CHAPTER 13 CODING DESIGN, CODING PROCESS, AND CODER RELIABILITY STUDIES

INTRODUCTION

The proficiencies of PISA respondents were estimated based on their performance on the test items administered in the assessment. In the PISA 2015 assessment, countries[1] taking part in the computer-based assessment (CBA) administered 18 clusters of trend items from previous cycles – six clusters each of mathematics, reading and science – and six clusters of new science items developed for 2015. Countries that chose to take part in the Financial Literacy assessment administered two additional clusters of financial literacy items. The tests in countries that used paper-based assessment (PBA) was based solely on the 18 clusters of items from previous PISA cycles.

The PISA 2015 tests consisted of both selected- and constructed-response items. Selected-response items had predefined correct answers that could be computer-coded. While some of the constructed-response items were automatically coded by computer, some elicited a wider variety of responses that could not be categorized in advance, thus requiring human coding. The breakdown of all test items by domain, item format, and coding method is shown in Table 13.1.

Table 13.1: Number of cognitive items by domain, item format, and coding method

Mode	Coding Method	Item Format	*Mathematics (Trend)*	*Reading (Trend)*	*Science (Trend)*	*Science (New)*	*Financial Literacy (Trend and New)*
CBA	Human	Constructed-response	18 (17)	40 (44)	28	30	16
	Automatic	Simple selected-response	16 (19)	31 (27)	29	25	10
		Complex selected-response	13 (13)	11 (10)	25	41	12
		Constructed-response	22 (20)	6 (6)	3	3	5
	Total		69 (69)	88 (87)	85	99	43
PBA	Human	Constructed-response	41 (38)	50 (51)	32	NA	
	Automatic	Simple selected-response	15 (18)	30 (27)	29		
		Complex selected-response	12 (12)	8 (9)	24		
		Constructed-response	3 (3)	0 (0)	0		
	Total		71 (71)	88 (87)	85		

Note:
1. Consistent with previous cycles, easier and standard forms were developed for mathematics and literacy. Number in the cell corresponds to the standard forms while the number in parenthesis corresponds to the easier form.
2. New Science and Financial Literacy are CBA domains only.
3. The six parts of the trend Reading unit, Employment, R219, were separately coded to achieve consistent and accurate scoring. Note that, in the final item counts, four parts related to completing an employment application form were counted as a single item.

The multiple coding design in PISA 2015 included all human-coded items for monitoring coder reliabilities within country as well as across countries. This chapter aims to describe coding procedures and preparation, coding design options, and coding reliability studies.

CODING PROCEDURES

For CBA participants, the coding designs for the CBA responses for Mathematics, reading, science, and Financial Literacy (when applicable) were greatly simplified through use of the Open-Ended Coding System (OECS). This computer system, developed for PISA 2015, supported coders in their work to code the CBA responses while ensuring that the coding design was appropriately implemented. Detailed information about the system was included in the OECS Manual. The OECS system worked offline,

[1] PISA participants can be a country, a region, an economy, or a subsample within the former three types of entities. In this chapter, the generic terms "countries" or "participants" are used for the purpose of simplicity.

meaning coders did not need a network connection. It organized responses according to the agreed-upon coding designs.

During the CBA coding, coders worked only with individual PDF files, one for each item, containing one page per item response to be coded. Each page displayed the item stem or question, the individual response, and the available codes for the item. The coder was instructed to click the circle next to the selected code, which was then saved within the file. Also included on each page were two checkboxes labeled “recoded” and “defer.” The recoded box was to be checked if the response had been recoded by another coder for any reason. The defer box was used if the coder was not sure what code to assign to the response. These deferred responses were later reviewed and coded by the coder. It was expected that coders would code the majority of responses for which they were responsible and defer responses only in unusual circumstances. When deferring a response, it was suggested that the coder enter comments into the box labeled “comment” to indicate the reason for deferring the given response. Coders worked on one file until all responses in that file were coded. The process was repeated until all items were coded. The approach of coding by item has been shown to improve reliability and was greatly facilitated by the OECS.

For PBA participants, the coding designs for the PBA responses for Mathematics, reading, and science were supported by the Data Management Expert System and reliability was monitored through the Open-Ended Reporting System (OERS), a computer tool that worked in conjunction with the Data Management Expert (DME) software to evaluate and report reliability for paper-based open-constructed responses. Detailed information about the system was provided in the OERS Manual. The coding process for PBA participants involved using the actual paper booklets, with some booklets single coded and others multiple coded by two or more coders. When single coded, coders marked directly in the booklets. When multiple coded, coders coded first on the coding sheets, while the last coder coded directly in the booklet.

National centres used the output reports generated by the OECS and OERS to monitor irregularities and deviations in the coding process. Careful monitoring of coding reliability plays an important role in data quality control. Through coder reliability monitoring, coding inconsistencies or problems within and across countries could be detected early in the coding process through OECS/OERS output reports, allowing action to be taken as soon as possible. The OECS/OERS worked in concert with the DME database to generate two types of reliability reports: i) proportion agreement and ii) coding category distributions. National Project Managers (NPMs) were instructed to investigate whether a systematic pattern of irregularities existed and was attributable to a particular coder or item. In addition, they were instructed not to carry out resolution (e.g. changing coding on individual responses to reach higher coding consistency). Instead, if systematic irregularities were identified, all responses from a particular item or a particular coder needed to be recoded, including those that showed disagreement as well as those that showed agreement. In general, inconsistencies or problems were due to misunderstanding of general scoring guidelines and/or a rubric for a particular item or misuse of OECS/OERS. Coder reliability studies also made use of the OECS/OERS reports submitted by national centres.

CODING PREPARATION

Prior to the assessment, a number of key activities were completed by National Centres to prepare for the process of coding responses to the human-coded constructed-response items.

Recruitment of National Coder Teams

National Project Managers were responsible for assembling a team of coders. Their first task was to identify a lead coder who would be part of the coding team and additionally be responsible for the following tasks:

- training coders within the country;
- organizing all materials and distributing them to coders;
- monitoring the coding process;
- monitoring the inter-rater reliability and taking action when the coding results were unacceptable and required further investigation;
- retraining or replacing coders if necessary;
- consulting with the international experts if item-specific issues arose; and
- producing reliability reports.

The lead coder was required to be proficient in English (as international training and interactions with the contractors were in English only) and to attend the international coder trainings in Malta in January 2014 and Portugal in January 2015. It was also assumed that the lead coder for the Field Trial would retain the role for the Main Survey. When this was not the case, it was the responsibility of the National Centre to ensure that the new lead coder received training equivalent to that provided at the international coder training prior to the Field Trial.

The guidelines for assembling the rest of the coding team included the following requirements:

- all coders should have more than a secondary qualification (i.e., high school degree); university graduates were preferable;
- all should have a good understanding of secondary level studies in the relevant domains;
- all should be available for the duration of the coding period, which was expected to last two to three weeks;
- due to normal attrition rates and unforeseen absences, it was strongly recommended that lead coders train a backup coder for their teams; and
- two coders for each domain MUST be bilingual in English and the language of the assessment.

International Coder Training

Detailed coding guides were developed for all the new science items that included coding rubrics as well as examples of correct and incorrect responses. For trend items, coding information from previous cycles was included in the coding guides. For new items, coding rubrics were defined for the Field Trial, and then information from Field Trial coding was used to revise the coding guides for the Main Survey.

Prior to the Field Trial, NPMs and lead coders were provided with a full item-by-item coder training in Malta in January 2014. The Field Trial training covered all the items across all domains. Prior to the Main Survey, NPMs and lead coders were provided with a new round of full item-by-item coder training in Portugal in January 2015. The Main Survey training covered all new items as well as a set of trend

science and trend reading items that required additional training based on the Field Trial experience. During these trainings, the coding guides were presented and explained. Training participants practiced coding on sample items and discussed any ambiguous or problematic situations as a group. By focusing on sample responses most challenging to code, training participants had the opportunity to ask questions and get the coding rubrics clarified as much as possible. When the discussion revealed areas where rubrics could be improved, those changes were made and included in an updated version of the coding guide documents available after the meeting. As in previous cycles, a "workshop" version of the coding guides was also prepared for the national training. This version included a more extensive set of sample responses, the official coding for each response, and a rationale for why each response was coded as shown.

To support the national teams during their coding process, a coder query service was offered. This allowed national teams to submit coding questions and receive responses from the relevant domain experts. National teams were also able to review questions submitted by other countries along with the responses from the test developers. In the case of trend items, responses to queries from previous cycles were also provided. A summary report of coding issues was provided on a regular basis and all related materials were archived in the PISA 2015 Portal for reference by national coding teams.

National Coder Training Provided by the National Centres

Each National Centre was required to develop a training package for their own coders. The training package consisted of an overview of the survey and their own training manuals based on the manuals and materials provided by the international PISA contractors. Coding teams were asked to work on the same schedule and at the same location in order to facilitate discussion about any items that proved challenging. Past experience has shown that if coders can discuss items among themselves and with their lead coder, many issues can be resolved in a way that results in more consistent coding. Each coder was assigned a unique coder ID that was specific to each domain and design.

The National Centres were responsible for organizing training and coding using one of the following two approaches and checking with contractors in the case of deviations:

a) Coder training took place at the "item" level. Under this approach, coders were fully trained on coding rules for each item and proceeded with coding all responses for that item. Once that item was done, training was provided for the next item, and so on.

b) Coder Training took place at the "item set" level. While coding was conducted at the "item" level, the coder training took place at the "item set" level, with each "item set" containing a few units of items. In this alternative approach, coders were fully trained on a set that varied from 13 to 18 items. Once the full training was complete, coding took place at the item level. However, to ensure that the coding rules were still fresh in the coders' memory, a coding refresher was recommended before the coding of each item.

CODING DESIGN[2]

In order to meet the unique characteristics of the CBA participants during the Main Survey while ensuring that the coding process was completed within a two-to-three week period, 10 possible coding

[2] For a better understanding of the PISA coding designs, it is recommended that the descriptions of the PISA assessment designs in Chapter 2 be read first as important background information.

designs (1 standard design and 9 variations) were offered to the CBA participants and four possible coding designs (1 standard design and 3 variations) were offered to the PBA participants. Those designs were developed to accommodate participants' various needs in terms of the number of languages assessed, the sample size, and the specified number of coders required in each domain.

The number of coders by domain in each CBA coding design is shown in Table 13.2. The design of multiple coding in the CBA standard coding design is shown in Table 13.3. In CBA coding designs, human-coded items were bundled into one item set or multiple item sets in each domain. For each common item, coders coded a set of 100 student responses that were randomly selected from all the student responses. Each domain had two bilingual coders who needed additionally to code 10 anchor responses for each item assigned to both of them. The rest of the student responses to each item were evenly split among coders to be single coded. The difference in multiple coding between the standard coding design and other CBA coding designs mainly lay in the number of coders in each domain and which item sets were assigned to each coder.

Table 13.2: Number of CBA coders by domain and coding design

Design Label	Sample Size Requirements	*Mathematics (Trend)*	*Reading (Trend)*	*Science (Trend and New)*	*Financial Literacy (Trend and New)*
Standard design	Countries with the standard sample size (4,000 – 7,000) for a given language	4	6	8	4
Alternative design 1	Countries with a sample between 7,000 and 9,000 for a given language	4	9	12	4
Alternative design 1a	Countries with a sample between 7,000 and 9,000 for a given language	16	9	12	16
Alternative design 2	Countries with a sample between 9,000 and 13,000 for a given language	6	9	16	6
Alternative design 2a	Countries with a sample between 9,000 and 13,000 for a given language	6	12	16	6
Alternative design 3	Countries with a sample between 13,000 and 19,000 for a given language	6	12	20	6
Alternative design 3a	Countries with a sample between 13,000 and 19,000 for a given language	12	27	32	12
Alternative design 4	Countries with a sample larger than 19,000 for the majority language	9	21	36	6
Minority Language Design 1	Countries with a sample less than 1,500 for the minority language	2	2	2	2
Minority Language Design 2	Countries with a sample between 1,500 and 4,000 for the minority language	3	3	4	3

Table 13.3: Multiple coding in CBA standard coding design

		Coder IDs							
Mathematics (Trend)	**Number of Responses for Multiple Coding**	**301 (Bilingual)**	**302**	**303 (Bilingual)**	**304**				
Item Set 1	100 student responses per item	✓	✓	✓	✓				
Item Set 1	10 anchor responses per item	♦		♦					
Reading (Trend)	**Number of Responses for Multiple Coding**	**201 (Bilingual)**	**202**	**203 (Bilingual)**	**204**	**205**	**206**		
Item Set 1	100 student responses per item	✓	✓			✓	✓		
Item Set 2	100 student responses per item	✓	✓	✓	✓				
Item Set 3	100 student responses per item			✓	✓	✓	✓		
Item Set 1	10 anchor responses per item	♦							
Item Set 2	10 anchor responses per item	♦		♦					
Item Set 3	10 anchor responses per item			♦					
Science (Trend and New)	**Number of Responses for Multiple Coding**	**101 (Bilingual)**	**102**	**103 (Bilingual)**	**104**	**105**	**106**	**107**	**108**
Item Set 1	100 student responses per item	✓	✓					✓	✓
Item Set 2	100 student responses per item	✓	✓			✓	✓		
Item Set 3	100 student responses per item			✓	✓	✓	✓		
Item Set 4	100 student responses per item			✓	✓			✓	✓
Item Set 1	10 anchor responses per item	♦							
Item Set 2	10 anchor responses per item	♦							
Item Set 3	10 anchor responses per item			♦					
Item Set 4	10 anchor responses per item			♦					
Financial Literacy (Trend and New)	**Number of Responses for Multiple Coding**	**401 (Bilingual)**	**402**	**403 (Bilingual)**	**404**				
Item Set 1	100 student responses per item	✓	✓	✓	✓				
Item Set 1	10 anchor responses per item	♦		♦					

Note: "✓" denotes the coder should code 100 student responses for each item in the item set. "♦" denotes the coder should code 10 anchor responses for each item in the item set.

Four variations of coding design were offered to PBA participants (See Table 13.4).The design of multiple coding in the PBA standard coding design is shown in Table 13.5. For PBA participants, all paper-and-pencil booklets were organized by form type into 27 different bundle sets: 9 bundle sets per domain. Bundle sets 1, 2, and 3 in each domain were composed of forms for multiple coding: Forms 13, 15, and 17 for mathematics; Forms 1, 3, and 5 for reading; and Forms 7, 8, and 9 for science. For each form, 100 student booklets were randomly selected from all the student responses. Each coder coded his or her assigned clusters on the sets of 100 student booklets until all items in the booklets were coded. Bundle sets 4-9 in each domain were composed of 6 or 7 types 3 of anchor forms. The forms were labelled 301-307 for mathematics; 201-207 for reading; and 101-106 for science (See Table 13.5). Differing from non-anchor forms, the anchor forms each contained only one cluster of items. For example, Form 301 contained all the items from the first cluster in Maths, and Form 202 contained all the items from the second cluster in reading. Each anchor form had 10 pre-filled English booklets that were coded by the bilingual coders from each domain. Each domain in the PBA standard design had two bilingual coders: 31 and 33 for mathematics, 21 and 23 for reading, and 11 and 13 for science.

CBA constructed-response items were organized by item set during multiple coding; by contrast, PBA constructed-response items were organized by bundle set during multiple coding. In other words, multiple coding in the PBA standard design was form- rather than item-set-based. Although coders conducted coding on the booklets, each coder only coded the clusters assigned to him or her for each booklet, leaving the rest of the clusters to other coders. This multiple coding design enabled the within-

[3] In mathematics, there was an additional cluster, as instead of M06 there was M06A and M06B. Since countries could only choose M06A or M06B, but not both, the actual number of clusters in each domain is six rather than seven. The same is true for clusters R06A and R06B in reading.

and across-country comparison. After the multiple coding was completed, all the clusters that remained uncoded were equally split among coders and coded only once. The difference in multiple coding between the PBA standard design and other PBA coding designs mainly lay in the number of coders in each domain, and which forms were assigned to each coder.

Table 13.4: Number of PBA coders by domain and coding design

Design Label	Sample Size Requirements	*Mathematics (Trend)*	*Reading (Trend)*	*Science (Trend and New)*
Standard design	Countries with the standard sample size (3,501 – 5,500)	4	6	6
Alternative design 1	Countries with a sample larger than 5,500 for the majority language	6	9	9
Minority language design 1	Countries a sample less than 1,500 for the minority language	2	2	2
Minority design 2	Countries with a sample between 1,501 and 3,500 for the minority language	3	3	4

Table 13.5: Multiple coding in PBA standard coding design

			Coder IDs					
Mathematics (Trend)	**Forms (Clusters)**	**Number of Booklets per Form**	**31 (Bilingual)**	**32**	**33 (Bilingual)**	**34**		
Bundle set 1	Form 13 (PM1&PM2)	100 student booklets	✓	✓	✓	✓		
Bundle set 2	Form 15 (PM3&PM4)	100 student booklets	✓	✓	✓	✓		
Bundle set 3	Form 17 (PM5&PM6a or PM5&PM6b)	100 student booklets	✓	✓	✓	✓		
Bundle set 4	Form 301 (PM1)	10 anchor booklets	♦		♦			
Bundle set 5	Form 302 (PM2)	10 anchor booklets	♦		♦			
Bundle set 6	Form 303 (PM3)	10 anchor booklets	♦		♦			
Bundle set 7	Form 304 (PM4)	10 anchor booklets	♦		♦			
Bundle set 8	Form 305 (PM5)	10 anchor booklets	♦		♦			
Bundle set 9	Form 306 (PM6a) or 307 (PM6b)	10 anchor booklets	♦		♦			
Reading (Trend)	**Forms (Clusters)**	**Number of Booklets per Form**	**21 (Bilingual)**	**22**	**23 (Bilingual)**	**24**	**25**	**26**
Bundle set 1	Form 1 (PR1&PR2)	100 student booklets	✓	✓			✓	✓
Bundle set 2	Form 3 (PR3&PR4)	100 student booklets	✓	✓	✓	✓		
Bundle set 3	Form 5 (PR5&PR6a or PR5&PR6b)	100 student booklets			✓	✓	✓	✓
Bundle set 4	Form 201 (PR1)	10 anchor booklets	♦					
Bundle set 5	Form 202 (PR2)	10 anchor booklets	♦					
Bundle set 6	Form 203 (PR3)	10 anchor booklets	♦		♦			
Bundle set 7	Form 204 (PR4)	10 anchor booklets	♦		♦			
Bundle set 8	Form 205 (PR5)	10 anchor booklets			♦			
Bundle set 9	Form 206 (PR6a) or 207 (PR6b)	10 anchor booklets			♦			
Science (Trend)	**Forms (Clusters)**	**Number of Booklets per Form**	**11 (Bilingual)**	**12**	**13 (Bilingual)**	**14**	**15**	**16**
Bundle set 1	Form 7 (PS1&PS2)	100 student booklets	✓	✓			✓	✓
Bundle set 2	Form 8 (PS3&PS4)	100 student booklets	✓	✓	✓	✓		
Bundle set 3	Form 9 (PS5&PS6)	100 student booklets			✓	✓	✓	✓
Bundle set 4	Form 101 (PS1)	10 anchor booklets	♦					
Bundle set 5	Form 102 (PS2)	10 anchor booklets	♦					
Bundle set 6	Form 103 (PS3)	10 anchor booklets	♦		♦			
Bundle set 7	Form 104 (PS4)	10 anchor booklets	♦		♦			
Bundle set 8	Form 105 (PS5)	10 anchor booklets			♦			
Bundle set 9	Form 106 (PS6)	10 anchor booklets			♦			

Note:
1. "✓" denotes the coder should code 100 student booklets for the specific form as a bundle set. "♦" denotes the coder should code 10 anchor booklets for the specific form as a bundle set.
2. Paper-based Mathematics, Reading and Science assessments are referred as PM, PR and PS in this table. The number following PM, PR, and PS is the Cluster number. For instance, PM1 represents Cluster 1 in Mathematics domain.
3. Mathematics and Reading domains have two versions of item cluster 06: 06A and 06B. Each PISA participant selected one or the other version to administer.
4. CBA participants' coder ID is three-digit; while PBA participants' coder ID is two-digit.

Within-Country and Across-Country Coder Reliability

Reliable human coding is critical for ensuring the validity of assessment results within a country, as well as the comparability of assessment results across countries. Coder reliability in PISA 2015 was evaluated and reported at both within- and across-country levels. The evaluation of coder reliability was made possible by the design of multiple coding - a portion or all of the responses from each human-coded constructed-response item were coded by at least two human coders.

The purpose of evaluating the **within-country coder reliability** was to ensure coding reliability within a country and identify any coding inconsistencies or problems in the scoring process so they could be addressed and resolved earlier in the process. The evaluation of within-country coder reliability was carried out by the multiple coding of a set of student responses—assigning identical student responses to different coders so those responses were coded multiple times within a country. To multiple code all student responses in an international large-scale assessment like PISA is not economical, so a coding design combining multiple coding and single coding was utilized to reduce national costs and coder burden. In general, a set of 100 responses per human-coded item was randomly selected from actual student responses to be multiple coded. The rest of the student responses needed to be evenly split among coders to be single coded.

Accurate and consistent scoring within a country does not necessarily mean that coders from all countries are applying the coding rubrics in the same manner. Coding bias may be introduced if one country codes a certain response differently than other countries. Therefore, in addition to within-country coder reliability, it was also important to check the consistency of coders across countries. The evaluation of **across-country coder reliability** was made possible by the multiple coding of a set of anchor responses. In each country, two coders in each domain had to be bilingual in English and the language of assessment. These coders were responsible for coding the set of anchor responses in addition to any student responses assigned to them. For each constructed-response item, a set of 10 anchor responses in English was provided. These anchor responses were answers obtained from real students and their authoritative coding were not released to the countries. Since countries using the same mode of administration coded the same anchor responses for each human-coded constructed-response item, their coding results on the anchor responses could be compared to each other.

CODER RELIABILITY STUDIES

Coder reliability studies were conducted to evaluate consistency of coding of human-coded constructed-response items within and across the countries participating in PISA 2015. The studies were based on 59 CBA countries (for a total of 72 country-by-language groups) and 15 PBA countries (for a total of 17 country-by-language groups) with sufficient data to yield reliable results.[4] The coder reliability studies were conducted for three aspects of coder reliability:

- the domain-level proportion agreement;
- the item-level proportion agreement; and
- the coding category distributions of coders on the same item.

Proportion agreement and coding category distribution are the main indicators of coder reliability used in PISA 2015.

[4] Coding data from Kazakhstan (Kazakh) and Kazakhstan (Russian) were not included in this analysis and all human-coded responses were excluded from the calculation of proficiency estimates.

- *Proportion agreement* refers to the percentage of each coder's coding that matched the other coders' coding on the identical set of multiple-coded responses for an item. It can vary from 0 (0% agreement) to 1 (100% agreement). Each country was expected to have an average within-country proportion agreement of at least 0.92 (92% agreement) across all items, with a minimum 85% agreement for any one item.

- *Coding category distribution* refers to the aggregation of the distributions of coding categories (such as "full credit", "partial credit", and "no credit") assigned by a coder to two sets of responses: a unique set of 100 responses for multiple coding, and responses randomly allocated to the coder for single coding. Notwithstanding that negligible differences of coding categories among coders were tolerated, the coding category distributions between coders were expected to be statistically equivalent based on the standard chi-square distribution due to the random assignment of the single-coded responses.

Domain-Level Proportion Agreement

The average within-country agreement by domain in PISA 2015 exceeded 92% in each domain across the 89 country-by-language groups with sufficient data (see Table 13.6 and 13.7). The difference between CBA and PBA participants' average proportion agreements in each of the Mathematics, reading, and trend science domain was less than 0.5%. Within each mode, the within-country agreements between domains was not significantly different, either. The mathematics domain had higher agreement (97.5% for CBA; 97.5% for PBA) than the other domains. The reading domain also had agreement higher than 95% (95.6% for CBA; 95.8% for PBA). The trend science domain had an average agreement of 94.2% for CBA and 94.7% for PBA. The new science domain for CBA also had an average agreement of 94.2%. The Financial Literacy domain had slightly lower agreement (93.7% for CBA) than the other domains.

Across-country agreement by domain in PISA 2015 exceeded 92% when averaged over all the 72 CBA country-by-language groups (see Table 13.6). The PBA participants had lower across-country agreement than the CBA participants on average (see Table 13.6 and 13.7). The difference in domain-level proportion agreement between CBA and PBA is 3.3% for mathematics, 3.9% for reading, and 5.0% for trend science. Domain-level agreement was the highest in the mathematics domain for both CBA and PBA responses (97.2% for CBA; 94.0% for PBA). For the CBA participants, the reading, trend Science, new Science, and Financial Literacy domain had across-country agreement at similar levels, ranging between 93.1% and 93.9%. For the PBA participants, the average across-country agreements of the reading and trend Science domains were 90.0% and 88.6%, respectively, slightly lower than the criterion but still acceptable.

Table 13.6: Summary of within-country and across-country agreement (%) per domain for CBA participants

Computer-Based Participants		Within-Country Agreement					Across-Country Agreement				
		Mathematics (Trend)	*Reading (Trend)*	*Science (Trend)*	*Science (New)*	*Financial Literacy (Trend and New)*	*Mathematics (Trend)*	*Reading (Trend)*	*Science (Trend)*	*Science (New)*	*Financial Literacy (Trend and New)*
OECD Members	**Languages**										
Australia	English	97.8	92.6	91.7	90.0	93.5	99.4	95.2	99.2	93.3	94.7
Austria	German	97.1	96.3	94.0	95.1		98.1	93.7	96.1	98.0	
Belgium (Flemish)	Dutch	97.5	96.3	93.4	93.5	93.4	97.8	95.9	93.2	94.0	91.6
Belgium (French)	French	98.5	96.8	96.7	96.9		98.9	97.9	97.9	97.0	
Canada	English	96.6	93.6	88.7	89.2	92.2	97.2	95.7	91.8	93.3	95.3
Canada	French	96.9	94.0	89.1	90.3	89.6	96.4	95.6	92.5	93.0	91.6
Chile	Spanish	96.5	94.1	95.1	94.7	92.6	97.9	92.0	94.3	95.7	92.8
Czech Republic	Czech	97.9	96.3	94.2	93.8		98.9	95.1	95.0	95.0	
Denmark	Danish	98.8	97.5	96.6	97.3		98.9	94.1	96.1	93.0	
Estonia	Estonian	95.7	95.4	94.1	93.4		97.8	94.2	94.3	94.3	
Estonia	Russian	95.8	94.1	93.0	92.2		97.8	93.8	94.6	95.2	
Finland	Finnish	99.4	98.2	94.9	94.9		98.6	95.2	95.9	96.0	
France	French	97.8	98.6	95.5	95.0		98.6	94.7	93.9	94.7	
Germany	German	96.5	94.8	93.4	92.3		96.9	94.4	92.9	95.7	
Greece	Greek	96.1	96.2	91.7	92.3		96.9	96.0	93.6	92.3	
Hungary	Hungarian	98.5	94.6	95.6	95.1		96.9	95.2	96.1	96.3	
Iceland	Icelandic	97.7	95.9	95.3	95.0		97.8	96.4	96.1	94.3	
Ireland	English	97.3	94.2	93.7	92.8		97.8	94.5	93.9	94.0	
Israel	Arabic	96.8	96.6	93.8	94.3		90.7	95.5	93.4	89.9	
Israel	Hebrew	96.3	95.3	94.3	94.3		98.3	91.3	95.2	93.2	
Italy	Italian	98.8	94.0	93.2	93.3	93.6	98.5	93.8	92.3	93.9	93.7
Japan	Japanese	97.6	97.4	94.9	96.0		98.1	91.7	92.9	93.0	
Korea	Korean	98.5	97.7	97.0	96.4		98.6	93.3	94.3	90.3	
Latvia	Latvian	95.7	92.5	92.3	94.0		95.3	93.6	93.6	90.7	
Latvia	Russian	96.3	93.4	91.5	92.1		96.1	93.7	93.0	90.7	
Luxembourg	German	97.6	97.1	96.6	97.4		97.8	96.4	96.1	95.3	
Luxembourg	French	98.1	97.3	97.2	97.1		98.3	97.7	96.3	96.2	
Mexico	Spanish	96.7	94.1	92.0	90.5		94.4	93.1	94.3	92.3	
Netherlands	Dutch	99.0	98.7	94.2	95.8	92.2	98.3	96.4	95.4	96.0	94.7
New Zealand	English	97.9	94.4	94.2	93.8		98.3	94.3	95.4	95.8	
Norway	Bokmål	98.0	95.7	96.0	96.4		97.8	95.6	96.1	95.3	
Poland	Polish	98.6	97.3	95.6	94.2	94.5	98.1	94.7	95.0	95.0	94.1
Portugal	Portuguese	97.9	97.5	95.7	95.6		99.4	96.2	95.0	95.0	
Slovak Republic	Slovak	97.5	97.7	95.3	95.4	95.3	98.6	96.6	92.9	92.3	94.7
Slovenia	Slovenian	96.4	96.2	94.5	94.1		98.1	96.0	93.6	95.7	
Spain	Catalan	96.0	95.8	93.6	94.0	96.3	97.2	89.2	93.2	91.0	95.6
Spain	Spanish	96.1	94.1	94.1	94.2	94.0	96.4	88.3	93.2	91.3	90.9
Spain	Basque	93.6	95.2	95.5	92.4	93.3	93.3	90.2	90.7	94.5	92.5
Spain	Galician	97.3	96.6	92.6	94.6		98.1	90.9	93.0	92.3	93.1
Sweden	Swedish	97.6	95.1	94.2	94.2		97.5	95.8	95.9	96.2	
Switzerland	German	97.6	98.0	95.3	95.9		98.1	94.9	94.5	93.2	
Switzerland	French	94.9	95.1	92.7	90.6		98.1	95.6	95.4	92.2	
Switzerland	Italian	96.8	95.9	95.3	94.9		96.9	95.0	96.4	93.7	
Turkey	Turkish	97.7	93.8	94.7	94.1		93.9	89.2	94.6	92.7	
United Kingdom excluding Scotland	English	98.1	95.7	92.7	92.5		98.1	95.6	93.9	94.7	
United Kingdom - Scotland	English	98.1	96.7	94.9	94.8		97.5	96.5	95.4	93.7	
United States excluding Puerto Rico	English	97.3	94.0	91.1	89.4	92.7	99.1	96.3	93.4	93.3	95.6
Mean - OECD Members		**97.3**	**95.7**	**94.1**	**94.0**	**93.3**	**97.5**	**94.4**	**94.5**	**93.9**	**93.6**
Median - OECD Members		**97.5**	**95.8**	**94.2**	**94.2**	**93.4**	**97.9**	**95.0**	**94.3**	**94.0**	**93.9**
OECD Partners	**Languages**										
Brazil	Portuguese	97.2	93.9	92.1	92.4	93.0	86.2	90.7	85.7	79.7	86.3
Bulgaria	Bulgarian	93.2	87.1	90.7	90.8		95.0	82.9	92.1	91.3	
China (B-S-J-G)	Chinese	97.4	96.8	93.1	93.9	94.4	96.9	95.8	93.6	93.7	90.6
Chinese Taipei	Chinese	97.3	96.3	96.2	95.8		99.4	95.1	95.0	96.7	
Colombia	Spanish	99.9	98.8	99.5	98.8		98.2	93.5	88.2	88.7	
Costa Rica	Spanish	97.6	95.3	93.3	93.3		95.6	94.1	82.9	82.0	
Croatia	Croatian	98.5	95.7	96.2	97.1		98.9	95.1	94.6	94.3	
Cyprus [2,3]	Greek	98.5	96.4	93.4	93.8		98.8	95.1	95.0	97.0	
Dominican Republic	Spanish	97.0	95.9	95.9	96.4		92.4	81.3	96.8	95.0	
Hong Kong	Chinese	98.0	95.7	95.6	94.4		98.8	94.8	95.7	95.3	
Lithuania	Lithuanian	98.0	96.7	96.5	96.5	95.7	98.6	95.1	95.0	96.3	94.4
Macao	Chinese	99.3	96.2	94.6	94.0		99.2	94.7	93.2	93.0	
Malaysia	English	97.9	95.3	95.6	95.5		98.6	92.8	90.5	91.2	
Malaysia	Malay	98.4	95.6	94.7	97.1		98.5	95.7	87.4	91.4	
Montenegro	Serb (Yekavian)	98.9	96.7	94.2	94.8		97.5	93.7	85.6	85.4	
Peru	Spanish	99.2	96.9	96.2	96.7	96.6	97.6	93.6	93.2	95.0	95.3
Qatar	Arabic	98.9	94.7	93.5	93.9		95.3	92.3	93.1	88.7	
Qatar	English	97.4	94.8	92.7	93.4		97.2	94.4	88.9	91.0	
Russian Federation	Russian	98.1	95.8	92.7	92.9	94.5	97.5	97.2	92.9	95.7	94.4
Singapore	English	98.2	95.5	95.5	94.8		96.9	95.9	95.0	94.7	
Thailand	Thai	98.3	97.3	95.5	96.5		98.9	95.3	95.7	95.7	
Tunisia	Arabic	99.5	97.0	95.5	95.2		95.3	90.0	87.9	86.7	
United Arab Emirates	Arabic	97.8	94.1	90.5	91.7		94.1	88.9	92.1	88.0	
United Arab Emirates	English	96.4	93.5	92.7	92.5		96.2	94.6	92.9	92.0	
Uruguay	Spanish	97.5	93.8	95.3	94.1		97.1	92.3	92.1	93.3	
Mean - OECD Partners		**97.9**	**95.4**	**94.5**	**94.6**	**94.8**	**96.8**	**93.0**	**91.8**	**91.7**	**92.2**
Median - OECD Partners		**98.0**	**95.7**	**94.7**	**94.4**	**94.5**	**97.5**	**94.4**	**92.9**	**93.0**	**94.4**
Mean - CBA Participants		**97.5**	**95.6**	**94.2**	**94.2**	**93.7**	**97.2**	**93.9**	**93.6**	**93.1**	**93.3**
Median - CBA Participants		**97.6**	**95.8**	**94.3**	**94.2**	**93.6**	**97.8**	**94.7**	**93.9**	**93.7**	**94.1**

1. PISA participants can be a country, a region, an economy, or a subsample within the former three types of entities.
2. Note by Turkey: The information in this table 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 recognizes 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".
3. Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognized 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.

Table 13.7: Summary of within-country and across-country agreement (%) per domain for

PBA participants

Paper-Based Participants		Within-Country Agreement			Across-Country Agreement		
		Mathematics (Trend)	*Reading (Trend)*	*Science (Trend)*	*Mathematics (Trend)*	*Reading (Trend)*	*Science (Trend)*
OECD Members	**Languages**						
United States - Puerto Rico	Spanish	98.0	94.8	95.5	95.8	94.4	95.6
OECD Partners	**Languages**						
Albania	Albanian	97.5	95.9	96.4	91.7	87.6	86.3
Algeria	Arabic	85.8	81.9	78.3	80.9	85.6	84.7
Argentina	Spanish	99.5	98.5	95.0	96.8	93.5	95.0
Georgia	Georgian	95.3	95.6	97.8	90.7	90.7	88.1
Indonesia	Indonesian	96.9	96.6	95.5	93.8	92.5	90.2
Jordan	Arabic	99.6	99.6	98.7	95.3	84.3	90.2
Kosovo	Albanian	98.2	92.5	87.8	97.0	89.9	89.1
Lebanon	English	99.3	97.6	98.8	96.5	86.8	93.8
Lebanon[1]	French	99.5	99.2	98.2	NA	NA	NA
Macedonia	Macedonian	97.8	98.8	98.9	95.9	91.7	74.5
Macedonia	Albanian	98.1	99.1	99.2	95.9	89.7	79.2
Malta	English	97.7	94.6	92.3	98.0	94.4	95.0
Moldova	Romanian	99.2	99.4	98.1	97.2	90.5	95.0
Romania	Romanian	99.4	97.4	98.2	85.2	87.6	85.6
Trinidad and Tobago	English	96.2	90.2	87.6	96.1	91.9	89.8
Vietnam	Vietnamese	99.3	97.0	94.1	96.4	89.6	85.5
Mean - OECD Partners		**97.5**	**95.9**	**94.7**	**93.8**	**89.7**	**88.1**
Median - OECD Partners		**98.2**	**97.2**	**97.1**	**95.9**	**89.9**	**89.1**
Mean - PBA Participants		**97.5**	**95.8**	**94.7**	**94.0**	**90.0**	**88.6**
Median - PBA Participants		**98.1**	**97.0**	**96.4**	**95.9**	**90.2**	**89.5**

Note:
1. Lebanon did not produce coded anchor responses in French.
2. New Science and Financial Literacy are CBA domains only in the Main Survey.
3. PISA participants can be a country, a region, an economy, or a subsample within the former three types of entities.
4. Note by Turkey: The information in this table 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 recognizes 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".
5. Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognized 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.

Item-Level Proportion Agreement

In terms of student responses, all CBA participants had only five or fewer items with proportion agreement lower than 85% in mathematics, new Science, and Financial Literacy (see Table 13.8). 96% of them had proportion agreement higher than 85% for every item in those three domains. More than 97% of CBA participants had five or fewer items with proportion agreement lower than 85% in the reading and trend Science domains. In terms of student responses, 94% of PBA participants had only five or fewer items with proportion agreement lower than 85% in mathematics; 83% did in reading and trend Science.

Table 13.8: Percentages of CBA and PBA participants with different number of items for which proportion agreement is lower than 85%

Mode	Number of Participants	N of Items with Proportion Agreements Lower than 85%	*Mathematics (Trend)*	*Reading (Trend)*	*Science (Trend)*	*Science (New)*	*Financial Literacy (Trend and New)*
CBA	72	N = 0	96%	83%	85%	86%	84%
		1 ≤ N ≤ 5	4%	14%	13%	14%	16%
		6 ≤ N ≤ 10	0%	1%	3%	0%	0%
		N > 10	0%	1%	0%	0%	0%
PBA	17	N = 0	76%	59%	59%	NA	
		1 ≤ N ≤ 5	18%	24%	24%		
		6 ≤ N ≤ 10	0%	6%	12%		
		N > 10	6%	12%	6%		

Note:
1. "Item" in the table refers to "human-coded constructed-response item".
2. PISA participants can be a country, a region, an economy, or a subsample within the former three types of entities.
3. Only 19 out of the 72 CBA participants administered the Financial Literacy domain.
4. New Science and Financial Literacy are CBA domains only in the Main Survey.
5. The summary in the table is based on student responses rather than anchor responses.

As shown in Table 13.9, not a single item had an international mean lower than 85% over the student responses in both CBA and PBA participants. The overall proportion agreement averaged over each item's international mean was 95% for CBA participants and 96% for PBA participants. Only three items had an international mean lower than 85% over the CBA anchor responses, while the international means of eight items were lower than 85% over the PBA anchor responses. The overall proportion agreement averaged over each CBA item's international mean and each PBA item's international mean was 94% and 91%, respectively.

Table 13.9: Summary of proportion agreement across the PISA participants

	Source of Response	CBA Participants						PBA Participants			
		Mathematics (Trend)	*Reading (Trend)*	*Science (Trend)*	*Science (New)*	*Financial Literacy (Trend and New)*	***Average***	*Mathematics (Trend)*	*Reading (Trend)*	*Science (Trend)*	***Average***
Number of items with average proportion agreement lower than 85% averaged across participants	Student Responses	0	0	0	0	0	**0**	0	0	0	**0**
	Anchor Responses	1	6	4	2	2	**3**	3	12	9	**8**
Overall proportion agreement averaged over items' international means	Student Responses	97%	95%	94%	94%	94%	**95%**	97%	96%	95%	**96%**
	Anchor Responses	97%	94%	94%	93%	93%	**94%**	94%	91%	89%	**91%**

Note:
1. "Item" in the table refers to "human-coded constructed-response item".
2. PISA participants can be a country, a region, an economy, or a subsample within the former three types of entities.

Coding Category Distributions

In mathematics, 10% of coders in an average CBA country and 27% of coders in an average PBA country had significantly different coding category distributions from other coders on more than 20% of items (see Table 13.10). In reading, it was 17% for CBA and 52% for PBA, while in trend Science, it was 20% for CBA and 66% for PBA. In new Science, 35% of coders in an average CBA country had significantly different coding category distributions from other coders on more than 20% of items. In Financial Literacy, the average was 44%. Although some of those percentages may appear high, all the participants reached an acceptable level of coder reliability which is the minimum of 85% for an item and the average of 92% across all items. For few PBA countries, dissimilar coding category distributions among coders could be occasionally observed along with high proportion agreement on an item. This largely resulted from the different pools of responses upon which coding category distribution and proportion agreement were measured. As mentioned earlier, proportion agreement per item across coders was only based on the unique set of 100 responses for multiple coding; while coding category distribution per item across coders also took into account the randomly assigned responses for single

coding. Compared to CBA countries, the randomization of responses was more challenging for PBA countries where the distribution of booklets were handled manually.

Table 13.10: Percentage of coders whose coding category distributions on more than 20% of coded items were significantly different from other coders, averaged across CBA and PBA participants

	Mathematics (Trend)	*Reading (Trend)*	*Science (Trend)*	*Science (New)*	*Financial Literacy (Trend and New)*
CBA Participants	10%	17%	20%	35%	44%
PBA Participants	27%	52%	66%	NA	NA

Note:
1. The summary in the table is based on both student responses and anchor responses.
2. "Item" in the table refers to "human-coded constructed-response item".
3. New Science and Financial Literacy are CBA domains only in the Main Survey.
4. PISA participants can be a country, a region, an economy, or a subsample within the former three types of entities.

Across all the CBA participants, the percentage of items over which more than two coders' coding category distributions were significantly different from other coders was 6% in mathematics, 14% in reading, 8% in trend Science, 13% in new Science, and 13% in Financial Literacy (see Table 13.11). Across all the PBA participants, the percentage of items over which more than two coders' coding category distributions were significantly different from other coders was 17% in mathematics, 38% in reading, and 23% in trend Science (see Table 13.11). Although some of those percentages for PBA participants may appear high, all the participants have reached an acceptable level of coder reliability which is the minimum of 85% for an item and the average of 92% across all items.

Table 13.11: Percentages of participant × item pairs that have more than two coders' coding category distributions significantly different from other coders

	Mathematics (Trend)	*Reading (Trend)*	*Science (Trend)*	*Science (New)*	*Financial Literacy (Trend and New)*
CBA	6%	14%	8%	13%	13%
PBA	17%	38%	23%	NA	NA

Note:
1. The summary in the table is based on both student responses and anchor responses.
2. "Item" in the table refers to "human-coded constructed-response item".
3. PISA participants can be a country, a region, an economy, or a subsample within the former three types of entities.

The scales on which the PISA statistical framework is built are only as good as the scores used to establish them. In sum, the results from the coder reliability studies revealed that the coding designs that were tailored to meet every PISA participant's specific survey needs and the availability of coders were executed well, especially for CBA human-coded responses. The management of the coding process went smoothly and efficiently, with less involvement from the National Project Managers than necessary in previous cycles. CBA participating countries produced more complete and consistent coding data, while PBA participants showed some errors in the handling of the booklets and less reliable human coding. However, PBA participants still achieved acceptable levels of coder reliability amid the challenge of handling the booklet bundles manually.