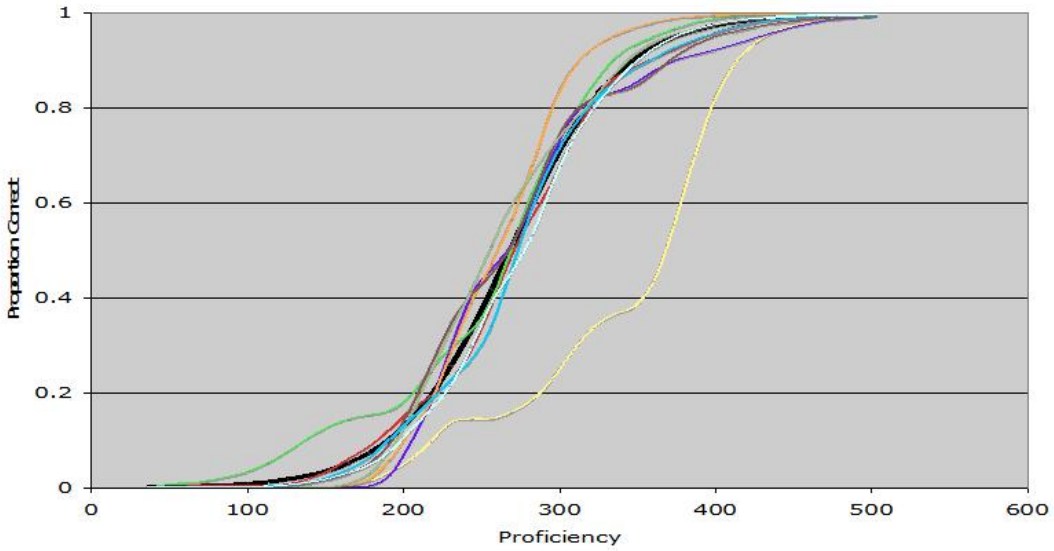
where $g = 1, \ldots, G$ is a country-by-language group; $p_g^{obs}(X|\theta)$ and $p_g^{exp}(X|\theta)$ are the observed and expected probability of correct response given proficiency θ; and $f_g(\theta)$ is the distribution for the specific group g.

Both MD and RMSD quantify the magnitude and direction of deviations in the observed data from the estimated common or group-specific item characteristic curves for each single item. However, while MD is more sensitive to the difference in observed and model-based item difficulty, RMSD is sensitive to the differences in both item difficulty and item discrimination (slope).

To demonstrate the use of item fit statistics (RMSD, MD), Figure 9.10 shows one example plot for a dichotomously scored item estimated via 2PLM. It illustrates how the common item parameter fits data from all groups, except for one group. In the figure, the solid black curve is the model-based 2PLM item response curve that corresponds to the international item parameters; the other lines are observed proportions of correct responses along the proficiency

scale (horizontal axis) for the data from each group. This plot indicates that the IRT model-based curve conforms to the observed data; proportions of correct responses given the proficiency that are quite similar for most countries. However, the data for one country, indicated by the yellow line, shows a noticeable departure from the common item characteristic curve and curves for other groups. This item is far more difficult in that particular country, conditional on proficiency level. Thus, a unique set of parameters would be estimated for this group for this item.

Figure 9.10 Item response curve (ICC) for an item where the common item parameter is not appropriate for one group



Reading MSAT item calibration and scaling

In Reading MSAT, higher performing students were likely to be assigned to more difficult items and lower performing students to easier items by selecting sets of items adaptive to the student's ability. As a result, it is no longer assumed that students' abilities responding to different items are randomly equivalent in the MSAT, unlike when non-adaptive designs were employed in previous cycles.

In most testing situations where MSAT is administered, item parameters are pre-calibrated and treated as known, then item fits are evaluated to examine the parameter drift (Glas, 2010). However, PISA estimates item parameters and evaluates item fit using the national representative sample in the main survey, not only for new items but also for trend items. Under these circumstances, the calibration procedures typically used with non-adaptive data may not be appropriate for MSAT data (Jewsbury & van Rijn, in press). Therefore, building on available research and experience with the successful implementation of MSAT in PIAAC (OECD, 2013), the reading MSAT design was developed to balance PISA's need for accurate measurement across a broader proficiency range with the need for sufficient item responses for all items in all countries to be accurately estimated. Calibration approaches were studied using simulation and main survey data to develop and confirm that the approach implemented for the PISA 2018 main survey reading MSAT produced the desired results.

First, unit order effects were examined in the field trial as a pre-requisite to introduce the MSAT for PISA in preparation for the 2018 main survey (Chapter 2). Second, in designing the Reading MSAT, certain linkage across different blocks were considered as a crucial factor. It was because item parameters are likely to be biased due to selection effects without a satisfactory level of item linkage at each stage. Finally, before the main survey implementation, the effectiveness of the reading MSAT design was confirmed. Results from simulation studies showed that the calibration of the simulated MSAT data produced item parameters with the desired level of accuracy (Yamamoto, Shin, & Khorramdel, in press).

The scaling procedure for the reading MSAT main survey data was further investigated. In particular, the model data fit from the same calibration approach used for other non-adaptive domains and alternatives that incorporated MSAT-specific information, such as routing outcomes to define the group in the multi-group calibration process, were evaluated. In the end, the same approach used for the calibration of the other domains (delivered in the regular non-adaptive mode) was retained. A recent study (Jewsbury, Lu, & van Rijn, under review) also provides theoretical justification for this choice.

Reading MSAT equated P+

Unlike the previous nonadaptive computer-based assessment (CBA) design in which students were randomly assigned to forms and items, MSAT selects testlets based largely on student performance. More specifically, the MSAT administers one testlet at the core stage, then routes most students to either easy or difficult testlets at stage 1 and stage 2 depending on students' performance and, to a smaller degree, a random factor (Yamamoto, Shin, & Khorramdel, 2018). Thus, the subsamples of students routed to the more difficult (or easier) MSAT testlets and the items that belong to these sections can be expected to be more (or less) proficient than the total sample of students. As a result, CBA and MSAT classical item statistics are not comparable. In particular, the observed proportion (or percent) correct (P+) for items included in the most difficult or easiest testlets will be different from values obtained in previous nonadaptive PISA cycles.

To facilitate comparisons between MSAT and nonadaptive CBA items, a new equated proportion correct statistic, P+ EQ, has been reported for the PISA 2018 main survey reading MSAT items. It is computed based on the item modelling and the mean deviation (MD) statistic described earlier.

Equated proportion correct statistics are not new. They have been used by many testing programs to verify and ensure the comparability and quality of operational items and provide the information needed to assemble new test forms. Typically, they are needed when the sampled population changes over repeated test administrations or when some form of adaptive testing is employed.

IRT that is used to model PISA data can easily provide model-based P+, and when the model fits the data, the observed and model-based P+ for the sample of students who answered the item are the same. IRT models can also be used to accurately estimate the proportion correct for any other sample of students with known proficiency. When the sample for which the model-based P+ is produced is a reference sample, the results are called equated item proportion correct, noted as P+ EQ. For the dichotomously scored and partial credit items used in PISA and modelled using the two-parameter logistic model and generalized partial

credit models (2PLM and GPCM described above), the P+ EQ can then be computed as follows (Ali & Walker, 2014):

9.7

$$P + EQ_g = \int wp_g(X|\theta)f_g(\theta)d\theta, \quad g = 1, \ldots G$$

$P + EQ_g$ is computed for each country-by-language group (g = 1, … ,G) with each student having a sample weight w; $p_g(X|\theta)$ is the probability of correct response given proficiency θ; and $f_g(\theta)$ is the distribution for the specific group g. To facilitate computations, the integral in the equation can be approximated using a proficiency point estimate $\hat{\theta}_s$:

9.8

$$\int wp_g(X|\theta)f_g(\theta)d\theta \approx \frac{1}{\sum_{s=1}^{N_g} w_s}\sum_{s=1}^{N_g} w_s p_g\left(X|\hat{\theta}_s\right).$$

However, because PISA assesses and scores many countries together on the same common scale, some relatively small degree of item model misfit may remain. The MD statistic quantifies this misfit as the difference in proportion correct between the model prediction and the observed responses. It can be added to the estimation to account for the misfit and provide more accurate P+ EQ values:

9.9

$$P + EQ_g = \frac{1}{\sum_{s=1}^{N_g} w_s}\sum_{s=1}^{N_g} w_s p_g\left(X|\hat{\theta}_s\right) + MD.$$

MD is defined as:

9.10

$$MD = \int\left[p_g^{obs}(X|\theta) - p_g^{exp}(X|\theta)\right]f_g(\theta)d\theta;$$

MD is computed based on the deviation between observed and expected item characteristics curves (ICCs) (von Davier, 2005). $p_g^{obs}(X|\theta)$ represents the observed ICC and $p_g^{exp}(X|\theta)$ the expected ICC given student ability θ. In addition, $f_g(\theta)$ is the group-specific weight on the students' ability scale. The observed ICC is obtained from the observed responses across students for each item, and the expected ICCs are computed based on the IRT model using the estimated item parameters. The integral in MD was approximated with Gaussian quadrature points q ranging from -5 to 5:

9.11

$$\int\left[p_g^{obs}(X|\theta) - p_g^{exp}(X|\theta)\right]f_g(\theta)d\theta \approx \sum_{q=1}^{Q}\left[p_g^{obs}\left(X|\theta_q\right) - p_g^{exp}\left(X|\theta_q\right)\right]f_g\left(\theta_q\right).$$

$P + EQ_g$ values were reported to be between 0 and 1, regardless of response type (dichotomous or polytomous). For polytomous items (partial credit items), this was done by averaging the probability of correct response $p_g(X|\theta)$ over the of partial credit categories:

9.12

$$p_g(X|\theta) = \frac{1}{Z}\sum_{z=0}^{Z} z p_g(X = z|\theta), \quad z = 0, 1, \ldots Z$$

where $p_g(X = z|\theta)$ is the probability of category z and is computed from the generalized partial credit model.

Figures 9.11a and 9.11b show examples of scatterplots of P+ versus P+ EQ for a PISA 2018 main survey participating country. Items are identified by the MSAT stage (core, stage 1 and stage 2) and the difficulty of the testlets to which they belong (Low, Low-High, High-Low, or High). The Figure 9.11a results were obtained without using MD in the equating and the Figure 9.11b results with MD. The figures show the effectiveness of the equating and the benefit of adjusting for misfit by including MD in the computations: core item P+ and P+ EQ were nearly identical (more so in Figure b), as they should be, since no adaptation had yet taken place; stage 1 and stage 2 results showed the expected pattern of adjustment. That is, focusing on Figure 9.11b, we see that as stage 1 and stage 2 testlets are adaptively selected for more (or less) proficient subsamples of students, the items in testlets belonging to High MSAT testlets have lower P+ EQ than P+, and items in Low testlets have higher P+ EQ than P+. These patterns, as well as the Low-High or High-Low patterns (for items that belong to both Low and High testlets) that show much smaller differences between P+ and P+ EQ, are expected given the PISA MSAT design. This confirmed the effectiveness of the equating and the choice of including MD in P+ EQ computations. Note that when the P+ values were extreme—outside of (0.025, 0.975) —the P+ EQ computations were carried out without the MD adjustment. This avoided some instances when estimates would have been negative or greater than 1. To avoid any possibility of estimates that were less than 0 or greater than 1, P+ EQ was also limited to (0, 1).

Figure 9.11a. Scatterplot of item proportion correct (P+) and equated proportion correct (P+ EQ) for one country; P+ EQ computed without adjusting for item misfit (mean deviation, MD)

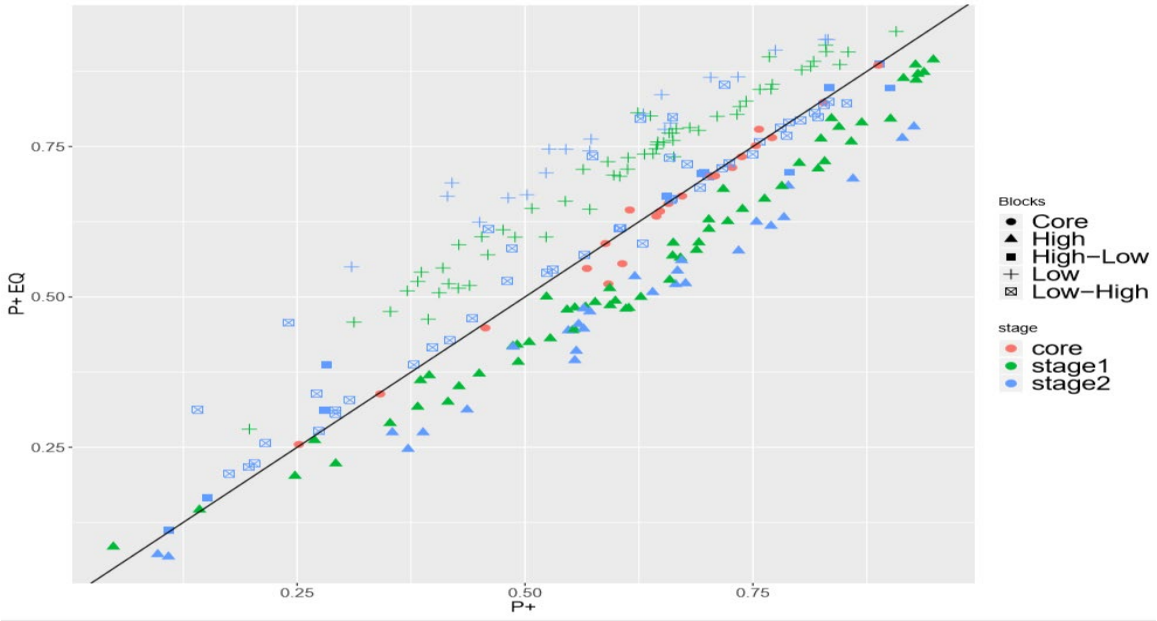


Figure 9.11b Scatterplot of item proportion correct (P+) and equated proportion correct (P+ EQ) for one country; P+ EQ computed adjusting for item misfit (mean deviation, MD)

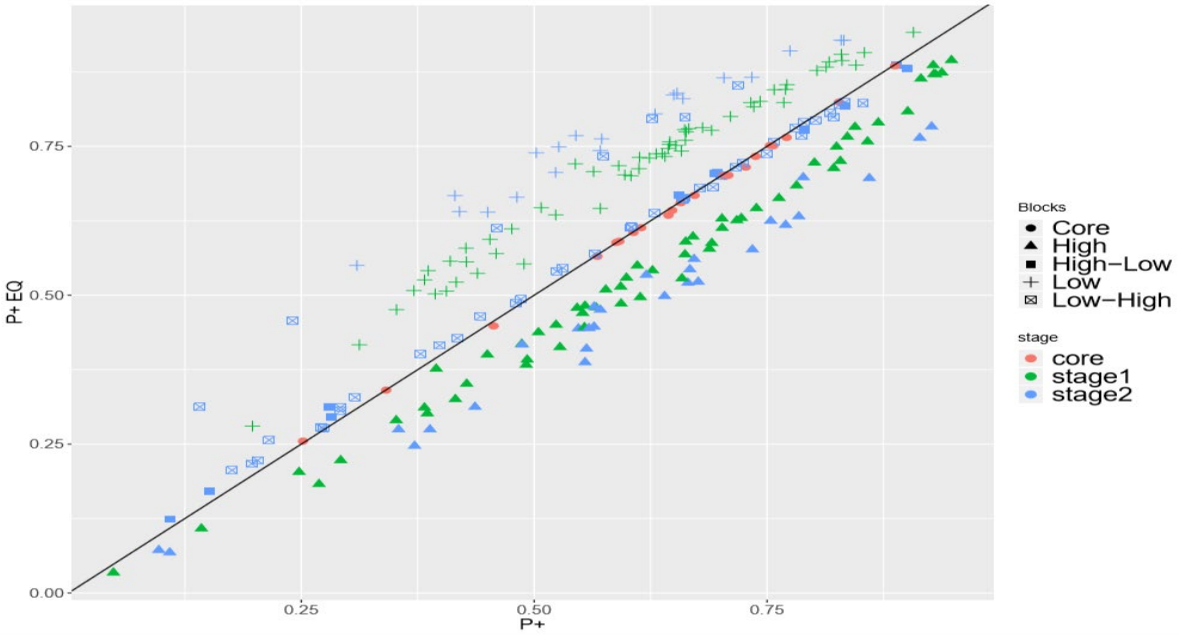


Table 9.14 summarizes the effect of the MSAT design on Reading P+ and shows the average differences between P+ and P+ EQ by testlet type across countries. On average, for stage 1 and stage 2 High testlets and Low testlets, the percent correct P+ EQ is further from the grand mean P+ by approximately 8-12%. With stage 2 High-Low or Low-High, the MSAT effect was less than 2% on average. As expected with the core, there was no noticeable effect.

Finally, Table 9.15 shows that the consistency of the MS 2018 Reading trend item P+ EQ with the Reading 2015 is similar to that of the Science trend items P+ between 2015 and 2018 (similar correlation values), whereas the P+ is less consistent (lower correlations). This further confirms that P+ EQ produces estimates that are comparable to observed P+ from previous nonadaptive PISA cycles. The same approach will be applicable in the future as the use of MSAT is extended to other domains.

Table 9.14 PISA 2018 Reading average item percent correct P+ and P+ EQ across countries by MSAT stage and testlets in which the items are included

Stage	Testlets	P+	P+ EQ
All	All	52.2	53.2
Core	Core	57.7	58.0
Stage 1	High	53.1	46.3
Stage 1	Low	53.7	62.0
Stage 2	High	51.8	41.9
Stage 2	High and Low	47.7	48.5
Stage 2	Low and High	48.0	50.7
Stage 2	Low	52.6	66.5

Table 9.15 PISA 2015 and 2018 correlations between Reading item P+ EQ and P+ by stage, and between science item P+'s overall

	Correlation (2015 P+, 2018 P+)	Correlation (2015 P+, 2018 P+ EQ)	Correlation (2018 P+, 2018 P+ EQ)

Reading trend by stage			
Core	0.98	0.97	1.00
Stage 1	0.89	0.97	0.93
Stage 2	0.92	0.96	0.95
Science trend			
All	0.97		

In short, these findings confirmed the effectiveness of the equating procedure in providing comparable item proportion correct across adaptive and nonadaptive designs. As expected, since core testlets were randomly assigned, the 2018 core MSAT P+ and P+ EQ were found to be equivalent and equally comparable to the P+ from 2015. Similar results were found for the stage 2 items included in both low and high or high and low testlets. For stage 1 and stage 2 items that belonged to either high or low MSAT testlets, results clearly showed differences between P+ and P+ EQ. This was expected because the items were assigned mostly to students performing consistently high or low. Overall, the 2018 MSAT P+ EQ and 2015 CBA P+ for the trend items were found to be comparable in the same way the P+ science trend items were comparable between 2015 and 2018. As PISA expands its use of adaptive designs to other domains, this same procedure will continue to be applicable.

LATENT REGRESSION MODEL AND POPULATION MODELLING

This section describes the population (or conditioning) model – a combination of an IRT model and a latent regression model – employed in the analyses of the PISA data and explains the multiple imputation or "plausible values" methodology that aims to increase the accuracy of the estimates of the multivariate proficiency distributions for various subpopulations and the population as a whole.

Tests that are reported at the individual-level are concerned with accurately assessing the performance of individual test-takers for the purposes of diagnosis, selection, or placement. The accuracy of these measurements can be improved by increasing the number of items administered to the individual. Thus, individual-level reporting tests containing more than 70 items are common. Because the uncertainty associated with each test-taker's proficiency θ is small, the distribution of proficiency, or the joint distribution of proficiency with other variables, can be approximated using point estimates. However, point estimates that are (in some sense) optimal for individual-level reporting could lead to seriously biased estimates of population characteristics, thus, not appropriate for group-level reporting such as PISA (Wingersky, Kaplan, & Beaton, 1987; Mislevy, 1991; Thomas 2002; von Davier, Sinharay, Oranje, and Beaton, 2006; von Davier, Gonzalez, Mislevy, 2009).

The prime goal of PISA is to compare the skills and knowledge of 15-year-old students across countries and economies, focusing on group-level scores (Kirsch, Lennon, von Davier, Gonzalez, & Yamamoto, 2013). In PISA, reporting outcomes does not entail consequences of any sort for the individual test taker, and test forms are kept relatively short to minimise individuals' response burden. At the same time, PISA aims to achieve broad coverage of the tested constructs. The full set of items is organised into different, but linked, test forms in assessment designs; each student receives only reasonable number of items that they can solve in two-hour testing period. For example, in the PISA 2018 Reading MSAT, students responded 33-40 items out of 245 items in total. Thus, the survey solicits relatively few

responses from each student on any one domain while maintaining a wide range of content representation when responses are aggregated. The advantage of estimating population characteristics more efficiently is offset by the inability to reliably measure and make precise inferences about individuals' performance on a single domain.

In the case of PISA, improved proficiency distributions are derived based on the (small) number of responses to cognitive assessment and contextual information collected from the PISA BQ. In addition, the covariance among skill domains (e.g. the PISA core domains mathematics, reading and science) is utilised to further improve the estimation of skill distributions. The *plausible value methodology* uses these proficiency distributions and accounts for error (or uncertainty) at the individual level by using multiple imputed proficiency values (plausible values) instead of single point estimates. Retaining this component of uncertainty in multiple imputed values requires that additional analysis procedures be used to estimate student proficiencies.

Population modelling for PISA 2018 followed the same general approach used in previous cycles. This approach incorporates the IRT scaling of the students' cognitive data from multiple domains and the students' background data specified as covariates (e.g. gender, country of birth, reading practices, academic and non-academic activities and attitudes) through multivariate latent regression models (von Davier et al. 2006). Data from multiple cognitive domains are modelled altogether to increase the accuracy of the estimates in each domain by borrowing information from the other cognitive domains. Plausible values are drawn from the posterior distributions constructed through the multiple latent regression model and the student data. Computations are carried through the following steps:

1. *IRT scaling of the cognitive responses:* estimates the item parameters that provide the comparable latent scales across countries and cycles in each content domain.
2. *Multivariate latent regression:* estimates the model's regression coefficients (Γ) and the residual variance-covariance matrix (Σ) ; the item parameter estimates from step 1 are taken as true (known) parameters.
3. *Plausible value generation:* draws 10 plausible values for each student and each domain from posterior distributions drawn from and the estimated Γ and Σ (Mislevy & Sheehan, 1987; von Davier, Gonzalez, & Mislevy, 2009).
4. *Variance estimation:* estimates the variance of the proficiency means and other statistics of interest for each country and subgroups from the plausible values through a replication approach (see Johnson, 1989; Johnson & Rust, 1992; Rust, 2014).

Regarding the step 2, more details are useful. First, all variables in the BQ are contrast coded. Contrast coding allows for the inclusion of refused responses, avoiding the necessity of linear assumption. Second, principal components analysis was conducted to accommodate a huge number of contrast-coded BQ variables. The use of principal components also serves to retain information for students with missing responses to one or more BQ variables. Principal components, accounting for a large proportion of the variation in the BQ variables, were used in the latent regression models instead of the observed responses of context questionnaire variables. This process is done at the country-level to accommodate common BQ variables collected across all participating countries and optional specific BQ variables of country's interest.

Mathematical expressions follow. The latent regression gives an expression for student's proficiency distribution conditional on covariates. For the multivariate latent regression models, multivariate latent proficiency distributions are assumed to follow multivariate normal distributions:

9.13

$$\boldsymbol{\theta} \sim N(\mathbf{y}\Gamma, \Sigma)$$

with the regression coefficients Γ and Σ that are estimated conditional on the previously determined item parameter estimates (from the item calibration stage). Γ is the matrix of regression coefficients and Σ is a residual variance-covariance matrix. Alternatively, the latent regression model of Θ on Y with $\Gamma = (\gamma_{sl}, s = 1,\ldots,S; 1 = 0,\ldots,L)$, $Y = (1, y_1,\ldots,y_L)^t$, and $\Theta = (\theta_1,\ldots, \theta_S)^t$ can be described as follows:

9.14

$$\theta_s = \gamma_{s0} + \gamma_{s1}y_1 + \ldots + \gamma_{sL}y_L + \varepsilon_s$$

where ε_s is an error term for the assessment skill s. As indicated, the covariates determine the mean of this density, often called as a prior distribution.

The residual variance-covariance matrix can then be estimated using the following formula:

9.15

$$\Sigma = \Theta\Theta^t - \Gamma(YY^t)\Gamma^t$$

The expectation-maximization (EM) algorithm is used for estimating Γ and Σ which is described in Mislevy (1985) for the unidimensional case. A multidimensional variant of the latent regression model based on Laplace approximation (Thomas, 1993) is applied in reporting PISA proficiencies on more than two skill dimensions (content domains).

Based on the Bayes' theorem, posterior distribution of skills given the observed item responses and covariates (i.e., contextual information) are constructed as follows:

9.16

$$P(\theta_j|\boldsymbol{x}_j, \boldsymbol{y}_j, \Gamma, \Sigma) \propto P(\boldsymbol{x}_j|\theta_j, \boldsymbol{y}_j, \Gamma, \Sigma)P(\theta_j|\boldsymbol{y}_j, \Gamma, \Sigma) = P(\boldsymbol{x}_j|\theta_j)P(\theta_j|\boldsymbol{y}_j, \Gamma, \Sigma)$$

where θ_j is a vector of scale values (these values correspond to performance on each of the skills) for student j. As shown, the posterior distribution of proficiency is proportional to the likelihoods of the data and prior distributions. $P(x_j|\theta_j)$ is the product over the scales of the independent likelihoods induced by responses to items within each scale, and $P(\theta_j|y_j, \Gamma, \Sigma)$, which is a prior distribution, is the multivariate joint density of proficiencies of the scales,

conditional on the principal components y_j derived from background responses, and parameters Γ and Σ. The item parameters are fixed and regarded as population values in the computation. This two-step approach is generally used to reduce computational burden ("divide-and-conquer", Patz & Junker, 1999). Then, plausible values for each student j are then drawn from the posterior distribution.

Plausible values are drawn following a three-step process. First, a value of Γ is drawn from a normal approximation to $P(\Gamma,\Sigma|x_j,y_j)$ that fixes Σ at the value $\hat{\Sigma}$ (Thomas, 1993). Second, conditional on the generated value of Γ (and the fixed value of $\Sigma = \hat{\Sigma}$), the mean m_j^p and variance Σ_j^p of the posterior distribution are computed using the same methods applied in the EM algorithm. In the third step, the θ are drawn independently from a multivariate normal distribution with mean vector m_j^p and posterior co-variance matrix Σ_j^p. These three steps were repeated 10 times, producing 10 imputations of θ for each sampled student.

ANALYSIS OF DATA WITH PLAUSIBLE VALUES

If the latent proficiencies θ were known for all students, it would be possible to directly compute any statistic t(θ,y), for example, subpopulation sample means, sample percentiles, or sample regression coefficients, to estimate a corresponding population quantity T. However, θ values are not observed, but estimated latent variables through measurement models. To overcome this problem, the approach taken by Rubin (1987) is taken and treat θ as missing data. Therefore, the value t(θ,y) is approximated by its expectation given the observed data, (x,y), as follows:

9.17

$$t^*(\bar{x},\bar{y}) = E[t(\bar{\theta},\bar{y})|\bar{x},\bar{y}] = \int t(\bar{\theta},\bar{y})p(\bar{\theta}|\bar{x},\bar{y})d\theta$$

It is possible to approximate t^* using plausible values (also referred to as multiple imputations) instead of the unobserved θ values. As described in the earlier section, plausible values are random draws from the posterior distribution of the proficiencies given the item responses x_j, background variables y_j, and estimated model parameters. For any student, the value of θ used in the computation of t is replaced by a randomly selected value from the student's posterior distribution. Rubin (1987) argued that this process should be repeated several times so that the uncertainty associated with imputation can be quantified. For example, the average of multiple estimates of t, each computed from a different set of plausible values, is a numerical approximation of t^* in the above formula (9.17); the variance among them reflects uncertainty due to not observing θ. It should be noted that this variance does not include any variability due to sampling from the population.

It cannot be emphasized strongly enough that the plausible values are not a substitute for point estimates (e.g., single test scores) for individuals. Plausible values are used to make accurate group-level inferences, they cannot be used to make any inferences about individuals. Plausible values are only intermediary computations in the calculation of the integrals in the above formula in order to estimate population characteristics such as subgroup means and standard deviations. When the underlying model is correctly specified, plausible values will provide consistent estimates of population characteristics, even though they are not generally unbiased

estimates of the proficiencies of the individuals with whom they are associated (see von Davier, Gonzalez, & Mislevy, 2009, for examples and a more detailed explanation). Unlike the plausible values, the more familiar ability estimates of educational measurement are, in a sense, optimal for each student (e.g. bias corrected maximum likelihood estimates, which are consistent estimates of a student's proficiency θ, and Bayesian estimates, which provide minimum mean-squared errors with respect to a reference population). Point estimates that are optimal for individual students have distributions that can produce decidedly non-optimal (inconsistent) estimates of population characteristics (Little & Rubin, 1983). Plausible values, on the other hand, are constructed explicitly to provide consistent estimates of population effects. For a further discussion of plausible values, see Mislevy et al. (1992).

After obtaining the 10 plausible values from the posterior distribution, they can be employed to evaluate the formula (9.17) for an arbitrary function T as follows:

1. Use the first vector of plausible values (out of ten) for each student, calculate the group estimator T as if the plausible values were the true values of θ. Denote the result T_1.
2. Estimate the sampling variance of T with respect to students' first vectors of plausible values. Denote the result VT_1.
3. Carry out steps 1 and 2 for each of the U vectors of plausible values (in PISA 2015, U=10), thus obtaining T_u and Var_u for u = 2,…,U.
4. The best estimate of the group estimator T obtainable from the plausible values is the average of T_u obtained from the different sets of plausible values:

9.18

$$T. = \frac{\sum_{u=1}^{U} T_u}{U}$$

5. An estimate of the variance of group estimator T is the sum of two components, which are the variance due to sampling of examinees and the variance due to latency of the proficiency θ (often called as measurement error):

9.19

$$Var(T.) = \frac{\sum_{u=1}^{U} VT_u}{U} + (1 + \frac{1}{U}) \frac{\sum_{u=1}^{U} (T_u - T.)^2}{U - 1}$$

The first component in *Var(T.)* reflects uncertainty due to sampling from the population because PISA samples only a portion of the entire population of 15-year old students. The second component reflects uncertainty due to measurement error because the students' proficiencies θ are only indirectly estimated on a finite number of item responses for each respondent.

Example for partitioning the estimated error variance

The following example illustrates the use of plausible values in one country for partitioning the error variance which is derived from the variance of group estimator T in formula 9.19.

Tables 9.13 through 9.15 present data for six subgroups of students differing in the context questionnaire variable "books at home" (variable ST013Q01TA: 1 = 0-10 books; 2 = 11-25 books; 3 = 26-100 books; 4 = 101-200 books; 5 = 201-500 books; 6 = more than 500 books). Ten plausible values were calculated for each student in the particular domain considered in this example. Each column in this table presents the means of these 10 plausible values ($T_{u,g}$, u=1,…10 columns) and the sampling standard error ($Var_{u,g}$, u=1,…10) for each subgroup *g* defined by the variable ST013Q01TA.

Table 9.16 Example for use of plausible values to partitioning the error

Plausible value	1		2		3		4		5		6	
	Mean	(s.e.)	Mean	(s.e.)	Mean	(s.e.)	Mean	(s.e.)	Mean	(s.e.)	Mean	(s.e.)
1	429.16	3.51	473.20	3.19	512.84	2.32	538.82	2.74	559.98	2.93	547.44	4.79
2	429.91	3.38	474.43	3.24	512.68	2.42	539.22	2.63	559.50	3.09	546.99	4.75
3	429.99	3.57	474.13	3.22	513.51	2.40	537.97	2.65	561.92	2.94	546.52	4.44
4	429.34	3.39	475.64	3.35	513.31	2.41	538.97	2.45	559.42	3.01	545.47	4.97
5	429.87	3.42	473.92	3.24	512.92	2.42	539.68	2.54	559.51	3.04	546.58	4.75
6	429.04	3.25	474.58	3.34	513.29	2.43	536.60	2.59	562.07	3.05	546.57	4.66
7	429.35	3.54	474.59	3.35	513.04	2.40	539.21	2.67	559.83	3.05	546.16	4.94
8	429.21	3.41	475.42	3.17	512.85	2.51	541.71	2.60	560.24	3.05	546.25	4.71
9	428.76	3.42	473.17	3.10	512.36	2.36	537.66	2.92	559.86	3.19	547.96	4.64
10	429.50	3.43	473.77	3.04	512.25	2.35	538.45	2.64	560.68	3.04	547.98	4.90

Table 9.17 Example for use of plausible values to partitioning the error – sample error, measurement error and standard error based on the 10 PVs

ST013Q01TA	Mean of 10 PVs	Sampling error	Measurement error	Standard error
1	429.41	3.43	0.43	3.46
2	474.29	3.23	0.87	3.34
3	512.90	2.40	0.42	2.44
4	538.83	2.64	1.42	3.00
5	560.30	3.04	1.02	3.20
6	512.90	2.40	0.42	2.44

Because the standard error associated with the group estimator T is comprised of sampling error and measurement error, it can be reduced by either increasing the precision of the measurement instrument (for example, increasing the number of items) or reducing the sampling error (increasing sample size). In PISA, a resampling method is used to estimate the variance due to sampling ($Var_{u,g}$), which uses a balanced repeated replication (BRR) approach (see Chapter 8 for details). This component of variance is similar across the ten plausible values; its values are influenced by the homogeneity of proficiencies among students in the subgroup. The sampling error is smaller when the subgroup consists of students with similar proficiencies.

APPLICATION TO PISA 2018 MAIN SURVEY

This section describes the implementation of IRT scaling and population modelling of the PISA 2018 main survey data. The IRT scaling of each of the CBA and PBA domains is described first. The dimensionality analyses conducted to verify the applicability of the unidimensional 2PL and GPCM models to the updated reading and the innovative global competency domains are described next. Then, the country-specific multivariate (i.e., multiple domains) population modelling analyses and computations of plausible values are detailed. Finally, the procedure utilised to estimate of the linking errors between the 2018 and the prior PISA cycles is explained.

IRT scaling

Scaling is the first step in the modelling of PISA data. It was conducted through a multi-group IRT concurrent calibration using the 2018 main survey data, with the trend items as linking items setting the scale to the PISA scale established in 2015. That is, item parameters for trend items were fixed to the ones obtained in PISA 2015 (either common or country-by-language group specific). Each domain was calibrated separately using the multidimensional discrete latent traits modelling software *mdltm* (von Davier, 2005; Khorramdel, Shin, & von Davier, 2019) setup to estimate item parameters with the unidimensional 2PL and GPCM.

The CBA reading and financial literacy assessments included both trend and new items. Mathematics and science included only trend items. As the innovative domain, global competence included only new items. All domains for PBA (mathematics, reading, and science) assessments included only trend items. Table 9.18 details the number of trend (linking) items and new items by domain and mode of assessment. Note that in Reading, 245 items in total was administered, but only one item (reading CBA DR563Q12C) had to be excluded from the analyses in all countries-by-language groups due to issues that could not be resolved. Because PBA assessment was the same assessment as used in PISA 2015, all items were trend items with known item parameters. Nevertheless, the PBA 2018 data was calibrated to estimate new parameters in cases where they no longer fit the data in particular groups or to estimate parameters for new participating countries, when the international parameters do not fit the data.

Table 9.18 Number of trend (linking) items and new items by domain and mode of assessment

	CBA			PBA
Domain	Trend	New	Total	Trend
Mathematics	82	NA	**82**	**83**
Reading	172	72	**244**	**103**
Reading Fluency	NA	65	**65**	NA
Science	115	NA	**115**	**85**
Financial Literacy	29	14	**43**	NA
Global Competence	NA	69	**69**	NA

Notes: For mathematics and reading PBA, countries chose “regular” or “easy” forms assembled from a subset of the total item pool.

Altogether, data from 623,276 students for mathematics, reading, and science (CBA and PBA); 119,983 students for financial literacy; and 253,274 students for global competence were used in the PISA 2018 scaling. In the IRT calibration, a senate weights (5,000 for each country) of the student sampling weights was used to ensure that each country contributed equally to the estimation process. Nonresponses prior to a valid response were considered omitted and treated as incorrect responses; whereas, non-responses at the end of each of the two one-hour test sessions in both PBA and CBA were viewed as not reached (and thus not administered) and were treated as missing in the scaling analyses.

Note that the mathematics, reading, science, financial literacy, and global competence scales are separate scale. Each scale was originally set to the same overall proficiency mean and standard deviation values of 500 and 100, respectively. Therefore, explorations of differences in subpopulation performance across these domains can be made.

Estimation of International and group-specific item parameters

Cultural and language differences between countries, different language versions of the assessment used in some countries, different modes of assessment, and population changes over PISA cycles could result in some items functioning differently across groups. These differences were considered in defining the groups specified in the multi-group IRT models. Minority languages within a country were considered as separate groups when their weighed sample sizes were greater than 250. In total, 116 country-by-language groups were used in the PISA 2018 main survey multiple-group IRT calibrations for mathematics, reading, and science. In financial literacy and global competence, 30 and 39 country-by-language groups were used, respectively.

To account for cultural and language differences, the stepwise calibration process described earlier was implemented to scale the 2018 data. The operational procedure is now described in more detail. In the first calibration and fit analyses run, for the trend items, item parameter estimates obtained from the PISA 2015 scaling were used as the fixed values. For the new items, common item parameters to all the groups were estimated. Given these parameter estimates, RMSD and MD fit statistics were then computed for all item-by-group, and cases with RMSD above a set threshold[2] were identified. In the relatively rare instances where quite large RMSD misfit was found (values above 0.4), the item was dropped in the specific group.

In the subsequent calibrations and fit analyses runs, the RMSD threshold used to identify misfitting items was gradually lowered to 0.12—a value that was found to be optimum. Although further lowering this threshold could improve the overall model-data fit, this would increase the proportion of unique item parameters; this may not be optimal for achieving the comparability across PISA participating countries and economies (Joo et al., under review).

In addition to ensuring appropriate model fit and reducing the measurement error, maintaining the comparability of scales through common item parameters across countries, assessment modes, and assessments over time is of prime importance. Therefore, when common (international) item parameters showed misfit in more than one group in a similar way (direction and magnitude of the misfit), its estimation was constrained to produce the same unique parameters for this subset of groups. For example, if two groups (e.g. two countries) showed poor item fit for the same item in the same direction, both groups received the same unique item parameter estimated for these two groups (note that the term *unique item parameters* in this report is used for both cases: 1) single group that receives a unique group-specific item parameter or 2) more than one group that receive the same unique item parameter that is different from the international/common item parameter). If an item showed poor fit to a different extent in different groups, different unique group-specific item parameters were used to reduce the measurement error further.

The software used for item calibration, *mdltm* readily provides the RMSD and MD fit statistics based on the formulas (9.18 and 9.19). The software implements an algorithm that monitors RMSD and MD across the specified groups and suggests a list of items to be re-estimated for specific group. This algorithm seeks to minimize the number of group-specific item parameters needed to fit the data. It does so, item by item, constraining the item parameters to be the same across the groups in which the item exhibits similar misfit. Thus, the same specific item parameters may be unique to one group or multiple groups (e.g., country-by-language groups) exhibiting the similar misfit patterns. By doing so, the communality of item parameters across

[2] Note that RMSD are always larger than absolute MD values. Therefore, unless one wishes to set different thresholds on RMSD and MD to identify misfit, it is sufficient to use a single threshold on RMSD.

country and cycles is increased, and stronger linking to the PISA scale is achieved. Ultimately, the PISA allowed for different sets of item parameters to improve model fit and optimize the comparability of groups and countries.

In most cases, the item responses across different countries and language groups were accurately described by the international (common) item parameters. For some items, misfit led to the estimation of unique parameters for certain groups, and in some cases the same unique parameters applied to more than one group. Outcomes of the PISA 2018 main survey analyses, including an overview of the percentage of common and group-specific item parameters across countries and across PISA 2015 and 2018 main surveys, are provided in Chapter 12.

Scaling of the reading fluency items

As discussed in Chapter 2, reading fluency items were newly included as a part of the reading scale which was assessed principally through the reading MSAT. These items have been introduced to provide additional information at the lower end of the reading scale not available in prior cycles. However, as their content and format tend to differ from that of the "regular" reading items, the reading fluency items could affect the existing reading scale. Therefore, to maintain the existing reading scale and avoid any potential issues that could weaken the comparability of reading scale across cycles, the calibration of the reading fluency was done after the reading items had been scaled. That is, after the scaling of reading was finalized, the reading fluency data was added to the reading data and the reading fluency items were scaled, with all the reading items parameters fixed to their final values. This approach was successfully implemented for scaling the PISA for Development Strand A data, as well. Thus, the 2018 reading assessment (reading and reading fluency) provided more information to the lower end of the scale without making substantial changes to the trend and comparability of existing reading scale.

Dimensionality analyses of the reading and global competency instruments

The results of the scaling analyses just described show that the IRT models used, with the unidimensionality and local independence assumptions, do fit the data quite well. However, further evaluations of these assumptions are important. In particular, local dependence among items, if strong, can be addressed, and the accuracy of measurement can be improved by combining the scoring of multiple dependent items into one. The major domain of reading included new items based on the revised framework. Thus, verification that the trend and the new items are measuring the same, or very closely related, latent traits is important to ensure the comparability of proficiency over PISA cycles. For that purpose, multidimensional IRT analyses were conducted for reading, which treated trend and new items as two different latent traits (confirmatory factor analysis). The overall model fit obtained from the two-dimensional model was found to improve only marginally over the unidimensional model, and the correlations between the unidimensional and multidimensional latent traits were very high, confirming that the use of unidimensional modelling is sufficient and appropriate. More details are provided below.

For the new innovative global competency domain—with all items being new—inter-item correlations, residual and principal component analyses of field trial and main survey data and multidimensional analyses of main survey data were conducted. A residual analysis was conducted in a same way as the PISA 2015 cycle (OECD, 2017), using the response residuals computed from the scaling software, *mdltm* (von Davier, 2005).

An item response residual quantifies the difference between the model expectation and the observed item response of a respondent to an item. Using the *mdltm* software (von Davier, 2005), response residuals are computed as follow up step after the calibration of the data. For dichotomous item responses, response residuals for a person *v* with estimated ability $\hat{\theta}_v$ for each item *i = 1, ..., K* were defined as below:

9.20

$$r(x_{iv}) = \frac{x_{iv} - P(X_i = 1 \mid \hat{\theta}_v)}{\sqrt{P(X_i = 1 \mid \hat{\theta}_v)\left[1 - P(X_i = 1 \mid \hat{\theta}_v)\right]}}$$

For polytomous item responses, response residuals were calculated using the conditional mean and variance defined below.

9.21

$$(x_{iv}) = \frac{x_{iv} - E(X_i \mid \hat{\theta}_v)}{\sqrt{V(X_i)}}$$

9.22

$$E(X_i^m \mid \hat{\theta}) = \sum_{x=1}^{max(X_i)} x^m \, P(X_i = x \mid \hat{\theta})$$

9.23

$$V(X_i \mid \hat{\theta}) = E(X_i^2 \mid \hat{\theta}) - \left[E(X_i \mid \hat{\theta})\right]^2$$

Once the item response residuals have been calculated, the item residual correlations (across respondents) can be computed to produce an item residual correlation matrix. Unidimensional and locally independent data are expected to show random residual correlations patterns around zero across all items and across items within each unit. Local item dependencies are found when an item pair show highly correlated response residuals and their item slope parameter estimates are high. In such cases, local item dependence may be addressed by converting these two items into a single polytomous item with partial correct scoring (Rosenbaum, 1988; Wilson & Adams, 1995).

As part of the residual analysis, principal component analysis was conducted using the correlation matrix of residuals. This was to evaluate the extent to which data are unidimensional. If the unidimensional assumption holds, little common variance among the response residuals is expected after the ability dimension is accounted for.

Reading Dimensionality analyses

Dimensionality analyses of the CBA reading field trial and main survey data were conducted. Because the field trial data did not lend themselves to the residual analyses described earlier, local dependencies were evaluated based on item-by-item correlations. When the local independence assumption is met, a similar level of correlation is expected among all the item pairs within a unit, and no distinctive pattern can be discerned. When exceptionally high correlation patterns among some items for a given unit were found consistently across many countries, conditional independence assumption of the IRT model could be threatened. Exceptionally large slope estimates can provide the similar information. In such cases, those highly correlated reading items could be combined into polytomous items with partial credit to remove local dependencies after the discussion with content experts and item developers.

Given that the PISA item responses are the mixture of dichotomous and polytomous item responses, the Spearman's rho statistic was used to estimate a rank-based measure of association. This statistic is known to be robust and has been recommended for data that does not necessarily follow a bivariate normal distribution. Based on the item-by-item correlations for all reading items, no item pairs were identified with exceptionally strong correlations Furthermore, the unidimensional IRT scaling analyses of the field trial data and later the main survey data (as described above) did not show any items with unusually large slope parameters. This provided evidence that the local item independence assumption was met. Two-dimensional IRT model of the field trial data, where new and trend items were assigned to two different latent traits, showed only marginal improvement in overall model fit over the unidimensional IRT model.

Using the main survey data, the unidimensional IRT model described earlier showed that the trend and new items assigned to the same proficiency scale fitted the data well. Two-dimensional IRT model for the reading main survey data, where trend and new items were assigned to two different latent proficiency scales provided an additional check of the unidimensional assumption. When the multidimensional IRT model was fitted, the trend item parameters were fixed to the unidimensional international (common) item parameters obtained from the PISA 2015 cycle, and the new items were constrained to estimate international parameters. Although AIC (Akaike, 1974) showed better fit for the two-dimensional model, BIC (Schwartz, 1978) and log-penalty improvement showed that the unidimensional model fits better and multidimensional model provides very little improvement over the unidimensional model (Table 9.19). In particular, we see that the unidimensional model reached 99.53% of the model fit improvement over the independence model compared to the gains expected from the multidimensional model. Moreover, the correlations of two sets of group means (the trend item only and the new items only) from the multidimensional model were very high, ranging from 0.91 to 0.99 across the different country-by-language groups. Additionally, the dimension-specific weighted likelihood estimates (WLEs) of student ability were very highly correlated with the unidimensional WLEs.

Table 9.19 Model selection criteria for the unidimensional and the two-dimensional IRT models for trend and new reading items in the main survey

Model	# Parameters	AIC	BIC	Log Penalty	% Improvement
Independence	NA	NA	NA	0.6506	0.00%
Unidimensional	566	13503721	13509811	0.5637	99.53%
Two-dimensional	875	13494467	13503881	0.5633	100.00%

Note: Log penalty (Gilula & Haberman, 1994) provides the negative expected log likelihood per observation, the % Improvement compares the log-penalties of the models relative to the difference between most restrictive and most general model.

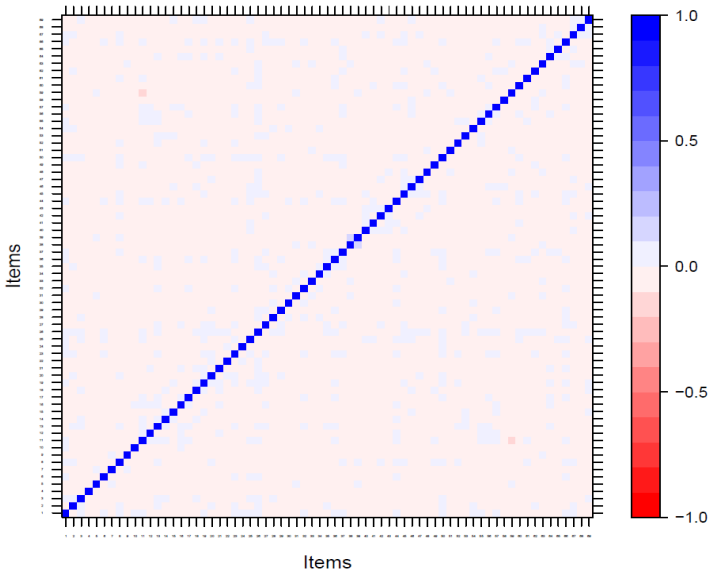
Considering all the evidence gathered from the field trail and main survey data analyses, there is a strong evidence that the new and trend reading items and scores can be placed on the same unidimensional scale.

Global competence dimensionality analyses

Global competence, as the innovative domain in 2018, was an entirely new instrument composed of new items. Preliminary classical item analyses and inter-item correlation analyses of the global competence field trial data provided information about local dependencies, leading to some items to be combined and an effective scoring rule for the main survey. After the response data were rescored, residual analyses were conducted. Two additional items were combined into one polytomous item with partial credit.

For the PISA 2018 main survey, 69 items were selected out of the 85 (partly combined) field trial items. Unidimensional IRT scaling was conducted and response residuals were calculated. Pairwise residual item correlations were then computed for each country-by language group, and averaged across groups. Figure 9.12 shows the residual correlation matrix obtained. Besides the darker blue squares on the diagonal that represent each item correlating with itself, there was no other noticeable item-pair correlation patterns.

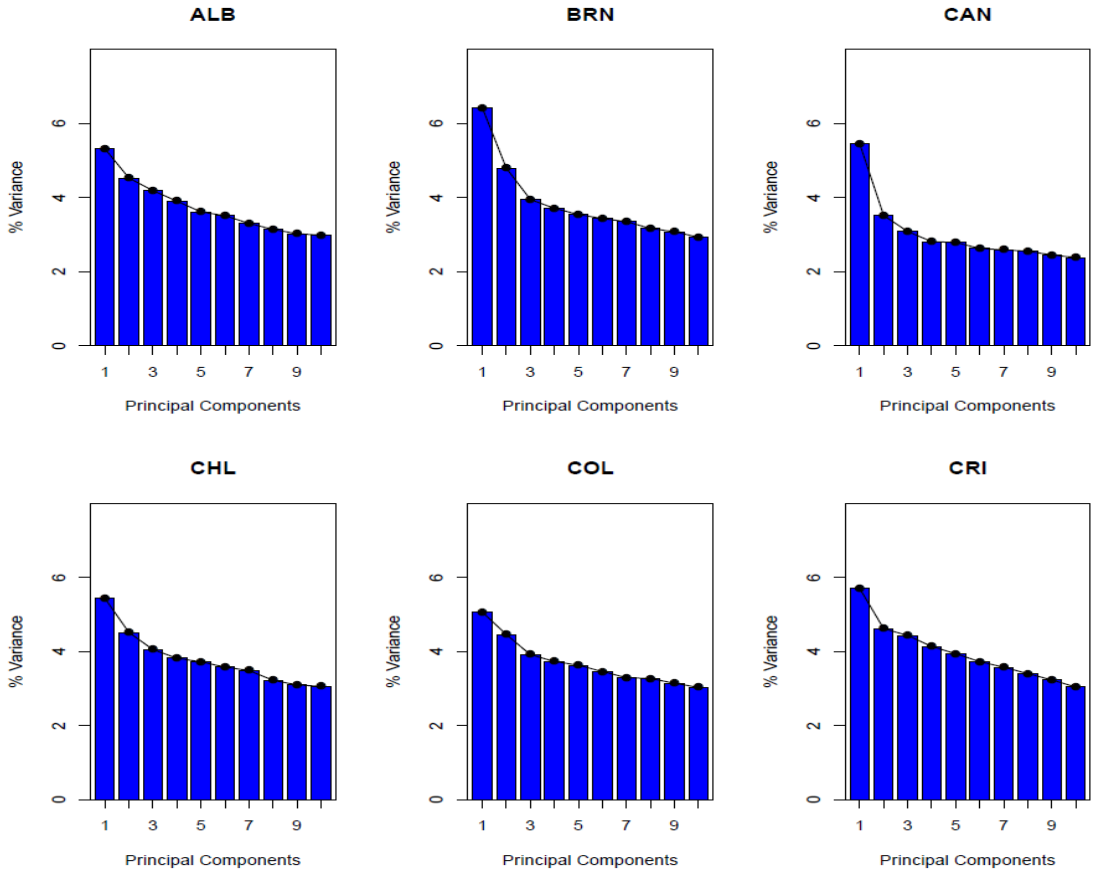
Figure 9.12 Residual correlation matrix for global competence main survey



As part of the residual analysis for the global competence, a principal components analysis was conducted using the residual correlation matrix. The principal components analysis was used to further evaluate the dimensionality of the new global competence items. Should the eigenvalue of the first principal component extracted from response residuals be large, an additional latent trait, other than the overall ability, could be assumed. When residuals for all items are included as variables, the percentage of variance adds up to 100%. The percentage of variance for the first principal component ranges from 4.55% to 7.87% with a mean of 5.73%. This number can be considered a small amount of common variance. When the percentages of variance for the first 10 principal components are summed up, the value ranges from 29.12% to 45.22%, with a mean of 37.67%, a value that is more typical for a substantial amount to be considered due to a common source of variability of response variables. The small amount of variance of the first, relative to the sum of the variances of the first ten

components also shows that another dimension is not needed to explain statistical dependencies between residuals. In other words, once the ability dimension is accounted for, there is very little common variance among the response residuals. The principal component analysis results for 6 countries, as examples, are presented in Figure 9.13. Altogether, item residual correlation patterns and principal component analyses results provided confirmation that the local independence and unidimensional IRT assumptions were met for the global competence items.

Figure 9.13 Percentage of variance from principal component analyses for 6 countries



Population modelling in PISA 2018

The population model described earlier section (*Latent regression model and population modelling*) was applied to the PISA 2018 data. Fixing the item parameters to their values obtained from the IRT scaling, multivariate latent regression models were fitted to the data at the country level, and 10 plausible values per domain were generated for each student. Plausible values for core and innovative domains (mathematics, reading, science and global competence when administered) were generated for all students participating in the assessment, regardless of whether they were administered items in that domain. That is, students received plausible values for each test domain administered in their country according to the test design implemented (PBA or CBA; regular or Une Heure (UH) test form; global competence or not; see Chapter 2 for more information on the test design). Students who did not participate or did not have responses in a particular domain were assigned model-dependent plausible values for that domain based on their responses to the BQ as well as the cognitive information coming from the other domains. Note also that, while most covariates used come directly from the student BQ, additional covariates derived from the BQ or process data from the cognitive

assessment such as the number of not-reached items and other variables relevant to predicting proficiency distributions within each country have been used in PISA 2018 as well as in previous PISA cycles.

Measurement errors have to be taken into account when dealing with the plausible values in the secondary analyses. The plausible values for the domain(s) students did not take (model-dependent) have a larger uncertainty (measurement error) than the plausible values for the other domains that were administered to them. By using repeated analysis with each of the 10 plausible values, the measurement error will readily be reflected in the analyses, and the final aggregation of results can be conducted in a way that the variability across the 10 analyses is properly reflected.

The following sections provide further information about how the population model was applied to PISA 2018 data, how plausible values were generated, and how plausible values can be used in further analyses. Changes introduced for PISA 2018 population modelling will be described in more detail in the Appendix.

Main sample and financial literacy sample models

The software called DGROUP (Rogers et al., 2006) was used to estimate the multivariate latent regression models and generate plausible values (von Davier et al. 2006; von Davier & Sinharay, 2014). During the estimation, the item parameters for the cognitive items were fixed at the values obtained from the multi-group IRT models described earlier in this chapter (see *Estimation of International and group-specific item parameters*). The results of the concurrent calibration are item parameter estimates that provide comparable scales across different PISA countries and cycles. As in other large-scale assessments and previous PISA cycles, nearly all student BQ variables as well as some school characteristics were included. As in 2015, the ratio of not-reached responses per country was categorized and included as covariates in the latent regression models. In 2018, additional conditioning variables were included (e.g., categorized response times data by item types), and school characteristics were incorporated in a different way. Appendix A provides more detailed information about these changes newly introduced for the PISA 2018 population modelling. Further, a description of the different sections of the BQ data can be found in Chapter 3 of this report. BQ variables in the PISA include international variables (collected by every participating country), as well as national variables (country-specific variables in addition to the international variables).

All the BQ variables were contrast-coded before they were processed further. Contrast coding provides a way to handle nonresponses and to satisfy the modelling' linearity assumption. The contrast coding scheme is reproduced in Annex B. However, contrast coding substantially increases the number of variables included in the model and increases the risk overparameterization resulting in inaccurate modelling. To address this issue, a principal component analysis was conducted as a dimensional reduction technique. Because each population can have unique associations among the BQ variables, a single set of principal components was insufficient for all countries. As such, the extraction of principal components was carried out separately by country to account for the differences in associations between the BQ variables and cognitive skills. In the PISA, the number of principal components y^c retained in each of the multivariate latent regression models was selected to be the smaller of 1) the number of principal components needed to explain 80% of the BQ variance, and 2) the number that corresponds to 5% of the raw sample size. This ensured that numerical instability

due to potential overparameterization of the model would not occur and that as much BQ variance as possible was modelled, given the amount of data available.

Unlike in PISA 2015 when financial literacy domain was administered as part of the *main sample*, financial literacy was administered separately to an additional *financial literacy sample* of (N=1,650) beyond the main sample (N=6,300). In addition to financial literacy, student in the financial literacy sample took mathematics or reading items, with reading administered in the same adaptive mode as in the main sample, including the fluency tasks (see Chapter 2). Because science was not included, it was not possible to estimate covariance between financial literacy and science and student did not receive plausible values in the science domain. Also, reading subscale information was insufficient to produce plausible values for the reading subscale. Thus, the financial literacy sample received plausible values in mathematics, reading, and financial literacy, but not in science and reading subscales. Note that during the analyses phase, a subset of main sample who took math and reading for two hours was added to the financial literacy data so that more robust estimates of the covariances among domains could be obtained.

Treatment of students with fewer than six test item responses

This section addresses the issue of students who provided background information but did not respond to enough cognitive items. A minimum of six completed cognitive items was considered necessary to assure sufficient information about the proficiency of students. In the PISA 2018, there were very few students[3] (0.03 %) with responses to fewer than six cognitive items in at least one of the domains assessed (mathematics, reading, science, and global competence for the main sample; financial literacy, mathematics and reading for the financial literacy sample). Although this number is small, students with responses to fewer than six cognitive items per domain were not included in the multivariate latent regression modelling to ensure stable estimations of the Γ and Σ. Nevertheless, these students along with the other students, received plausible values. For each of the two reading subscales (*cognitive scaling* and *text structure required for the item*), students had to respond to at least six items in one of the subscales to be included in the multivariate latent regression model.

In PISA 2018 (as in 2015), for all CBA and PBA domains except reading MSAT, all consecutively missing responses at the end of a cluster were treated as "not reached" and thus coded as missing response (similar to "not administered" items); hence, they were ignored in the model. For Reading MSAT, all consecutively missing responses throughout the student's entire path (core, stage 1, stage 2) were treated as "not reached". This scoring method is important with regard to both the scaling of the cognitive item and the population model since they are both rely, at least in part, on responses to the cognitive items.

Plausible values

Plausible values for the core domains of mathematics, reading, science, for the optional domain of financial literacy, and for the innovative domain of global competence were drawn from the posterior distributions estimated from the multivariate latent regression models. To accommodate country's selection of testing domains, different types of multivariate latent regression models were fitted. For the main sample: the three dimensions of mathematics, reading, and science were included for countries that did not select global competence and for PBA countries; the four dimensions of mathematics, reading, science, and global competence

otherwise. For the financial literacy sample, the three dimensions of mathematics, reading, and financial literacy were included.

The plausible value variables for the domains follow the naming convention PV1xxxx through PV10xxxx, where "xxxx" took on the following form:

- MATH for mathematics
- READ for reading
- SCIE for science
- GLCM for global competency
- FLIT for financial literacy

Population modelling for the reading subscales

There were two sets of subscales reported for reading. These were cognitive scaling subscales (locate information, understand, evaluate and reflect) and text structure required for the item subscales (single source, multiple sources). Reading subscale proficiencies were provided for the CBA only. The population modelling for of the cognitive process subscales and the population modelling for the text structures subscales were done separately.

Note that while reading fluency items were included in the population model for reading (section above), they were not included in population modelling for the reading subscales. Also, note that item parameters used for population modelling for the reading subscales are consistent with the population model for overall reading scales described above, which were obtained from the unidimensional multi-group IRT model for reading. Therefore, the reading subscales and the reading (reading fluency and reading MSAT) scale proficiencies and their distributions can be compared using the plausible values. However, the reading scale is not the weighted average of the reading subscales, which do not include any of the reading fluency items. Thus, it is possible for a country's mean proficiency to be higher than all of the country's means subscore proficiencies, if its reading fluency performance is higher than its performance on the on reading subscales. The opposite can happen as well when the performances on reading fluency is poorer in comparison to the performances on subscales.

To generate 10 plausible values for each of the reading subscales, two multidimensional population models were fitted for each country, in addition to the core and innovative domains as discussed above. These two models include:

- model 1: mathematics, science, global competence, and the three subscales of reading cognitive process, thus, 5 or 6 dimensions in total, depending on whether global competence was assessed;
- model 2: mathematics, science, global competence, and the two subscales of reading text structure subscales, thus, 4 or 5 dimensions in total, depending on whether global competency was assessed.

The aim of generating plausible values for the different reading subscales is to represent a more nuanced picture of the important aspects within the overall reading framework. These subscales allow for investigations of different aspects within the reading domain.

Table 9.20 gives an overview of the distributions of the 72 trend and 172 new items by the cognitive process and the test structure. It should be noted that the two reading subscales types

are based on a two-way classification of the same 244 items (distributed into the 3+2=5 subscales). Thus, each item contributed to one of the cognitive scaling subscales as well as one of the text structure required for the item subscales. As a result, reading subscale plausible values can be correlated among themselves under the same category: either within cognitive process or within the text structures. However, the reading subscale plausible values cannot be correlated with the overall reading scale or other domains (mathematics, science, financial literacy, global competence) produced from different population models.

Table 9.20 Distribution of 72 trend and 172 new items to the science scales and subscales

Cognitive Scaling			**Text Structure Required for the Item**		
Subscales	Trend	New	Subscales	Trend	New
Locate Information	14	36	Single text structure required for the item	70	123
Understand	41	90	Multiple text structures required for the item	2	50
Evaluate and Reflect	17	47			
Total:	**72**	**173**	**Total:**	**72**	**173**

The plausible values for the reading subscales follow the naming convention PV1xxxx through PV10xxxx, where "xxxx" takes on the following form:

- RCLI Cognitive Process Subscale of Reading - Locate Information
- RCUN Cognitive Process Subscale of Reading - Understand
- RCER Cognitive Process Subscale of Reading - Evaluate and Reflect
- RTSN Text Structure Subscale of Reading - Single
- RTML Text Structure Subscale of Reading - Multiple

Linking PISA 2018 to previous PISA cycles

PISA accounts for measurement errors due to student sampling, the reliability of the assessment, and the linking of different instruments across assessment cycles.

Following the approach implemented in 2015 (OECD, 2017, Chapter 9), an evaluation of the magnitude of linking error was conducted by considering differences between reported country results from previous PISA cycles and the transformed results from rescaling. This variability over time and over different PISA assessment designs (minor/major, etc.), as well as the fact that we do not "know" the difficulty of items exactly, introduces a source of uncertainty in the results. It becomes apparent as soon as there are multiple samples that were collected successively that the item difficulty parameter estimates tend to be (slightly) different every time new data is collected. This, in turn, has an effect on the results reported to countries, and

it is (and was in previous cycles) quantified in the linking error. This linking error is a part of the variability of country means that is due to the tests not being exactly the same and having different samples of students in the estimation of item parameters.

In summary, the uncertainty due to linking can result from changes in the assessment design or the scaling procedure used, such as:

1. different calibration samples used to estimate parameters in different cycles,
2. the inclusion of items that are unique to each cycle in addition to common items,
3. changes in the cluster position within the assessment (PISA 2000 was an unbalanced design; later designs balanced cluster positions),
4. changes in the model used for scaling, and
5. the particular set of trend items that are common to all assessment cycles of interest and which can be seen as one among an infinite set of possible trend items.

In PISA, it is important to note that the composition of the assessment in any two cycles are different due to Major-minor-minor (M-m-m) domain changes, cluster changes and units released and recombined, framework changes, assessment mode changes, and test design changes. Although the reporting model remains a unidimensional IRT model, which fits quite well, trend items are modelled based on data collected in different contexts (M-m-m or mode, etc.). Thus, estimating linking error for trend measures is a key tool to account for cycle-to-cycle differences. Note again that linking error estimates quantify the uncertainty about the link of a scale value compared between two assessment cycles.

As in past cycles, scale-level differences across countries for adjacent calibrations are considered as the target of inference. The effect of the variability of two calibrations is evaluated at the cross-country level, while within-country sampling variability is not targeted. Moreover, sampling variance and measurement error are two separate variance components that are accounted for by plausible values and replicate weights-based variance estimation. Taken together, the focus lies on the expected variability on the country level over calibrations, which is the highest reporting level. The calibration differences incorporate scaling differences, model differences, and different sets of unique items that may lead to somewhat different estimates in the two calibrations that can be used to characterize linking error.

The definition of calibration differences starts from the ability estimates of a respondent v from country g in a target cycle under two separate calibrations (e.g. the original calibration of a particular PISA cycle and its recalibration), C1 and C2. We can write for calibration C1:

9.24

$$\tilde{\theta}_{v,C1,g} = \theta_{v,true} + \hat{u}_{C1,g} + \tilde{e}_v$$

where $\hat{u}_{C1,g}$ denotes the estimated country specific error term in C1 and $\tilde{e}_v$ is the respondent specific measurement error; and for calibration C2 accordingly:

9.25

$$\tilde{\theta}_{v,C2,g} = \theta_{v,true} + \hat{u}_{C2,g} + \tilde{e}_v$$

Defined in this way, there may be country level differences in the expected values of respondents based on the calibration. These are a source of uncertainty and can be viewed as adding variance to country level estimates. Given the assumption of a country level variability of estimates due to C1 and C2 calibrations, for the differences between estimates we find:

9.26

$$\tilde{\theta}_{v,C1,g} - \tilde{\theta}_{v,C2,g} = \hat{u}_{C1,g} - \hat{u}_{C2,g}$$

and the expectation can be estimated by:

9.27

$$E\left(\hat{u}_{C1,g} - \hat{u}_{C2,g}\right) = \tilde{\mu}_{g,C1} - \tilde{\mu}_{g,C2} = \hat{\Delta}_{C1,C2,g}$$

Across countries, the expected differences of country means ($\tilde{\mu}$) can be assumed to vanish, since the scales are transformed after calibrations to match moments. That is, we may assume:

9.28

$$\sum_{g=1}^{G} E\left(\hat{u}_{C1,g} - \hat{u}_{C2,g}\right) = 0 = \sum_{g=1}^{G} \hat{\Delta}_{C1,C2,g}$$

The variance of the differences of country means based on C1 and C2 calibrations can then be considered the linking error of the trend comparing the Y2 cycle means that were used to obtain calibration C2 estimates, and the Y1 cycle estimates. The link error can be written as:

9.29

$$V\left[\hat{\Delta}_{C1,C2,g}\right] = \frac{1}{G}\sum_{g=1}^{G}\left(\tilde{\mu}_{g,C1} - \tilde{\mu}_{g,C2}\right)^2$$

The main characteristics of this approach can be summarised as follows:

- Scale-level differences across countries from adjacent-cycle IRT calibrations C1 and C2 are considered.
- The effect of the variability of scale-level statistics between two calibrations is evaluated at the country level.
- Within-country sampling variability is not targeted.
- Sampling variance and measurement error are two separate variance components that are accounted for by plausible values and replicate weights-based variance estimation.

The use of this variance component is analogous to that of previous cycle linking errors. The variance calculated in the formula (9.29) is a measure of uncertainty due to re-estimation of the model when using additional data from subsequent cycles, obtained with potentially different assessment designs, estimation methods, and underlying databases. To avoid the possibility that some data points (countries) have excessive influence on the results, as in PISA 2015, the robust S_n statistic was used. The S_n statistic was proposed by Rousseeuw and Croux (1993) as a more efficient alternative to the scaled median absolute deviation from the median (1.4826*MAD) that is commonly used as a robust estimator of standard deviation. It is defined as:

9.30

$$S_n = 1.1926 * med_i\left(med_j\left(|x_i - x_j|\right)\right)$$

The differences defined above are plugged into the formula, that is, $x_{i=\hat{\Delta}_{C1,C2,i}}$ are used to calculate the linking error for comparisons of cycles Y1 and Y2 based on calibrations C1 (using only Y1 data) and C2 (using Y2 data and additional data including Y1). The robust estimates of linking error between cycles by domain are presented in Chapter 12.

The S_n statistic is available in SAS as well as the R package "robustbase." See also *https://cran.r-project.org/web/packages/robustbase/robustbase.pdf*.

REFERENCES

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

Ali, U. S., and Walker, M. E. (2014), "Enhancing the equating of item difficulty metrics: Estimation of reference distribution", Research Report No. RR–14-07, Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12006

Bock, R. D. and M. F.Zimowski (1997), "Multiple group IRT", In W. J. van der Linden and R. K. Hambleton (eds.), Handbook of Modern Item Response Theory (pp. 433-448), Springer-Verlag, New York, NY.

Fischer, G. H. and Molenaar, I. W. (Eds.), Rasch models: Foundations, recent developments, and applications (pp. 69-95). New York, NY: Springer.

Gilula, Z. and S. J. Haberman (1994), "Conditional log-linear models for analyzing categorical panel data", Journal of the American Statistical Association, Vol. 89/426, pp. 645-656.

Glas, C. A. W. and K. Jehangir (2014), "Modelling country specific differential item functioning", In L. Rutkowski, M. von Davier, and D. Rutkowski (eds.), Handbook of International Large-Scale Assessment, CRC Press, Boca Raton, FL.

Glas, C. A. W. and N. D. Verhelst (1995), "Testing the Rasch model", in G. H. Fischer and I. W. Molenaar (Eds.), Rasch Models: Foundations, Recent Developments, and Applications (pp. 69-95), Springer, New

Glas, C. A. W. (2010), "Item parameter estimation and item fit analysis", In W. J. van der Linden and C. A. W. Glas (Eds.), Elements of adaptive testing (pp. 269–288). New York: Springer

Jewsbury, P. A., and van Rijn, P. W. (in press), "Item calibration in multistage tests," In D. Yan (Ed.), Research for Practical Issues and Solutions in Computerized Multistage Testing. London, England: Chapman & Hall.

Jewsbury, P., Lu, R. and van Rijn, P. (under review), "Modeling multistage and targeted testing data with item response theory".

Johnson, E. G. (1989), "Considerations and techniques for the analysis of NAEP data", Journal of Educational Statistics, Vol. 14/4, pp. 303-334.

Johnson, E. G., and K. F. Rust (1992), "Population inferences and variance estimation for NAEP data", Journal of Educational Statistics, Vol. 17, pp. 175-190.

Joo, S. H., Khorramdel, L., Yamamoto, K., Shin, H. J., and Robin, F., (under review), "Evaluating item fit statistic thresholds in PISA: The analysis of cross-country comparability of cognitive items".

Khorramdel, L., Shin, H. and von Davier, M. (2019), "GDM software mdltm including parallel EM algorithm", In M. von Davier and Y.-S. Lee (Eds.), Handbook of Psychometric Models for Cognitive Diagnosis (pp. 603-628). New York: Springer.

Kirsch, I., Lennon, M. L., von Davier, M., Gonzalez, E., and Yamamoto, K. (2013), "On the Growing Importance of International Large-Scale Assessments", In In: von Davier M., Gonzalez E., Kirsch I., Yamamoto K. (eds) The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research. Springer, Dordrecht. doi:10.1007/978-94-007-4629-9_1

Little, R. J. A. and D. B. Rubin (1983), "On jointly estimating parameters and missing data", American Statistician, Vol. 37, pp. 218-220.

Masters, G. N. (1982), "A Rasch model for partial credit scoring", Psychometrika, Vol. 47, pp. 149-174.

Meredith, W (1993), "Measurement invariance, factor analysis and factorial invariance", Psychometrika, Vol. 58, pp. 525-543.

Meredith, W., and Teresi, J. A. (2006), "An Essay on Measurement and Factorial Invariance", Medical Care. 44(11):S69-S77. doi:10.1097/01.mlr.0000245438.73837.89

Mislevy, R. J. (1991), "Randomization-based inference about latent variables from complex samples", Psychometrika, Vol. 56/2, pp. 177-196.

Mislevy, R. J. (1985), "Estimation of latent group effects", Journal of the American Statistical Association, Vol. 80/392, pp. 993-997.

Mislevy, R. J. et al. (1992), "Estimating population characteristics from sparse matrix samples of item responses", Journal of Educational Measurement, Vol. 29, pp. 133-161.

Mislevy, R. J. and K. M. Sheehan, (1987), "Marginal estimation procedures", in A. E. Beaton (Ed.), Implementing the New Design: The NAEP 1983-84 Technical Report, (Report No. 15-TR-20), Educational Testing Service, Princeton, NJ.

Muraki, E. (1992), "A generalized partial credit model: Application of an EM algorithm", Applied Psychological Measurement, Vol. 16(2), pp. 159-177.

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

Oliveri, M. E. and von Davier, M. (2011), "Investigation of model fit and score scale comparability in international assessments", Psychological Test and Assessment Modelling, Vol. 53/3, pp. 315-333, Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011_20110927/04_Oliveri.pdf.

Organisation for Economic Co-Operation and Development (2013), "Technical report of the Survey of Adult Skills (PIAAC)", Ch. 17 (pp. 406-438). Paris, France. Retrieved from: http://www.oecd.org/site/piaac/_Technical%20Report_17OCT13.pdf

Organisation for Economic Co-Operation and Development (2017), "Technical report of the Programme of International Student Assessment". Retrieved from: https://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf

Patz, R. J., and Junker, B. W. (1999), "A Straightforward Approach to Markov Chain Monte Carlo Methods for Item Response Models", Journal of Educational and Behavioral Statistics, Vol. 24, No. 2, pp. 146-178.

Rasch, G. (1960), Probabilistic Models for Some Intelligence and Attainment Tests, Copenhagen, Denmark: Nielsen and Lydiche (Expanded edition, Chicago, University of Chicago Press, 1980).

Rogers, A. et al. (2006), DGROUP (computer software), Educational Testing Service, Princeton, NJ.

Rosenbaum, P. R. (1988), "Permutation tests for matched pairs with adjustments for covariates", Applied Statistics, Vol. 37, pp. 401-411.

Rousseeuw, P. J. and C. Croux (1993), "Alternatives to the median absolute deviation", Journal of the American Statistical Association, Vol. 88/424, pp. 1273-1283, doi:10.2307/2291267, JSTOR 2291267.

Rubin, D. B. (1987), Multiple Imputation for Nonresponse in Surveys, John Wiley and Sons, New York, NY.

Rust, K. F. (2014), "Sampling, weighting, and variance estimation in international large-scale assessments", in L. Rutkowski, M. von Davier, and D. Rutkowski (eds.), Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis, pp. 117-154, CRC Press, Boca Raton, FL.

Schwarz, G. E. (1978), "Estimating the dimension of a model", Annals of Statistics, 6 (2): 461–464, doi:10.1214/aos/1176344136.

Thomas, N. (2002), "The role of secondary covariates when estimating latent trait population distributions", Psychometrika, Vol. 67/1, pp. 33-48.

Thomas, N. (1993), "Asymptotic corrections for multivariate posterior moments with factored likelihood functions", Journal of Computational and Graphical Statistics, Vol. 2, pp. 309-322.

van der Linden, W. J. and R. K. Hambleton (2016), Handbook of Modern Item Response Theory, 2nd ed. Springer, New York, NY.

van der Linden & Hambleton (1997), "Item Response Theory: Brief History, Common Models, and Extensions", In: van der Linden W.J., Hambleton R.K. (eds) Handbook of Modern Item Response Theory, Springer, New York, NY.

von Davier, M. (2005), A General Diagnostic Model Applied to Language Testing Data (Research Report No. RR-05-16), Educational Testing Service, Princeton, NJ.

von Davier, M. and Sinharay, S. (2014), "Analytics in international large-scale assessments: Item response theory and population models", in L. Rutkowski, M. von Davier and D. Rutkowski eds.), Handbook of International Large-Scale Assessment: Background, Technical Issues, And Methods Of Data Analysis, CRC Press, Boca Raton, FL.

von Davier, M., E. Gonzalez and R. Mislevy (2009), "What are plausible values and why are they useful?" In: IERI Monograph Series: Issues and Methodologies in Large Scale Assessments, Vol. 2, Retrieved from IERI website: http://www.ierinstitute.org/IERI_Monograph_Volume_02_Chapter_01.pdf.

von Davier, M., Sinharay, S. Oranje, and Beaton (2006), "Statistical procedures used in the National Assessment of Educational Progress (NAEP): Recent developments and future directions", in C. R. Rao and S. Sinharay (eds.), Handbook of Statistics,: Psychometrics, Vol. 26, Elsevier, Amsterdam, Netherlands.

von Davier, M. and K. Yamamoto (2004), "Partially observed mixtures of IRT models: An extension of the generalized partial credit model", Applied Psychological Measurement, Vol. 28/6, pp. 389-406.

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, 26(4), 466-488. https://doi.org/10.1080/0969594X.2019.1586642

Wilson, M. and R. J. Adams, (1995), "Rasch models for item bundles", Psychometrika, Vol. 60, pp. 181-198.

Wingersky, M., B. Kaplan and A.E. Beaton (1987), "Joint estimation procedures", in A. E. Beaton (ed.), Implementing the new design: The NAEP 1983-84 technical report (pp. 285-292), Educational Testing Service, Princeton, NJ.

Yamamoto, K., Shin, H., & Khorramdel, L. (2018). Multistage adaptive testing design in international large-scale assessments. Educational Measurement: Issues and Practice, 37, 16-27.

Yamamoto, K., Shin, H., & Khorramdel, L. (in press), "Introduction of multistage adaptive testing design in PISA 2018", OECD Education Working Papers, OECD Publishing, Paris, https://dx.doi.org/10.1787/19939019