Chapter 13
Coding design, coding process, coding reliability studies, and machine-supported coding in the main survey

INTRODUCTION

The proficiencies of PISA respondents were estimated based on their performance on the test items administered in the assessment. In the PISA 2018 assessment, countries and economies taking part in the computer-based assessment (CBA) administered six clusters each of science and mathematics *trend items* (items administered in previous cycles). The reading domain was a multi-stage adaptive assessment (MSAT), which included three stages (core, stage 1, and stage 2) consisting of both new and trend items. Countries that chose to take part in the financial literacy assessment administered two clusters of financial literacy items, and countries choosing to take part in the global competence assessment received four clusters of global competence items. Countries and economies participating in the paper-based assessment (PBA) administered 18 clusters of trend items across the domains of reading, mathematics, and science from previous PISA cycles.

The PISA 2018 assessment consisted of both multiple choice (MC) and constructed-response (CR) items. Multiple choice items (simple multiple choice [S-MC], with a single response selection, and complex multiple choice [C-MC], with multiple response selections) had predefined correct answers that could be computer-coded. While a few of the CR items were automatically coded by computer, most of them elicited a wider variety of responses that could not be categorised in advance and, therefore, required 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 Type	Mathematics (trend)	Reading (new)	Reading (trend)	Science (trend)	Financial Literacy	Global Competence
CBA	Human Coded	CR	21	46	36	32	13	13
	Computer Scored	S-MC	20	104	22	32	12	24
		C-MC	15	23	9	48	13	32
		CR	26	0	5	3	5	0
	Total		**82**	**173**	**72**	**115**	**43**	**69**
PBA	Human Coded	CR	48	NA	59	32	NA	NA
	Computer Scored	S-MC	19		35	29		
		C-MC	13		9	24		
		CR	3		0	0		
	Total		**83**		**103**	**85**		

Notes: CBA stands for computer-based assessment and PBA stands for paper-based assessment; CR refers to constructed-responses, S-MC is simple multiple choice, and C-MC is complex multiple choice.

New items were developed only for the CBA Reading, Financial Literacy, and the new innovative domain of Global Competence.

From, the 2018 cycle, the CBA coding teams were able to benefit from the use of a machine-supported coding system (MSCS). While an item's response field is open-ended, there is a commonality among students' raw responses, meaning that we can expect to observe the same responses (correct or incorrect) regularly throughout coding (Yamamoto, He, Shin, von Davier, 2017; 2018). High regularity in responses means that variability among all responses for an item is small, and a large proportion of identical responses can receive the same code when observed a second or third time. In such cases, human coding can be replaced by machine coding, thus reducing the repetitive coding burden performed by human coders.

This chapter describes the coding procedures, preparation, and multiple coding design options employed in CBA. Then it follows with the coding reliability results and reports the volume of responses coded through the MSCS from the 2018 PISA main survey.

CODING PROCEDURES

Since 2015 cycle, the coding designs for the CBA item responses for mathematics, reading, science, and financial literacy (when applicable) were greatly facilitated through use of the Open-Ended Coding System (OECS). This computer system 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. Coders could easily access to the organised responses according to the specified coding design through the OECS platform that was available offline.

The CBA coding is done online on an item-by-item basis. Coders retrieve a batch of responses for each item. Each batch of responses included the anchor responses in English that were coded by the two bilingual coders, the students' responses to be multiple coded as part of the reliability monitoring process, and the students' response to be single coded. Each web-page displays the item stem or question, the individual student response and the available codes for the item. Also included on each web-page were two checkboxes labelled *defer* and *recoded*. The defer box was used if the coder was not sure which code to assign to the response. These deferred responses were later reviewed and coded either by the coder or lead coder. The recoded box was checked to indicate that the response had been recoded for any reason. It was expected that coders would code most responses assigned to them and defer responses only in unusual circumstances. When deferring a response, coders were encouraged to note the reason for deferral into an associated comment box. Coders generally worked on one item at a time until all responses in that item set were coded. The process was repeated until all items were coded. The approach of coding by item was greatly facilitated by the OECS, which has been shown to improve reliability by helping coders to apply the scoring rubric more consistently.

For the paper-based assessment (PBA), the coding designs for the PBA responses for mathematics, reading and science were supported by the data management expert (DME) system, and reliability was monitored through the Open-Ended Reporting System (OERS), additional software that worked in conjunction with the DME to evaluate and report reliability for CR items. Detailed information about the system was provided in the OERS manual. The coding process for PBA participants involved using the actual paper booklets, with sections of

some booklets single coded and others multiple-coded by two or more coders. When a response is single coded, coders mark directly in the booklets. When a response is multiple-coded, the final coder codes directly in the booklet while all others code on coding sheets; this allows coders to remain independent in their coding decisions and provides for the accurate evaluation of coding reliability.

Careful monitoring of coding reliability plays an important role in data quality control. National Centres used the output reports generated by the OECS and OERS to monitor irregularities and deviations in the coding process. Through coder reliability monitoring, coding inconsistencies or problems within and across countries could be detected early in the coding process, and action could be taken quickly to address these concerns. The OECS and OERS generate similar reports of coding reliability: i) proportion agreement and ii) coding category distribution (see later sections of this chapter for more details). National Project Managers (NPMs) were instructed to investigate whether a systematic pattern of irregularities exist and if the observed pattern is attributable to a particular coder or item. In addition, NPMs were instructed not to carry out *coding resolution* (changing coding on individual responses to reach higher coding consistency). Instead, if systematic irregularities were identified, coders were retrained and all responses from a particular item or a particular coder needed to be recoded, including codes that showed disagreement as well as those that showed agreement. In general, if happened, inconsistencies or problems were found to be coming from a misunderstanding of general coding guidelines and/or a rubric for a particular item or misuse of the OECS/OERS. Coder reliability studies conducted by the PISA contractors also made use of the OECS/OERS reports submitted by National Centres.

CODING PREPARATION

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

Recruitment of national coder teams

NPMs 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/economy,
- organising all materials and distributing them to coders,
- monitoring the coding process,
- monitoring 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 for PISA contractors to review.

Additionally, the lead coder was required to be proficient in English (as international training and interactions with the PISA contractors were in English only) and to attend the international coder trainings in Athens in January 2017 and in Malta in January 2018. 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 main survey.

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.
- 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 items (in the domains of Reading, Financial Literacy, and Global Competence), which included coding rubrics and examples of correct and incorrect responses. Coding rubrics for new items were defined for the field trial, and this information was later used to revise the coding guides for the main survey. Coding information for trend items from previous cycles was also included in the coding guides.

Prior to the field trial, NPMs and lead coders were provided with a full item-by-item coder training in Athens in January 2017. The field trial training covered all reading items - trend and new. Training for the trend items were provided through recorded training followed by Webinars. Prior to the main survey, NPMs and lead coders were provided with a full round of item-by-item coder training in Malta in January 2018. The main survey training covered all items – trend and new –in all domains. During these trainings, the coding guides were presented and explained. Training participants practiced coding on sample responses and discussed any ambiguous or problematic situations as a group. During this training, participants had the opportunity to ask questions and have the coding rubrics clarified as much as possible. When the discussion revealed areas where rubrics could be improved, those changes were made and were 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 coding 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 stored on the PISA 2018 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 and replicate as much as possible of the international training 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 facilitate discussion about any items that proved challenging. Past experience has shown that when coders discuss items among themselves and with their lead coder, many issues could be resolved, and more consistent coding could be achieved.

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

1. 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.
2. Coder training took place at the item set (CBA) or booklet (PBA) level. In this alternative approach, coders were fully trained on a set of units of items. Once the full training was complete, coding could take place at the item level; however, to ensure that the coding rules were still fresh in coders' minds, a coding refresher was recommended before coding each item.

CODING DESIGN

Coding designs for CBA and PBA were developed to accommodate participants' various needs in terms of the number of languages assessed, the sample size, and selected domains. In general, it was expected that coders would be able to code approximately 1,000 responses per day over a two- to three-week period. Further, a set of responses for all human-coded CR items were required to be multiple-coded to monitor coding reliability. *Multiple coding* refers to the coding of the same student response multiple times by different coders independently, such that inter-rater agreement statistics can be evaluated for the purpose of ensuring the accuracy of scores on human-coded CR items. For each human-coded CR item in a standard sample, a fixed set of 100 student responses were multiple-coded, which provided a measure of within-country coding reliability. Regardless of the design each participating country/economy chose, a fixed set of anchor responses were also coded by two designated bilingual coders. *Anchor coding* refers to the coding of ten to thirty (in PBA and CBA, respectively) anchor responses per item in English for which the correct code for each response is already known by the PISA contractor (but not provided to coders). The bilingual coders independently code the anchor responses, which are compared to the known code in the anchor key, to provide a measure of across-country coding reliability.

Each coder was assigned a unique coder ID that was specific to each domain and design. The OECS platform offers some flexibility for CBA participants so that a range of coding designs were possible to meet the needs of the participants. For PBA participants, four coding designs were possible given the sample size of each assessed language.

Table 13.2 shows the number of coders by domain in the CBA coding designs. CBA participants were able to determine the appropriate design for their country/economy with a provided calculator template, which could then be used to set-up the OECS platform with the designated number of coders by design.

Table 13.2 **CBA coding designs: Number of CBA coders by domain**

Design	Sample Size	Reading	Science	Mathematics	Financial Literacy	Global Competence

Minority language design	< 4,500	2 - 8	2 - 3	2 - 3	2 - 3	2 - 3
Standard design	4,501 - 8,000	12 - 16	4 – 5	4 - 5	4 - 5	4 - 5
Alternative design	8,001 - 13,000	16 - 24	6 - 9	6 - 9	6 - 9	6 - 9
Over-sample design	> 19,000	24 - 32	10 - 12	10 - 12	10 - 12	10 - 12

Note: The number of students is based on the test options (reading, science, and mathematics, as well as the additional options of financial literacy and global competence) and the number of languages assessed; the designed number of coders is exclusive by assessment language.

The design for multiple coding in the CBA is shown in Table 13.3. The first digit of the coder ID identified the domain and the remaining digits the coder number. In the CBA, human-coded CR items may be bundled into item sets when the number of items and/or volume of responses to be coded in a particular domain is high. There were four item sets for the major domain of reading and one item set for each of the other domains. For each item, multiple coders coded the same 100 student responses that were randomly selected from all student responses. Each domain had two bilingual coders – always coders 01 and 03 – who additionally coded thirty anchor responses in English for each item. Following multiple coding, the OECS evenly distributed the remaining student responses for each item among coders to be single coded.

Table 13.3 Organization of multiple coding for the CBA designs

		Coder IDs																														
Science		101 (bilingual)	102	103 (bilingual)	104	105	106	107	108	109	110	111	112																			
Item set 1	100/item	●	●	●	●	●	●	●	●	●	●	●	●																			
Anchor set	30/item	○		○																												
Reading		201 (bilingual)	202	203 (bilingual)	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231	232
Item set 1	100/item	●	●					●	●	●			●	●			●	●			●	●			●	●			●	●		●
Item set 2	100/item	●	●			●	●			●		●		●		●		●		●		●		●		●		●		●		
Item set 3	100/item			●	●	●	●				●	●			●	●			●	●			●	●			●	●			●	
Item set 4	100/item			●	●			●	●		●		●		●		●		●		●		●		●		●		●		●	●
Anchor set	30/item	○		○																												
Mathematics		301 (bilingual)	302	303 (bilingual)	304	305	306	307	308	309	310	311	312																			
Item set 1	100/item	●	●	●	●	●	●	●	●	●	●	●	●																			
Anchor set	30/item	○		○																												
Financial Literacy		401 (bilingual)	402	403 (bilingual)	404	405	406	407	408	409	410	411	412																			
Item set 1	100/item	●	●	●	●	●	●	●	●	●	●	●	●																			
Anchor set	30/item	○		○																												
Global Competence		501 (bilingual)	502	503 (bilingual)	504	505	506	507	508	509	510	511	512																			
Item set 1	100/item	●	●	●	●	●	●	●	●	●	●	●	●																			
Anchor set	30/item	○		○																												

Notes: "●" denotes that the coder codes 100 student responses per item in the item set. "o" denotes the coder codes 30 anchor responses in English per item.

Four coding design variations were offered to PBA participants (see Table 13.4). All thirty unique paper-and-pencil booklets contain four clusters from at least two different domains; only one domain from each set of booklets is multiple-coded.

Table 13.4 PBA coding designs: Number of PBA coders by domain

Design	Sample Size	Reading	Science	Mathematics
Minority language design 1	< 1,500	2	3	2
Minority language design 2	1,500 - 3,500	3	6	3
Standard design	3,501 - 5,500	4	9	4
Alternative design	> 5,501	5	12	5

Note: The number of students is based on the languages assessed; the designed number of coders is exclusive by language.

In the first step of the PBA coding, the bilingual coders code the anchor responses, enter the data into the project database using the Data Management Expert (DME, data management system) and evaluate the across-country coding reliability in the OERS reliability software. In the second step, 100 student responses are multiple-coded for each item. While CBA human-coded CR items were organised by item set during multiple coding, by contrast, PBA human-coded CR items were organised by bundle set rather than item set. The PBA standard coding design is shown in Table 13.5. When the National Centre receives student booklets, they are first sorted by booklet number (1-30): Booklets 7-12 comprise bundle 1, for which science items are multiple-coded; similarly, booklets 1-6, 13-18, and 25-30 comprise bundles 2-4, for which reading items are multiple-coded; finally, booklets 19-24 comprise bundle 5, for which mathematics items are multiple-coded. For multiple coding, all but the final coder code responses on coding sheets; the final coder codes responses directly in the booklet. All multiple-coded response codes are entered into the project database using the DME and run the OERS reliability software for review. Any coding issues identified by the OERS are investigated and corrected before moving forward. The final step is the single coding when, any remaining uncoded responses are equally distributed among coders and are coded directly into the booklets. This step is also when items from the second domain in each of the booklets are single coded. Codes are recorded into the project database using the DME; the distribution of single codes are reviewed in the OERS as a quality check. This coding design enabled the within- and across-country comparisons of coding.

Table 13.4 Organization of multiple coding for the PBA standard coding design

			Coder IDs			
	Science	**Multiple-Coding**	101 (bilingual)	102	103 (bilingual)	104
Bundle 1	Booklet 7	50 booklets from each type (100 student responses for each item)	●	●	●	
	Booklet 8			●	●	●
	Booklet 9		●		●	●
	Booklet 10		●	●		●
	Booklet 11		●		●	●
	Booklet 12			●	●	●

	Anchor Booklet	10 anchor responses for each item	○		○						
	Reading		201 (bilingual)	202	203 (bilingual)	204	205	206	207	208	209
Bundle 2	Booklet 1	50 booklets from each type (100 student responses for each item)	●	●	●	●			●		
	Booklet 2			●	●	●			●	●	
	Booklet 3		●		●	●			●	●	
	Booklet 4		●	●	●	●				●	
	Booklet 5		●	●		●			●	●	
	Booklet 6		●	●	●				●	●	
Bundle 3	Booklet 13	50 booklets from each type (100 student responses for each item)	●	●			●	●			●
	Booklet 14		●	●				●	●		●
	Booklet 15		●	●			●		●		●
	Booklet 16		●	●			●	●	●		
	Booklet 17			●			●	●	●		●
	Booklet 18		●				●	●	●		●
Bundle 4	Booklet 25	50 booklets from each type (100 student responses for each item)			●	●	●	●		●	
	Booklet 26					●	●	●		●	●
	Booklet 27				●		●	●		●	●
	Booklet 28				●	●	●	●			●
	Booklet 29				●	●		●		●	●
	Booklet 30				●	●	●			●	●
	Anchor Booklet	10 anchor responses for each item	○		○						
	Mathematics		301 (bilingual)	302	303 (bilingual)	304					
Bundle 5	Booklet 19	50 booklets from each type (100 student responses for each item)	●	●	●						
	Booklet 20			●	●	●					
	Booklet 21		●		●	●					
	Booklet 22		●	●		●					
	Booklet 23		●		●	●					
	Booklet 24			●	●	●					
	Anchor Booklet	10 anchor responses for each item	○		○						

Note: "●" denotes that the coder codes 100 student booklets for the specific form as a bundle set. "○" denotes that the coder codes 10 anchor responses in English per item.

Within-country and across-country coder reliability

Reliable human coding is critical for ensuring the validity of assessment results within a country/economy, as well as the comparability of assessment results across countries (Shin, von Davier, & Yamamoto, 2019). Coder reliability in PISA 2018 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 CR item were coded by at least two human coders.

The purpose of monitoring and evaluating the within-country coder reliability was to ensure accurate coding within a country/economy and identify any coding inconsistencies or problems in the scoring process so they could be addressed and resolved early 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/economy. Multiple coding 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 used to reduce national costs and coding burden. In general, a set of 100 responses per human-coded CR item was randomly selected from actual student responses and a set of coders in that domain scored those responses. The rest of the student responses needed to be evenly split among coders to be single coded.

Accurate and consistent scoring within a country/economy 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/economy codes a certain response differently than other countries (Shin, et al., 2019). 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 coding of a set of anchor responses. In each country/economy, 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 human-coded CR item, a set of thirty CBA (or ten PBA) anchor responses in English were provided. These anchor responses were answers obtained from real students and their authoritative coding was not released to the countries. Because countries using the same mode of administration coded the same anchor responses for each human-coded CR item, their coding results on the anchor responses could be compared to the anchor key and, thereby, to each other.

CODER RELIABILITY STUDIES

Coder reliability studies were conducted to evaluate the consistency of coding of human-coded CR items within and across the countries participating in PISA 2018. The studies included 70 CBA countries/economies (for a total of 112 country/economy-by-language groups) and nine PBA countries/economies (for a total of 14 country/economy-by-language groups) with sufficient data to yield reliable results. The coder reliability studies were conducted in three aspects:

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

Score agreement (domain-level and item-level) and coding category distribution were the main indicators used by National Centres for the purpose of monitoring and PISA contractors for the purpose of evaluating the coder reliability. The domain-level proportion scoring agreement was the average of item-level proportion scoring agreement across items. Note that only the exact agreement was considered as the scoring agreement.

- *Proportion scoring agreement* refers to the proportion of scores from one coder that matched exactly the scores of other coders on an identical set of multiple-coded responses for an item. It can vary from 0 (0% agreement) to 1 (100% agreement). Each country/economy 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 or coder. One-hundred responses for each item were multiple-coded for the calculation of within-country score agreement while ten or thirty responses (PBA or CBA, respectively) for each item were coded for the calculation of across-country score agreement.

- *Coding category distribution* refers to the distributions of coding categories (such as "full credit", "partial credit" and "no credit") assigned by a coder to two sets of responses: a 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.

Country-level score agreement

The average within-country score agreement based on 100 multiple coding set in PISA 2018 exceeded 92% (pre-defined threshold) in each domain across the 112 country/economy-by-language groups with sufficient data (see Tables 13.6 and 13.7).

During coding, the formula used to by the OECS to calculate ongoing interrater agreement is:

$$R_{ix} = \frac{G_{ix}\left(\frac{N-A}{N}\right)}{D_{ix}(C-1)} + \frac{A}{N}$$

where R_{ix} is the calculated agreement rate for coder C_i for item x, N is the total number of responses for this item x, A is the number of machine coded responses (see the section on the *Machine-Supported Coding System* at the end of this chapter) for item x, C is number of coders for this item, G_{ix} is the number of agreed codes for coder C_i for item x (max = (C-1)), and D_{ix} is the number of multiple-coded responses for item x coded by coder i so far (at the end of coding, this will equal 100 in a standard sample).

Follwing coding, the difference between CBA and PBA participants' average proportion agreements in each of the mathematics, science, and trend reading domains were less than 0.5%. Within each mode, the within-country score agreement between domains was not significantly different, either. The mathematics domain had highest agreement (98.9% for CBA; 98.2% for PBA). Trend reading items showed the second highest agreement of 97.7% for CBA and 97.9% for PBA; new reading items in the CBA showed similarly high agreement at 97.3%. The science domain had an inter-rater agreement of 97.3% for CBA and 96.7% for PBA. The optional CBA domains of financial literacy and global competence had inter-rater agreements of 95.1% and 96.6%, respectively.

Across-country score agreement based on 10/30 anchor coding set in PISA 2018 was slightly lower than within-country score agreement. Domain-level agreement was again the highest in mathematics, with 95.0% for CBA and 97.7% for PBA. Trend reading had anchor agreement of 92.9% in CBA and 93.0% in PBA; new reading in CBA had similar anchor agreement of 92.1. The science domain showed anchor agreement at 89.4% in CBA and 96.1% in PBA. The

optional CBA domains of financial literacy and global competence had anchor agreements of 88.8% and 86.3%, respectively.

Table 13.6 (1/3) **Summary of within- and across-country (%) agreement for CBA participants for the main domains of Reading, Science, and Mathematics**

		Within-country Agreement				Across-country Agreement			
	Country/Economy - Language	Mathematics (trend)	Reading (new)	Reading (trend)	Science (trend)	Mathematics anchor (trend)	Reading anchor (new)	Reading anchor (trend)	Science Anchor (trend)
OECD	Australia - English	99.7	96.7	97.5	98.1	94.9	94.2	94.4	92.2
	Austria - German	99.2	97.4	98.3	97.5	96.2	94.1	94.7	91
	Belgium - German	98.9	98.5	98.7	97.4	96	95.9	95.2	92.9
	Belgium - French	98.4	97.6	98.3	97.5	93.2	96.5	97.4	96.2
	Belgium - Dutch	98.7	96.8	97.9	97.6	94.8	93.2	95.2	91.6
	Canada - English	99.4	96.4	98	97.4	96.2	92.2	93.6	91.4
	Canada - French	99.1	95.7	96.6	95.4	95.6	93.6	92.4	91.4
	Chile - Spanish	99.3	97.6	98.1	96.6	92.2	92.2	92.6	90.6
	Colombia - Spanish	99.9	99.9	99.9	99.9	97.1	94.8	95.4	90.6
	Czech Republic - Czech	99	99	99.4	98.3	96.1	95.6	95.1	91.1
	Denmark - Danish	99	97.1	97.9	97.7	94.6	90.9	93.8	90.5
	Denmark - Faroese	97.5	97	95.6	97.4	93.9	92.3	91.3	84.3
	Estonia - Estonian	98.7	96.1	96.7	95.3	96.7	91.3	94	87.2
	Estonia - Russian	97.8	95.6	95.5	95.4	97.1	92.3	93.4	92.3
	Finland - Finnish	99.9	99.9	99.9	97.7	97.2	95.4	94.6	93.6
	Finland - Swedish	99.9	99.7	99.9	99.9	97.4	95.6	94.6	95.2
	France - French	99.4	99.8	99.8	98.7	95.9	91.9	92.7	95
	Germany - German	98.8	97.1	97.5	96.4	93.2	91.8	91.6	86.5
	Greece - Greek	99.5	98	98.6	98	97.6	94.6	94.9	89.2
	Hungary - Hungarian	99.2	97	97.6	97.7	95.8	93.8	94.3	93.3
	Iceland - Icelandic	98.1	94.8	96.4	96.2	95.1	93.4	95.3	91.6
	Ireland - English	99.4	96	96.6	96.4	96.2	94.9	95.3	94
	Ireland - Irish	99.6	94.2	92.7	98.6	96.8	95.2	95	93.7
	Israel - Arabic	98.8	96.4	97.4	97	95.4	88.5	90.2	90.7
	Israel - Hebrew	98.6	97	97.3	96.1	96.3	92.1	94.5	92.2
	Italy - German	98	98.7	98.2	99.5	95.3	94.2	94.3	86.3
	Italy - Italian	99.6	98.9	99.2	98.8	96.6	94.1	94	92.5
	Japan - Japanese	99	97.8	98.1	97.9	95.8	95.4	95	90.6
	Korea - Korean	99.3	98.7	98.8	96.9	96.4	93.2	93.4	89.8
	Luxembourg - German	99.2	97.9	98.3	97.2	96	94.9	95.1	93.4
	Luxembourg - English	98.4	97.6	97.7	96.5	95.6	94.9	95.6	93.6
	Luxembourg - French	98.3	97.8	97.8	97.2	95.8	95.1	95.5	93.1
	Mexico - Spanish	99.1	95.7	96.7	95.6	94.3	89.2	93.5	87.2
	Netherlands - Dutch	98.5	95.7	96.5	95.7	96.5	94.2	94.2	91.9
	New Zealand - English	98.9	96.9	97.6	96.3	94.8	95.2	94.9	90.5
	Norway - Nynorsk	97.5	95.6	96.9	95.6	92.9	93	94.6	90.6
	Norway - Bokmål	99	97.1	98.1	97.3	96	94.4	95	90.5
	Poland - Polish	99.4	97.2	97.9	97.6	96	93	93.2	92.9
	Portugal - Portuguese	99.5	98.8	99.2	97.6	95.9	94.4	94.9	94.8
	Slovak Republic - Hungarian	98.8	99.3	99.5	99.2	96.2	92.5	91.1	92.9
	Slovak Republic - Slovak	99.2	99.1	99.2	98.6	97.3	93.9	94.2	93.4
	Slovenia - Slovenian	99.6	97.5	97.8	98.6	95.4	93.2	93.8	91.2
	Spain - Catalan	98.2	96.1	96.8	95.3	94.5	88.6	88.8	87.7
	Spain - Spanish	99.5	98.6	99	98.7	90.9	92	93.8	84.3
	Spain - Basque	96.4	94.9	96.9	93.1	91.5	85.1	86.6	84.3
	Spain - Galician	96.7	94.9	94.5	93.6	91.9	91.1	93.2	81.5
	Spain - Valencian	97.4	96.1	96.5	96.2	94.7	87.9	89.4	89.7
	Sweden - English	100	94.4	87	94.6	96	94.3	92.9	94
	Sweden - Swedish	99.2	96.4	97	97.1	96.3	92.5	94.8	92
	Switzerland - German	99	98.9	99.4	96.5	95.8	93.9	95	92.4
	Switzerland - French	97.7	96	96.8	96.7	95.1	94.4	94.5	89
	Switzerland - Italian	98	98.3	99.2	99.8	96.6	94.1	93.8	92.3
	Turkey - Turkish	99.5	99.2	99.5	98.5	96.2	93.1	92.1	86.2

United Kingdom (Excl. Scotland) - Welsh	99.3	99	99.8	100	95.2	95.4	95.1	90.7
United Kingdom (Excl. Scotland) - English	99.7	97.4	98.3	97.7	95.6	95.4	95.6	92.2
United Kingdom (Scotland) - English	95.3	95.2	98.9	96.6	95	92.8	97	95.7
United States - English	99.2	96.3	97	95.2	95.9	95.7	94.8	93.3
Mean - OECD	**98.8**	**97.2**	**97.6**	**97.2**	**95.4**	**93.3**	**93.9**	**91**
Median - OECD	**99**	**97.1**	**97.9**	**97.4**	**95.8**	**93.9**	**94.4**	**91.6**

Table 13.6 (2/3) **Summary of within- and across-country (%) agreement for CBA participants**

Partners

Country/Economy - Language	Within-country Agreement				Across-country Agreement			
	Mathematics (trend)	Reading (new)	Reading (trend)	Science (trend)	Mathematics anchor (trend)	Reading anchor (new)	Reading anchor (trend)	Science anchor (trend)
Albania - Albanian	97.8	93.6	94.6	92.5	93.6	85.5	87.4	86.3
Baku (Azerbaijan) - Azeri	99.4	98.6	98.6	98.7	90.1	88.6	90.9	82.9
Baku (Azerbaijan) - Russian	98.3	96.6	97.6	96.7	90.4	92	89.9	85.8
Belarus - Belarusian	98	98.6	99	98.1	92.8	94.2	94	90.7
Belarus - Russian	97.9	97	97.4	96.7	91.7	94.1	94	90.5
Bosnia and Herzegovina - Bosnian	99.4	97.8	98.1	98.2	93.1	87.2	84.7	79.1
Bosnia and Herzegovina - Croatian	99.1	97.5	97.2	97.7	95.3	89.9	84	78.3
Bosnia and Herzegovina - Serbian	99.1	97.7	97.4	98.3	95.4	86.8	86.8	78.2
Brazil - Portuguese	99.7	97.8	98.3	98.1	95.3	87.9	89.2	82.7
Brunei Darussalam - English	99	96.6	97.5	97.2	94.5	92.2	93.7	87.9
Bulgaria - Bulgarian	99.1	98.1	98.6	98.2	93.4	89.6	93.2	90.9
B-S-J-Z* (China) - Chinese	99.2	97.6	97.3	97.5	96.6	91.1	93.3	89
Chinese Taipei - Chinese	99.3	97.9	98.5	96.8	97	92.2	94.2	90.6
Costa Rica - Spanish	99.4	96.2	97.1	99.6	94.3	93	93.5	71.2
Croatia - Croatian	99.4	97.5	98	98.7	97.7	92.7	93.5	90.8
Cyprus - Greek	99.1	96.7	97.4	97.4	95.8	92.2	93.7	91.1
Cyprus - English	98.8	94.4	94.7	94.2	95.9	93.6	94.4	92
Dominican Republic - Spanish	99.3	97.6	97.9	98.4	93.4	89.1	89.9	86.8
Georgia - Azerbaijani	98.8	98.4	98.9	98.2	90.2	88.7	92	85.7
Georgia - Georgian	99.2	97.8	98.1	98.1	95.1	90.7	93	91.1
Georgia - Russian	98.2	97.9	96.7	97.2	95.3	92.5	93.8	85.6
Hong Kong (China) - English	98.4	96	96.8	95.7	96.4	91.8	93.8	92.2
Hong Kong (China) - Chinese	98.7	96.8	97.2	96.6	96.6	93	93.7	91.4
Indonesia - Indonesian	97.5	96.1	96.3	98.7	87.9	88.1	91.5	78.2
Kazakhstan - Kazakh	99	98	98.4	98.6	94.7	94.3	95.6	92.6
Kazakhstan - Russian	99	96.7	97.8	96.4	94.6	94.4	94.7	92.9
Kosovo - Albanian	99.1	96	96.9	97.2	93.6	88.4	90.1	76.9
Latvia - Latvian	97.8	96.2	96.1	94.8	95.4	90	91.1	87.9
Latvia - Russian	97	93.7	95.5	95.5	96.6	87.2	90.4	84.2
Lithuania - Lithuanian	99.4	98.6	98.8	97.8	96.5	93.8	94.4	92.8
Lithuania - Polish	98.9	99.6	99.6	97.7	96.8	94.9	94.6	93.2
Lithuania - Russian	98.7	98.6	99	97.3	97	95.3	95.5	93.4
Macao (China) - English	97.3	97.6	98.6	95.5	95.9	94.3	93.4	92.1
Macao (China) - Portuguese	100	100	100	100	96.2	93.9	92.9	89.5
Macao (China) - Chinese	97.8	96	96.9	95.7	95.7	91.9	93.2	93.6
Malaysia - English	98.4	94.2	95	96.8	96.5	91.4	91.6	88.8
Malaysia - Malay	99.2	95.6	96.6	95.3	96.2	85.7	90.7	88.3
Malta - English	98.4	96.9	97.5	94.2	95.4	92.5	93.8	89.5
Montenegro - Serb (Yekavian)	99.5	99.2	99.5	98.4	95.7	94.2	94.5	86.8
Montenegro - Albanian	99.5	100	100	99.3	94.3	93.5	93	88.2
Morocco - Arabic	99.2	97.3	98	98.2	95.3	84.9	87.5	86
Panama - English	98.3	97.6	96.6	96.7	89.9	90.1	92.4	81.4
Panama - Spanish	98.8	97.5	98.1	97.1	90.1	83.3	87.4	85.3
Peru - Spanish	99.4	98	98.6	97.2	95.2	94	94	92.6
Philippines - English	99.4	96.5	98.3	98.2	95.1	87.9	84.8	86.3
Qatar - Arabic	99.6	98.8	99.2	98	94.9	89.4	89.3	86.2
Qatar - English	99.1	97.4	97.5	97.1	95.3	90.7	90.2	87.9
Russian Federation - Russian	99.6	98.7	99	98.2	92.7	93.7	94.6	88.8
Serbia - Hungarian	99.7	99.9	98.4	97.9	95.7	89.1	91.3	88.5
Serbia - Serbian	99.3	98.7	99.1	98.4	93.2	87.7	92.2	90.1
Singapore - English	99.1	98	98.1	98.3	96.2	94.9	95.5	93
Thailand - Thai	99.7	96.3	96.9	96.2	96.2	91.2	92.7	89.8
United Arab Emirates - Arabic	99.3	98.3	98.9	97.8	94.5	87.1	88.1	88.7
United Arab Emirates - English	99.7	97.1	98.1	98.1	95	89.1	92.5	85.8
Uruguay - Spanish	99.3	96.9	97.7	98.3	95.6	90.6	91.3	93

Mean - Partners	**98.9**	**97.4**	**97.8**	**97.3**	**94.6**	**90.8**	**91.9**	**87.7**
Median - Partners	**99.1**	**97.6**	**98**	**97.7**	**95.3**	**91.2**	**92.9**	**88.7**
Mean - All CBA	**98.9**	**97.3**	**97.7**	**97.3**	**95**	**92.1**	**92.9**	**89.4**
Median - All CBA	**99.1**	**97.5**	**97.9**	**97.4**	**95.5**	**92.7**	**93.7**	**90.6**

*B-S-J-Z (China) refers the four PISA-participating Chinese provinces: Beijing, Shanghai, Jiangsu, and Zhejiang.

Table 13.6 (3/3) **Summary of within- and across-country (%) agreement for Financial Literacy and Global Competence domains**

		Within-country		Across-country	
	Country/Economy - Language	**Financial Literacy**	**Global Competence**	**Financial Literacy (anchor)**	**Global Competence (anchor)**
OECD	Australia - English	98.4		91.5	
	Canada - English	97.9	95.6	94.0	87.4
	Canada - French	95.1	94.1	93.5	88.8
	Chile - Spanish	98.3	96.0	86.0	84.4
	Colombia - Spanish		97.9		88.7
	Estonia - Estonian	96.5		89.6	
	Estonia - Russian	96.9		91.7	
	Finland - Finnish	99.8		93.8	
	Finland - Swedish	99.9		94.5	
	Greece - Greek		97.9		80.6
	Israel - Arabic		96.1		87.6
	Israel - Hebrew		96.0		91.7
	Italy - German	97.6		92.1	
	Italy - Italian	99.5		91.7	
	Korea - Korean		97.9		88.8
	Netherlands - Dutch	96.2		93.3	
	Poland - Polish	98.6		93.1	
	Portugal - Portuguese	98.2		92.9	
	Slovak Republic - Hungarian	98.4	99.3	93.2	92.2
	Slovak Republic - Slovak	99.2	98.3	92.7	91.9
	Spain - Catalan	96.2	96.5	86.2	82.6
	Spain - Spanish	98.5	98.5	89.2	83.2
	Spain - Basque	93.0	92.9	85.4	78.8
	Spain - Galician	97.0	93.4	89.2	83.8
	Spain - Valencian	96.4	95.6	88.5	82.3
	United Kingdom (Scotland) - English		94.8		90.3
	United States - English	97.9		94.2	
Partners	Albania - Albanian		89.4		77.7
	Brazil - Portuguese	99.2		91.5	
	Brunei Darussalam - English		100.0		85.9
	Bulgaria - Bulgarian	99.2		92.3	
	Chinese Taipei - Chinese		96.8		86.2
	Costa Rica - Spanish		92.7		87.7
	Croatia - Croatian		95.5		88.1
	Georgia - Azerbaijani	98.4		91.2	
	Georgia - Georgian	98.4		87.6	
	Georgia - Russian	94.4		87.7	
	Hong Kong (China) - English		95.0		87.3

Hong Kong (China) - Chinese		94.5		84.6
Indonesia - Indonesian	97.8	97.4	85.3	85.0
Kazakhstan - Kazakh		98.0		91.4
Kazakhstan - Russian		96.5		91.3
Latvia - Latvian	96.9	96.1	92.7	80.8
Latvia - Russian	99.6	94.9	93.5	84.4
Lithuania - Lithuanian	98.6	98.9	94.0	89.1
Lithuania - Polish	97.0	99.6	93.8	89.9
Lithuania - Russian	97.0	99.7	93.3	89.5
Malta - English		96.5		86.3
Morocco - Arabic		97.0		84.7
Panama - English		97.5		78.2
Panama - Spanish		98.2		76.8
Peru - Spanish	98.3		95.0	
Philippines - English		97.2		80.6
Russian Federation - Russian	99.3	98.7	88.2	90.9
Serbia - Hungarian	96.8	97.9	91.0	85.8
Serbia - Serbian	99.1	99.0	93.8	93.2
Singapore - English		97.3		91.4
Thailand - Thai		94.2		89.1
Mean - OECD	**97.6**	**96.3**	**91.3**	**86.5**
Median - OECD	**97.9**	**96.0**	**92.1**	**87.5**
Mean - Partners	**98.0**	**96.7**	**91.4**	**86.2**
Median - Partners	**98.4**	**97.2**	**92.3**	**86.3**
Mean - All	**97.8**	**96.6**	**91.3**	**86.3**
Median - All	**98.3**	**96.8**	**92.2**	**87.3**

Table 13.7 **Summary of agreement (%) per domain for PBA participants**

	Country/Economy - Language	Within-country Agreement			Across-country Agreement		
		Mathematics (trend)	Reading (trend)	Science (trend)	Mathematics (anchor)	Reading (anchor)	Science (anchor)
Partners	Argentina – Spanish	96.7	92.5	90.2	96.8	93.2	92.8
	Jordan – Arabic	97.8	93.4	92.6	96.8	91.2	93.0
	Lebanon – English	98.8	99.7	98.9	98.7	92.8	97.8
	Lebanon – French[2]	99.6	99.9	98.7	NA	NA	NA
	Republic of Moldova – Romanian	99.5	99.4	99.0	97.7	91.0	97.0
	Republic of Moldova – Russian	99.3	98.0	99.1	97.5	92.0	97.3
	Former Yugoslav Republic of Macedonia – Macedonian	96.8	94.7	92.7	97.6	93.6	93.6
	Former Yugoslav Republic of Macedonia – Albanian	94.8	99.1	92.1	97.7	93.6	94.0
	Romania – Hungarian[1]	NA	NA	NA	97.5	95.4	98.0
	Romania – Romanian	99.4	99.7	99.7	97.6	94.4	99.0
	Saudi Arabia – Arabic	97.5	96.7	95.5	98.5	83.2	95.1
	Saudi Arabia – English[2]	97.8	100.0	100.0	NA	NA	NA
	Ukraine – Russian	99.1	97.9	98.6	98.4	96.8	95.8
	Ukraine – Ukrainian	99.0	98.9	97.3	98.8	96.5	97.6
	Viet Nam – Vietnamese	98.9	100.0	99.3	97.0	95.2	98.0
	Mean – PBA Partners	**98.2**	**97.9**	**96.7**	**97.7**	**93.0**	**96.1**
	Median – PBA Partners	**98.8**	**99.0**	**98.6**	**97.6**	**93.6**	**97.0**

1. Romania did not multiple-code responses in Hungarian.

2. Lebanon did not code anchor responses in French; Saudi Arabia did not code anchor responses in English.

Notes: New reading, financial literacy, and global competence are computer-based assessment domains only in the main survey.

Item-level scoring reliability

The number of items in a domain with score agreement below 85% was further monitored for each country/economy. Table 13.8 shows the number of country/economy-language groups that had either no items in a domain (N = 0) below 85% inter-rater score agreement, between one and five items ($1 \leq N \leq 5$), or up to ten items ($6 \leq N \leq 10$) in a domain below 85% score agreement. Across CBA participants, it can be observed that there were no mathematics items or financial literacy items, in any country/economy-language group, below 85% score agreement, although a few country/economy-language groups did tend to have more than one but less than five items in a domain under 85% score agreement.

Table 13.8 Number of country-language groups with score agreement < 85% on N multiple-coded items

Mode	Country-language groups	N of items with score agreement < 85%	Mathematics (trend)	Reading (new)	Reading (trend	Science (trend)	Financial Literacy (trend, new)	Global Competence (new)
CBA	112*	N = 0	112	109	107	110	36	38
		1 ≤ N ≤ 5	0	2	3	2	0	3
		6 ≤ N ≤ 10	0	1	2	0	0	0
PBA	14	N = 0	12		11	10		
		1 ≤ N ≤ 5	2		3	3		
		6 ≤ N ≤ 10	0		0	1		

* There was a total of 112 country/economy-language groups in the CBA main survey sample participating in the domains of reading (new and trend), mathematics, and science; only 36 of those country/economy-language groups participated in the financial literacy assessment, and only 41 country/economy-language groups participated in the global competence assessment.

Notes: N refers to the number of items with inter-rater score agreement below 85%. CBA stands for computer-based assessment and PBA for paper-based assessment.
Notes: N = 0 items indicates that a country/economy-language group had no items in the domain with less than 85% inter-rater score agreement, $1 \leq N \leq 5$ indicates that a country/economy-language group had 1-5 items in the domain with less than 85% inter-rater score agreement, and $6 \leq N \leq 10$ indicates that a country/economy-language group had 6-10 items with less than 85% inter-rater score agreement.

While Table 13.8 presented a breakdown of coding reliability across country/economy-language groups, Tables 13.9 and 13.10 present a breakdown of coding reliability at the item-level. Table 13.9 shows that all domains had an average item-level inter-rater score agreement (item-level multiple-coding score agreement averaged across participating country/economy-language groups) was quite high, above 96% in all domains. The average item-level anchor score agreement was slightly lower for every domain in both CBA and PBA. Table 13.10 shows exactly how many items in each domain had an average score agreement below the desired threshold of 85%. Again, multiple-coding score agreement was quite high, and no items in either CBA or PBA had an average inter-rater score agreement below 85%. Anchor items, however, were shown to have slightly lower agreement rates, which may be attributed to most domains having some items with an average anchor score agreement below 85%.

Table 13.9 Average item-level score agreement (across country-language groups) by domain

Mode	Source	Mathematics (trend)	Reading (new)	Reading (trend)	Science (trend)	Financial Literacy	Global Competence
CBA	Multiple-coded	98.7	97.3	97.7	97.1	97.6	96.5
	Anchor	94.7	92.1	92.9	89.4	91.3	86.3
PBA	Multiple-coded	98.4		97.9	96.7		
	Anchor	98.1		93.7	96.1		

Table 13.10 Number of items with average item-level score agreement (across country-language groups) below 85% by domain

Mode	Source	Mathematics (trend)	Reading (new)	Reading (trend)	Science (trend)	Financial Literacy	Global Competence
CBA	Multiple-coded	0	0	0	0	0	0
	Anchor	3	6	4	5	1	8
Out of total human-coded:		*21*	*46*	*36*	*32*	*13*	*13*
PBA	Multiple-coded	0		0	0		
	Anchor	0		3	0		
Out of total human-coded:		*48*		*59*	*32*		

1. "Item" in the table refers to "human-coded constructed-response item."
Notes: CBA stands for computer-based assessment and PBA for paper-based assessment; PISA participants can be a country, a region, an economy, or a subsample within the former three types of entities; new reading, financial literacy, and global competence are computer-based assessment domains only in the main survey.

Coding category distributions across coders

Coding category distributions provide more detailed information about interactions between coders and items. Tables 13.11 and 13.12 summarize overall coder quality and the impact of coder quality by item. In general, all coders should agree with their fellow coders at least 85% of the time on each item. Table 13.11 shows the percentage of coders who struggled to reach the 85% agreement

threshold on 20% or more items assigned to them. In mathematics, 2.3% of coders across all CBA countries/economies and 4.3% in PBA fell below the 85% inter-rater agreement threshold on more than 20% of their assigned items. In reading, this was 3.1% for CBA and 0.0% for PBA; in science, it was 3.3% for CBA and 4.3% for PBA.

Table 13.11 **Percentage of coders whose coding was below 85% inter-rater agreement on 20% or more of items, averaged across countries**

Mode	Mathematics (trend)	Reading (new, trend)	Science (trend)	Financial Literacy	Global Competence
CBA	2.3%	3.1%	3.3%	4.9%	2.0%
PBA	4.3%	0.0%	4.3%	NA	NA

Notes: CBA stands for computer-based assessment and PBA for paper-based assessment; "Items" in the title refers to "human-coded constructed-response items." The summary in the table is based on multiple-coded responses; new reading, financial literacy, and global competence are computer-based assessment domains only in the main survey.

Notes: Percentages were calculated by summing the number of coders in each country/economy-language group (by mode of assessment) with inter-rater score agreement below 85% on 20% or more of the items assigned to them; this value was divided by the total number of coders across all country/economy-language groups by mode.

Because coder quality is reflected at the item level, the percentage of items over which two or more coders showed less than 85% score agreement was also evaluated. For mathematics, 1.7% CBA and 1.1% PBA items across country/economy-language groups had two coders with less than 85% agreement with other coders within the same country/economy. For trend reading, this was 3.5% of items in the CBA and 2.2% of items in the PBA; 5.1% of new reading items across country/economy-language groups had at least two coders that were not able to achieve at least 85% agreement with other coders within the country/economy. This was similar for the domain of science (4.2% in the CBA and 1.6% in the PBA). The CBA domains of financial literacy and global competence showed 6.0% and 4.5%, respectively.

Table 13.12 **Percentage of items with at least two coders with less than 85% coding agreement, averaged across countries**

Mode	Mathematics (trend)	Reading (new)	Reading (trend)	Science (trend)	Financial Literacy	Global Competence
CBA	1.7%	5.1%	3.5%	4.2%	6.0%	4.5%
PBA	1.1%	NA	2.2%	1.6%	NA	NA

Notes: CBA stands for computer-based assessment and PBA for paper-based assessment. "Item" in the table refers to "human-coded constructed-response item." PISA participants can be a country, a region, an economy, or a subsample within the former three types of entities.

Notes: Percentages were calculated by summing the number of occurrences in which two or more coders within a country/economy-language group had less than 85% score agreement on a single item; these occurrences were pooled by domain and divided by the total number items administered in that domain across countries/economies by mode of assessment.

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 each PISA participant's specific survey needs, and the availability of coders were executed well. The management of the coding process went

smoothly and efficiently, with less involvement from the NPMs than necessary in previous PBA-only cycles.

MACHINE-SUPPORTED CODING SYSTEM

The shift in PISA 2015 from PBA to CBA enabled digitalizing students' raw responses and associated codes. A machine-supported coding system (MSCS) was newly developed and introduced as a result of such technological advances to improve the efficiency and accuracy of the coding process. Unlike commonly used automated scoring systems that generally involve algorithms, the MSCS relies soley on data that has already been human-coded in past PISA cycles and is specific to each country/economy-language group (responses that are identified for automatic coding are not shared among country/economy-language groups; rather, each country/economy-language group generates its own set of responses and associated codes to be automatically coded when the response appears again). In brief, the MSCS allows for the exactly same response to receive the appropriate code automatically so that scoring the same responses could be minimized (Yamamoto, et al., 2017 ; 2018). More specifically, the MSCS approach parallels automated scoring in the sense that a scoring model is first trained on existing historic data (2015 main survey and 2018 field trial) and then applied to future data (2018 main survey).

The MSCS capitalizes on the regularity and commonality of students' raw responses. Regularly observed responses are identified and verified, then the MSCS automatically applies the appropriate code, releaving coders from the burden of repeatedly coding the same response. For instance, combining 500 identical strings into the same code would eliminate 499 instances of repetitive and verified coding (or 99.8% of coding work for this particular example). The proportion of workload reduction is item dependent, as it is related to the level of response complexity and the consistency of codes assigned to that unique response. For instance, straightforward responses to short CR items (such as "30 minutes" as the response to a question about finding a gap between two time points) would more likely result in more consistent codes and, hence, lead to a larger workload reduction than moderately complex responses (such as explanations of how a medicine functions).

Further, MSCS can reduce inaccuracy caused by human coder's error (e.g., not understanding the coding rubric, fatigue, carelessness, etc.). The MSCS applies only verified codes taken from the coded unique responses (CUR) previously coded in the historical data. Raw responses can generally be categorized into three types: (a) nonresponse, (b) responses with verified coding, and (c) unique responses that require human judgment. The MSCS can be applied to the first two types (a and b). Human coding would only be required for unique responses (c). When the verified correct and incorrect codes could be assigned automatically for identical responses, coding the CR items is much more efficient and accurate, as well as less resource intensive for each participating country/economy.

The workflow of the MSCS can be divided into two phases: (a) learning from the codes assigned to past responses and (b) applying the learned coding to a new set of responses. In the first phase, historical data—for example of PISA 2018 main survey, the coded raw responses from PISA 2015 main survey and PISA 2018 field trial—are combined and analyzed together, and a simple algorithm sorts each of unique raw responses by code categories (e.g., 0, 1, 2, 9, 7). If there is a common code that applies to the sets of at least identical responses and is exclusive (i.e., if the same response exists in only "correct" category, but not in "incorrect" category), a CUR pool is generated based on the equivalent code and the code is considered as verified. In the second phase, MSCS assigned the verified code to new uncoded responses as much as applicable. If a new respondent's answer to a CR item is found in the CUR pool for

the same item in the given country/economy-language, the stored response code in CUR pool is directly applied to the new respondent's answer. Nonresponses, such as blanks, are assigned the appropriate nonresponse code for all items, including new items, in all participating countries/languages.. Only those responses that cannot be matched to an identical response stored in the CUR pool are assigned to human coders.

Development of the MSCS

The development of the MSCS was initiated through the pilot study using a set of sample items and became fully operationalized for the PISA 2018 field trial and main survey. In the pilot study conducted by Yamamoto et al. (2017), the sample item with the most instances of repeated raw responses resulted in a 94-98% workload reduction across country/economy-language groups, whereas the sample item with the fewest repeated responses reduced coding workload by as low as 5-29%. When all PISA 2015 main survey items were examined across country/economy-language groups, the percentages of identical responses among all responses constituted, approximately, 40% in mathematics, 28% in reading, 22% in science, and 18% in financial literacy, meaning the human workload could potentially be reduced that amount depending on the domains. Furthermore, when items were categorized into three groups in terms of regularities – high, medium, and low – there was a fairly consistent pattern in item categorization across many country/economy-language groups. These results suggest that it is feasible to increase the use of MSCS for PISA, which has more than 80 countries and 100 language versions, and all of the participants could be potentially benefited from this system.

In preparation for the PISA 2018 field trial, the MSCS was applied to all the CR items across all domains based on the harvested data from the PISA 2015 main survey (i.e., data-driven verification). Raw responses from a total of 146 items (trend items between 2015 main survey and 2018 field trial; 21 items from mathematics, 58 from science, 51 from reading, and 16 from financial literacy) across 59 countries/economies were used to prepare the PISA 2018 field trial CUR pool. The CUR pool was built to be country/economy-language-specific; within the CUR pool, coded unique responses are stored separately by domains and language groups. In the current CUR pool, each unique response was associated with a verified code (i.e., 0, 1, 2, 9 and etc.) that is consistent with the coding guidelines from the PISA 2015 and 2018 main surveys.

Two major rules were used when the unique responses were extracted and entered into the CUR pool. First, the response to an item in a specific country/economy-language group should occur at least five times. In order to ensure that the CUR pool contained accurate and verified codes for each unique response, only unique responses with the identical and exclusive codes were included. The second rule was set for the nonresponse category. An empty response was added to each item regardless of the frequency of nonresponses. This approach ensures at least one unique raw response (i.e., empty response) could be found in each CR item in the CUR pool. That is, the nonresponse can be directly filtered and coded by the machine rather than being assigned to human coders. It is expected that nonresponses are automatically assigned appropriate nonresponse code for all items, including new items in reading, and new items in financial literacy. Furthermore, automatic filtering of nonresponses is applicable for new participants without any historic data, thus no available CUR pool. Finally, some known types of responses for new items (i.e., responses where a student selected a radio button option but typed no text in the text box was coded as "incorrect" automatically) were added to the CUR pool.

In preparation for the PISA 2018 main survey, the PISA 2015 main survey data and PISA 2018 field trial data were combined together to extract an expanded CUR pool. The expanded CUR pool was constructed based on the same two major rules, particularly, expecting larger gain for new items from 2018 field trial. During the PISA 2018 main survey, the responses that could

not be matched with the existing CUR pool as well as the responses collected for the new items were assigned to human coders. The moving toward CBA will be continued and expected going forward, thus, the CUR pool can be expanded, further verified, and prepared for future cycles with the accumulated historical data.

Reduction of human-coding burden as the result of the MSCS

Table 13.13 and Table 13.14 summarize the efficiency of the MSCS with the reduction of human-coding burden in the PISA 2018 field trial and main survey. The tables summarize the percentage of responses coded by the MSCS and by human coders across all items in four domains (mathematics, reading, science, and financial literacy) and across country/economy language groups using mean and median. Given that the distribution of proportions for each item per group can be skewed, medians are reported in addition to the mean values.

The first two columns under the "machine-coded" header, *nonresponse* and *CUR*, indicate the average and median percentage of responses across CBA participants that were automatically coded by the MSCS as either nonresponse or a verified response (correct [full and partial] and incorrect). The total of these values is also presented, which can be compared to the percentage of human-coded responses, noted in the first column. Note that without the MSCS, all of the responses to CR items had to be coded by humans including nonresponses. On average, the coding burden for human coders was reduced for the 2018 field trial from a low of approximately 13% in Financial Literacy to a high of 34% in Mathematics. For the 2018 main survey, the coding burden was reduced a low of approximately 19% in financial literacy to a high of 33%, in mathematics.

For both field trial and main survey, approximately ten to seventeen percent of the total responses across all domains were empty responses on average, and automatically coded by the system. In particular, the MSCS was efficient for new items in reading, where no historic data were available, and reduced coding burden for human coders by 11-15% on average just by excluding blank responses. More specifically, comparing Tables 13.13 and 13.14 shows that relative gains of adding the 2018 field trial data is mostly noticeable for new reading items (4.6% to 8.5%) and financial literacy items (0.8% to 6.2%). Exceptionally, proportions of human-coded responses became slightly larger (less than 1% point increase) between field trial and main survey for the mathematics, mainly due to the observed inconsistencies between CUR pool when 2018 field trial data was combined. With the 2018 main survey data included for the 2021 CUR pool, more efficiency gains are expected.

Table 13.13 **Percentage of responses coded by the MSCS and by human coders across countries in the 2018 field trial**

		Human-coded	Machine-coded		
			Nonresponse	CUR	*Total*
Mathematics	mean	66.1%	17.4%	16.5%	33.9%
	median	70.3%	14.4%	7.0%	29.7%
Reading (new)	mean	84.3%	11.1%	4.6%	15.7%
	median	86.7%	8.3%	0.0%	13.3%
Reading (trend)	mean	79.3%	10.0%	10.7%	20.7%
	median	87.2%	7.6%	0.0%	12.8%
Science	mean	75.3%	12.4%	12.3%	24.7%
	median	77.5%	8.8%	3.0%	22.5%
	mean	87.1%	12.1%	0.8%	12.9%

Financial Literacy	median	88.9%	10.6%	0.0%	11.1%

Notes: "CUR" stands for "coded unique responses."

Table 13.14 Percentage of responses coded by the MSCS and by human coders across countries in the 2018 main survey

		Human-coded	Machine-coded		
			Nonresponse	CUR	*Total*
Mathematics	mean	67.5%	18.3%	14.2%	*32.4%*
	median	71.5%	15.2%	4.9%	*28.5%*
Reading (new)	mean	76.7%	14.7%	8.5%	*23.3%*
	median	81.7%	10.5%	0.0%	*18.2%*
Reading (trend)	mean	71.2%	18.4%	10.3%	*28.7%*
	median	78.9%	15.4%	0.0%	*20.9%*
Science	mean	74.6%	12.6%	12.7%	*25.3%*
	median	77.2%	8.9%	3.4%	*22.6%*
Financial Literacy	mean	79.9%	13.4%	6.2%	*19.6%*
	median	83.3%	10.3%	0.0%	*16.3%*

Notes: "CUR" stands for "coded unique responses."

NOTES

For a better understanding of the PISA coding designs, it is recommended that the descriptions of the PISA assessment designs in Chapter 2 be reviewed for important background information.

REFERENCE

Shin H.J., von Davier M., Yamamoto K. (2019) Investigating Rater Effects in International Large-Scale Assessments. In: Veldkamp B., Sluijter C. (eds) Theoretical and Practical Advances in Computer-based Educational Measurement. Methodology of Educational Measurement and Assessment. Springer, Cham.

Yamamoto, K., He, Q., Shin, H.J., & von Davier, M. (2018). Development and Implementation of a Machine-Supported Coding System for Constructed-Response Items in PISA. *Psychological Test and Assessment Modeling, 60*(2), 145-164.

Yamamoto, K., He, Q., Shin, H.J., & von Davier, M. (2017). Developing a machine-supported coding system for constructed-response items in PISA (ETS Research Report No. RR-17-47). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12169