9

Scaling PISA data

Overview 128

Data yield and data quality 128

The IRT models for scaling 141

Latent regression model and population modelling 145

Analysis of data with plausible values 147

Application of IRT and population models to PISA 149

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

OVERVIEW

The test design for PISA was based on a variant of matrix sampling (using different sets of items and different assessment modes) where each student was administered a subset of items from the total item pool. That is, different groups of students answered different yet overlapping sets of items. That makes it inappropriate to use any statistic based on the number of correct responses in reporting the 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 difficulty of the test forms. Unless one makes very strong assumptions – for example, that the different test forms are perfectly parallel – the performance of two groups assessed in a matrix sampling arrangement 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 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 characterise students as well as items in terms of a common scale, even if not all students take identical sets of items. This makes it possible to describe distributions of performance in a population or subpopulation and to estimate the relationships between proficiency and background variables.

To increase the accuracy of the measurement, PISA uses plausible values – which are multiple imputations – drawn from a posteriori distribution by combining the IRT scaling of the test items with a latent regression model using information from the student context questionnaire in a population model.

In the following section, an overview of the data yield, data preparation, and data quality is given. Then the population model used for PISA (IRT analysis, latent regression model and computation of plausible values) is described formally, followed by demonstrating its application to the PISA data describing the national and international item calibration, as well as the computation of plausible values. The procedures utilised for the linking, with the aim to obtain equivalent scales, are further described.

DATA YIELD AND DATA QUALITY

Before data were used for scaling and population modelling, different analyses were carried out to examine the quality of data and to ensure that data met the test design criteria. The following subsections give an overview of these analyses and their results. Overall, the data quality could be confirmed and data could be approved for scaling.

Targeted sample size, routing and data yield

Targeted sample size

The main survey assessment design for PISA 2015 covered the domains of science, reading and mathematics, as well as financial literacy as an optional domain, as computer- and paper-based designs. The computer-based design also included the collaborative problem-solving (CPS) domain. The computer-based design for countries that opted out of the collaborative problem solving assessment is described in Chapter 2 of this technical report. These designs required participating countries to sample a minimum of 150 schools representing their national population of 15-year-old students. Countries taking the computer-based assessment (CBA) with collaborative problem solving needed to sample 42 students from each of 150 schools for a total sample of 6 300 students, while countries taking the computer-based assessment without collaborative problem solving or the paper-based assessment (PBA) needed to sample 35 students from each of 150 schools for a total sample of 5 250. It is important to understand that 88% to 92% of students received a form that consists of four 30-minute clusters, or sets of tasks, assembled from two domains, resulting in one hour of assessment time per domain with a total of two hours of testing time per student. An additional 8% to 12% of students received forms consisting of four 30-minute clusters covering three of the four core domains; science was included in each of these forms (see Chapter 2 for more details).

Data yield

Table 9.1 shows the sample sizes and assessment languages for all 72 participating countries. Note that a student was only considered a "respondent" and included in the analysis if the student responded to at least half of the test items. When less than half of the test items were answered, the student had to respond to at least one test item and have at least one non-missing response to a part of context questionnaire items ST012 or ST013 (ST012 has 8 questions that ask about how many TV's cars, etc. are in the household; ST013 asks how many books are in the house).

[Part 1/2]

Table 9.1 **Test mode, sample size per country and language**

Country/economy	Language	Test mode	Financial literacy	N of subsample	N of schools	N total
Albania	Albanian	PBA		5 215	230	5 215
Algeria	Arabian	PBA		5 519	161	5 519
Argentina	Spanish	PBA		6 349	234	6 349
Australia	English	CBA/CPS	X	14 530	758	14 530
Austria	German	CBA/CPS		7 007	269	7 007
Belgium	Dutch French German	CBA/CPS	X	5 675 3 594 382	288	9 651
Brazil	Portuguese	CBA/CPS	X	23 141	841	23 141
B-S-J-G (China)*	Chinese	CBA/CPS	X	9 841	268	9 841
Bulgaria	Bulgarian	CBA/CPS		5 928	180	5 928
Canada	English French	CBA/CPS	X	15 444 4 614	759	20 058
Chile	Spanish	CBA	X	7 053	227	7 053
Colombia	Spanish	CBA/CPS		11 795	372	11 795
Costa Rica	Spanish	CBA/CPS		6 866	205	6 866
Croatia	Croatian	CBA/CPS		5 809	160	5 809
Cyprus[1]	English Greek	CBA/CPS		775 4 796	126	5 571
Czech Republic	Czech	CBA/CPS		6 894	344	6 894
Denmark	Danish	CBA/CPS		7 161	333	7 161
Dominican Republic	Spanish	CBA		4 740	194	4 740
Estonia	Estonian Russian	CBA/CPS		4 338 1 249	206	5 587
Finland	Finnish Swedish	CBA/CPS		5 534 348	168	5 882
France	French	CBA/CPS		6 108	252	6 108
FYROM	Albanian Macedonian Turkish	PBA		1 338 3 895 91	106	5 324
Georgia	Azerbaijani Georgian Russian	PBA		205 4 954 157	262	5 316
Germany	German	CBA/CPS		6 504	256	6 504
Greece	Greek	CBA/CPS		5 532	211	5 532
Hong Kong (China)	Chinese English	CBA/CPS		5 238 121	138	5 359
Hungary	Hungarian	CBA/CPS		5 658	245	5 658
Iceland	Icelandic	CBA/CPS		3 371	124	3 371
Indonesia	Indonesian	PBA		6 513	236	6 513
Ireland	English Irish	CBA		5 638 103	167	5 741
Israel	Arabian Hebrew	CBA/CPS		1 683 4 915	173	6 598
Italy	German Italian Slovenian	CBA/CPS	X	1 581 9 914 88	474	11 583
Japan	Japanese	CBA/CPS		6 647	198	6 647
Jordan	Arabian	PBA		7 267	250	7 267
Kazakhstan	Kazakh Russian	PBA		4 808 3 033	232	7 841
Korea	Korean	CBA/CPS		5 581	168	5 581
Kosovo	Albanian	PBA		4 826	224	4 826
Latvia	Latvian Russian	CBA/CPS		3 584 1 285	250	4 869
Lebanon	English French	PBA		1 850 2 696	270	4 546
Lithuania	Lithuanian Polish Russian	CBA/CPS	X	5 153 624 748	311	6 525

[Part 2/2]

Table 9.1 **Test mode, sample size per country and language**

Country/economy	Language	Test mode	Financial literacy	N of subsample	N of schools	N total
Luxembourg	English French German	CBA/CPS		215 1 440 3 644	44	5 299
Macao (China)	Chinese English Portuguese	CBA/CPS		3 651 779 46	45	4 476
Malaysia	English Malaysian	CBA/CPS		1 433 7 428	225	8 861
Malta	English	PBA		3 634	59	3 634
Mexico	Spanish	CBA/CPS		7 568	275	7 568
Moldova	Romanian Russian	PBA		4 258 1 067	229	5 325
Montenegro	Serbian	CBA/CPS		5 665	64	5 665
Netherlands	Dutch	CBA/CPS	X	5 385	187	5 385
New Zealand	English	CBA/CPS		4 520	183	4 520
Norway	Bokmål Nynorsk	CBA/CPS		5 007 449	229	5 456
Peru	Spanish	CBA/CPS	X	6 971	281	6 971
Poland	Polish	CBA	X	4 478	169	4 478
Portugal	Portuguese	CBA/CPS		7 325	246	7 325
Qatar	Arabian English	CBA		7 341 4 742	167	12 083
Romania	Hungarian Romanian	PBA		414 4 462	182	4 876
Russian Federation	Russian	CBA/CPS	X	6 036	210	6 036
Singapore	English	CBA/CPS		6 115	177	6 115
Slovak Republic	Hungarian Slovak	CBA/CPS	X	402 5 948	290	6 350
Slovenia	Slovenian	CBA/CPS		6 406	333	6 406
Spain	Basque Catalan Galician Spanish Valencian	CBA/CPS	X	141 1 202 161 5 092 140	201	6 736
Sweden	English Swedish	CBA/CPS		71 5 387	202	5 458
Switzerland	French German Italian	CBA		1 307 3 531 1 022	227	5 860
Thailand	Thai	CBA/CPS		8 249	273	8 249
Chinese Taipei	Chinese	CBA/CPS		7 708	214	7 708
Trinidad and Tobago	English	PBA		4 692	149	4 692
Tunisia	Arabian	CBA/CPS		5 375	165	5 375
Turkey	Turkish	CBA/CPS		5 895	187	5 895
United Arab Emirates	Arabian English	CBA/CPS		7 436 6 731	473	14 167
United Kingdom	English Welsh	CBA/CPS		13 818 339	288	14 157
United States	English	CBA/CPS	X	5 712	177	5 712
Uruguay	Spanish	CBA/CPS		6 062	220	6 062
Viet Nam	Vietnamese	PBA		5 826	188	5 826
All Countries	N/A	N/A	N/A	N/A	17 429	509 032

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

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".

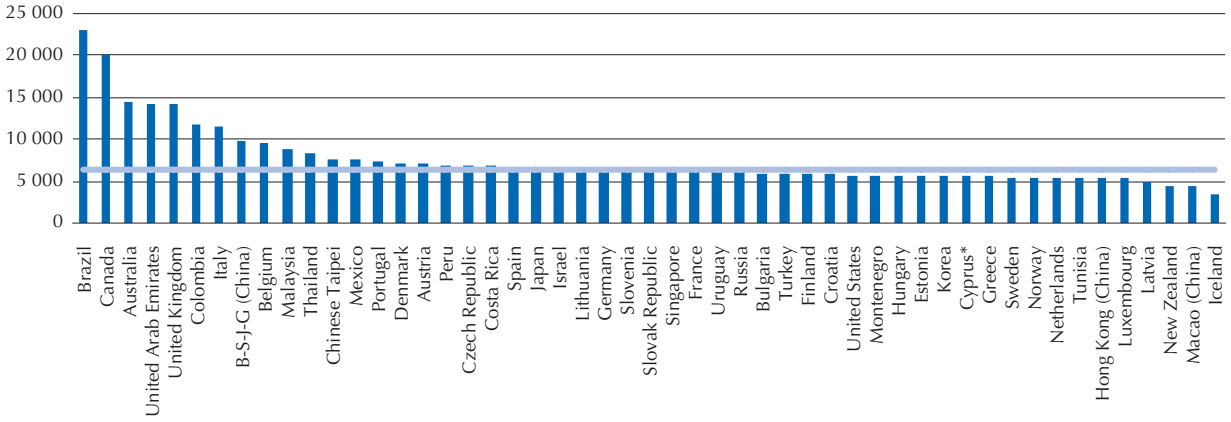
Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of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

Note: Only students taking assessment in Dutch took financial literacy.

Due to population size and operational issues, not all countries satisfied the sample size requirement for the assessments they chose. Figures 9.1 and 9.2 show the sample yields for each participating country. Two charts are used because the sample size requirement is 6 300 for computer-based testing and collaborative problem solving and is 5 250 for both computer-based (without collaborative problem solving) and paper-based testing.

■ Figure 9.1 ■

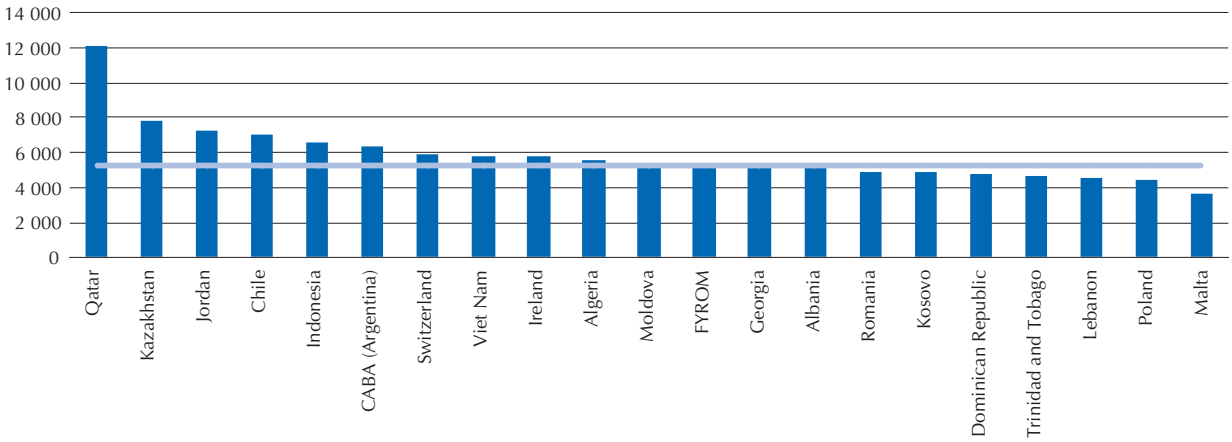
Sample yield for the participating countries with CBA/CPS format



* See note 1 below Table 9.1.

■ Figure 9.2 ■

Sample yield for the participating countries with CBA or PBA format



Since the sample sizes changed greatly from country to country, the numbers of schools and the sample sizes from each school changed as well. As seen in Table 9.1, number of schools runs from 44 (Luxembourg) to 841 (Brazil). But most countries met the requirement for the number of schools (a minimum of 150 schools).

Classical test theory statistics: item analysis

Item analyses were conducted on all computer and paper-based testing items at both the national and international levels to identify outliers, as well as human- and machine-scoring issues and other technical issues with regard to the CBA-collected data. All descriptive statistics were provided for observed responses as well as the various missing response codes and they were compared across modes and cluster positions for each item. Statistics were shared with countries and the OECD.

The following statistics were computed:

- item difficulties (proportion of correct responses, or P+)
- frequencies of scores (number of students attempted, correct and incorrect responses, omitted items, not-reached items)
- cluster scores (that is the total score within a cluster) of students with specified response types for a given item

- point biserial correlations
- response time information within each domain per item and item cluster were examined in the PISA 2015 main survey.

Proportion correct and missing rates of trend items were compared to results from all prior PISA cycles when relevant. Statistics were compiled separately for the paper-based and computer-based assessments and also examined at the aggregate level across countries. The analyses were also performed separately for each country to identify outliers (single items that seem to work differently across assessment cycles and countries). Comparisons were made at a language-by-country level, and irregular cases, such as outliers as well as cases with obvious scoring rule deviations, were identified.

The PBA results included only paper-based student responses for the core domains of science, reading and mathematics (trend items only). The CBA results included computer-based student responses for the core domains of science, reading and mathematics (both trend and new items), as well as financial literacy and collaborative problem solving, where applicable. In addition, the results were disaggregated by language within a country (Note that *une-heure* (UH) booklet results are provided for countries where applicable).

Table 9.2 Example output for examining response distributions

BLOCK M01 (UNWEIGHTED)
Response Analysis

A View Room

ITEM 1	1	NOT RCH	OFF TSK	OMIT	0	1				TOTAL	R BIS =	0.5707
CM033Q01S	N	0	0	9	184	664				857	PT BIS =	0.4100
	Percent	0.00	0.00	1.05	21.47	77.48				100.00	P+ =	0.7748
TRN_MATH	Mean Score	0.00	0.00	0.89	2.55	5.47				4.79	DELTA =	9.98
	Std. Dev.	0.00	0.00	1.20	2.29	2.92				3.05		
	RESP WT	0.00	0.00	0.00	0.00	1.00					Item WT =	1.00

Running Time

ITEM 2	2	NOT RCH	OFF TSK	OMIT	0	1				TOTAL	R BIS =	0.6124
CM474Q01S	N	2	0	8	403	444				855	PT BIS =	0.4882
	Percent	0.23	0.00	0.94	47.13	51.93				100.00	P+ =	0.5193
TRN_MATH	Mean Score	0.00	0.00	2.25	3.28	6.24				4.80	DELTA =	12.81
	Std. Dev.	0.00	0.00	2.17	2.44	2.85				3.05		
	RESP WT	0.00	0.00	0.00	0.00	1.00					Item WT =	1.00

Population Pyramids

ITEM 3	3	NOT RCH	OFF TSK	OMIT	00	11	12	13	21	TOTAL	R BIS =	0.8725
DM155Q02C	N	10	0	227	163	71	59	11	316	847	PT BIS =	0.7445
	Percent	1.17	0.00	26.80	19.24	8.38	6.97	1.30	37.31	100.00	P+ =	0.4563
TRN_MATH	Mean Score	1.50	0.00	2.33	2.88	4.99	4.97	5.55	7.55	4.83	DELTA =	13.44
	Std. Dev.	1.12	0.00	1.63	1.99	1.98	2.13	2.46	2.26	3.05		
	RESP WT	0.00	0.00	0.00	0.00	0.50	0.50	0.50	1.00		Item WT =	2.00

Table 9.2 is an example of the response analysis output for a country using computer-based testing for the first three items in block/cluster M01. The first item, CM033Q01S, is the scored version of item CM033Q01 – a multiple-choice item. More details are given below for this item in the table.

The first column says CM033Q01S is the first item in the trend maths scale (TRN_MATH).

In the second column, the first is the number of the item in the list, which is 1. All others are statistics for the response types, which are in the first row, starting from the third cell. They are:

1. N = Number of responses for the given type
2. Percent = Percent of responses for the given type
3. Mean Score = Mean score of the cluster (TRN_MATH) for the given type
4. Std. Dev. = Standard deviation of the cluster (TRN_MATH) for the given type
5. RESP WT = Response weight for the given type.

The response types are:

1. NOT RCH (not reached) = Students did not answer the given item nor the subsequent items within that cluster.
2. OFF TSK (off task) = Students did not answer the question in the expected manner.
3. OMIT (omit) = Students did not answer the given question but answered at least one subsequent question.
4. 0 = Wrong responses.
5. 1 = Correct responses.

The values in the TOTAL column (third to the last column) are based on all categories except "NOT RCH". For example, for Item 2, Total is the sum of OMIT, 0 (Wrong) and 1 (Correct), i.e. 855 = 8 + 403 + 444, which does not include NOT RCH, whose value is 2.

The statistics shown in the last two columns of Table 9.2 are ETS-developed indices. They are:

1. R-biserial (R BIS) and R-polyserial (R POLY): R BIS is used for dichotomous items and is a statistic used to describe the relationship between performance on a single test item and a continuous criterion variable (total score on the cluster). It is an estimate of the correlation between the criterion cluster score and an unobserved normally-distributed variable assumed to determine performance on the observed categorical item score. R POLY is used for polytomous items and is a generalisation of the biserial correlation for use with either dichotomous or polytomous items. At ETS, it is the generalised form of the correlation with 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 (total score on the cluster).
2. Point biserial (PT BIS) and Point-polyserial (PT POLY): PT BIS is used for dichotomous items and is the pearson product moment correlation coefficient between the dichotomous item score and the total cluster score. For polytomous items PT POLY is used.
3. P+: This is the usual percent correct for a given item.
4. Delta: This statistic is an index of item difficulty associated with the percent correct (P+). The P+ values are converted to z-scores, and then linearly transformed to an expected value 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.
5. Item WT: This value is the sum of RESP WT values of all response type except NOT RCH.

Table 9.3 provides an example of the breakdown of item score categories and biserial correlations by category as well as a summary of items that were flagged for surpassing certain thresholds (the thresholds are shown in Table 9.4). In this example, the third item is flagged for having an omit rate of greater than 10%, which prompts that further review is needed.

Table 9.3 Example table providing summary item statistics

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

	Category	N	Pct. At	Pct. Below	Mean	Std. Dev.	Biserial	B *
ITEM 1	0	193	22.52	0.00	2.47	2.28		
CM033Q01S	1	664	77.48	22.52	5.47	2.92	0.5707	-1.3220
ITEM 2	0	411	48.07	0.00	3.26	2.44		
CM474Q01S	1	444	51.93	48.07	6.24	2.85	0.6124	-0.0788
ITEM 3	0	390	46.04	0.00	2.56	1.81		
DM155Q02C	1	141	16.65	46.04	5.02	2.09	0.6728	0.3992
	2	316	37.31	62.69	7.55	2.26	0.6133	-0.1780

BLOCK M01 (UNWEIGHTED)
Item Analysis Flag Summary

Item ID	Num Resp	Type	R BIS	P+	% NOTRCH	% OFFTSK	% OMIT	% MISS	Flags
CM033Q01	2	SCR	0.5707	0.7748	0.00	0.00	1.05	1.05	.....
CM474Q01	2	SCR	0.6124	0.5193	0.23	0.00	0.94	1.17	.....
DM155Q02	5	ECR	0.8725	0.4563	1.17	0.00	26.80	27.65	...O...

Table 9.4 **Flagging criteria for items in the item analyses**

Magnitude	Criteria for flagging items
Min rbis/rpoly	0.3
Min P+	0.2
Max P+	0.9
Omit % greater than	10
Off task % greater than	10
Not-Reached % greater than	10

The delta statistic, polyserial correlation, and B* are part of the standard output from the software used for the classical item analysis; however, they may not be as familiar as other statistics such as P+, R-Bis, percent not reached, and percent of omitted responses. Countries were therefore advised to use the latter statistics when evaluating the quality of items for their sample.

The PISA 2015 computer delivery platform successfully delivered, captured, and exported information for more than 900 items, with problems encountered in less than 1% of the items. Most of these items showed no obvious problems, yet there were a few items that had to be excluded from the analyses (in all countries/language groups) due to either almost no response variance, technical issues or very low item total correlations. These excluded items are shown in Table 9.5.

Table 9.5 **Items excluded from the IRT scaling based on classical item analyses or technical problems**

Domain	Item	Mode of administration
Maths (1 item)	CM192Q01	CBA
Science trend (7 item)	S327Q02/DS327Q02C*	PBA/CBA
	PS456Q01S	PBA
	PS456Q02S	PBA
	PS133Q01S	PBA
	PS133Q03S	PBA
	PS133Q04S	PBA
Collaborative problem solving (4 items)	CC104104	CBA
	CC104303	CBA
	CC102208	CBA
	CC105405	CBA

*Five of the listed science items were dropped based on field-trial performance and content review. The items were not administered in the Microsoft computer-based instruments but were included in the paper-based assessments, as the booklets had been prepared before the decision was made to exclude the items. These items were excluded from the IRT scaling and population modelling. One item (DS327Q02C) was excluded from the main survey analysis, as it was discovered it had been dropped from the international analysis in 2003 and therefore could not be considered a trend item. Coders were instructed not to code this item and it was not included in the IRT scaling and population modelling. However, these six should have impacted the timing information on the clusters that contain them.

Response time analyses

The computer-based platform captured response time information for all computer-based items. This information was used to compute the amount of time spent by the student on each item cluster at each cluster position within the spiral design. This information was also used to examine within- and between-country differences in response time and potential administration issues. The data for these analyses included item cluster response times and plausible values from the PISA 2015 main survey.

Detailed timing information is one of the two key features of the computer delivery platform (obviously) not available in paper-based assessments; another is process sequence information. Response times are recorded for each item in milliseconds; hence, they allow for precise, timing-related analyses. For instance, these data can be used to identify rapid guessing (e.g. Wise and DeMars, 2005) and/or potential administration issues (e.g. groups of students who take substantively longer to complete the assessment than expected). Timing information can also be used to address issues of speediness and fatigue, between-country differences in allocated time, position effects, and interaction effects with variables such as student performance. Sequence information, on the other hand, can provide insights into how students progress through a set of items, including the number of times that an aspect or an item component is revisited, item sets that are skipped, and items that are truly not reached. Further, sequence information can be used in conjunction with the timing data to identify potentially problematic items, units, and/or clusters.

Timing and process data were successfully recorded for all data collections in the CBA countries in the field trial and the main survey. The available timing data were instrumental in evaluating the level of student engagement and effort over the course of the four 30-minute clusters in addition to identifying response time outliers. Very little time spend on the items/assessment was interpreted as low effort; too much time spend on the items/assessment could be an indication of technical problems or low ability. Results from the analyses indicate that the CBA data provide valid information that can be used to evaluate student performance within and across countries.

Outliers

Students were generally expected to complete each cluster within 30 minutes, but they had 60 minutes for the first two clusters and 60 minutes for the last two clusters with a break in between. In line with this expectation, an examination of the data shows that students rarely exceeded this maximum time. This was the case in the vast majority of cases; however, it was possible for some students to take additional time on the first and third clusters and less time on the second and fourth clusters, respectively, as the clusters were administered in pairs – before and after the mid-test break given at the 60-minute mark. Response times were identified greater than 4.4478*(MAD) (MAD = median{$|x_i -$ median$(x_j)|$}, where $\{x_i\}$ is the collection of all sample values) above the sample mean within each cluster as outliers (Rousseeuw and Croux, 1993; Leys et al., 2013).

On average, 55 000 students took each cluster in the assessment; about 850 of them were labelled as outliers. Not surprisingly, all clusters have outliers. Table 9.6 shows the percentages of outliers by domain (science is split into science trend and science new).

Table 9.6 Percentage of response time outliers in domains of PISA 2015 Main Survey

Domain	Mathematics	Reading	Science trend	Science new	CPS*	FL**
Number of clusters	7	7	6	6	3	2
Percent of outliers	1.78%	1.89%	1.30%	1.21%	1.37%	2.49%

* CPS = Collaborative problem solving
** FL = Financial literacy
Note: Argentina, Malaysia and Kazakhstan were not included in this analysis due to adjudication issues (inadequate coverage of either population or construct).

Descriptive Statistics

Table 9.7 presents descriptive statistics for the item cluster response times, by domain, with outliers excluded. These values are aggregated across countries and cluster positions. On average, students completed the items within each cluster in around 18 minutes, with 75% of the students completing the cluster in less than 22 minutes. With the outliers removed no student in any country took longer than 60 minutes to finish a given 30-minute cluster. Note that some variability in assessment time was expected as test administrators had to log off the computer-based assessment 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 sufficient time to finish the subsequent cluster with which it was paired. That is, for the pair of clusters administered before or after the mid-test break, the use of up to 60 minutes for the first of the two clusters left no time to finish the second cluster. These long response times point to potential administration issues. On the other hand, there were also recorded cluster response times of less than one minute. It seems highly unlikely that a student could have completed a given cluster in under a minute; hence, this may indicate a technical problem with the data collection/time coding, or a breakoff, or input reflecting rapidly advancing through the items. It should be noted that 152 students had response times equal to 0 minutes due to technical issues (with 149 of these cases coming from Qatar); these values were excluded for all response time analyses.

Table 9.7 Item cluster response time (in minutes) descriptive statistics

Domain	Min	Q1	Median	Mean	Q3	Max	SD
Maths	0.95	13.53	17.38	17.40	21.25	36.93	5.88
Reading	0.81	13.47	17.09	17.18	20.86	36.79	5.78
Science trend	0.93	12.96	16.69	16.77	20.53	35.86	5.85
Science new	0.78	14.94	19.42	19.42	23.93	41.79	6.82
CPS*	3.04	19.24	22.52	22.77	26.18	42.14	5.62
FL**	1.17	14.77	19.28	19.12	23.82	38.70	6.56

* CPS = Collaborative problem solving.

** FL = Financial literacy.

Notes: Q1 is the 25th percentile and Q3 is the 75th percentile; all zero times were removed from the analyses. Argentina, Malaysia, and Kazakhstan were not included in this analysis due to adjudication issues (inadequate coverage of either population or construct).

The median item cluster response time is similar across all domains for all countries taking the computer-based testing with the exception of collaborative problem solving, which is 3-5 minutes longer than the other domains. The standard deviation is almost the same across all domains, with science new and financial literacy items having slightly higher standard deviations.

To address the relationship between response time and student performance, median item response times grouped by proficiency levels were examined. Table 9.8 reports median response times by proficiency levels (both science and reading have Level 1a and 1b, instead of Level 1; both collaborative problem solving and financial literacy have only 5 levels). It is evident that the least able students (below Level 1) tended to complete a cluster in less time than other groups. Across all domains, more able students generally spent more time on each cluster. Except for collaborative problem solving, the differences between below Level 1 students and the highest level students exceeded around 7 minutes in all domains.

Table 9.8 Cluster level response time by PV1 proficiency level (min)

	Below Level 1	Level 1[1]		Level 2	Level 3	Level 4	Level 5	Level 6
Mathematics	12.53	15.02		17.01	18.58	19.53	19.69	19.30
Reading	9.95	12.50*	15.22	17.20	18.12	18.32	18.20	17.96
Science trend	10.53	12.45	14.75	16.69	17.78	18.01	17.88	17.47
Science new	11.33	13.39	16.32	19.26	21.04	21.80	21.95	21.84
CPS	19.41	21.34		23.29	23.77	23.67	N/A	N/A
Financial literacy	14.88	19.38		21.33	22.52	23.17	N/A	N/A

1. Reading and science have 1a and 1b on Level 1.

Note: Argentina, Malaysia and Kazakhstan were not included in this analysis due to adjudication issues (inadequate coverage of either population or construct).

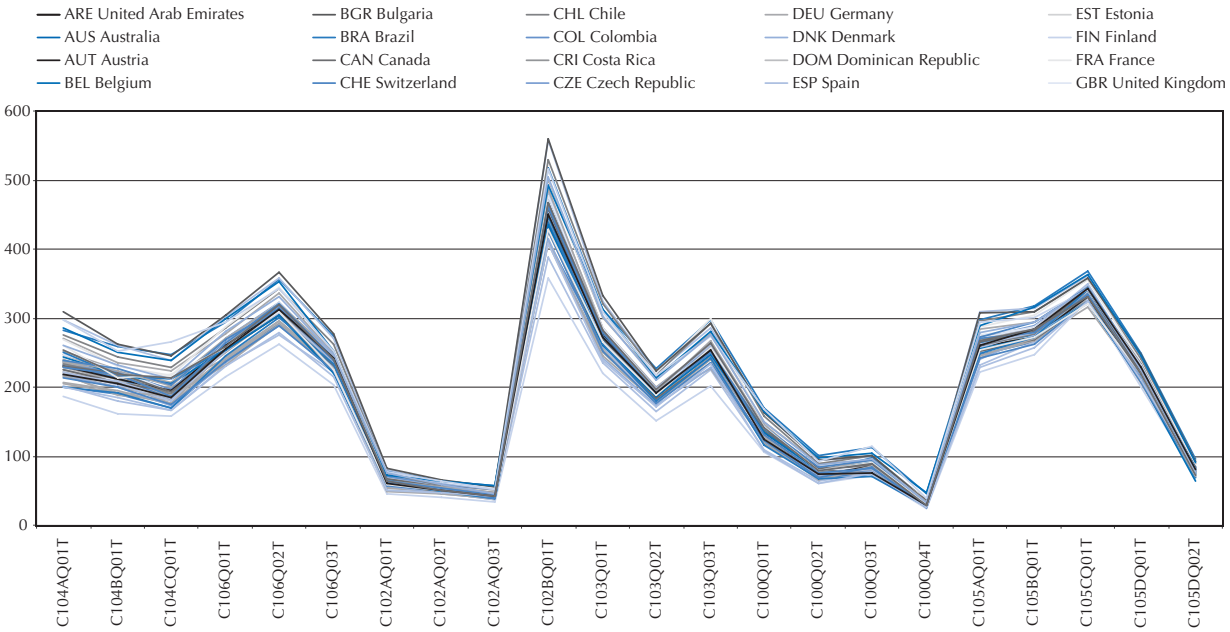
Response time was not only explored at the cluster level but also at the item level. The median response time for all items are similar across all countries. Figure 9.3 illustrates the median time of items across all countries using the CPS domain as an example.

Figures 9.4 and 9.5 show the median response time of science trend items and science new items based on the performance level across all countries (using weighted P+ and response times). The charts are sorted by the item response time. It can be seen that low performance students have almost identical response time patterns for both science trend items and science new items. The interaction between response time and ability (PV1) by items is greater for high performing students than for low performing students.

While the more able students generally need more time to complete the test, this is not true at the country level (see Figure 9.6). For example, Singapore has the highest average score in science, but its median response time is fairly close to the overall median time. Korea on the other hand has an unusually short median response time while its performance is relatively high.

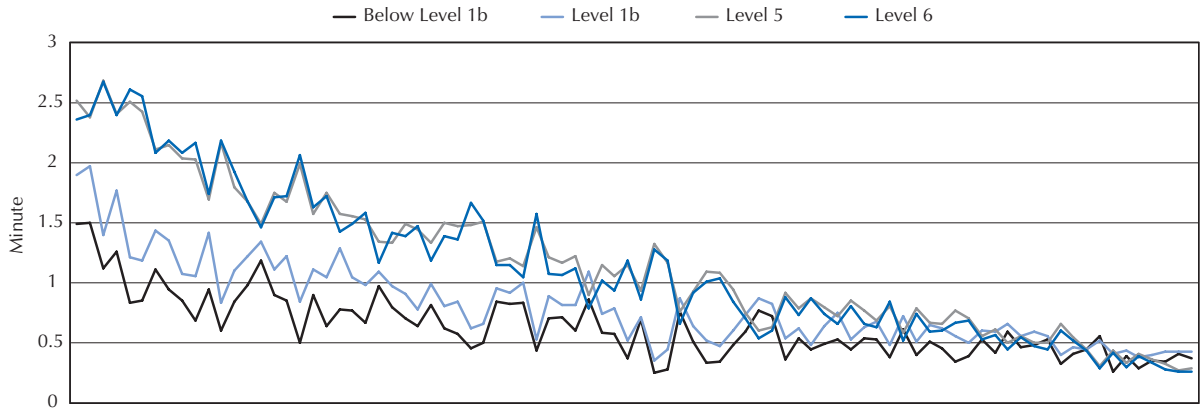
■ Figure 9.3 ■

Median response time by item – Collaborative problem solving



■ Figure 9.4 ■

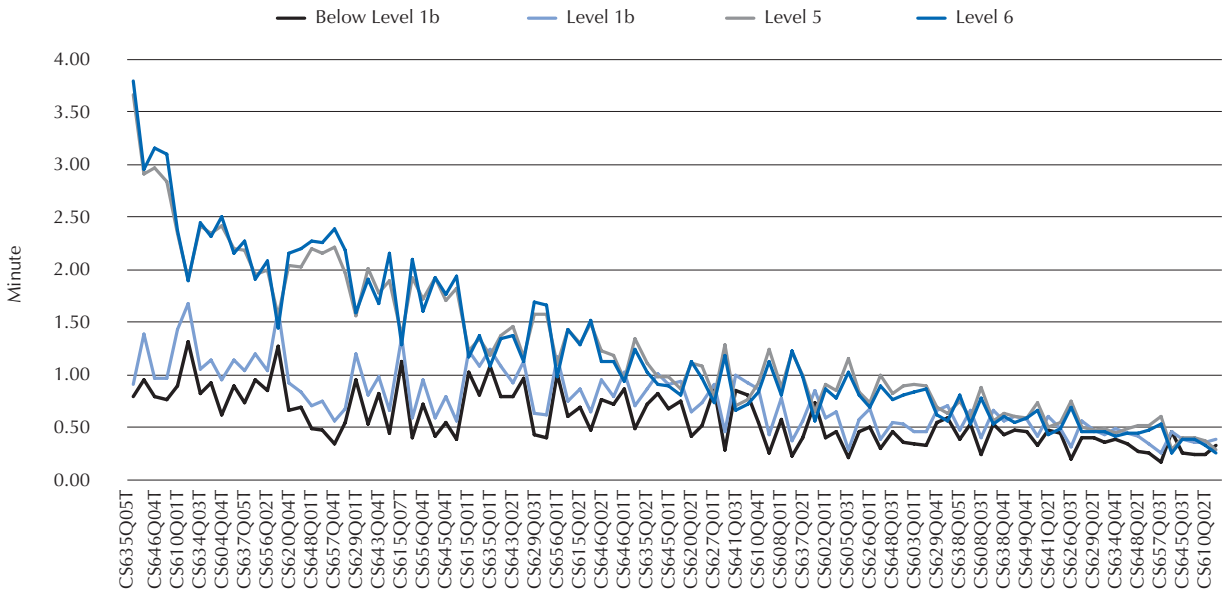
Median response time by PV1 proficiency level – Science trend items



Note: Argentina, Malaysia and Kazakhstan were not included in this analysis due to adjudication issues (inadequate coverage of either population or construct).

■ Figure 9.5 ■

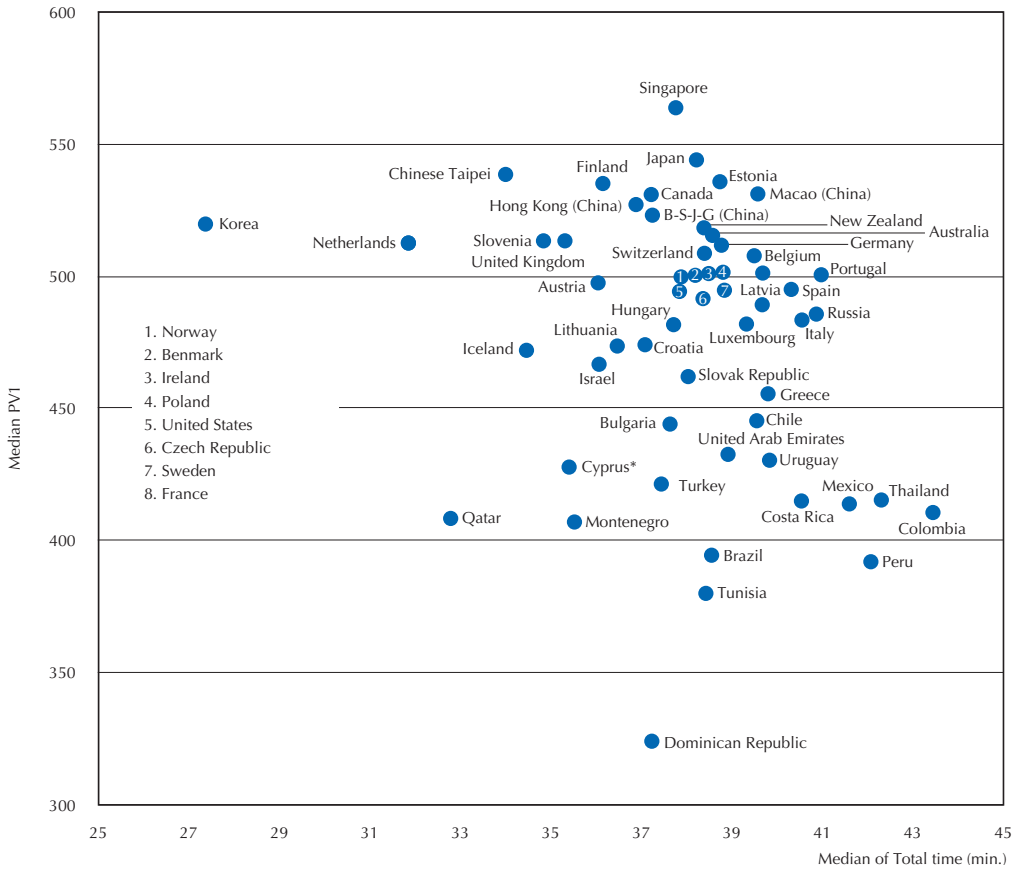
Median response time by PV1 proficiency level – Science new items



Note: Argentina, Malaysia, and Kazakhstan were not included in this analysis due to adjudication issues (inadequate coverage of either population or construct).

■ Figure 9.6 ■

Median response time vs. country median score (PV1) – All science items (2 clusters)

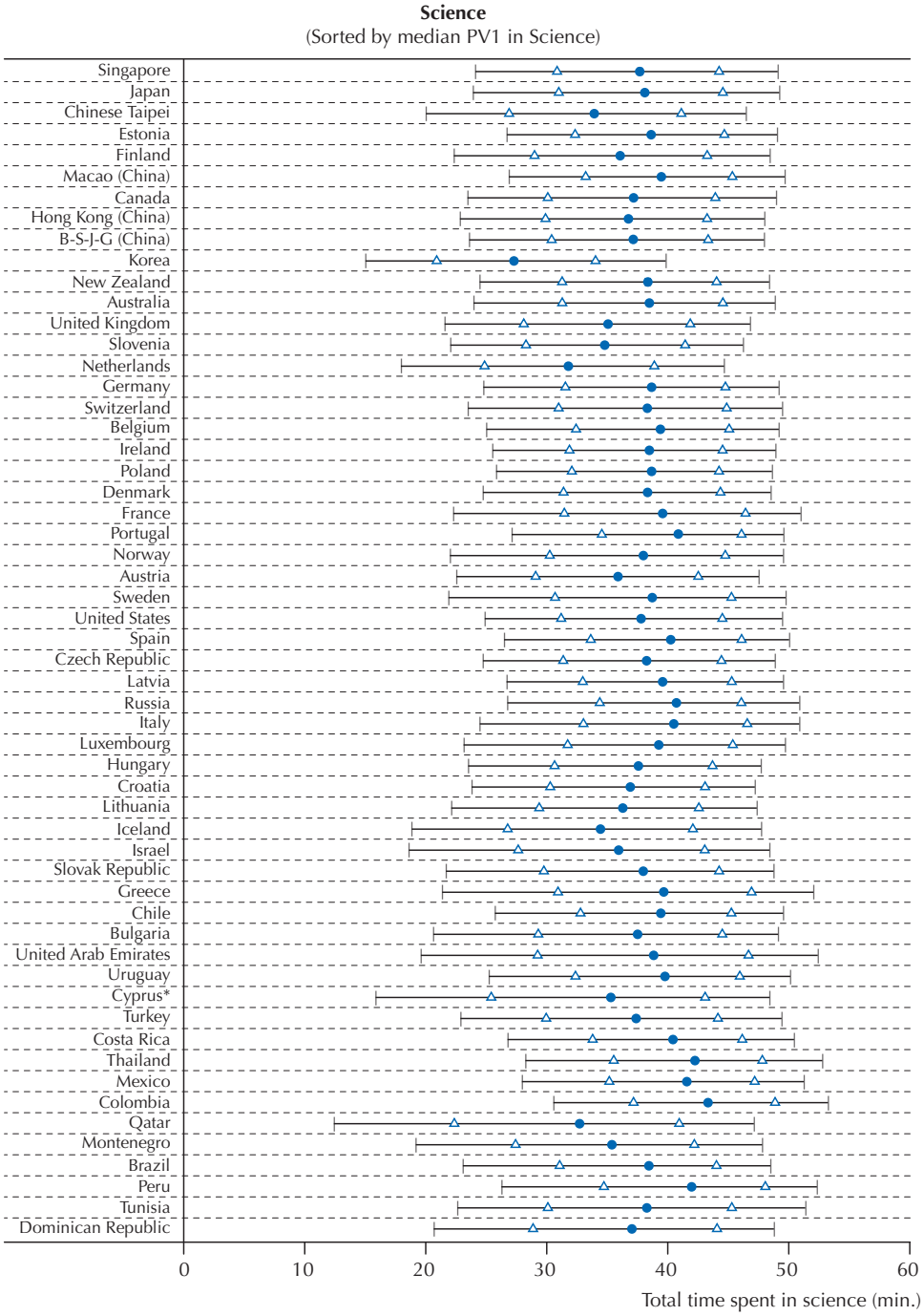


* See note 1 below Table 9.1.

As part of this analysis, the within-country variability of response times was examined for all countries. Since science is the major domain for PISA 2015, with every student taking two clusters, results are presented for this domain only. Figure 9.7 shows the distribution of time spent on science for all countries sorted by their performance using the median of the first plausible value (PV1). The middle red solid dot is the median response time, and hollow triangles indicate the 25th and 75th percentiles of the response time, respectively, for a given country. The grey horizontal bars range from the 10th percentile of the response time to the 90th percentile of the response time for a given country. The figure suggests that the within-country variability is quite similar across countries.

■ Figure 9.7 ■

Variability of time used in science



* See note 1 below Table 9.1.

Administration (and possible student motivation) issues

Results from the previous subsection suggest that there are few problematic patterns in the response times within and between countries. On average, students completed the entire test in 77.97 minutes (SD = 20.36), with 1% of the students across countries taking 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. Students in Peru, Colombia, Thailand, and Tunisia took the longest median time to complete the test in 95.09, 90.12, 89.16, and 89.01 minutes, respectively. Students in Korea took the shortest median time to complete the test in 59.28 minutes.

There were five countries where 3% or more of the students exceeded the time limit: Tunisia (8.1%), Thailand (4.9%), United Arab Emirates (4.1%), Colombia (3.4%), and the Russian Federation (3.3%). On the other end of the distribution, 1.3% of the students completed the four clusters of the test in less than 30 minutes. These students were found in nearly all countries. The results for the students with very long or short total response times suggest that there were no systematic administration and/or motivation issues in specific schools. That is, in general, these students appear to be randomly distributed across schools and countries.

Position effects

Item position effects are a common issue of concern in large-scale assessment programmes because substantial position effects can increase measurement error and introduce bias. The PISA 2015 main survey design balanced cluster position in order to control for the impact of item position and to monitor its impact of the item position on various item statistics. The cluster position effects were examined in terms of: 1) proportion of correct responses by cluster (average P+), 2) median response time by cluster and 3) rate of omitted responses by cluster (omission rate).

In order to establish a reference point for examining the magnitude of position effects, average P+ values were computed at the cluster level using both PISA 2009 and 2012 data. These values are shown in Table 9.9. We can see in this table that across the content domains there is a decrease of 0.04 to 0.08 points in the average P+ metric between cluster positions 1 and 4. For the PISA 2015 main survey data (see Table 9.10), the decrease is about 0.02 to 0.06 points in P+ values between cluster positions 1 and 4, which are smaller than the earlier cycles' values.

Table 9.9 PISA 2009 and 2012 PBA proportion correct across clusters and across countries

		Position 1	Position 2	Position 3	Position 4	Position 4 - Position 1
2009	Mathematics	0.411	0.402	0.385	0.371	-0.040
	Reading	0.584	0.559	0.534	0.501	-0.083
	Science	0.490	0.478	0.457	0.435	-0.055
2012	Mathematics	0.443	0.435	0.413	0.397	-0.046
	Reading	0.595	0.561	0.551	0.512	-0.083
	Science	0.526	0.515	0.493	0.468	-0.058

Note: Malaysia was not included in this analysis due to adjudication issues (inadequate coverage of either population or construct).

Table 9.10 PISA 2015 CBA proportion correct across clusters and across countries

	Position 1	Position 2	Position 3	Position 4	Position 4 - Position 1*
Mathematics	0.426	0.416	0.411	0.403	-0.023
Reading	0.587	0.548	0.554	0.522	-0.065
Science trend	0.493	0.465	0.476	0.452	-0.042
Science new	0.459	0.428	0.445	0.415	-0.044
CPS	0.536	0.508	0.517	0.482	-0.054
FL	0.480	0.433	NA	NA	-0.047

* For financial literacy, the difference is taken between positions 1 and 2 because these instruments only had two clusters.

Note: Argentina, Malaysia and Kazakhstan were not included in this analysis due to adjudication issues (inadequate coverage of either population or construct).

Table 9.11 shows the median cluster time averaged over all clusters at each position for all five domains. There are notable drops in median response times for all students from the first cluster to the second (3-6 minutes) and from the third cluster to the fourth (2-5 minutes); however, increases in the median response times for cluster 2 to cluster 3 (1-4 minutes) are relatively small compared to the drops. In addition to a decrease in P+ values from position 1 to position 4 for the 2015 main survey data (6-10%), there is a notable decrease in the median response times (around 4-6 minutes, i.e. nearly 20% reduction) for clusters administered in each of the four positions.

Table 9.11 **PISA 2015 CBA median cluster timing averaged across countries (in minutes)**

	Position 1	Position 2	Position 3	Position 4	Position 4 - Position 1*
Mathematics	19.81	16.91	17.34	15.71	-4.10
Reading	20.01	16.16	17.48	15.36	-4.65
Science trend	19.75	15.26	17.67	14.76	-4.98
Science new	23.38	17.40	20.73	16.89	-6.49
CPS	25.96	20.59	24.48	19.98	-5.98
FL	23.03	17.69	NA	NA	-5.33

* For financial literacy, the difference is taken between positions 1 and 2 because these instruments only had two clusters.

Note: Argentina, Malaysia and Kazakhstan were not included in this analysis due to adjudication issues (inadequate coverage of either population or construct).

The omission rates at different positions for all countries using computer-based assessments were analysed to further examine the quality of data affected by position. The omission rates for the PISA 2015 main survey in all domains and cluster positions are shown in Table 9.12. These rates do not include 'not reached' items.

Table 9.12 **PISA 2015 CBA omission rates across clusters and across countries**

	Position 1	Position 2	Position 3	Position 4	Position 4 - Position 1*
Mathematics	0.051	0.064	0.063	0.075	0.025
Reading	0.039	0.053	0.052	0.067	0.028
Science trend	0.029	0.046	0.038	0.052	0.023
Science new	0.027	0.039	0.035	0.045	0.018
FL	0.043	0.071	NA	NA	0.029

* For financial literacy, the difference is taken between positions 1 and 2 because these instruments only had two clusters.

Note: Please note that Argentina, Malaysia and Kazakhstan were not included in this analysis due to adjudication issues (inadequate coverage of either population or construct).

The omission rate for collaborative problem solving is 0% as students were forced to choose a response at each decision point in the tasks. Hence, omission rates for collaborative problem solving are not shown in the table.

Although no omission rate for any domain in any position exceeds 10%, the omission rates in Positions 2 and 4 are higher than those in Positions 1 and 3, respectively. Further, for reading, mathematics, and science, the omission rates in Position 3 are lower than those in Position 2, respectively. This is an indication that some students spent considerably more time on clusters 1 and 3, leaving them with less time for clusters 2 and 4.

THE IRT MODELS FOR SCALING

Moving from the Rasch model and partial credit model to the two-parameter logistic model and generalised partial credit model

The analysis of the PISA 2015 main survey data follows best practices outlined in, for example, Adams, Wilson, Glas and Verhelst (1995), Mislevy and Sheehan (1987), Yamamoto and Mazzeo (1992) and Wu (1997). More recent overviews of the different aspects of the methodology can be found in Glas and Jehangir (2014), Mazzeo and von Davier (2014), von Davier and Sinharay (2014), Weeks, von Davier and Yamamoto (2014), and von Davier (2006). The methods used in PISA as well as other assessments are based on models originally developed within the framework of IRT that have evolved into very flexible approaches for the analysis of large-scale, multilevel categorical data (e.g., Adams, Skrondal and Rabe-Hesketh, 2004; von Davier and Yamamoto, 2007, 2004; Wu and Carstensen, 2007).

In prior PISA cycles (2000-2012), the Rasch model (1960) and the partial credit model (PCM; Masters, 1982) were used to estimate item difficulty parameters (calibrate/scale the items). The Rasch model is a mathematical model for the probability that an individual will respond correctly to a particular item, given the individual's location in a reference domain or dimension. The model postulates that the probability of response x to item i by a respondent depends on only two parameters, the difficulty of the item (β_i) and the respondent's ability or trait level (θ), where:

9.1

$$P\left(x_i = 1 \middle| \theta, \beta_i\right) = \frac{exp\left(\theta - \beta_i\right)}{1 + exp\left(\theta - \beta_i\right)}$$

The probability of a positive response (e.g. solving an item) is strictly monotonically increasing in θ and decreasing in $ß_i$. If a respondent's ability matches the item difficulty, the expected probability of a correct response is equal to .50. Stated differently, item difficulty under the Rasch model can be interpreted as the location along the ability continuum at which a person is just as likely to answer the item correctly or incorrectly.

The partial credit model is an extension of the Rasch model to model the probability of responses to items with more than two ordered response categories. For a comprehensive review of the Rasch model, please refer to Chapter 3 (von Davier, 2016) of the *Handbook of Modern Item Response Theory* (2nd Ed.) edited by van der Linden (2016). For a review of the partial credit model, please refer to Chapter 7 of the same volume (Masters, 2016). Alternatively, von Davier and Sinharay (2014) review the use of IRT models in the context of international comparative assessments.

Concerns over the insufficiencies of the Rasch model to adequately address the complexity of the PISA data have been raised in the past (Kreiner and Christensen, 2014; Oliveri and von Davier, 2011, among others). Other national and international studies utilise more general IRT models (Mazzeo and von Davier, 2014; von Davier and Sinharay, 2014). The National Assessment of Educational Progress (NAEP), for example, uses the three-parameter IRT model and the generalised partial credit model (GPCM; Allen, Donoghue and Shoeps, 2001) as does the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS) (Martin, Gregory and Stemler, 2000).

To address the concerns about usage of the Rasch model, PISA 2015 implemented the two-parameter-logistic model (2PLM; Birnbaum, 1968) for dichotomously scored responses and the generalised partial credit model (Muraki, 1992) for items with more than two ordered response categories.

The two-parameter logistic model is a generalisation of the Rasch model. Similar to the Rasch model, the 2PLM assumes that the probability of response x to item i by a respondent depends on the difference between the respondent's proficiency θ and the difficulty of the item difficulty, β_i. But in addition, the 2PML allows that for every item, the association between this difference and the response probability can depend on an additional item discrimination parameter (α_i), characterising its sensitivity to proficiency. Under the 2PLM the response probability to an item is given as a function of this person parameter and the two item parameters; and it can be written as follows:

9.2

$$P\left(x_{ij}=1\middle|\theta,\beta_i,\alpha_i\right)=\frac{\exp\left(D\alpha_i\left(\theta-\beta_i\right)\right)}{1+\exp\left(D\alpha_i\left(\theta-\beta_i\right)\right)}$$

where D is a constant of arbitrary size, often either 1.0 or 1.7, depending on the parameterisation used in the software implementation. Note that, for $\alpha_i > 0.0$ this is a monotone increasing function with respect to θ; that is, the conditional probability of a correct response increases as the value of θ increases. One important special case is when $\alpha_i = 1.0/D$ for all items, in which case the Rasch model can be recognised as a special case of the two-parameter logistic model (2PLM). This means that the 2PLM does not force a difference from the Rasch model; it only differs from the model if the optimal estimates for the slope parameter are different across the items.

A central assumption of the Rasch model, the two-parameter logistic model, and most IRT models is conditional independence (sometimes referred to as local independence). Under this assumption, item response probabilities depend only on 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. Moreover, the 2PLM assumes unidimensionality, that is, a single latent variable, θ, that accounts for performance on the full set of items. This enables the formulation of the following joint probability of a particular response pattern $x = (x_1, ..., x_n)$ across a set of n items:

9.3

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

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 $f(\theta)$. The likelihood function is therefore characterised as:

9.4

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

The item parameter estimates obtained by maximising this function are used in the subsequent analyses.

The generalised partial credit model (Muraki, 1992), like the two-parameter logistic model, is a mathematical model for responses to items with two or more ordered response categories. While the two-parameter logistic model is suitable for dichotomous responses only, the generalised partial credit model can be used with polytomous and dichotomous responses. The generalised partial credit model reduces to the two-parameter logistic model when applied to dichotomous responses. For an item i with $m_i + 1$ ordered categories, the model formula of the generalised partial credit model can be written as:

9.5

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

where d_i is the category threshold parameter.

The approach that was taken for the PISA 2015 analysis is a model that combines features of the Rasch model/partial credit model and the two-parameter logistic model/generalised partial credit model. This more general model was applied to the PISA 2015 field trial and main survey data. As a first step, the Rasch and partial credit models were applied to all trend items. The two-parameter logistic model or generalised partial credit model were used for items that showed poor fit to the Rasch model or partial credit model. Moreover, in order to account for cultural and language differences in the multiple populations tested, procedures outlined in Glas and Verhelst (1995), Yamamoto (1997), Glas and Jehangir (2014), as well as Oliveri and von Davier (2014, 2011) were applied. The specific procedure used for PISA 2015 is described below in more detail. Based on the research studies just cited, the approach can be expected to help to retain linking items across modes or to prior assessments that would otherwise be excluded from the trend measure (the more link items with good fit across groups, the more stable the link becomes).

In order to ensure that the IRT model used provides adequate fit to the observed data, different types of model checks are customarily applied. One of these checks is the evaluation of differential item functioning (DIF), which checks to determine whether items are harder or easier for a particular group compared to other groups of equal or similar ability. While the item parameters were estimated, empirical conditional percentage-correct statistics were monitored across the samples to test for differential item functioning between countries. More precisely, for each item, the empirical item characteristic curves (ICC) for each country-by-language group were compared to the expected ICC, given an estimate of the item parameter based on the total sample. If the empirical item characteristic curves for a certain group differed noticeably from the expected ICC, this would be evidence of differential item functioning. In order to examine the difference between the empirical and expected item characteristic curves, item fit statistics were calculated. More specifically, the approach for identifying differential item functioning in PISA 2015 is based on the mean deviation (MD) and the root mean square deviation (RMSD) fit statistics. Both measures quantify the magnitude and direction of deviations in the observed data from the estimated item characteristic curves for each single item. While mean deviation is more sensitive to deviations of observed item difficulty parameters from the estimated item characteristic curves, the root mean square deviation is sensitive to the deviations of both the item difficulty parameters and item slope parameters. In contrast to other measures for the evaluation of model data fit, such as INFIT and OUTFIT measures under the Rasch model, the mean deviation and root mean square deviation indices are not affected by sample size. Moreover, mean deviation and root mean square deviation statistics are available for a range of IRT models, while INFIT and OUTFIT measures are typically only provided for the Rasch model.

Group-specific item parameters (i.e. national item parameters) for items exhibiting group-level differential item functioning in the international calibration were estimated to reduce potential bias introduced by these deviations. This approach was favoured over dropping the group-specific item responses for these items from the analysis in order to retain the information from these responses. While the items with country differential item functioning treated in this way no longer contribute to the international set of comparable responses, they continue to contribute to the reduction of measurement uncertainty for the specific country-by-language group.

The software used for item calibration, *mdltm* (von Davier, 2005), implements an algorithm that monitored **differential item functioning** measures and that automatically generated a suggested list of group-specific item treatments. This algorithm grouped similar deviations of subgroups so that unique parameters were assigned to either an individual country-by-language group or multiple country-by-language groups that showed the same level and direction of deviation.

Measurement invariance (mode effect) model

Beginning in 2015, PISA became a computer-based assessment with a paper option for a small number of countries, while it was a paper-based asesssment with optional computer-based scales in prior cycles. To address possible effects associated with this change, a mode effect study was conducted in the PISA 2015 field trial. The goal was to examine whether tasks presented in one mode (e.g. paper-based assessment) function differently when presented in another mode (e.g. computer-based assessment). A detailed description of the study and the results can be found in the section *Developing Common Scales for the Purpose of Trends* below. A comparison of different IRT models (extensions of the two-parameter logistic model assuming different mode effect parameters) in the field trial showed that the best fitting model is one that assumes item-specific mode effects for a subset of items, where items are affected differentially (i.e. some items could be more difficult, some could be at the same difficulty level, and some could become easier). This leads to a model that adds an item-specific effect for a subset of items to the difficulty parameter quantifying the item-specific difficulty difference between assessment modes, namely:

9.15

$$P\left(X = 1 \mid \theta, \alpha_i, \beta_i, \delta_m\right) = \frac{\exp\left(\alpha_i\theta + \beta_i - 1_{\{i>I\}}\delta_{mi}\right)}{1 + \exp\left(\alpha_i\theta + \beta_i - 1_{\{i>I\}}\delta_{mi}\right)}$$

Please note that this model is described again in the section *Developing Common Scales for the Purpose of Trends*; to avoid confusion the same numbering (9.15) is used in both sections. The computer-based difficulties are indexed with reference to the paper mode (computer-based items are indexed $j = I + 1 \ldots 2I$ and paper-based items $i = 1 \ldots I$. Then, difficulty parameters are decomposed into two components, that is, $\beta_i = \beta_{i+1}$ with an optional mode effect parameter δ_{mj} for $j = i + I$, while it is assumed that the slope $\alpha_i = \alpha_{i+1}$. This decomposition is formulated so the difficulties are shifted by some item-dependent amount associated with the item or item feature. For other items, we may further assume that $\delta_{mi} = 0$ (e.g. items for which the response mode differs but does not have a significant effect). As will be discussed below, for most items, there is no mode effect, that is $\delta_{mj} = 0$.

When the model given in formula (9.15) includes constraints across both modes on slope parameters, as well as potential constraints on the differential item functioning parameters δ_{mi}, this establishes a measurement invariance (e.g. Meredith, 1993) IRT model that can be viewed as representing metric invariance. The more constraints of the type that $\delta_{mj} = 0$ we have, the more we approach a model with strong or scalar invariance. Note that we already assume the equality of means and variances of the latent variable within groups in both modes because it is assumed that students receiving the test in computer or paper mode are randomly selected from a single population.

Using this model (9.15), it was possible to identify a subset of items that showed mode effects in the field trial. To account for these mode effects in the main survey, different item parameters were estimated for paired paper-based and computer-based items with substantive mode effects in the 2015 field trial; the paper-based and computer-based item parameters for items with no substantive mode effects were constrained to be the same (see *National and International Item Calibration and Handling of item-by-country/language and item-by-mode interactions* below for more information about the application of the IRT scaling approach to the PISA 2015 main survey data). This established an invariance model that assumes scalar or strong invariance for the majority of items and metric invariance for a minority of items for which difficulty differences were detected.

LATENT REGRESSION MODEL AND POPULATION MODELLING

This section reviews 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.

Individual test skills tests are concerned with accurately assessing the performance of individual students for the purposes of diagnosis, selection, or placement. The accuracy of these measurements can be improved (i.e. reducing the amount of measurement error) by increasing the number of items administered to the individual and that measure the same skill. Thus, individual achievement tests containing more than 70 items are common. Because the uncertainty associated with each estimated proficiency θ is negligible, the distribution of proficiency or the joint distribution of proficiency with other variables can be approximated using individual proficiency estimates. When analysing the distribution of proficiencies for populations or subpopulations, more efficient estimates can be obtained from a matrix-sampling design.

In international large scale assessments (ILSAs) such as PISA, test forms are kept relatively short to minimise individuals' response burden. This is important since ILSAs are low-stakes assessments that do not provide feedback and do not entail consequences of any sort for the individual test taker. At the same time, ILSAs aim to achieve broad coverage of the tested constructs. The full set of items is organised into different, but linked, test forms; each individual receives only one booklet. 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 statements about individuals' performance on a single domain. As a consequence, point estimates of proficiency that are (in some sense) optimal for each student could lead to seriously biased estimates of population characteristics (Wingersky, Kaplan and Beaton, 1987). In the case of ILSAs, improved proficiency distributions are derived that are based on both the (small) number of responses to items in the booklet and responses to background questions administered in the PISA student questionnaire. In addition, the covariance between skill domains (e.g. the PISA core domains mathematics, reading and science) is utilised to further improve the estimates of skill distributions. This approach allows estimation of proficiency distributions given responses to all domains received in the test booklet and the student questionnaire. 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) rather than assuming that this type of uncertainty is zero. Retaining this component of uncertainty requires that additional analysis procedures be used to estimate student proficiencies.

The population model used for PISA 2015 incorporated test responses (responses to the test items) as well as variables measured by the student context questionnaire (e.g. academic and nonacademic activities, and attitudes), which serve as covariates, in the computation of plausible values (von Davier et al. 2006). For each student, 10 plausible values are computed. The combined model requires the estimation of the IRT measurement model, which provides information about test performance, and the latent regression, which provides information about the extent to which student background information can predict proficiency. The estimation of this combined model is carried out as follows:

1. *Item calibration based on IRT (scaling):* The responses consist of dichotomously and polytomously scored values. These responses are used to calibrate the test and provide item parameter estimates for the test items. The two-parameter logistic model is fitted for dichotomous item responses and the generalised partial credit model is fitted for polytomous item responses. Note that for a subset of trend items, the Rasch model and the partial credit model continue to be fitted for dichotomous and polytomous responses, respectively, to maintain consistency with prior PISA cycles.

2. *Population modelling using latent regressions and plausible value generation:* The population model assumes that item parameters are fixed at the values obtained in the calibration stage. Taking the item parameters estimates from Step 1, a latent regression model is fitted to the data to obtain regression weights (Γ) and a residual variance-covariance matrix for the latent regression (Σ). Next, 10 plausible values (Mislevy and Sheehan, 1987; von Davier, Gonzalez and Mislevy, 2009) are drawn for all students using the item parameter estimates from the item calibration stage and the estimates of Γ and Σ from the latent regression model.

3. *Variance estimation:* To obtain a variance estimate for the proficiency means of each country and other statistics of interest, a replication approach (see Johnson, 1989; Johnson and Rust, 1992; Rust, 2014) is used to estimate the sampling variability as well as the imputation variance associated with the plausible values.

As stated above, the population model used for PISA is a combination of the IRT model and a latent regression model. In the latent regression model, the distribution of the proficiency variable θ is assumed to depend on the test item responses X, as well as background variables, Y, derived from responses obtained from the context questionnaire (e.g. gender, country of birth, reading practices, etc.). The item parameters from the calibration stage and the estimates from the regression analysis are both needed to generate plausible values.

A considerable number of background variables (predictors) are usually collected in international large scale assessments. Principal components accounting for a large proportion of the variation in the context questionnaire variables were used in the latent regression instead of the observed context questionnaire variables. For PISA it was decided to use the components for each country that accounted for 80% of the variance in order to avoid numerical instability due to potential overparameterization of the model. The use of principal components also serves to retain information for students with missing responses to one or more background variables. For the regression of the background variables on the proficiency variable it is assumed that:

9.6

$$\theta \sim N(y\Gamma, \Sigma)$$

The latent regression parameters Γ and Σ 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 common residual variance-covariance matrix.

The latent regression model of Θ on Y with $\Gamma = (\Upsilon_{sl}, s = 1,\ldots,S; 1 = 0,\ldots,L)$, $Y = (1, y_1, \ldots, y_L)^t$, and $\Theta = (\theta_1, \ldots, \theta_S)^t$ can be described as follows:

9.7

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

where ε_s is an error term for the assessment skill s.
The residual variance-covariance matrix can then be estimated using the following formula:

9.8

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

Plausible values for each student j are drawn from the conditional distribution:

9.9

$$P(\theta_j | x_j, y_j, \Gamma, \Sigma)$$

Using standard rules of probability, the conditional probability of proficiency can be represented as follows:

9.10

$$P\left(\theta_j \middle| x_j, y_j, \Gamma, \Sigma\right) \propto P\left(x_j \middle| \theta_j, y_j, \Gamma, \Sigma\right) P(\theta_j | y_j, \Gamma, \Sigma) = P\left(x_j \middle| \theta_j\right) P(\theta_j | y_j, \Gamma, \Sigma)$$

where θ_j is a vector of scale values (these values correspond to performance on each of the skills), $P(x_j|\theta_j)$ is the product over the scales of the independent likelihoods induced by responses to items within each scale, and $P(\theta_j|y_j, \Gamma, \Sigma)$ 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 described in this section.

The basic method for estimating Γ and Σ using the expectation-maximization (EM) algorithm is described in Mislevy (1985) for the single scale case. The EM algorithm requires the computation of the mean and variance of the posterior distribution in the formula above.

After the estimation of Γ and Σ is complete, plausible values are drawn from the joint distribution of the values of Γ for all sampled students in 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.

The software DGROUP (Rogers et al., 2006) was used to estimate the latent regression model and generate plausible values (von Davier et al. 2006; von Davier and Sinharay, 2014). A multidimensional variant of the latent regression model based on Laplace approximation (Thomas, 1993) was applied as PISA reports proficiencies on more than two skill dimensions.

ANALYSIS OF DATA WITH PLAUSIBLE VALUES

If the scale proficiency values θ were known for all students, it would be possible to directly compute any statistic $t(\theta,y)$, for example, a scale or composite subpopulation sample mean, a sample percentile point, or a sample regression coefficient to estimate a corresponding population quantity T.

However, because the scaling models are latent variable models, θ values are not observed. To overcome this problem, we follow the approach taken by Rubin (1987) and treat θ as "missing" data. The value $t(\theta,y)$ is approximated by its expectation given the observed data, (x,y), as follows:

9.11

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

It is possible to approximate t^* using plausible values (also referred to as imputations) instead of the unobserved θ values. Plausible values are random draws from the conditional distribution of the scale proficiencies given the item responses x_j, background variables y_j, and 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 conditional distribution. Rubin (1987) argues 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; 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 emphasised too strongly that the plausible values are not a substitute for test scores for individuals. Plausible values incorporate responses to test items and information about the background of responses; therefore, they cannot be used to compare 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 (von Davier, Gonzalez and Mislevy, 2009, provided examples and a more detailed explanation). The key idea lies in a contrast between plausible values and the more familiar ability estimates of educational measurement that 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 and 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 formula (9.11) for an arbitrary function T as follows:

1. Use the first vector of plausible values (out of ten) for each student, calculate T as if the plausible values were the true values of θ. Denote the result T_1.
2. In the same manner as in step 1 above, estimate the sampling variance of T, or $Var(T_1,)$, with respect to students' first vectors of plausible values. Denote the result Var_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 T obtainable from the plausible values is the average of the U values obtained from the different sets of plausible values:

9.12

$$T_{\cdot} = \frac{\sum_{u=1}^{U} T_u}{U}$$

5. An estimate of the variance of T is the sum of two components: an estimate of Var_u obtained as in step 4 and the variance among the T_us:

9.13

$$Var(T_{\cdot}) = \frac{\sum_{u=1}^{U} Var_u}{U} + \left(1 + \frac{1}{U}\right)\frac{\sum_{u=1}^{U}(T_u - T_{\cdot})^2}{U-1}$$

The first component in *Var(T.)* reflects uncertainty due to sampling from the population; the second component reflects uncertainty due to measurement error, in other words because the students' proficiencies θ are only indirectly observed through the item responses *x* and the background variables *y*.

Example for partitioning the estimated error variance:

The following example illustrates the use of plausible values in one particular country for partitioning the error variance. 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 science domain. Each column in this table presents the means of these 10 plausible values and the sampling standard error for each subgroup defined by the variable ST013Q01TA.

Table 9.13 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 9.14 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

The standard error reflects a component of error associated with the lack of precision of the measurement instrument and a component of error associated with sampling. The standard error can be reduced by either increasing the precision of the measurement instrument (for example, increasing the number of items) or reducing the sampling error. A resampling method is used to estimate the variance due to sampling. This component of variance is similar across the ten plausible values; the size is influenced by the homogeneity of proficiencies among students in the subgroup or by the precision of the survey instruments. The sampling error is smaller when the subgroup consists of students with similar proficiencies.

APPLICATION OF IRT AND POPULATION MODELS TO PISA

This section describes the implementation of the different steps of IRT and population modelling using the PISA main survey data. First, the national and international item calibration is described. Then the implementation of the population model and the computation of plausible values are described. More specifically, the procedures utilised for the linking, with the aim to obtain equivalent scales, are illustrated. It is also described how common scales were developed for the purpose of trends and an overview of the linking design and linking error is given.

Scaling and analyses of the PISA data were carried out separately for each of the domains: reading, mathematics, science, financial literacy and collaborative problem solving. By creating a separate scale for each domain, it remains possible to explore potential differences in subpopulation performance across these skills. The population model was then carried out separately for each country.

National and international item calibration

Item calibration is the first step in population modelling and provides the item parameters for the test items that are needed as one of the inputs for the population model used to calculate the plausible values. All analyses were carried out using the software *mdltm* (von Davier, 2005) for multidimensional discrete latent traits models. The software provides marginal maximum likelihood estimates obtained using customary expectation maximisation methods, with optional acceleration. Trend items were initially calibrated using the Rasch Model (Rasch, 1960) for dichotomous data and the partial credit model (Masters, 1982) for polytomous data by fixing the slope (a) parameters to 1. Item fit was examined for all country-by-language-by cycle groups using a concurrent calibration. In cases of item misfit (root mean square deviation and mean deviation), the fixation of the slope parameters was released and the two-parameter logistic model (Birnbaum, 1968) for dichotomous data or the generalised partial credit model (Muraki, 1992) for polytomous data were estimated. In the case of new items the two-parameter logistic model and the generalised partial credit model were used for calibration. The result of the calibration is that all item parameters in each domain are located on a common scale.

Omitted responses prior to a valid response are treated as incorrect responses; whereas, omitted responses at the end of each of the two one-hour test sessions in both paper-based and computer-based assessments are treated as not reached/not administered. In the latter case, these responses have no impact on the IRT scaling. However, the number of not-reached items was introduced as a covariate in the latent regression model, so it is part of the proficiency estimation in the generation of plausible values (see sections *Population Modelling in PISA 2015* and *Generating Plausible Values*).

In total 83 maths items (83 items in the paper-based and 82 in the computer-based assessments), 103 reading items (in both paper- and computer-based assessments), 85 science items (in both paper- and computer-based assessments) and 43 financial literacy items (in the computer-based assessments only) were used as linking items between PISA 2015 and past PISA cycles. In addition, the PISA 2015 main survey contained 99 new science items and 121 collaborative problem solving items. Each domain was calibrated separately with a unidimensional IRT model. The item calibration included historical PISA data (PISA 2006-2012) in addition to the 2015 PISA data. This was done for the purpose of producing a linked scale for trend measurement reaching back to the last major domain cycle (in science 2006). Table 9.15 provides an overview of the distribution of the test items across the different PISA cycles and assessment modes (paper-based, computer-based) used for the calibration of PISA 2015.

Table 9.15 **Distribution of the test items across PISA cycles and assessment modes by domain used in PISA 2015 item calibration (main survey)**

		2006 only	2009 only	2012 only	2015 only	Items linked through 2 cycles	Items linked through 3 cycles	Items linked through 4 cycles	Total items in calibration across cycles	Total items in calibration across modes
Mathematics	PBA	12	–	26	–	52	2	30	122	82
	CBA	–	–	–	82	–	–	–	82	
Reading	PBA	–	30	–	–	36	64	3	133	103
	CBA	–	–	–	103	–	–	–	103	
Science trend	PBA	23	–	–	5	27	–	53	108	85
	CBA	–	–	–	85	–	–	–	85	
Science new	CBA	–	–	–	99	–	–	–	99	NA

Note: Each item is counted only once to avoid duplication.

Altogether, data from 536 177 students for reading, mathematics, and science; 140 074 students for financial literacy; and 418 808 students for collaborative problem solving were available for the PISA 2015 international IRT calibration together with PISA data coming from past PISA cycles (2006-2012)[1]. During the item calibration, sample weights standardised to represent each country equally were used.

As the samples for each PISA cycle came from somewhat different populations with different characteristics, the calibration procedure needed to take into account the possibility of any systematic interaction between the samples and the items that were used to produce estimates of the item parameters and sample distributions. For this reason, a multiple-group IRT model using country-by-language groups over different cycles and assessment modes was estimated using a mixture of normal population distributions (one for each sample) where item parameters were generally constrained to be equal across groups with a unique mean and variance for each country (concurrent calibration). The moments of these distributions were updated for every step in the iterations of the item parameter estimation.

The item calibration was completed in two consecutive steps. First, the data from all participating countries in 2015 and from the 2006-2012 cycles were analysed in an international calibration under the assumption that the common item parameters are the same across all countries and administration cycles. To account for mode effects for a subset of items identified in the PISA 215 field trial mode effect study, different item parameters were estimated for the paired paper- and computer-based assessments; the item parameters for items in which no mode effects were found were constrained to be the same between the paper-based assessments and computer-based assessments.

In the subsequent step, unique item parameters were estimated to account for specific deviations for a subset of items. This involved a close monitoring of the IRT scaling for item-by-group interactions (group refers to country-by-language-by-cycle groups across modes) and allowing group-specific item parameters only in instances where deviations were identified. The following section describes this scaling step and the handling of item deviations from the model in more detail.

Handling of item-by-country/language and item-by-mode interactions

Given that international assessments are translated into multiple target languages, item-by-country interactions are a potential threat to validity (e.g. some terms may be harder to translate into a specific target language. As such, some items in some countries or country-by-language groups may function somewhat differently from how the item generally functions in the majority of countries or groups. The same issue occurs when changing modes from a paper- to computer-based assessment or when comparing items across different assessment cycles over years. Some items may function differently in different assessment modes or in different cycles. For this reason, an analysis step was added that investigates item-by-country, item-by-cycle, and item-by-mode interactions, to identify cases in which an item may exhibit such deviant functioning in one or more groups.

The consistency of item parameter estimates across groups and countries was of particular interest to achieve common and unbiased measures of proficiencies that are comparable across countries, assessment modes, and assessments over time. If a test measures the same latent trait in a given domain in all groups, the items should have the same relative difficulty or, more precisely, would fall within the interval defined by the standard error on the item parameter estimate (i.e. the confidence interval). In cases where common item parameters are not appropriate for certain items in certain groups (item-by-country, item-by-mode, or item-by-cycle interactions) as determined by group-specific item-fit statistics (mean deviation, MD; and root mean square deviation, RMSD), unique item parameters were estimated in a stepwise procedure. By allowing unique item parameters for items that show item-by-group interactions – in contrast to excluding such items, or forcing a common parameter – the measurement error is reduced without introducing bias. This approach

follows best practices described in the research literature on IRT and item fit assessment (Glas and Jehangir, 2014; Glas and Verhelst, 1995; Yamamoto, 1997; Oliveri and von Davier, 2014, 2011).

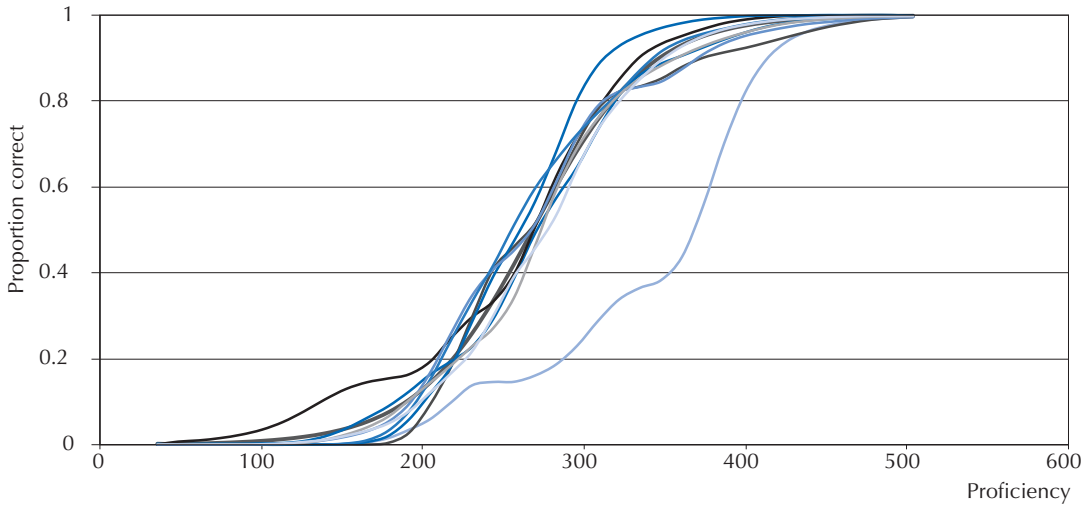
An algorithmic approach that automatically identified those group-by-item combinations requiring unique parameters based on differential item functioning detection was applied. Items not exhibiting appropriate fit using an international/common parameter received a group-specific parameter. However, if more than one group deviated from the international/common parameters in the same way (that is they showed similar differential item functioning), the algorithm assigned item parameters such that multiple groups share the same parameters, while differing from the international parameter estimate. For example, if two groups (e.g. two countries, or the same country in two PISA cycles) showed poor item fit for the same item in the international/common calibration, and 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: one group that receives a unique group-specific item parameter, and 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, unique group-specific item parameters were used for further analysis. Thus, PISA allowed for different sets of item parameters to improve model fit and optimise the comparability of groups and countries.

To identify ill-fitting items, fit statistics were estimated using the mean deviation and the root mean square deviation (see *The IRT models for scaling* below for more information on these statistics). Poorly fitting items were revealed using a root mean square deviation > 0.12 criterion and an mean deviation > 0.12 and < -.12 criterion (a value of 0 indicates no discrepancy; in other words, a perfect fit of the model). The identification of poor fitting items and the replacement of international item parameters with group-specific (unique) parameters was carried out using an automatic algorithm in *mdltm*. Thus, the international and national calibrations were conducted simultaneously for all groups so all of the estimated item parameters (international and unique) are located on one common scale.

In most cases, the item responses across groups and countries were accurately described by the international/common item parameters. For a subset of items, there was evidence of misfit for certain samples; however, this pattern was not consistent for any one particular group or country. Given this estimation and optimization approach, only a few items were dropped from the analysis in the PISA 2015 main survey. In all other cases, unique item parameters were estimated for items with substantial deviations from the international/common item parameters (poor fitting items). Figure 9.8 illustrates how the data from one group might not support the use of international item parameters.

■ Figure 9.8 ■

Item response curve for an item where the international item parameter is not appropriate for one group (example from a different ILSA)



The solid black line is the fitted two-parameter logistic item response curve that corresponds to the international item parameters; the other lines are observed proportions of correct responses at various points along the proficiency scale for the data from each subpopulation. The horizontal axis represents the proficiency scale. This plot indicates that the observed proportions of correct responses, given the proficiency, are quite similar for most countries and agree well with the IRT model-based curve. However, the data for one country indicated by the yellow line shows a noticeable departure from the common item characteristic curve. This item is far more difficult in that particular country, conditional on proficiency level. Thus, a unique set of item parameters was estimated for this country for this item.

Typically, only a small number of unique item parameters are assigned. The vast majority of items are expected to fit well for all, or nearly all, countries using international/common item parameters. Chapter 12 provides an overview of the percentage of group-specific item parameters per country.

Mode effect study in the 2015 field trial: identifying items with mode effects

To evaluate the stability of the link between paper- and computer-based assessments, a mode effect study was conducted with the PISA 2015f Field trial data where every country that later adopted a computer-based assessment in the main survey administered all trend items in both modes, thereby enabling a direct comparison between paper- and computer based assessment item parameters. The term "mode effect" refers to the observation that tasks presented in one mode (for example, paper-based) may function differently when presented in another mode (computer-based).

This section will first present a summary of the findings of the mode effect study and then illustrate in more detail the different approaches that were tested. In addition to some initial explorations (graphical model tests, correlations) of the similarity of item parameters across all domains, different formal conceptualisations of a "mode effect" were evaluated through statistical models (IRT model extensions) that contain parameters to quantify and compare potential differences between paper-based and computer-based assessments in an objective manner. This is followed by a description of how the best fitting model can be used to account and adjust for potential mode effects.

Mode effect analyses and scaling approach for the main survey

The mode effect study conducted in the PISA 2015 field trial showed that within mode, the item parameters are consistent across countries (and over time). Moreover, high correlations between item parameters across modes for all domains (0.94) was found. These findings indicate that the assessments administered in the two modes measure the same constructs. In the study with extended item response models that include different types of mode effect parameters, it was shown that the majority of items exhibit scalar or strong measurement invariance, while the remaining items exhibit metric invariance. Thus, a sound statistical link can be established, meaning computer-based and paper-based countries' results can be reported on the same scales for 2015 and inferences about the scales are comparable.

For the subset of items with evidence of metric, but not scalar invariance, this meant that some items were somewhat harder while others were easier when delivered on the computer. That is, even among the subgroup that was identified and not fully invariant, the direction of the mode effect was not uniform. This finding discounted the hypothesis of a uniform mode effect that would somehow allow an overall scale adjustment.

For the subset of items that showed a difference of difficulty parameters between modes, separate item difficulties were calculated by mode. Slope parameters were the same across computer- and paper-based assessment modes.

Trend items that showed mode effects were identified in the field trial mode effect study. These items were re-examined in the main survey using population specific item-fit statistics (root means square deviation, mean deviation) in a concurrent calibration to confirm that the same invariance model can be applied to the main survey data. The items identified as exhibiting metric invariance were treated with mode-specific item difficulty parameters. Thus, possible mode effects are unlikely to impact the proficiency estimation, as the link between modes and cycles is established on a large number of trend items that show scalar (strong) invariance.

Chapter 12 provides information about which trend items are scalar invariant, sharing all characteristics across modes, and which items are partially or metric invariant, sharing a common slope parameter.

Graphical Model Tests and Correlations

The comparison of mode differences in the current section is based on an approach that was first described by Rasch (1960). Parameter invariance across groups can be examined by applying the same identification constraints, and then estimating the parameters of a model in these groups separately and evaluating the level of agreement among the two sets of parameters. This "graphical model test" is useful to spot systematic differences between modes of administration, but it provides less statistical rigor than other model-based approaches. A graphical model test was conducted as a first step to examine the overall agreement of parameters of items administered in both modes and to explore potential drivers of any differences; the IRT models presented later (*IRT models to assess measurement invariance and mode differences*) were used to evaluate mode differences with a higher level of statistical rigor.

The PISA 2015 field trial incorporated an equivalent groups design that was implemented to aid the transition from paper- to computer-based assessment. This means that students were sampled in each country from a number of schools and then assigned randomly to one of two treatment conditions, taking the PISA field trial instruments on the computer or on paper. They were assigned independent of proficiency, prior experience, or other student variables.

This equivalent groups design allowed us to test the null hypothesis of "no mode effect". The comparison was based on estimating parameters for the computer-based assessment mode and comparing them with parameters obtained from the (smaller) paper-based field trial sample, which was strengthened by combining it with data from prior paper-based PISA assessments ranging the 2000-2012 cycles. Due to the random assignment of students to modes, the underlying ability distributions of the paper- and computer-based field trial samples are assumed to be identical. As such, the computer-based parameters should not differ significantly, or systematically, from the parameters obtained in the 2000-2012 reanalysis (see *Developing common scales for the purpose of trends* later on in chapter) and verified using the paper-based field trial sample.

The following figures (9.9 and 9.10) show parameter comparisons between the mode-based samples. The IRT analyses for estimating these parameters are based on data from 68 field trial countries that submitted their data through November 2014 (reading, mathematics, and science: n = 150,983; financial literacy: n = 34,443).

Note that the paper-based item parameters were taken from the PISA 2000-2012 linking study that aimed at finding common parameters across five cycles of historical PISA data, and derived under the guiding principle of retaining as many Rasch model-based parameters as possible. More precisely, the paper-based item parameters were fixed to the estimates obtained from the linking study (where there were only paper-based assessment items), while the item parameters for the computer-based items were freely estimated (but constrained to be equal across countries). This was done simultaneously in the software *mdltm* (when fixing item parameters in a calibration, no additional constraints are needed since the fixation of parameters already takes care of the indeterminacy of the scale). Therefore, the paper-based set contains a number of slope values that are not estimated but fixed to 1 (retained Rasch Model items), which produces fewer pairs of freely estimated parameters. However, the difficulty parameters can be compared for all items that were administered in paper- and computer-based modes.

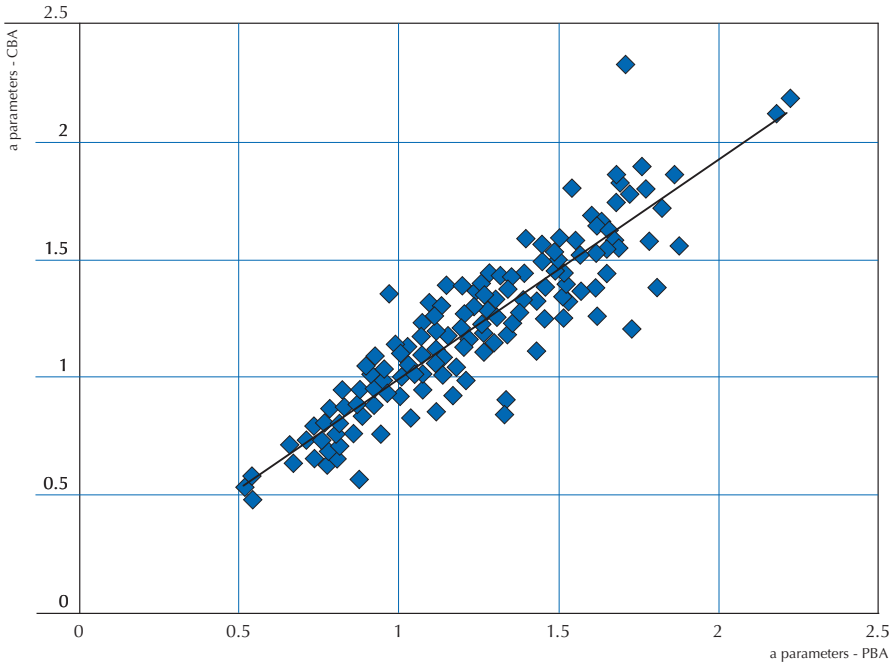
The distinction among the domains of reading, mathematics and science, as well as financial literacy was ignored because the parameters obtained across modes appeared to vary consistently across all domains.

These figures provide evidence of overall general agreement between the parameters based on the paper- and computer-based assessment modes. While there are differences, it appears that the level of difficulty of an item remains largely the same between paper-based parameters – based on historical data – and computer-based estimates. The same holds for the freely estimated slope parameters.

Moreover, correlations between the difficulty parameters for paper- and computer-based trend items are high within each domain, ranging from 0.92 to 0.95; the correlations between the discrimination parameters (slopes) range from 0.90 to 0.94 (note that only the two-parameter-logistic-model-based slopes were used to calculate correlations). The correlation of item difficulty parameters across modes and domains is 0.94, and the correlation of item slope parameters is 0.91. Table 9.16 presents an overview of these correlations. These high correlations as well as the Figures 9.9 and 9.10 suggest that the same constructs are being measured under both modes. The results from these field trial analyses suggested that a statistical link can be established whereby the computer- and paper-based countries' results can be reported on the same scales for 2015 (for more information about the impact on mode effects on country means see *The impact of mode effects on country means in the field trial* later on in this chapter).

■ Figure 9.9 ■

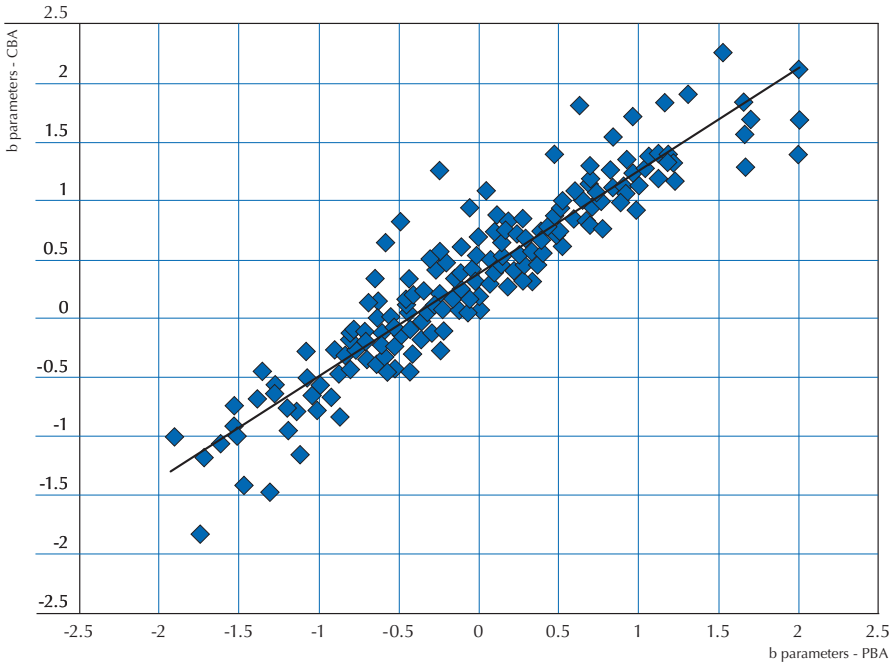
Comparison of slope parameter estimates across paper-based (horizontal axis) and computer-based (vertical axis) assessment modes for the PISA 2015 field trial data



Note: All domains with trend items (reading, mathematics and science, as well as financial literacy) are included.

■ Figure 9.10 ■

Comparison of difficulty parameter estimates across paper-based (horizontal axis) and computer-based (vertical axis) assessment modes for the PISA 2015 field trial data



Note: All domains with trend items (reading, mathematics and science, as well as financial literacy) are included.

Table 9.16 **Correlations of item difficulty and item slope parameters between paper-based and computer-based trend items within and across domains**

Domain	Correlation of difficulty parameters (PBA,CBA)	Correlation of slope parameters (PBA,CBA)
Mathematics	0.95	0.91
Reading	0.95	0.90
Science	0.92	0.94
Financial literacy	0.94	0.92
All Domains	0.94	0.91

IRT models to assess measurement invariance and mode differences

Several mode-effect models that can be used to account for differences across groups were tested. More specifically, we tested whether mode differences are present on a global level, that is, whether the difference between paper and computer modes just adds or subtracts a level of difficulty to all assessment tasks, or whether the effect is person-specific, that is, whether some people are more affected by mode differences than others. Finally we tested a model that examines whether some items show mode effects, while others do not – that is, whether items are affected differently by mode effects.

Strong measurement invariance holds if the same item parameters fit the items independent of the mode of administration. A mode effect that homogeneously applies to all items in a test when changing the mode can be modelled by adding the same constant to all difficulty parameters in the case of the affected mode. Consider the two-parameter logistic model in formula (9.2) for greater ease of exposition. The notation in (9.2) can be transformed to the customary two-parameter logistic model notation via the transformation $a = \alpha /1.7$ and $b = -\beta/\alpha$.

If item *i* is presented in two different modes of administration, say paper and computer, a common (but arguably simple) assumption is that all items are "shifted" by a certain amount with respect to their difficulty. The reason for this could be that reading or, more generally, processing the item stem or stimulus is generally harder (by the same amount for all items and stimuli) on the computer, or entering a response on the computer is more tedious than filling in a bubble on an answer sheet of a paper-based instrument.

In order to represent this, we assumed a logistic IRT model with a general mode effect parameter $-\delta_m$ that represents how much more difficult (or easy) solving an item is when presented in a given mode relative to a reference mode. For items presented in the reference mode, we assumed that model (9.2) holds; for items in the "new" model, we assume that:

9.14

$$P\left(X = 1 \mid \theta, \alpha_i, \beta_i, \delta_m\right) = \frac{\exp\left(\alpha_i\theta + \beta_i - 1_{\{i>I\}}\delta_m\right)}{1 + \exp\left(\alpha_i\theta + \beta_i - 1_{\{i>I\}}\delta_m\right)}$$

The expression $1_{\{i>I\}}$ denotes the indicator function which returns 1 if $i > I$. This shift by a mode effect in the same direction for all items in a specific mode can be thought of as a model with items (instead of items for each delivery mode separately) in which the difficulty parameters for items presented in one mode (say paper) are assumed to be β_i for $i = 1,\ldots, I$ and the item parameters for computer mode are appended as parameters $j = I + 1,\ldots,2I$ and arranged in the same order and constrained to be $\beta_j = \beta_{(j-I)} - \delta_m$. That is, all computer-based item difficulties are simply shifted by a certain amount compared to paper-based items. Note that all IRT models illustrated in this section are based on the assumption of equivalent groups.

To explain why such an approach may be needed, or why it would be considered to estimate a mode effect in this way, the question of transitioning from paper- to computer-based testing can be used as a prototypical application. In such a setting, the same test items would exist in two modes, and information on how the test behaves (and more specifically, about the item parameters) may be available from large samples drawn from the reference population. In this setting, estimating completely new β_j may not be advisable, while estimating an overall mode effect $-\delta_m$ could be considered for the purpose of adjusting for the effect of moving the items to computer administration.

In contrast to the assumptions made in model (9.14), one could argue that not all items become more difficult when moving them to the computer; some could be more difficult, some could be at the same difficulty level, and some could

even become easier. This leads to a model that adds an item-specific effect to the difficulty parameter. In model (9.15) we write this as a DIF parameter, which quantifies the difference from the paper-based assessment, namely:

9.15

$$P\left(X = 1 \mid \theta, \alpha_i, \beta_i, \delta_m\right) = \frac{\exp\left(\alpha_i\theta + \beta_i - 1_{\{i>I\}}\delta_{mi}\right)}{1 + \exp\left(\alpha_i\theta + \beta_i - 1_{\{i>I\}}\delta_{mi}\right)}$$

As outlined above, the difference in comparison to the model of metric (or "weak") factorial invariance (Meredith, 1993) is that the computer-based difficulties are written in reference to the paper mode and are decomposed into two components, that is, $\beta_j = \beta_{i+I} - \delta_{mj}$, while it is assumed that the slopes $\alpha_j = \alpha_{i+I}$. Again, this is written as a model with items, of which the first I items are presented in the reference mode, while the second I items are presented in the "new" mode. This decomposition is formulated so the difficulties are shifted by some item-dependent amount associated with the item or item feature. For paper-based items $i \leq I$ we can assume $\delta_{mi} = 0$. In addition, there may be other items for which we may further assume that $\delta_{mi} = 0$ (e.g., items for which the response mode differs but does not have a significant effect on item difficulty). These unaffected items are the basis for linking across modes, and below we show that these can indeed be assumed to be the majority of items.

The model given in formula (9.15) with constraints across both modes on slope parameters, as well as potential constraints on the DIF parameters, establishes a measurement invariance (e.g., Meredith, 1993) IRT model that can be viewed as representing a mixture of items with strong and weak factorial invariance. The more constraints of the type $\delta_{mi} = 0$ we have, the more we approach a model with strong factorial invariance. Note that the equivalent groups design allows us to assume the equality of means and variances of the latent variable in both modes because it is assumed that students receiving the test via computer or paper mode are randomly selected from a single population.

Finally, if it cannot be assumed that the mode effect is a constant shift in difficulty for all students, one may assume that an additional ability ϑ is required to predict the response probabilities in the new mode accurately. We still assume the same average in the paper- and computer-based ability distribution for the domain specific dimension; the additional mode dimension is independent. This leads to Model (9.16) in which a second latent variable was assumed, that is, another random effect was added to the item function for items administered in the new mode. The expression $\alpha_{mi}\vartheta$ in Model (9.16) below indicates that there is a second slope parameter α_{mi} for items administered in the new mode ($i = I,\ldots,2I$) and that the effect of the mode is person dependent and quantified by the second latent variable ϑ. We obtain:

9.16

$$P\left(X = 1 \mid \theta, \alpha_i, \beta_i, \delta_m\right) = \frac{\exp\left(\alpha_i\theta + \beta_i - 1_{\{i>I\}}\alpha_{mi}\vartheta\right)}{1 + \exp\left(\alpha_i\theta + \beta_i - 1_{\{i>I\}}\alpha_{mi}\vartheta\right)}$$

Note that the slope parameters and item difficulties, α_i, β_i, are as before in models (9.14) and (9.15) equal across modes. Only the additional "mode slope" parameter α_{mi} needs to be estimated for all items administered in the "new" mode, plus the joint distribution $f(\theta,\vartheta)$ for which we can assume that the variables are uncorrelated, that is, $\text{cov}(\theta,\vartheta) = 0$.

In formula (9.16) it is assumed that the effect of the person "mode" variable varies across items, which is likely the more plausible variant, even though a mode in which person-dependent but item-homogenous effects $\alpha_m\vartheta$ (a Rasch variant of a random mode effect) could also be defined. Models (9.14), (9.15), and (9.16) can be applied to multiple populations, that is, by assuming one population per participating country or language group in PISA.

We conducted an empirical comparison of the models based on the field trial data. Table 9.17 below shows the results of models (9.14), (9.15), and (9.16) for a multiple population mode effects analysis using the PISA field trial data. All analyses were conducted with the software *mdltm* (von Davier, 2005). As a general rule, lower values for the statistics (Akaike information criterion, AIC; Bayesian information criterion, BIC; Consistent Akaike Information Criterion, CAIC, log-penalty, and Akaike) indicate better fit. However, when the magnitude of the statistics is similar, the more parsimonious model should be preferred. In all cases, Model (9.16) has the lowest values for these statistics, yet they do not differ appreciably from the fit for Model (9.15). To provide additional evidence for this interpretation we examined

the marginal reliability of scores under each model as well as the correlation between estimates of student ability obtained from both models. The median reliability for scores in all domains for each of the models was quite similar across groups, with median values ranging from 0.8 to 0.85. There were a few groups where the reliabilities were notably lower (less than 0.6). The inclusion of these data had some influence on the model fit, but there was insufficient evidence based on the reliability to suggest that Model (9.16) should be preferred over Model (9.15). Additionally, the correlation between estimated scores for Models (9.15) and (9.16) in each domain was r = 0.999, which suggests that there was little added utility in using Model (9.16). We can conclude based on these results that model (9.15) describes the data sufficiently well.

This means that there is a need to specify item-specific, but not person- (or country[2]-) specific, mode effect parameters.

Table 9.17 Measurement invariance assessment using mode effect models for the PISA field trial data, analysed separately for the domains of financial literacy, maths, reading and science

Domain	Model	Penalty AIC	AIC	Penalty BIC	BIC	Penalty CAIC	CAIC	Log Penalty	Akaike
Financial literacy	(9.14)	192	253996	1003	254807	1099	254903	0.564498	0.564925
Financial literacy	(9.15)	236	251899	1233	252896	1351	253013	0.559736	0.560260
Financial literacy	(9.16)	248	251744	1295	252792	1419	252916	0.559365	0.559917
Maths	(9.14)	620	1416987	3697	1420064	4007	1420374	0.526304	0.526534
Maths	(9.15)	674	1409948	4019	1413293	4356	1413630	0.523668	0.523919
Maths	(9.16)	714	1409235	4257	1412778	4614	1413135	0.523388	0.523654
Read	(9.14)	818	1770885	4877	1774944	5286	1775353	0.534144	0.534391
Read	(9.15)	990	1760709	5903	1765622	6398	1766117	0.531022	0.531320
Read	(9.16)	1104	1758594	6583	1764073	7135	1764625	0.530349	0.530682
Science	(9.14)	1694	5378045	10100	5386451	10947	5387298	0.586249	0.586433
Science	(9.15)	1984	5361306	11830	5371152	12822	5372144	0.584392	0.584608
Science	(9.16)	2180	5356556	12998	5367374	14088	5368464	0.583852	0.584090

An evaluation of the log-penalty shows that the simple item-independent mode-effect model does not fit as well as the item-specific Model (9.15) and the Model (9.16) with an additional latent variable. Models (9.15) and (9.16) appear to fit the data almost equally well, both accounting for item-specific effects in slightly different ways. Therefore, it can be assumed that a mixture of strong and weak factorial invariance holds and that the computer-based version of the test measures the same construct as the paper-based version. Clearly, the mode effect is not a homogenous shift of difficulties, but rather one that affects some items more than others; a large percentage of items show strong invariance and are not affected in a significant way by mode differences. Further, the results of estimating Model (9.15) for each domain showed that most mode effects on individual tasks were positive, although some were negative. This result shows that a common linear adjustment-based equating method would not be appropriate, and it opens opportunities to optimise the linking between paper- and computer-based assessments by means of item selection, and equality constraints for those items that are least affected by changes in presentation mode.

The distribution of the mode-effect sizes indicated that we can identify a set of items for which strong measurement invariance holds. Those items for which no significant mode effect could be detected formed the basis for linking the computer-based assessment to past PISA cycles, while all trend items can be used, if retained in future studies, to measure the construct due to the invariance properties established in this section.

In summary, the model that balances complexity and model data fit for evaluating and accounting for item mode effects among those considered here was the model that assumes the same parameters for the paper-based assessment as for the computer-based assessment and adjusted the paper-based item difficulty parameters by a differentional item-functioning parameter for a subset of items, without the introduction of an additional mode-specific skill. This indicated that *strong measurement invariance can be established for the majority of items* while weak factorial invariance could be assumed *for the remaining trend items* administered in the computer-based PISA field trial.

It is important to point out that these results indicate that the computer- and paper-based trend items for PISA 2015 can be linked using this approach based on established measurement invariance. The adjustment, if necessary, for a number of items appears to be small compared to the range of difficulty parameters in the trend item set, while the direction of adjustment points to added difficulty.

The impact of mode effects on country means in the field trial

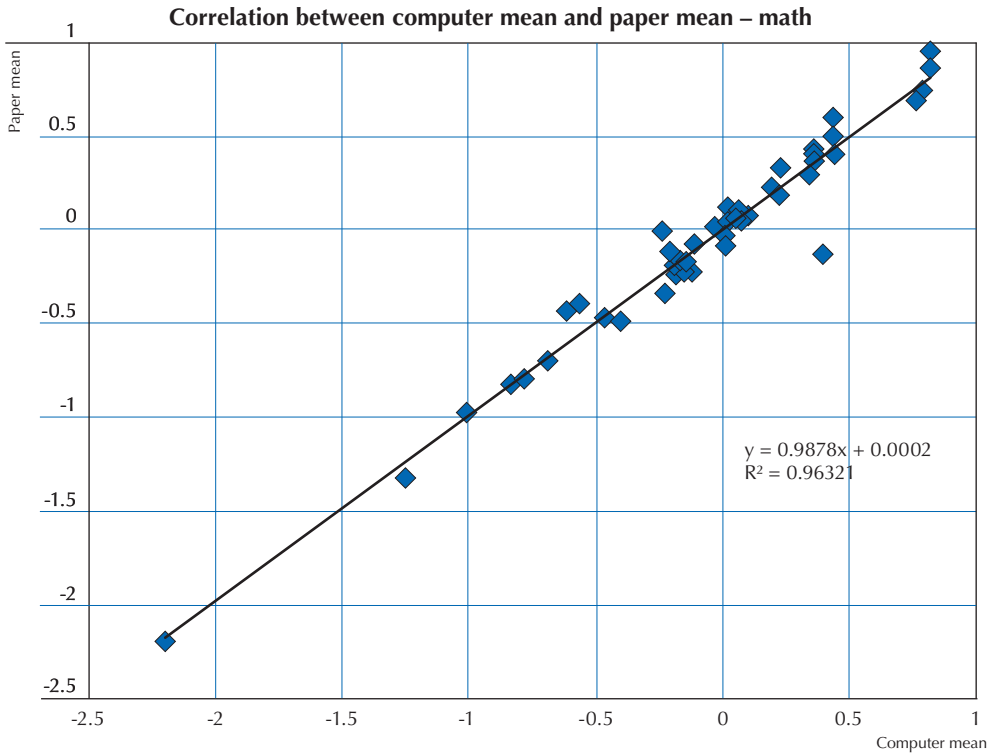
To evaluate the impact of mode effects relative to other variables of interest, country means based on the domain-specific skills obtained from a simplified version of Model (9.16) were split by three variables and compared to one another: gender, mode and a random split of schools within each country. Model (9.16) was simplified for this analysis so that it incorporates scalar invariance for those items that showed little or no mode difficulty differences and assumes metric invariance for the remaining items. There were no country-specific mode effects needed or applied in these analyses. This ensures comparability across countries while accounting for item-specific difficulty differences for a subset of items only, with these differences applied across all countries in the same way. This approach ensured that comparability is maximised, while mode effects that affected different items in different directions were accounted for so that potential effects on scale comparisons were minimised.

The comparisons are illustrated in Figures 9.11 to 9.19 separately for the domains of reading, mathematics and science. These figures show that for each domain, good agreement between country means by assessment mode could be achieved. The largest differences between means were observed based on a random school split, not based on mode. Thus, differences between countries might be due more to differences between students and schools than to differences based on the mode of assessment.

In summary, the differences and variability between gender groups and also the two groups formed by randomly splitting the 25 schools in the field trial were at the same level or larger than the differences obtained by splitting the sample by mode (in other words: mode effects do not seem to be the biggest problem). The apparent mode differences that may be observed if individual countries split their data by mode have to be viewed in the light of these results. Given the sample size of the field trial, differences that one may be tempted to attribute to mode differences are at the same order of magnitude as what could be observed if we split the field trial sample randomly by some other criterion.

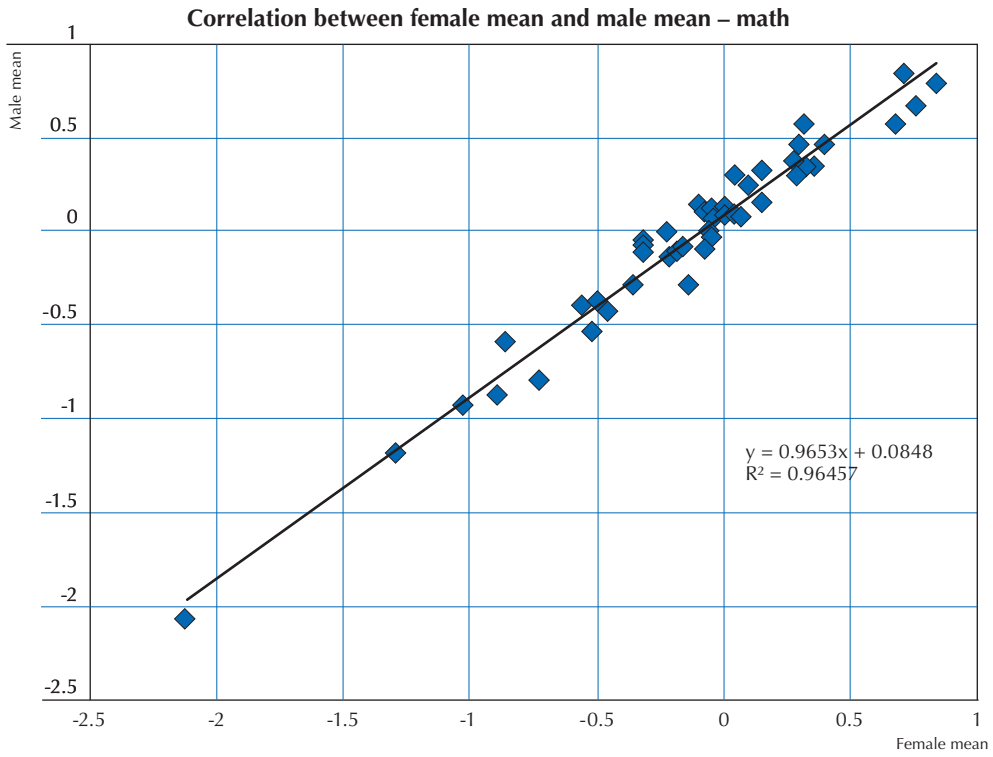
■ Figure 9.11 ■

Split of country means by assessment mode for mathematics



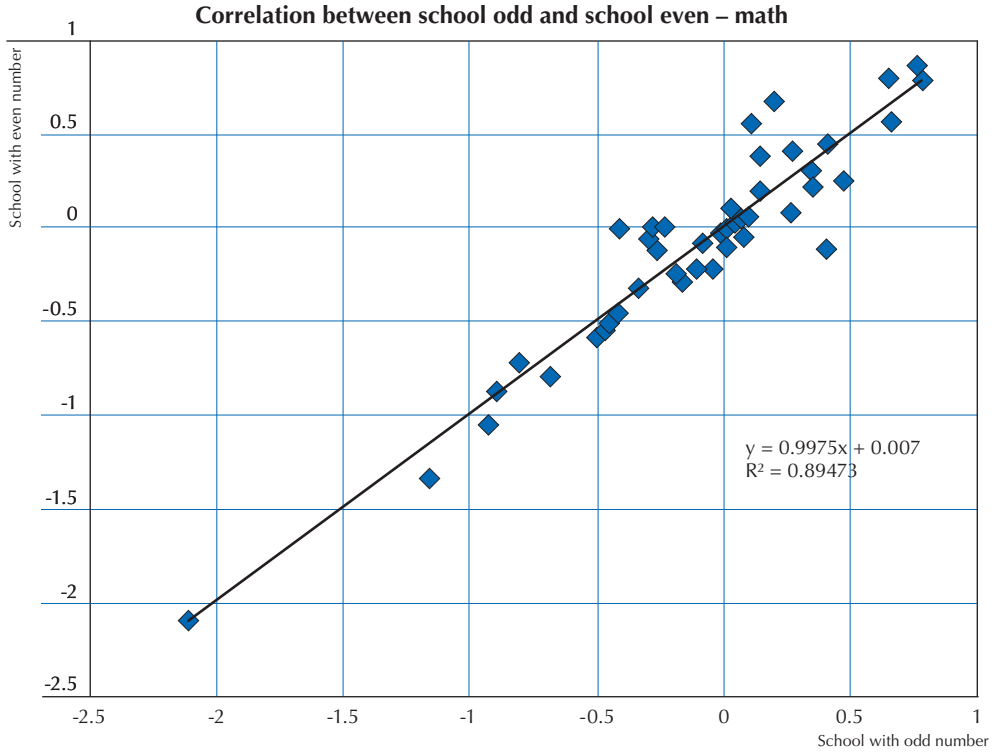
■ Figure 9.12 ■

Split of country means by gender for mathematics



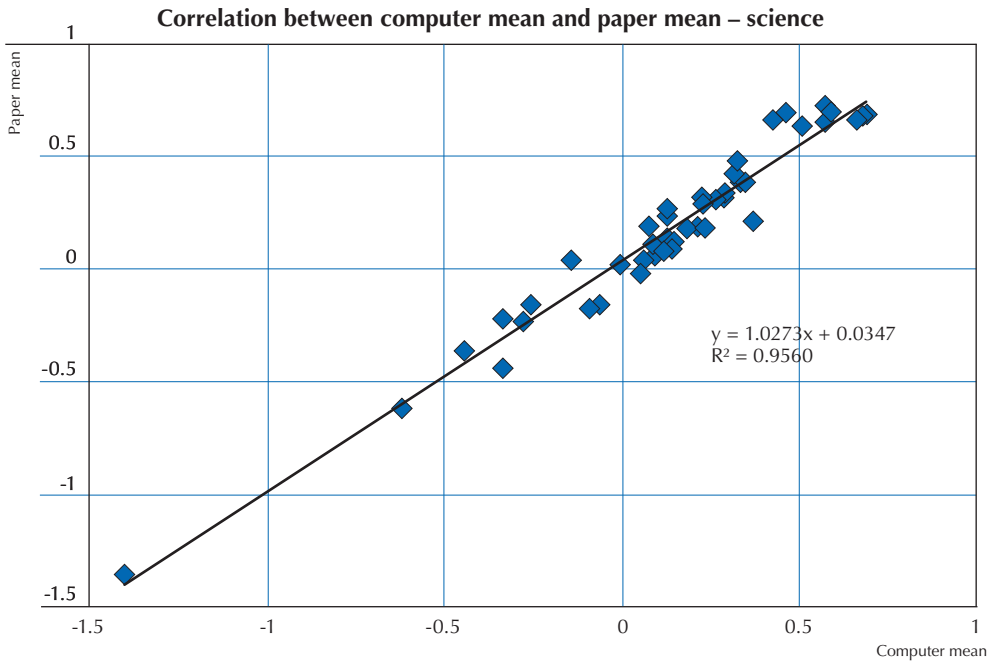
■ Figure 9.13 ■

Split of country means by random school split for mathematics



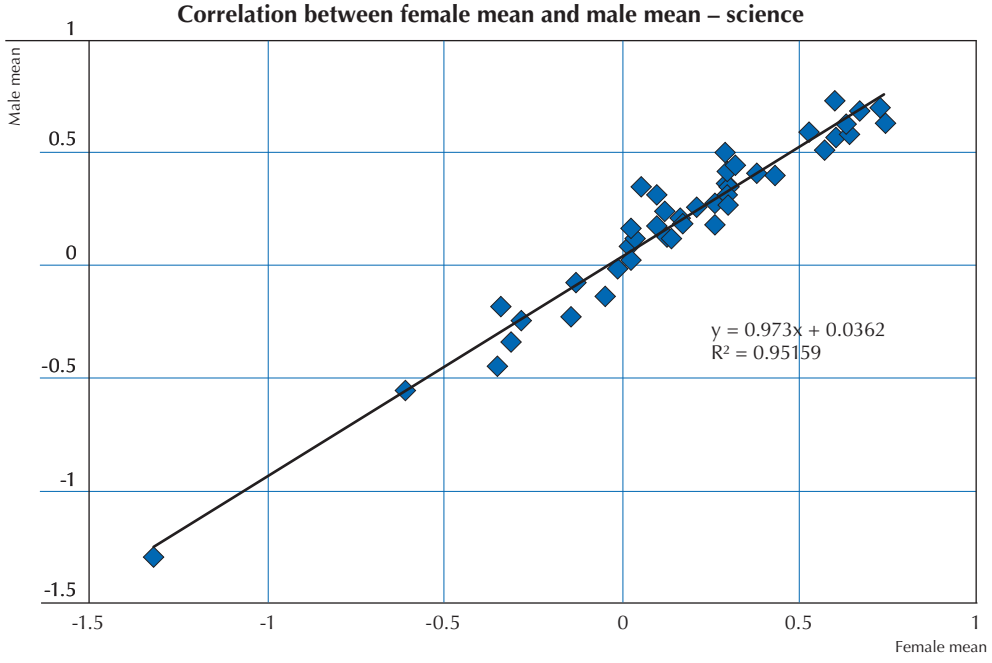
■ Figure 9.14 ■

Split of country means by assessment mode for science



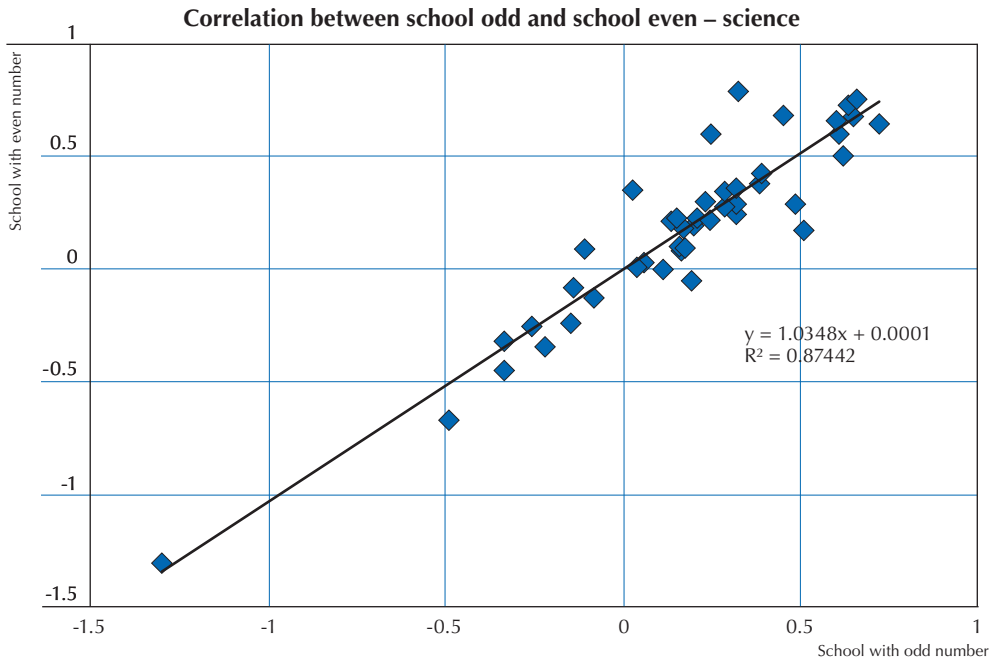
■ Figure 9.15 ■

Split of country means by gender for science



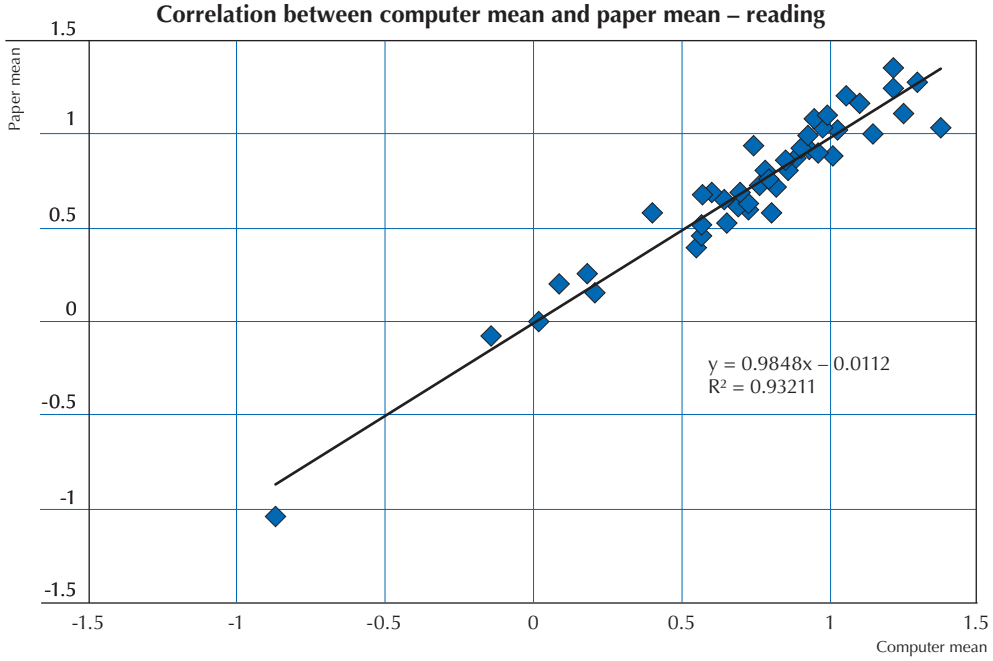
■ Figure 9.16 ■

Split of country means by random school split for science



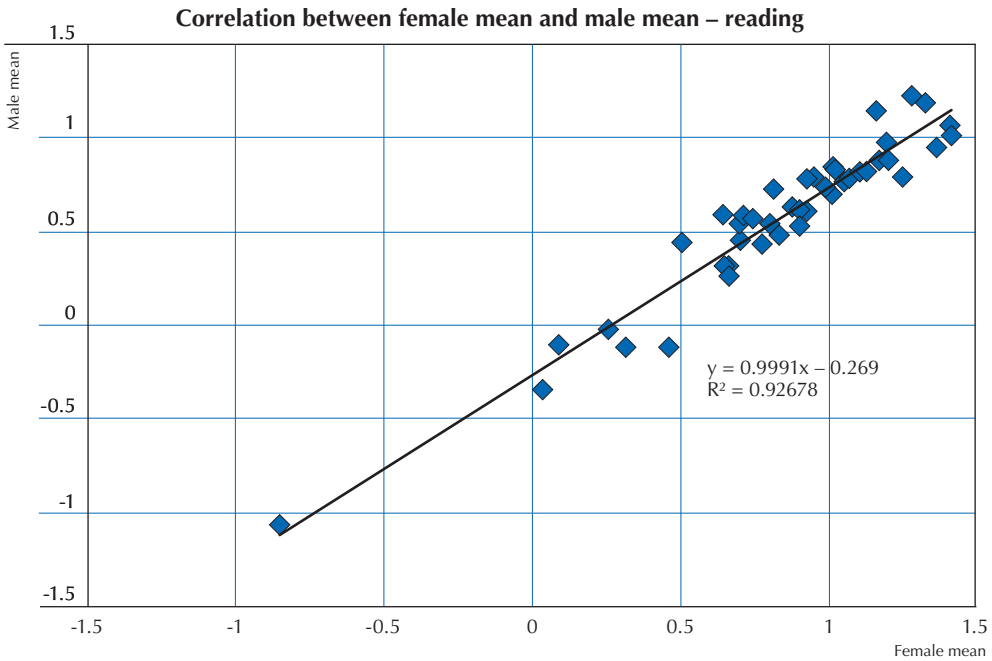
■ Figure 9.17 ■

Split of country means by assessment mode for reading



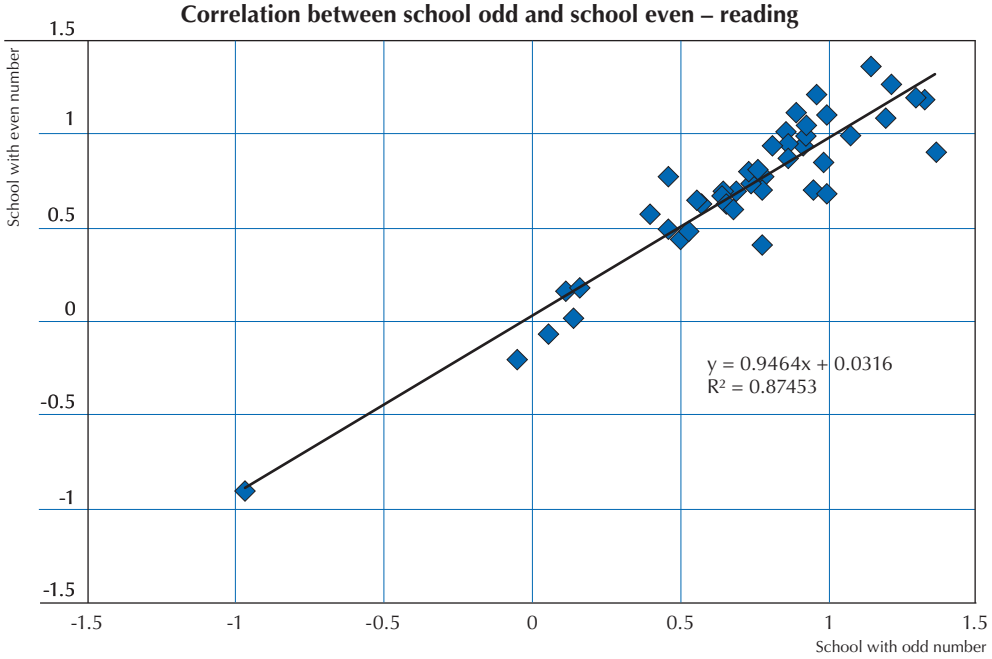
■ Figure 9.18 ■

Split of country means by gender for reading



■ Figure 9.19 ■

Split of country means by random school split for reading



Dimensionality and scaling of science trend and new items

Dimensionality of the science scale

The new science items developed for 2015 are based on a revised assessment framework for this domain. These new items exist in the computer-based assessment mode only because PISA 2015 represents a shift from a paper- to a computer-based survey. In addition to the 85 trend science items from previous PISA rounds, the science domain in the main survey consists of 99 new items resulting in a total of 184 overall. The scales for all PISA content domains have historically been based on the assumption that all underlying constructs are unidimensional. With the revised framework for science it is important to evaluate whether the unidimensionality assumption still holds before new and trend items can be scaled together.

This assumption was tested by comparing a unidimensional model (where new and trend items were assigned to the same unidimensional factor) and a 2-dimensional (multidimensional) confirmatory IRT model (where new and trend items were assigned to two different factors). In addition, a Rasch model for the unidimensional science scale was provided as comparison. All models, the Rasch, the two-parameter logistic /generalised partial credit model and the 2-dimensional (multidimensional) confirmatory IRT model two-parameter logistic/generalised partial credit model were estimated as multiple group models using country-by-language groups. The data used for this analysis came from the subset of computer-based assessment countries that was available at the end of March 2015; please note that due to the potential on the analysis of the PISA 2015 data, this analysis had to be completed prior to analysing the data from all PISA computer-based assessment countries.

Results based on overall model selection criteria show that the unidimensional two-parameter logistic/generalised partial credit model should be preferred over the 2-dimensional model (see Table 9.18). The difference in model fit improvement based on the Gilula and Haberman (1994) log penalty measure is negligible. The two-parameter logistic/generalised partial credit model reaches 99.91% of the model fit improvement compared to the 2-MIRT model, both in reference to improvement over the independence (baseline) model. Moreover, model-based correlations obtained from the 2-dimensional model show high correlations between the two factors (new and trend items) ranging from 0.83 to 0.96 across the different groups, suggesting there is a single identifiable underlying latent variable. Additionally, the dimension-specific weighted likelihood estimates (WLEs) of student ability are very highly correlated with the unidimensional WLEs. Hence it is reasonable to assume that new and trend science items and scores can be placed on the same unidimensional scale.

Table 9.18 Model selection criteria for the unidimensional and the two-dimensional IRT models for trend and new science items

	AIC	BIC	Log penalty	% improvement
Independence	NA	NA	0.6479	0.00%
Rasch model	8021282.185	8024639.114	0.5720	90.88%
2PL/GPCM	7916247.615	7922743.894	0.5645	99.91%
MIRT 2-dimensions	7915262.270	7922400.924	0.5644	100.00%

Note: Log penalty (Gilula and 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. The two-parameter logistic/generalised partial credit model reaches 99.91% of the likelihood improvement compared to the 2-dimensional MIRT model, while the Rasch model reaches 90.88%.

Residual Analysis for Science

As additional evidence in support of the unidimensionality assumption for the science scale, a residual analysis was conducted for the new science items. Due to the nature of the new science items (simulation-based tasks, including different steps for the students to follow) the goal was to investigate possible local dependencies among items. If such dependencies are present, this would pose a threat to the assumption of a unidimensional scale.

First, response residuals were calculated for each item response and correlations among residuals (across respondents) were computed. A principal component analysis using the resulting correlation matrix was then conducted. The principal components analysis was used to evaluate the dimensionality of the scale. Should the first component among residuals be much larger than the second component, an additional latent trait other than the overall ability would be assumed.

Response residuals were computed after the item calibration process in each domain using the *mdltm* software (von Davier, 2005). For dichotomous item responses, response residuals for a person v with estimated ability $\hat{\theta}_v$ for each item $i = 1, \ldots, K$ were defined as below:

9.17

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

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

9.18

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

9.19

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

9.20

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

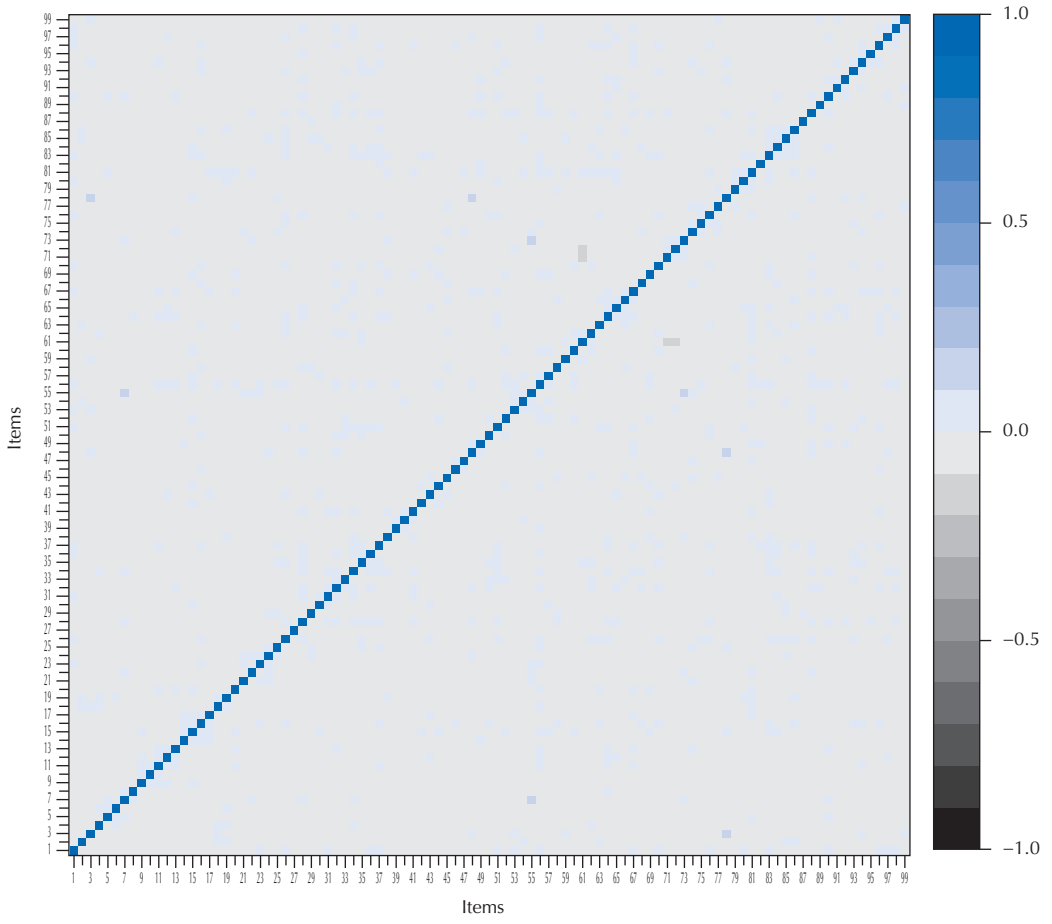
Response residuals were calculated for the 99 new science items using data from a subset of computer-based assessment countries (46 countries). Note again that due to the timeline of PISA 2015, this analysis was completed prior to receiving the data from all PISA countries.

In PISA 2015, no student responded to all of the questions. Given this missing-at-random design, Pearson correlations among items were calculated via pairwise deletion. The visual representations of the correlation matrices were evaluated for remaining dependencies. When a pair of items showed higher correlations, the pattern was checked to determine if it was consistent across countries. Findings from the correlation matrix were interpreted in connection with the item slope parameter estimates and item-total correlations. If an item pair showed highly correlated response residuals and the item slope parameter estimates were high as well for both items, converting these two item scores into a sum score and treating this score as one polytomous item was considered (Rosenbaum, 1988; Wilson and Adams, 1995).

Figure 9.20 shows a heat map plot of the correlations among item level response residuals for the new science items, averaged across countries. Highly-positive correlations between item pairs would be indicated by blue diamonds, highly-negative correlations would be indicated by red diamonds. Since there are none apart from the expected perfect correlation of each residual with itself, this plot suggests that there are no remaining local dependencies among the items after controlling for the latent ability. This pattern was consistently observed across countries. These findings, as well as the results of the principal component analyses, show that there are no local dependencies among the items. Hence, no further treatment (combination or exclusion of items) was needed for new science items.

■ Figure 9.20 ■

Correlation plot among new science items averaged across countries (46 countries)



Final scaling of science in the main survey

After confirming that all science items can be calibrated unidimensionally and without the need to change the scoring of the new simulation-based items, all items were calibrated using a single-scale multiple-group IRT model. No item had to be excluded from the calibration. The IRT scaling was conducted using the 2015 data together with the historical PISA data (2006-2012). The estimation of international/common item parameters and unique item parameters, in case of item misfit, and the treatment of items with identified mode effects followed the procedure described earlier.

The IRT calibration results show very good fit of the international item parameters. The international/common item parameters for both new and trend items were retained for 89.7% of trend items and for 93.3% of the new science items (see Chapter 12 for more information about scaling outcomes).

Scaling of reading and mathematics

In the PISA 2015 main survey, the domains reading and mathematics consisted of trend items only. Mathematics comprised 83 trend items in the paper-based assessment (PBA) and 82 equivalent trend items in the computer-based assessment (CBA). Reading consisted of 103 trend items in the PBA and 103 equivalent trend items in the CBA. Both domains were scaled separately using unidimensional multiple-group IRT models (see *The IRT models for scaling above*). The IRT scaling was conducted using the 2015 data together with the historical PISA data (2006-2012). The estimation of international/common item parameters and unique item parameters, in case of item misfit, and the treatment of items with identified mode effects followed the procedure described in the sections *National and international item calibration* and *Handling of item-by-country/language and item-by-mode interactions* earlier in this chapter. One mathematics item had to be excluded from the scaling (see Table 12.1 in Chapter 12); no items were excluded for reading.

The IRT calibration shows very good fit of international/common item parameters. The international parameters were retained in 89% of cases for common item parameters for reading items and in 94.5% of cases for items from the mathematics scale (see Chapter 12 for more information about scaling outcomes). The results illustrate high comparability of the results across different countries and languages, and across different assessment cycles and assessment modes.

Dimensionality and scaling of collaborative problem solving

Dimensionality of collaborative problem solving in the field trial

The collaborative problem solving (CPS) scale in the 2015 PISA field trial consisted of 7 units that comprised 188 items. The units are based on simulated conversations with one or more computer-based agents that are designed to provide a virtual collaborative conversation. Students have to choose an optimal sentence from a multiple-choice list to go through the conversation with agents, or choose one or more actions programmed in the unit.

For two of the seven units (unit 101 and unit 105) changes to the scoring of responses were necessary before the data could be used for IRT scaling. Using path analyses, it was found that – due to the nature of the collaborative problem solving items – data from the two mentioned units showed item dependencies in the responses. This was because of different paths that could be taken by students through the simulated chat, resulting in negative residual correlations. Since such dependencies have the potential to introduce bias into the results, the collaborative problem solving chat items exhibiting dependencies were combined into polytomous "composite items" by summing the responses for the different paths students could take. Table 9.19 provides an overview of the combination rules used for these composite items. Given these combinations, the number of items available for the IRT scaling was 164.

Table 9.19 Combination of collaborative problem solving items of Units 101 and 105 to achieve fair scoring in the PISA 2015 field trial

New item ID for composite items	Combinations of CPS items
CC101201C	CC101201+CC101202
CC101203C	CC101203+CC101204+CC101205
CC101206C	CC101206+CC101207
CC101301C	CC101301+CC101302+CC101303
CC101304C	CC101304+CC101305
CC101307C	CC101307+CC101308+CC101309A+CC101309B+ CC101310+CC101311+ CC101312A
CC101312BC	CC101312B+CC101313
CC101317C	CC101317+CC101318+CC101319
CC105103C	CC105103+CC105104
CC105105C	CC105105+CC105106+CC105107
CC105201C	CC105201+CC105202
CC105208C	CC105208+CC105209+CC105210
CC105212C	CC105212+CC105213
CC105304C	CC105304+CC105305

Dimensionality analysis of collaborative problem solving field trial data

The different units were combined into four clusters presented as C1 to C4 in the assessment design. The correlations between the clusters in the Field Trial were generally reasonable, with a range from 0.76 to 0.81 except for those involved with C1. Cluster 1, which contained only a single unit, had lower correlations with the other clusters, ranging from 0.69 to 0.73.

The specific structure of the CPS units and response types, as well as the results from the IRT analysis of the CPS using the unidimensional models, prompted the need to conduct additional analyses (discussed below). However, the unidimensional IRT models showed acceptable fit in terms of item mean deviation and root mean square deviation.

The structure of the CPS units was such that there were a relatively large number of response variables within a unit, while the number of units was small. The contextual coherence of the chat selections that made up these responses followed a common theme within a unit; the conjecture thus could follow that what is measured is more the understanding of what a particular topic requires and might therefore be very specific to each unit.

In order to examine this question, the collaborative problem solving data from the PISA 2015 field trial were analysed using multidimensional IRT models, more specifically with a bifactor model (Holzinger and Swineford, 1937). This model allows an evaluation of whether there is a single source of common variance shared across units, or whether the observed responses are additionally driven by unit-specific response tendencies. In other words, the bifactor model, when compared to a unidimensional model, allows a test of whether unit-specific factors have to be taken into account.

Table 9.20 Comparison of two-parameter logistic/generalised partial credit models and bifactor model for 164 CPS items

	Likelihood	A-penalty	AIC	B-penalty	BIC
2PLM/GPCM	-971208	1000	1943417	5652	1948069
Bifactor	-962224	2206	1926653	12468	1936915

The results in Table 9.20 suggest that a bifactor model including a latent variable for each unit fitted the Field Trial data better than the unidimensional two-parameter logistic/generalised partial credit models. The bifactor model indicates that unit response variance was due to unique factors that are not fully measured by a latent variable defined across response variables without looking at their association with a specific content or unit.

It turned out that this result was mainly due to a single unit, presented as C1. As a consequence of these findings, one unit (unit 101) was not included in the PISA 2015 main survey. Additional dimensionality analyses (residual analysis, principal component analysis) were conducted with the main survey data in order to further examine and treat local dependencies of collaborative problem solving items. The next section describes these additional analyses and findings based on the main survey data.

Dimensionality and residual analysis of collaborative problem solving in the main survey

For the PISA 2015 main survey, 134 items were selected out of the 164 (partly combined) items for the collaborative problem solving domain (unit 101 was not selected). The multidimensional structure of these items was examined residual analyses revealed further dependencies among items that led to further combinations of items into polytomous items (composite items). The residual analyses for CPS followed the same procedure as described earlier for science (*Final scaling of scientific literacy in the main survey*). Item-level response residuals were calculated for each item by respondent interaction for all observed responses, and pairwise correlations among these residuals were computed for the different country samples. Note again that due to the timeline of PISA 2015, this analysis was completed prior to receiving the data from all participating PISA countries. Several pairs of items were identified with highly correlated residuals; the pattern was quite consistent across countries. Figure 9.21 shows the correlations among collaborative problem solving items averaged across countries. Relatively highly correlated item pairs are indicated by blue diamonds and were mainly found near the diagonal line. This indicates that the dependencies (high item-pair correlations of response residuals) were mainly localised and taking place within a few selections. Rather than accounting for these in generalised latent traits measured through all responses in a unit, these localised dependencies were treated by item combinations as described above.

Based on the findings from the residual analyses, additional items were combined into composite items to remove the remaining local dependencies. Table 9.21 shows the combination of these items into composite items. Details about the items included in this rescoring can be found in the databases containing country-specific data as well as variable and value labels.

■ Figure 9.21 ■

Correlation plot among collaborative problem solving items averaged across countries before treating them as composite items (31 countries)

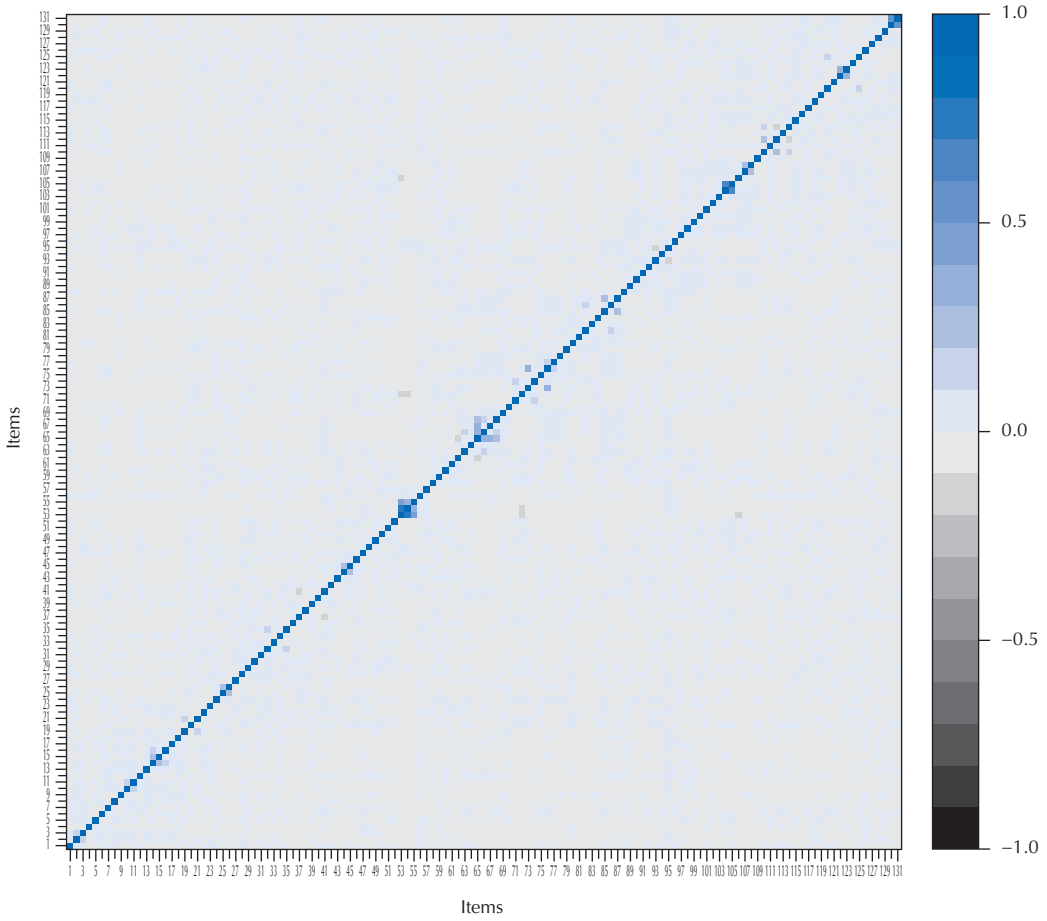


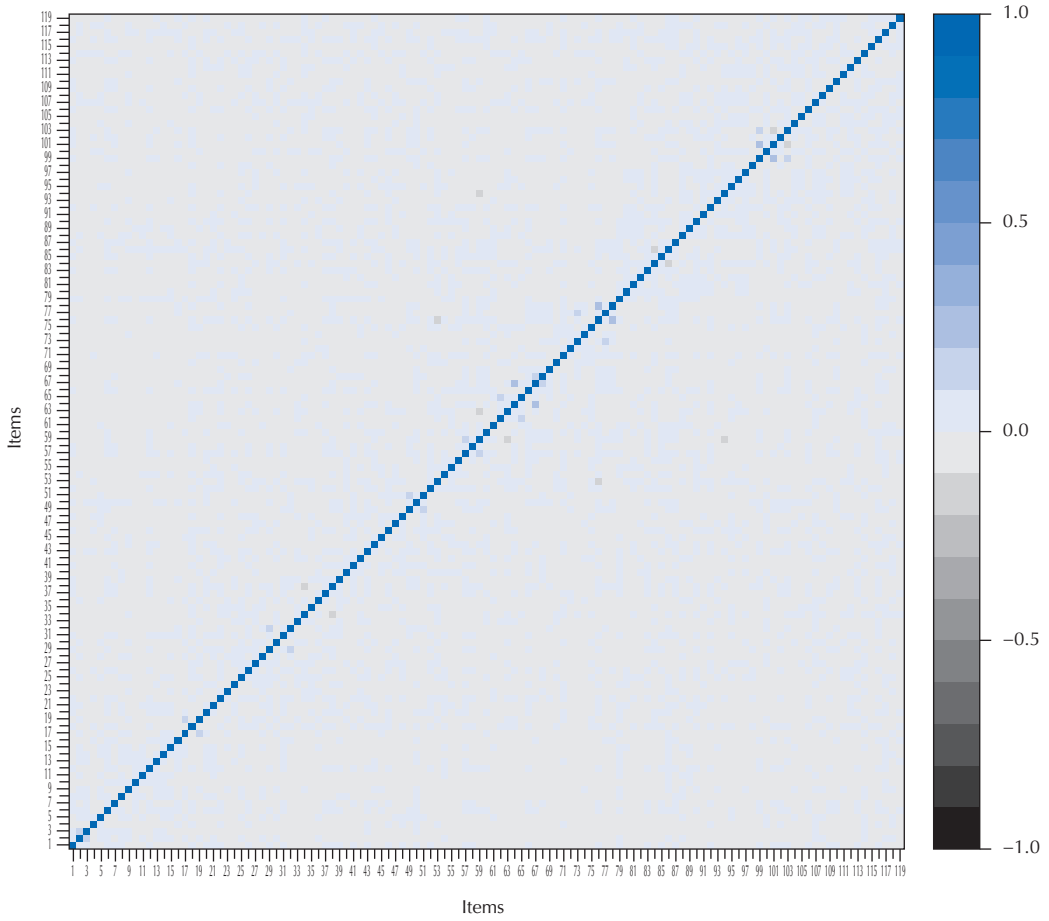
Table 9.21 List of composite items based on residual analyses

New item ID for composite items	Combinations of collaborative problem solving items
CC104301C	CC104301+CC104302+CC104304
CC106107C	CC106107+CC106108
CC102102C	CC102102+CC102103
CC102209C	CC102209+CC102210+CC102211
CC103108C	CC103108+CC103109+CC103110+CC103111
CC105108C	CC105108+CC105109
CC105203C	CC105203+CC105204
CC105308C	CC105308+CC105309
CC105408C	CC105408+CC105409

After the combination into additional composite items, the number of collaborative problem solving items was reduced to 121 (from the initial set or 134 items) for inclusion in the IRT scaling. In order to evaluate the performance of the composite items, residual analyses were repeated using the 31 countries and 11 additional countries for which data were later received (42 countries in total). Visual representation of the correlation matrix in Figure 9.22 confirmed that remaining local dependencies among items were successfully treated. In contrast to Figure 9.21 that shows several blue diamonds (highly correlated items) near the diagonal line, Figure 9.22 shows no blue diamonds off the diagonal line.

■ Figure 9.22 ■

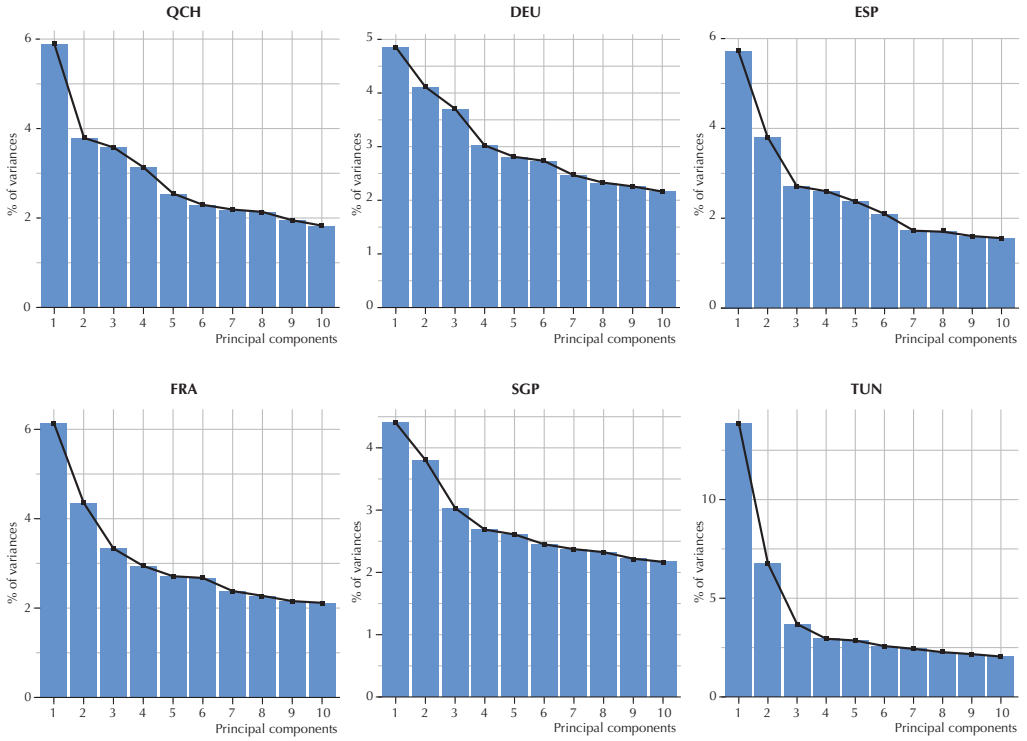
Correlation plot among collaborative problem solving items averaged across countries after treating them as composite items (42 countries)



In addition to the collaborative problem solving residual analysis, a principal components analysis was conducted using the residual correlation matrix. The principal components analysis was used to evaluate the dimensionality of the collaborative problem solving 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 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 4.4% to 13.9% with a mean of 6.9%. 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 26.2% to 41.5%, with a mean of 32.5%, 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 shows that one cannot justify the assumption of another dimension that may be able 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.

■ Figure 9.23 ■

Percentage of variance from principal component analyses (6 example countries)



Operational scaling of the collaborative problem solving main survey data

After removing all observed local dependencies by combining certain items into polytomous items, the resulting 121 collaborative problem solving items were calibrated using a unidimensional IRT model. Four items had to be excluded from the IRT scaling (due to low item total correlations, too few response in one response category, or technical issues; see Chapter 12), resulting in 117 CPS items on which the item parameter estimations are based. Note that all omitted responses in the CPS domain were scored as not reached (missing) due to differences in the administration of this domain. Omissions in reading, mathematics and science may be the result of intentional skipping of items, as students have the ability to move to the next item without interacting with the current one. In collaborative problem solving, however, students must make a sequence of successive choices and cannot skip forward to avoid a choice. Thus, unobserved responses in CPS items are a result of students taking different paths while working on an item, meaning some paths are not taken. Therefore, unobserved responses do not reflect student skill and need to be treated as not administered. The estimation of international/common item parameters and unique item parameters, in case of item misfit, followed the procedure described in the sections *National and international item calibration* and *Handling of item-by-country/language and item-by-mode interactions* earlier in this chapter.

The IRT calibration shows good fit of the international/common item parameters. International parameters were retained in 95% of the item parameters (see Chapter 12 for more information about scaling outcomes) and, thus, a high comparability of the scale across different countries and languages.

Scaling of financial literacy

In PISA 2015, financial literacy had a data collection design that provides stronger connections to data collected in other domains, compared to the PISA 2012 design. That is, every student who took financial literacy also took reading, mathematics, or both, in addition to science. Therefore, PISA 2015 provides a better estimate of the covariance between the core domains and financial literacy. However, because not every country took financial literacy in PISA 2015, there are only a few countries that have data available in both years. As such, the 2015 main survey calibration required data from PISA 2012 as well as the 2015 field trial. This approach provides a sound link for PISA 2015 because, in the 2015 field trial data, a larger group of countries took both the computer- and paper-based assessments (for the mode-

effect study). This is also important since the 2015 administration of financial literacy is based on data collection for a subset of students in a second (afternoon) testing session. All available financial literacy data (2012 main survey, 2015 field trial, and 2015 main survey) were combined for the IRT scaling using a multiple-group IRT model based on an equivalent-groups (for the field trial samples) design for the linking. This particular linking method provides a sound link and is robust against changes in the percent correct observed in the 2015 main survey; the inclusion of the field trial data allows the assumption of equivalent groups since students were randomly assigned in the field trial to paper- versus computer-based assessments.

The equivalent groups design is a method of linking that is common in test equating. While it provides a consistent linking approach, it does not provide information on which items are directly comparable. Neither does it require or assume that the items be invariant across assessment modes, since the comparability is established based on the premise that the distribution of student ability is equivalent across groups. The link in financial literacy is established through common populations, while for the other scales (reading, mathematics and science) it was possible to link across modes and assessment cycles using common items.

In the PISA 2015 main survey, the financial literacy domain consists of 43 trend items. No items were excluded from the scaling. The estimation of international or common item parameters and unique item parameters, in case of item misfit, and the treatment of mode effects followed the procedure described in earlier sections.

The IRT calibration shows a very good fit of the international/common item parameters. The scaling was able to retain common/international item parameters for 92.9% of the items (while for 7.1% of the items unique item parameters had to be estimated) and, thus, a high comparability of the scale across different countries and languages (see Chapter 12 for more information about scaling outcomes).

Developing common scales for the purpose of trends

The new modelling approach in PISA 2015 using a hybrid model (the combined Rasch /partial credit model and two-parameter logistic/generalised partial credit model) necessitated a reanalysis of data from prior cycles (2000-2012) with the aim of studying the effect of the more general model applied over multiple cycles on stabilizing the trend measure and to ensure its quality. With the introduction of computer-based assessments as the main mode of assessment in PISA 2015, there was concern that the mode might influence item parameter estimates for the linking items. Moreover, some linking items might not work equally well for all of the populations assessed in PISA 2015. Using these items reduces the comparability of the trend measure; hence, there may be a need to exclude them from the main survey item pool. However, given the new scaling approach for PISA 2015, it might be possible to retain a larger share of these items, since the model used is more flexible and contains the previous scaling approach as a special case.

Results from prior analyses (PISA 2000-2012) were replicated and then re-examined using the hybrid Rasch/partial credit model and two-parameter logistic/generalised partial credit model. The reanalysis produced a common parameter for each of the previously used items in the databases from PISA 2000 to 2012. These parameters were treated as fixed parameters for the PISA 2015 field trial scaling. This was done to establish a stable link between the field trial items and the international scale based on past frameworks of each domain. Parameter constraints for various items were released in subsequent rounds in case of item misfit. The common item parameters in the field trial generally fit well; thus, the same item parameter can be assumed over cycles for a large number of trend items.

The overall item fit for each domain was very good, with small numbers of items misfitting for reading (2.5%), mathematics (1.8%), and science (3.9%). Financial literacy showed the highest percentage of misfit (4.1%). Note that item misfit was defined for root mean square deviation values larger than 0.2 in the field trial then later in the main survey analysis. All of the main scales showed sufficient IRT-based (marginal) reliabilities (Sireci, Thissen and Wainer, 1991; Wainer, Bradlow and Wang, 2007, 76) with 0.83 for reading, 0.81 for mathematics, 0.80 for science (based on trend and new items), and 0.85 for financial literacy. These results illustrate the quality of trend measure across different assessment cycles (2015 data versus 2000-2012), different assessment modes (paper- versus computer-based assessments), and even across different countries and languages as the multi-cycle scaling with common item parameters assures the equivalence of inferences of trend assessment.

In the PISA 2015 main survey a comprehensive rescaling was carried out including the 2015 main data. This was done to ensure that the main survey data equally contributed to the estimation of item parameters, while establishing the link to past PISA rounds by including previous cycles. Instead of fixing the item parameters for trend items

obtained from past (historical) data to the 2015 data, item parameters were estimated based on all available data from 2006 through 2015. The historic data were only included back to 2006 because this was the last cycle when science was the major domain, and because there were very few items left in the 2015 round that dated back to the early (2000 and 2003) rounds of PISA. This approach ensured that domains tested in 2015 with a new design that improved minor domain coverage and broadened the assessment of the revised framework were contributing to the estimates that established the common scale linking across prior PISA cycles. The IRT calibration for each domain showed good fit of the items to the international/common item parameters. The results also showed high comparability in the item parameters across different countries and languages, and across different assessment cycles and assessment modes.

Rescaling PISA 2000-2012

The PISA 2015 field trial and main survey design were premised on the availability of a quality set of the linking items across the previous PISA cycles. These designs incorporated all previously used trend items from all previous cycles in the field trial so that the best possible link could be established.

This increase in scope also required that prior analyses be revisited because the integration of all previously used trend items required a full re-estimation of the scaling model on which prior PISA cycles were based. There is strong evidence in favour of a joint model for linking the cycles across multiple populations (von Davier and von Davier, 2007; Mazzeo and von Davier, 2008, 2014). This also allows different trend clusters containing items sets not previously used in a single assessment to be linked together within a comprehensive modelling approach.

PISA has collected data in representative samples of 15-year-old students around the world every three years since 2000. In each of the first five cycles (2000, 2003, 2006, 2009 and 2012), both OECD and partner countries participated, resulting in almost 300 cohorts defined by assessment year and country. Many of the OECD countries, as well as a substantial number of partner countries, had participated in each of the first five PISA cycles and continued to do so in 2015.

In work leading up to the 2015 main survey analysis, an effort to utilise all available evidence on item functioning and scale coverage of the task material used in PISA was made. ETS compiled a database that merged all five cycles and all countries. This yielded a file that contains roughly 2 million student records. ETS utilised a multiple group IRT model approach to link all items, by domain, across all PISA cycles by country combinations (Bock and Zimowski, 1997; Mazzeo and von Davier, 2014, 2008; von Davier and von Davier, 2007; von Davier and Yamamoto, 2004; Weeks, von Davier, and Yamamoto, 2014; Yamamoto and Mazzeo, 1992).

Several analytical steps were performed. More specifically, in order to find the best fitting model, different and increasingly complex IRT models were specified and estimated; model-data fit was compared using both AIC and BIC as well as measures of item fit. The analyses were carried out separately for each of the main PISA domains of reading, mathematics and science.

In a first step, the model used in the operational reporting of PISA 2000-2012 was recreated in order to ensure that the results obtained in the previous analyses could be replicated. Previous cycles of PISA utilised the mixed-coefficients multinomial logit model (MCMLM; Adams, Wilson, and Wu, 1997), which is a generalisation of the Rasch model (Rasch, 1960) that allows for category weights, multiple populations, and predictors of ability as well as polytomous response data. This was followed by an approach that utilised model-data fit indicators to relax model assumptions of the Rasch model where needed. More specifically, the Rasch model assumption of equal slopes was relaxed if it was found that the item discrimination was markedly different in the group of countries by cycles in which an item was used. ETS compared this analysis with an estimation of the two-parameter logistic /generalised partial credit model (Birnbaum, 1968; Muraki, 1992) for multiple populations (von Davier and Yamamoto, 2004).

This initial analytic step allowed the estimation of slope parameters for those items that were found to discriminate more (or less) well than the items that follow the Rasch model. In the next step, model assumptions were relaxed further. Given that international assessments are translated into multiple target languages, item-by-country interactions are a potential threat to validity. As such, some items in some countries may function differently from how the item generally functions in the majority of countries. For this reason, we added an analysis step that investigates item-by-country (by cycle) interactions in order to catch cases in which an item deviates substantively in one or the other cycles of PISA. This approach follows best practices (Glas and Jehangir, 2014; Glas and Verhelst, 1995; Oliveri and von Davier, 2011, 2014; Yamamoto, 1998). All analyses were carried out using the software *mdltm* (von Davier, 2005). The next

section describes the results of the rescaling with the Rasch model, followed by a description of the model that combines features of the Rasch model and the two-parameter logistic/generalised partial credit model and the model for country-by-item interactions.

Results for the Rasch model

In this subsection, we examine the comparability of rescaled and reported results from previous analyses. We initially fit the data with the Rasch model since it has been the operational model used for reporting PISA results by cycle for the past five assessments. The results from our reanalysis of the data were compared with published results available online. Note that for our analysis, we obtained item parameters and country means by estimating one model across all cycles and all participating countries. This approach differs from the operational approach used in past cycles in that it incorporates all data into the item calibration in order to link the results across cycles. The operational approach, on the other hand, uses only the mean of trend items in two adjacent cycles to find transformation constants in order to put the new scaling results on the old scale. If the fit of the model is perfect (i.e. if item parameters stay the same over cycles), and if the item functions can indeed be fitted by the Rasch model, both methods should produce identical results. In this case, however, the use of all cycles in a single comprehensive estimation of the Rasch model should lead to the most accurate item parameters possible, given the data at hand.

The comparison was carried out using two independent rescaling approaches. In contrast to the operational approach implemented by the contractor responsible for the 2000-2012 cycles, we did not use a random selection of 500 cases per country. Instead we used all data from every country participating in these five PISA cycles. The re-estimation of parameters was conducted either per assessment cycle using the ConQuest software (Wu, Adams and Wilson, 1997) or using all data from all five PISA cycles in a concurrent calibration using *mdltm* (von Davier 2005). The replication effort was done to ensure that we could recover the previously estimated item parameters.

In summary, the reproduction of the original reporting scale was fully successful under both estimated approaches. The correlations between country means as reported by PISA and those reproduced by calibrating all available data in a comprehensive scaling was above 0.998 and, in many cases, especially for the *mdltm* calibrations that used all available data across cycles (0.999 and above). This suggests that there were no issues with the data used to estimate the item parameters. However, the estimation of a comprehensive model using data from all cycles leads to the most consistent item parameter estimates and a scale that is linked in the most rigorous way (see also Chapter 2) across all available PISA cycles.

Results for the hybrid 'partial Rasch, partial two-parameter logistic/generalised partial credit'model

Given that we were able to replicate the Rasch model results, we moved on to an approach that combined features of the Rasch model and more general IRT models. Among these models are the two-parameter logistic/generalised partial credit model, which estimates a slope parameter for all items, a hybrid combination of the Rasch model and two-parameter logistic/generalised partial credit model that estimates slope parameters only for items that do not fit the Rasch model, and a model that additionally accounts for item-by-country interactions (IBCI) and estimates unique item parameters for countries and/or country-groups for items that cannot be fitted well using a common international parameter (Glas and Jehangir, 2014; Glas and Verhelst, 1995; Oliveri and von Davier, 2014, 2011; Yamamoto, 1998; Yamamoto and Mazzeo, 1992). Note that all model extensions are exponential family models, and that the operational model, the Rasch model, used in the first five rounds of PISA, is a special case of the extended approach. If the Rasch model indeed fits the data, the extended model will just reflect that, namely by fitting the data with something that very closely resembles the fit of the Rasch model. However, if the extended approach statistically fits the data substantially better than the approach used in previous rounds, this will be visible in model selection criteria.

This hybrid combination of item functions from either the Rasch model or the two-parameter logistic/generalised partial credit model allowed for fitting of a wider range of items compared to using the Rasch model alone. In contrast to the two-parameter logistic /generalised partial credit model being applied to all items, we were able to retain a number of slope parameters that are fixed across items, and hence were able to provide a model that makes the same assumption (an equal slope across items) as past PISA cycles for a subset of items. Table 9.22 gives an overview of the number of items that were retained as "Rasch" items using a common slope parameter of 1.0 in the hybrid model (Rasch/ two-parameter logistic /generalised partial credit model) accounting for IBCI (hybrid/IBCI model).

Table 9.22 Number of Rasch model items retained in the hybrid/IBCI model

	Total number of items	Rasch # retained	Rasch % retained
Mathematics	179	77	43%
Reading	223	42	19%
Science	133	19	14%
Financial literacy	40	15	38%

Table 9.23 summarises the improvement in model fit for the domains of reading, mathematics and science. The table shows the results for the Rasch /partial credit model, the two-parameter logistic /generalised partial credit model, and the "hybrid" model (Rasch/two-parameter logistic/generalised partial credit model), with one set of item parameters for all countries, and a model that accounts for IBCI by releasing some country-specific parameters. These results are based on all cycles from 2000-2012 combined for the three domains. In each domain, the IBCI model fits best (as characterised by the BIC), followed by the two-parameter logistic /generalised partial credit model, the hybrid model, and the Rasch /partial credit model. This can also be seen in the concomitant decrease in the number of items-by-country-by-cycle with root mean square deviation values greater than 0.15. Approximately 3% of the items in mathematics, 7% of the items in reading and 6% of the items in science did not fit the Rasch model in one or more countries. On the other hand, around 1% of the items exhibit misfit in reading for the IBCI model and less than 0.1% of the items exhibit misfit in mathematics and science under the hybrid/IBCI model. For all subsequent analyses, the item parameter estimates from the hybrid/IBCI model were used.

Table 9.23 Changes in model fit summary

		Rasch/PCM	2PLM/GPCM	Hybrid	IBCI
Maths	# of item-country-cycle deviations	549	397	415	4
	BIC	26400730	26118134	26175012	25946516
Reading	# of item-country-cycle deviations	1233	960	962	250
	BIC	30968125	30675531	30691983	30472304
Science	# of item-country-cycle deviations	921	717	708	8
	BIC	29908518	29585732	29591677	29302806

Total item-country-cycle values: maths = 15,795, reading = 18,603, science = 16,223
Deviations defined as RMSD values > 0.15

Fit of the Rasch Model and two-parameter logistic model for new science and collaborative problem solving items in the field trial

After examining the fit of the new modelling approach developed for PISA 2015 to data from past PISA cycles (2000-2012), described in the sections above, the fit of the Rasch/partial credit model versus the two-parameter logistic/ generalised partial credit model was tested for new science and collaborative problem solving items using data from the 2015 field trial (note that this comparison was done in the field trial in preparation for the main study; hence, no similar comparison was needed in the main study). The aim was to investigate whether the two-parameter logistic / generalised partial credit model shows a better fit, as would be expected.

While the item parameters for trend items in the field trial were fixed to those obtained from the reanalysis of previous PISA cycles (2000-2012), the new science and CPS items had to be scaled based solely on the field trial data. For these new scales, both a multigroup Rasch/partial credit model was estimated as well as a multigroup two-parameter logistic /generalised partial credit model. The concurrent calibration (multiple-group IRT model) was used to evaluate whether items were working in the same way across country-by-language groups or if there were item-by-group interactions. Both model approaches were compared (see Table 9.24) and it was found that the two-parameter logistic /generalised partial credit modelshowed better overall model fit than the Rasch/partial credit model. The item selection for the main survey was based on the two-parameter logistic /generalised partial credit model due to the improved model fit and because more information about each single item was provided.

Table 9.24 Comparison of the Rasch/ partial credit model and the two-parameter logistic /generalised partial credit model for new items in the PISA 2015 field trial

	Likelihood	A-penalty	AIC	B-penalty	BIC
CPS					
RM/PCM	-985477.57	686	1971641.15	3877.09	1974832.24
2PLM/GPCM	-971208.69	994	1943411.38	5617.83	1948035.21
Science					
RM/PCM	-2215483.30	1266	4432232.60	7406.46	4438373.06
2PLM/GPCM	-2192778.99	1698	4387255.97	9933.78	4395491.75

Linking PISA 2015 to previous PISA cycles

The goals of the PISA 2015 linking design centred on linking different test forms and assessments modes (paper- and computer-based) within the PISA 2015 cycle for comparability across countries and linking previous PISA cycles to PISA 2015 for comparability across assessment cycles and trend reporting.

To obtain comparable test results across the years in each domain, it was important that all items in a given domain were calibrated on one common scale. To establish a common scale, the items had to be linked together across test forms (subset of items), assessment modes (paper- and computer- based), and PISA cycles. This was achieved by using common sets of items in the different booklets and assessment modes. Moreover, the PISA 2015 linking design included items from the previous studies and links all PISA cycles (2000 through 2015). Note that for the scaling in the 2015 main survey, combined PISA data sets from 2006-2015 were used for parameter estimation. The new part of the science scale and collaborative problem solving as a new domain comprised only computer-based items (due to the nature of the items); because collaborative problem solving is a new domain, there are no linking items. Financial literacy, as an optional domain, was only linked back to 2012 (the first time financial literacy was assessed) in the 2015 main survey scaling.

In summary, the computer-based assessments included all domains and all linking/trend items (providing a link between paper- and computer-based testing and between the current and previous PISA cycles) as well as new items for science and collaborative problem solving. The computer-based assessments comprised the following item sets:

- reading, mathematics and financial literacy: intact clusters of paper-based items from previous cycles, reauthored for computer delivery
- science: intact clusters of paper-based items from previous cycles, reauthored for computer delivery, plus new items developed for computer delivery only
- collaborative problem solving: new items developed for computer delivery only.

Thus, all trend items were administered in both the paper- and computer-based assessments as well as in different test forms (across the different assessment modes). Within both assessment modes, all items were linked together in a booklet design, which relates to trend items in the paper-based assessments and the trend and new items in the computer-based assessments. The mode effect study allowed identification of scalar and metric invariant items across computer- and paper-based testing and thus allowed linking across modes. The inclusion of all non-released items in the new assessment design strengthened the construct coverage of the major and minor assessment domains and allowed linking the new science domain against all trend material dating back to the last major domain round in science, assessed in 2006.

The improved linking design established in 2015 (see Chapter 2) made it possible to calibrate all trend and new items answered by different students in different test forms and assessment modes on one common scale for each domain. This was done within the item calibration utilizing the approach described in the sections *The IRT models for scaling, national and international item calibration* and *Handling of item-by-country/language and item-by-mode interactions*.

To place the PISA 2015 results and the historic PISA results from cycles 2012 to 2006 on the same scale, a concurrent item calibration was used. This linking approach is different from the mean/mean IRT linking approach used in prior PISA cycles. For trend items that did not show mode effects, item difficulty, and slope parameters in the main survey were constrained to have the same parameters as the corresponding paper-based items and items found in the historical data, establishing scalar invariance for a majority of items in each domain. For the remaining items, metric invariance was established so that a common slope parameter is shared across cycles and across modes in 2015. This approach created a scale that allowed for the comparison of PISA 2015 main survey and historic PISA results.

For financial literacy, a slightly different approach was taken by linking the 2015 main survey data not only to the data from 2012 but from the 2015 field trial. The reason is that not every country took financial literacy in 2015, and only a few countries took the assessment in both cycles (2012 and 2015). Moreover, the administration of financial literacy in 2015 was based on the data collection from a subset of students in a second (afternoon) testing session. Consequently, linking through the 2015 field trial data, where a larger group of countries took both the computer- and paper-based versions, provides a more defensible scale.

More detailed information about the test design for PISA 2015 can be found in Chapter 2 and more information about the linking and IRT scaling in general and for each domain is given in the relevant sections of *Application of IRT* and *population models to PISA* above.

Linking error in PISA 2015

PISA accounts for student sampling error, measurement error of ability estimates and linking error. An evaluation of the magnitude of linking error can be accomplished by considering differences between reported country results from previous PISA cycles and the transformed results from the rescaling. Recall that prior PISA rounds used a separate item calibration for each cycle. That is, the same items, if repeatedly used in 2000, 2003, …, 2012 received slightly different statistical quantities as estimates of their difficulty, especially because they would be tested together with other sets of items, or as part of a smaller (minor domain) or larger (major domain) set of items.

This variability over time and different PISA assessment designs (minor/major, etc.), and also 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, as 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.

The extended analytical approach used in 2015 allows us to revisit the linking error and to reduce it when moving forward with the new design, which reduces construct coverage differences between minor and major domains and with the concurrent calibration used in the IRT scaling.

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
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.

In practice, not all of the sources of uncertainty around scale comparability were quantified or could be accounted for in past PISA cycles (2000-2012). The linking error estimated in past PISA cycles accounted only for differences across trend items observed for the re-estimated difficulty parameter from one cycle to the next. This approach of linking scales by "separate calibrations" includes the following steps. First, calibrations of data from assessment cycle one (Y1) and assessment cycle two (Y2) were run separately with trend items and items unique to each assessment cycles; this produces two sets of trend item parameter estimates (one set for each cycle). The mean of Y2 trend item difficulties was then transformed to the mean of Y1 trend items, in order to link the scales. This mean-only transformation is only valid for the Rasch model, if it is indeed fitting the data. Because the same "shift" parameter is added to all participating countries in order to equate results to previous assessments, any uncertainty that is introduced through this shift is common to all students and all countries. This is a form of mean/mean IRT linking, a method that operates on independently estimated item parameters. This method was applied in past PISA cycles (before 2015), and it relied directly on parameter invariance assumptions in the trend item set comparing estimates from two separate calibrations across adjacent PISA cycles. In this approach, the variance of differences between trend item estimates from the Y1 and Y2 calibration was used to characterise linking error; it can be written as:

9.21

$$LE_{<'15} = \frac{1}{k}\sum_{i=1}^{k}\left(\hat{b}_{Y1,i} - \hat{b}_{Y2,i}\right)^2$$

When we assume that each item parameter estimate can be written as the sum of the true parameter and an error term unique to the cycle, we can write for both cycles:

9.22

$$\hat{b}_{Y1,i} = b_{true,i} + \hat{u}_{Y1,i}$$

and

9.23

$$\hat{b}_{Y2,i} = b_{true,i} + \hat{u}_{Y2,i}$$

Assuming that the parameter estimates are unbiased yields that both error terms are vanishing in expectation. A final step combines the two separate cycle calibration errors, that is:

9.24

$$\hat{e}_{Y1,Y2,i} = \hat{u}_{Y1,i} - \hat{u}_{Y2,i}$$

Then, the pre-2015 linking error in (9.21) can be written as:

9.25

$$LE_{Y1,Y2} = \frac{1}{k}\sum_{i=1}^{k}\left(\hat{e}_{Y1,Y2,i}\right)^2 = V\left(\hat{e}_{Y1,Y2}\right)$$

The expression in (9.25) characterises linking error as the sum of the combined errors of item difficulty estimates obtained from two independently calibrated cycles Y1 and Y2 in which the trend items occur (potentially together with a set of items unique to each cycle). In other words, the linking error quantifies the item-by-cycle interactions, not the item-by-country-by-cycle interactions. The rationale for this approach was that the Rasch model is "symmetric," which means an increase in difficulty of items can be compensated by the same increase in average ability.

This approach to estimating linking error assumes that the variability of item parameters over cycles directly translates into variability of person estimates, and that the average effect of parameter differences is zero, since the scales between Y1 and Y2 are linked. Thus, all country measures are affected in the same way by linking errors, which results in scale-level linking error. Moreover, note that there are two sets of trend item parameter estimates for each cycle, but neither is correct because both differ from the expected true parameters.

Other contributing factors to linking error are limited sample sizes and the number of unique items in each assessment cycle (unique means only administered in a particular cycle). In turn, this variability stems from differences in the calibration sample and the sampling variability associated with choosing a calibration sample, and from the presence of items that are unique to each cycle. This uncertainty is also related to the particular sample of trend items that were used in both cycles.

The above approach is only possible for the Rasch model, as there is only one parameter type incorporated in the linking error. In addition, it does not directly take into account the differences due to model error, for example, differential item functioning across countries that is not fully accounted for in modelling. Therefore, a new approach to characterise linking error was implemented in PISA 2015 that provides an estimate of the expected uncertainty due to differences between older and newer calibrations with more data.

The premise underlying the new approach is consistent with previous PISA cycles, yet it makes a different set of assumptions that can also be applied to more general IRT model-based linking. 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 compared with regard to 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:

9.26

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

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:

9.27

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

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:

9.28

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

And the expectation can be estimated by:

9.29

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

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:

9.30

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

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:

9.31

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

The main characteristics of the new 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 (9.31) 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. In the application to linking error estimation for the 2015 PISA trend comparisons, a robust version of the scale was used. The robust measure of standard deviation that was used is the S_n statistic (Rousseeuw and Croux, 1993). It is defined as:

9.32

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

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*. 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.

Population modelling in PISA 2015

The population model described earlier (*Latent regression model and population modelling*) was applied to the PISA 2015 data after the IRT scaling in order to generate 10 plausible values for each student. Plausible values for students reflect the information contained in responses to the items of domains that respondents actually took and the context questionnaire variables. Plausible values in all the major domains were generated for all students participating in the assessment, regardless of whether they were administered items in that domain. In addition, in countries where collaborative problem solving was administered plausible values were generated for all students, regardless of the test form they took. That is, respondents will be assigned plausible values for domains in which they did not participate, borrowing statistical information from students similar in performance on other domains, and in their responses to background data. This is enabled through the use of the population model, which uses the covariance information among all domains and also nearly all context questionnaire variables, as well as data about the number of not-reached items and other variables relevant to predicting proficiency distributions within each country.

Students who received plausible values for the domain(s) they did not take, but these values have a larger uncertainty (measurement error) than the plausible values for the other domains (that were administered to them). The measurement error has to be taken into account when dealing with the plausible values in secondary analyses. By using repeated analysis with each of the 10 plausible values, the measurement error will already 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 information about how the population model was applied to PISA 2015 data, how plausible values were generated, and how plausible values can be used in further analyses.

Treatment of students with fewer than six test item responses

This section addresses the issue of students who provided background information but did not completely respond to the test items. A minimum of six completed items per domain was necessary to assure sufficient information about the proficiency of students. In general, there are very few students[3] (0.04%) with responses to fewer than six test items in at least one of the core test domains (reading, mathematics, science and collaborative problem solving). These cases, identified across the core domains, were initially removed from the first round of the population modelling for the core domains as well as for financial literacy. More precisely, students with responses to fewer than six test items per domain were not included in a first run of the population modelling (with regard to the regression model) in order to obtain unbiased Γ and Σ. In a second analysis step of population modelling, the regression parameters were treated as fixed to obtain plausible values for all cases, including those with fewer than six responses to test items.

For the science domain, students had to respond to at least six items in one of the subscales within a science dimension or subscale group (competency, system, knowledge) to be included in the latent regression model (note that a population modelling was done on the level of scientific subscale dimensions).

In PISA 2015 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. This scoring method is important with regard to the population model described (*Data yield and data quality*) since the population model is based on responses to the background questions and the test items.

Handling of item-by-country/language interactions

The population model was estimated separately for each country, with the exceptions of Belgium (Dutch, French), Canada (English, French), Israel (Hebrew, Arabic), and Qatar (Arabic, English) where the model was estimated separately by language. Item parameter files for test items, including common and unique item parameters, were obtained from the IRT scaling (described earlier in this chapter. Because the IRT scaling used a multiple-group (concurrent) calibration method, an item parameter file was created for each country. If there were larger language groups that allowed separate evaluation of item fit, these item parameter files were merged so that one file resulted for each country, except for Belgium, Canada, Israel, and Qatar, which received two separate item parameter files each (one for each main language); the language groups of those countries were introduced separately in the population modelling. By incorporating country-by-language group item parameter files into the analyses, the population modelling accounted for unique item parameters and thus for item-by-country and item-by language interactions.

The country-specific conditioning model assures that the latent regression is based only on data obtained within the same country version for background questionnaire and test (country-by-language where feasible). This ensures that the unique relationship between background variables and proficiency variables can be represented for each country without bias. The use of country-specific item parameter files that contain a large number of common/international item parameters ensures the comparability of the plausible values.

Population model for the domains

To generate plausible values for the domains of reading, mathematics, science, financial literacy and collaborative problem-solving, multidimensional population models were used. The multidimensional models included reading, mathematics and science, collaborative problem solving (computer-based assessment mode only) and financial literacy (if available).

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

- READ for reading
- MATH for mathematics
- SCIE for science
- CLPS for collaborative problem solving
- FLIT for financial literacy

Population model for the science subscales

There were several subscales reported for Science. These were knowledge scales (content; and procedural and epistemic), competency subscales (explain phenomena scientifically, evaluate and design scientific enquiry, and interpret data and evidence scientifically) and system subscales (physical; living; and earth and space).

To generate plausible values for the science subscales, multidimensional population models were used. In total, three different multidimensional population models were used within each country:

- model 1: reading, mathematics, collaborative problem solving (computer based assessment mode only) and the science knowledge subscales
- model 2: reading, mathematics, collaborative problem solving (computer-based assessment mode only) and the science competency subscales
- model 3: reading, mathematics, collaborative problem solving (computer-based assessment mode only) and the science system subscale.

The aim of generating plausible values for the different science subscales, is to represent a more nuanced picture of important aspects within the overall science framework. These subscales allow for investigations of different aspects within the science domain, thus, exploring further the variability of skills across participating countries. Table 9.25 gives

an overview of the distributions of 85 trend and 99 new items (184 in total) to the three scales knowledge, competency, and system as well as the eight subscales: content; procedural and epistemic; explain phenomena scientifically; evaluate and design scientific inquiry; interpret data and evidence scientifically; physical; living; earth and space. It should be noted that the three science subscales types are based on a three-way classification of the same 184 items (distributed into the 2+3+3=8 subscales) and thus cannot be compared among each other, since these contain the same items, classified in three different ways.

Table 9.25 Distribution of 85 trend and 99 new items to the science scales and subscales

Knowledge			Competency			System		
Subscales	Trend	New	Subscales	Trend	New	Subscales	Trend	New
Content	51	47	Explain phenomena scientifically	41	47	Physical	28	33
Procedural and epistemic (merged)	34 (24+10)	52 (36+16)	Evaluate and design scientific enquiry	16	23	Living	39	35
			Interpret data and evidence scientifically	28	29	Earth and space	18	31
Total no. of trend/ new items	85	99	Total no. of trend/new items	85	99	Total no. of trend/new items	85	99
Total no. of items	184		Total no. of items	184		Total no. of items	184	

Note: After the population modelling was finished and results reported to countries, the science experts recommended the reclassification of one item from the subscale "interpret data and evidence scientifically" to the subscale "explain phenomena scientifically" (see Chapter 2 for an updated item table). This change will be addressed in future PISA cycles but is not reflected in the PISA 2015 analyses.

The information about the eight subscales (2+3+3 subscales) was included in the population modelling. For example, the population model for scientific knowledge included the information about which items belonged to the two subscales "content" and "procedural and epistemic." Please note that for science, three additional population models (one for each of the three classifications of items) were computed in addition to science as a main scale. However, 10 plausible values were generated for each of the eight subscales.

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

- SKCO Science subscale – Content (knowledge)
- SKPE Science subscale – Procedural and epistemic (knowledge)
- SCEP Science subscale – Explain phenomena scientifically (competency)
- SCED Science subscale – Evaluate and design scientific inquiry (competency)
- SCID Science subscale – Interpret data and evidence scientifically (competency)
- SSPH Science subscale – Physical (system)
- SSLI Science subscale – Living (system)
- SSES Science subscale – Earth and science (system)

Generating plausible values

Plausible values are multiple imputations of proficiency values based on information from the test items and information provided by the students in the background context questionnaire (BQ). Plausible values are used to obtain more accurate estimates of group proficiency than would be obtained through an aggregation of point estimates. A more detailed description is given in *Latent regression model and population modelling* above as well as in Mislevy (1991), Thomas (2002), and von Davier, Sinharay, Oranje, and Beaton (2006).

In PISA, the computation of group-level reporting statistics – involving scores in each of the domains (reading, mathematics, science, financial literacy and collaborative problem solving) as well as science subscales – is based on 10 independently drawn plausible values for each of the test domains and subscales for each student. Each set of plausible values is equally well designed to estimate population parameters; however, multiple plausible values are required to represent the uncertainty in the domain measures appropriately (von Davier, Gonzalez, and Mislevy, 2009). The statistics based on scores are always computed at population or subpopulation levels. They should never be used to draw inferences at the individual level. Detailed information on the computation and the use of plausible values in analyses is given earlier in this chapter (in *Latent regression model and population modelling* and *Analysis of data with plausible values*).

For the population modelling and the calculation of plausible values for the scales of PISA, the computer programme DGROUP (Rogers et al., 2006) was used.

A normal multivariate distribution was assumed for $P(\theta_j|x_j, y_j, \Gamma, \Sigma)$, with a common variance, Σ, and with a mean given by a linear model with slope parameters, Γ, based on the principal components of several hundred selected main effects from the vector of context questionnaire variables.

The item parameters for the test items were obtained from the concurrent item calibration described earlier in this chapter (see *The IRT models for scaling, National and international item calibration and Handling of item-by-country/language and item-by-mode interactions*) using the data from past PISA cycles (2006-2012) and PISA 2015 as described above. The result of the concurrent calibration is a scale that provides comparable results across the different PISA cycles. To calculate the plausible values for PISA 2015 only, the item parameters for items administered in PISA 2015 were used in the population modelling.

The background variables included nearly all student questionnaire data, school ID, gender, and the number of not-reached items, among others. A description of the different sections of the background data can be found in Chapter 3 of this report. All variables in the context questionnaire were contrast coded before they were processed further in the population model. Contrast coding allows for the inclusion of codes for refused responses, avoiding the necessity of linear coding. The contrast coding scheme is reproduced in Annex B. The increased number of variables obtained through contrast coding is substantial. To capture most of the common variance in the contrast-coded background questions with a reduced set of variables, a principal component analysis was conducted. Because each population can have unique associations among the background variables, a single set of principal components was not sufficient for all countries included in PISA. As such, the extraction of principal components was carried out separately by country to take into account the differences in associations between the background variables and cognitive skills. In PISA, each set of principal components y^c (or conditioning variables) was selected to include 80 percent of the variance, or not to exceed a number of principal components greater than 5% of the raw sample size, with the aim of explaining as much variance as possible while at the same time avoiding over parameterization of the model.

Principal component scores based on nearly all (contrast coded) background variables were used in PISA, including international variables (collected by every participating country) as well as national background variables (country-specific variables in addition to the international variables).

Students with responses to fewer than six test items per domain were not included in a first run of the regression model in order to obtain unbiased Γ and Σ. In a second analysis step of population modelling, the regression parameters were treated as fixed to obtain plausible values for all cases, including those with fewer than six responses to items (see earlier section *Treatment of students with fewer than six cognitive item responses* for more information).

The financial literacy plausible values for the students who took this domain are based on a latent regression model that included the general background questionnaire variables plus the additional financial literacy background questions that were administered together with the financial literacy test items. A separate latent regression model based on the general background questionnaire variables alone was used for the remaining students who did not take the financial literacy test items as well as the financial literacy background questions.

Students received plausible values for each test domain administered in their country according to the test design that applied in a particular country (paper- versus computer-based assessment, financial literacy selected or not; collaborative problem solving selected or not; see Chapter 2 for more information on the test design). This means, students also received plausible values for test domains that were not administered to them. The same applies to students who took the Une Heure (UH) test design.

Note

1. A subset of cases from certain countries had to be excluded from the IRT calibration due to adjudication and data quality issues (please see Chapter 14 for more information).

2. Note that the random effect in Model 9.16 could be adjusted for each country separately, so this model picks up country differences as well. The similarity between the fit of models 9.16 and 9.15 shows, that no country-specific constraints are needed.

3. Note that a student was only considered a "respondent" and given an analysis weight if he or she responded to at least one test item and a certain amount of the context questionnaire items, or if he or she responded to at least half of the test items in cases of providing no context questionnaire information.

References

Adams, R. J., M. L. Wu, and **C. H. Carstensen** (2007), "Application of multivariate Rasch models in international large-scale educational assessments", in M. von Davier, and C. H. Carstensen (eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications*, pp. 271-280, Springer, New York, NY.

Adams, R. J., M. R. Wilson and **M. L. Wu** (1997), "Multilevel item response models: An approach to errors in variables regression", *Journal of Educational and Behavioural Statistics*, Vol. 22, pp. 46-75.

Allen, N. L., J. R. Donoghue and **T. L. Schoeps** (2001), *The NAEP 1998 Technical Report*, NCES 2001-509, Office of Educational Research and Improvement, National Center for Education Statistics, U.S. Department of Education, Washington, DC: National Center for Education Statistics.

Birnbaum, A. (1968), "Some latent trait models and their use in inferring a student's ability", in F. M. Lord and M. R. Novick (eds.), *Statistical theories of mental test scores,* Addison-Wesley, Reading, MA.

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.

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 York, NY.

Holzinger, K. and **F. Swineford** (1937), "The bi-factor method", *Psychometrika*, 2, 41-54.

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.

Kreiner, S. and **K. B.Christensen** (2014), "Analyses of model fit and robustness: A new look at the PISA scaling model underlying ranking of countries according to reading literacy", *Psychometrika*, Vol. 79/2, pp. 210-231.

Leys, C. et al. (2013), "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median", *Journal of Experimental Social Psychology*, Vol. 49, pp. 764-766.

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

Martin, M. O., K. D Gregory and **S. E. Stemler, (eds.)** (2000), *TIMSS 1999 Technical Report*, International Study Center, Boston, MA.

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

Mazzeo, J. and **M. von Davier** (2014), "Linking scales in international large-scale assessments", In L. Rutkowski, M. von Davier and D. utkowski (eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, CRC Press, Boca Raton, FL.

Mazzeo, J. and **M. von Davier** (2008), *Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results*, Doc. ref. EDU/PISA/GB(2008)28, Retrieved from *http://www.oecd.org/dataoecd/44/49/41731967.pdf*.

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

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*.

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.

Sireci, S. G., D. Thissen, and **H. Wainer** (1991), "On the reliability of testlet-based tests", *Journal of Educational Measurement*, Vol. 28, pp. 237-247.

Skrondal, A. and **S. Rabe-Hesketh** (2004), *Generalized Latent Variable Modelling: Multilevel, Longitudinal and Structural Equation Models*, Chapman and Hall/CRC, Boca Raton, FL.

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.

von Davier, M. (2016), "The Rasch Model: Chapter 3", in van der Linden, W. (ed.) *Handbook of Item Response Theory*, Vol. 1, *Second Edition*, CRC Press, pp. 31-48.

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 **S. Sinharay** (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. and **A. von Davier** (2007), "A unified approach to IRT scale linking and scale transformations", *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, Vol. 3/3, pp. 115-124.

von Davier, M. and **K. Yamamoto,** (2007), "Chapter 6: Mixture distribution Rasch models and Hybrid Rasch models", in: M. von Davier and C.H. Carstensen, *Multivariate and Mixture Distribution Rasch Models*, Springer, New York.

von Davier, M. et al. (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.

Wainer, H., E. T. Bradlow and **X. Wang** (2007), *Testlet response theory and its applications*. Cambridge University Press, New York, NY.

Weeks, J., K. Yamamoto and **M. von Davier** (2014), Design considerations for the Program for International Student Assessment, 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.

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.

Wise, S. L. and **C. E. DeMars** (2005), "Low student effort in low-stakes assessment: Problems and potential solutions", *Educational Assessment*, Vol. 10/1, pp. 1-17.

Wu, M. L., R. J. Adams and **M. R. Wilson,** (1997), ConQuest: Multi-Aspect Test Software [computer program]. Camberwell, Australia: Australian Council for Educational Research.

Yamamoto, K. (1998) "Scaling and scale linking", in T. S. Murray, I. S. Kirsch and L. B. Jenkins (eds.), *Adult Literacy in OECD Countries: Technical Report on the First International Adult Literacy Survey* (pp. 161-178), National Center for Education Statistics, Washington, DC.

Yamamoto, K. (1997), Scaling and scale linking. *International Adult Literacy Survey Technical Report*, Statistics Canada, Ottawa, Canada.

Yamamoto, K. and **J. Mazzeo** (1992), "Item response theory scale linking in NAEP", *Journal of Educational Statistics*, Vol. 17/2, pp. 155-174.