PISA 2022 Technical Report

15 Coding Design, Coding Process, and Reliability Studies

Introduction

The PISA 2022 assessment consisted of both constructed-response (CR) and multiple-choice items (MC). MC items could be simple multiple choice, with a single correct response selection, or complex multiple choice, with multiple correct response selections required. MC items had a predefined correct answer that could be computer coded. While a few CR items were designed to be coded by computer, most required a person to read the response and provide a code or score. These items are referred to as human-coded constructed response items.

This chapter describes the design, preparation, and processing of coding human-coded constructed-response (CR) items, and reports the reliability statistics and volume of responses that could be automatically coded for these items. A summary of all test items by domain, item format, and coding method is shown in Table 15.1.

The CBA mathematics assessment was administered within each country/economy as both a linear test and a Multistage Adaptive Test (MSAT). The CBA reading assessment was also administered as an MSAT with three stages, while the science assessment was administered using a linear design. Countries participating in the CBA also had the option of administering a Financial Literacy assessment and a Creative Thinking assessment. One country chose to participate in the paper-based assessment (PBA), which has been administered since 2015, and three countries participated in the new paper-based assessment. More on the PISA 2022 test design is presented in Chapter 2 of this Technical Report.

Coding design

Coding designs for CBA, PBA, and the new PBA were developed to accommodate the various needs of countries/economies in terms of the number of languages assessed, sample size, and assessed domains (i.e., meaning whether Financial Literacy or the innovative domain were to be coded in the country/economy). 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. The number of expected student responses per domain was based on the sample size completing the assessment in each assessed language in the core domains and in the optional domains of Financial Literacy and Creative Thinking.

Table 15.2 shows the number of coders recommended by domain in the CBA coding designs based on the sample size. This design is exclusive by language of assessment. CBA participants were able to determine the appropriate design for their country/economy and language(s) with a coding estimation tool, which estimated the coding workload for each coder (duration of coding and the number of responses to be coded by each coder). Table 15.2 also includes an example of this estimated workload.

PBA and new PBA countries' sample sizes had little variation, and there were no additional domain options; therefore, all countries participating in these assessments were advised to recruit six coders for each domain. Table 15.3 shows the estimated workload for six coders in each domain.

Designs for within-country and across-country scoring 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 (Shin, von Davier and Yamamoto, 2019[1]). Throughout the chapter, we use the term *coding* to refer to the assignment of a numerical value to a student text response, which indicates the type of response provided by the student, and the term *scoring* to refer to the assignment of full credit, partial credit, or no credit, which is derived from the codes applied. Scoring reliability in PISA 2022 was evaluated and reported at both within- and across-country levels.

The purpose of monitoring and evaluating within-country scoring reliability is to ensure accurate scoring of student responses across coders in the same county-by-language group and identify any coding inconsistencies or problems in the scoring process throughout the process so that they can be promptly addressed and resolved. Within-country scoring reliability was evaluated by reviewing the codes assigned by two or more human coders on the same student responses in a process called multiple coding. *Multiple coding* refers to the coding of the same student response data by different independent coders, such that inter-rater agreement statistics can be calculated and evaluated.

It was also important to check the consistency of coders across countries and language groups. Accurate and consistent *scoring* (full credit, partial credit, no credit) within a country does not necessarily mean that coders from all countries and language groups are applying the coding rubric in the same manner. Coding bias may be introduced if, for example, one country codes a certain type of response differently than other countries (Shin, von Davier and Yamamoto, 2019[1]). Across-country scoring reliability was evaluated by checking the correctness of the codes assigned by two bilingual human coders on a set of English anchor responses in a process called anchor coding. *Anchor coding* refers to the coding of a set of common (across-country-by-language groups) responses in English for each item, for which the correct code for each response is already known by the PISA international contractor (but not provided to coders). Because countries 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. For each human-coded CR item, a set of thirty anchor responses in CBA, and ten in PBA and New PBA, were distributed to the designated bilingual coders for coding.

In CBA, item responses were randomly selected from all student responses and gathered into coding sets for multiple coding. In the domains of Mathematics, Science, Financial Literacy, and Creative Thinking, one coding set was compiled, such that all coders contributed to the multiple-coding agreement for all items. In the domain of Reading, items were distributed among four coding sets, such that each coder only saw responses to half of the items and thus contributed only to the scoring reliability for the items in their assigned coding set. Each domain had two bilingual coders – always coders 01 and 03 – who additionally coded thirty anchor responses in English for each item in their coding set. The design for multiple coding for the CBA is shown in Figure 15.1.

For multiple coding in the paper-based designs, student test booklets are first sorted by booklet number. Because each test booklet contains responses from two administered domains (for example, Mathematics and Science were administered in booklets 1-6 in PBA), coding sets are first multiple coded by coders in one administered domain and then single coded by the coders in the other administered domain. A specified number of booklets (52 booklets of each booklet number in PBA and 90 of each number in the new PBA) are designated for multiple coding. These booklets are distributed equally among six coding sets and distributed to coders. In PBA, all coders code all coding sets, whereas a subset of coders code each coding set in new PBA. The PBA and new PBA coding designs are shown in Figure 15.2.

Figure 15.1. Organisation of multiple coding for the CBA designs

Mathematics

		Coder IDs															
	Number of Coded Responses	301 (bilingual)	302	303 (bilingual)	304	305	306	307	308	309	310	311	312	313	314	315	316
Coding Set	128/item	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Anchor Set	30/item	●		●													

Reading

		Coder IDs																																		
	Number of Coded Responses	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
Coding Set 1	100/item	●	●					●	●	●			●	●			●	●			●	●			●	●			●	●			●
Coding Set 2		●	●			●	●			●		●		●		●		●		●		●		●		●		●		●		●	
Coding Set 3				●	●	●	●				●	●			●	●			●	●			●	●			●	●			●	●	
Coding Set 4				●	●			●	●		●		●		●		●		●		●		●		●		●		●		●		●
Anchor Set	30/item	●		●																																

Science

		Coder IDs											
	Number of Coded Responses	101 (bilingual)	102	103 (bilingual)	104	105	106	107	108	109	110	111	112
Coding Set	128/item	●	●	●	●	●	●	●	●	●	●	●	●
Anchor Set	30/item	●		●									

Financial Literacy

		Coder IDs											
	Number of Coded Responses	401 (bilingual)	402	403 (bilingual)	404	405	406	407	408	409	410	411	412
Coding Set	100/item	●	●	●	●	●	●	●	●	●	●	●	●
Anchor Set	30/item	●		●									

Creative Thinking

		Coder IDs																							
	Number of Coded Responses	501 (bilingual)	502	503 (bilingual)	504	505	506	507	508	509	510	511	512	513	514	515	516	517	518	519	520	521	522	523	524
Coding Set	100/item	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Anchor Set	30/item	●		●																					

Figure 15.2. Organisation of multiple coding for the PBA and New PBA standard coding design

PBA

Mathematics

		Number of Coded Responses	Coder IDs: 301 (bilingual)	302	303 (bilingual)	304	305	306
Booklets 7-12 52 of each booklet number (312 total)	Coding Set 13	120-138/item	●	●	●	●	●	●
	Coding Set 14		●	●	●	●	●	●
	Coding Set 15		●	●	●	●	●	●
	Coding Set 16		●	●	●	●	●	●
	Coding Set 17		●	●	●	●	●	●
	Coding Set 18		●	●	●	●	●	●
Booklet 92	Anchor Coding Set	10/item	●		●			

Reading

		Number of Coded Responses	Coder IDs: 201 (bilingual)	202	202 (bilingual)	204	205	206
Booklets 13-18 52 of each booklet number (312 total)	Coding Set 25	104/item	●	●	●	●	●	●
	Coding Set 26		●	●	●	●	●	●
	Coding Set 27		●	●	●	●	●	●
	Coding Set 28		●	●	●	●	●	●
	Coding Set 29		●	●	●	●	●	●
	Coding Set 30		●	●	●	●	●	●
Booklet 91	Anchor Coding Set	10/item	●		●			

Science

		Number of Coded Responses	Coder IDs: 101 (bilingual)	102	102 (bilingual)	104	105	106
Booklets 1-6 52 of each booklet number (312 total)	Coding Set 1	104/item	●	●	●	●	●	●
	Coding Set 2		●	●	●	●	●	●
	Coding Set 3		●	●	●	●	●	●
	Coding Set 4		●	●	●	●	●	●
	Coding Set 5		●	●	●	●	●	●
	Coding Set 6		●	●	●	●	●	●
Booklet 93	Anchor Coding Set	10/item	●		●			

New PBA

Mathematics

		Number of Coded Responses	Coder IDs: 301 (bilingual)	302	303 (bilingual)	304	305	306
Booklets 5-8 90 of each booklet number (360 total)	Coding Set 13	180/item	●	●	●		●	
	Coding Set 14			●	●	●		●
	Coding Set 15		●		●	●	●	
	Coding Set 16			●		●	●	●
	Coding Set 17		●		●		●	●
	Coding Set 18		●	●		●		●
Booklet 93	Anchor Coding Set	10/item	●		●			

Reading

		Number of Coded Responses	Coder IDs: 201 (bilingual)	202	202 (bilingual)	204	205	206
Booklets 1-4 90 of each booklet number (360 total)	Coding Set 1	180/item	●	●	●		●	
	Coding Set 2			●	●	●		●
	Coding Set 3		●		●	●	●	
	Coding Set 4			●		●	●	●
	Coding Set 5		●		●		●	●
	Coding Set 6		●	●		●		●
Booklet 92	Anchor Coding Set	10/item	●		●			

Science

		Number of Coded Responses	Coder IDs: 101 (bilingual)	102	102 (bilingual)	104	105	106
Booklets 9-12 90 of each booklet number (360 total)	Coding Set 25	180/item	●	●	●		●	
	Coding Set 26			●	●	●		●
	Coding Set 27		●		●	●	●	
	Coding Set 28			●		●	●	●
	Coding Set 29		●		●		●	●
	Coding Set 30		●	●		●		●
Booklet 91	Anchor Coding Set	10/item	●		●			

Coding preparation

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

Recruitment of national coder teams

The first task of National Project Managers (NPMs) on the coding workflow was to assemble a national coder team. The size of the coding teams varied in each country, but the following criteria were used for selecting members of the team:

- All coders should have more than an upper secondary education qualification (i.e., high school degree); university graduates were preferred.
- All should have a good understanding of secondary education 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 in the language(s) of the assessment.

After the national coding team was assembled, the next task was to identify a *lead coder* who was part of the coding team but also responsible for the following tasks:

- training coders within the country,
- 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 or required further investigation,
- Producing reliability reports
- retraining or replacing coders if necessary, and
- consulting with the international experts if item-specific issues arose.

Additionally, the lead coder was required to be proficient in English, as international trainings and interactions with the PISA international contractors were in English only, and was encouraged to attend the international coder trainings. 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.

Coder training materials

Detailed coding guides were developed for all the new items in the domains of Mathematics, Financial Literacy, and Creative Thinking. These coding guides included coding rubrics for each item and example responses corresponding to each level (i.e., correct, partially correct, and incorrect) of the rubric. Coding rubrics for new items were revised for the main survey based on information learned from the field trial. Coding guides for trend domains were also prepared, but changes were limited to the correction of errors.

In addition to the coding guides, a separate workshop-materials file was either created for new domains or updated for trend domains. Unlike the coding guides which remain relatively static across cycles, the workshop-materials file can be updated. The workshop materials files contain additional example responses and annotations, which could be used to supplement the coder trainings. The additional example responses better illustrate the depth and breadth of the coding levels, and the lines between levels. Following the international trainings, final versions of all materials were prepared and released to participating countries/economies.

International coder trainings

Prior to the field trial, NPMs and lead coders were provided with a full item-by-item coder training for CBA, PBA, and new PBA participants in Athens, Greece in January 2020. The field trial training covered all items in all domains. Due to the one-year delay caused by the COVID-19 pandemic, a second international field trial training was held in January and February 2021. The second field trial training took place over several sessions and was conducted virtually. Additionally, the second field trial training covered only new material. That is, the sessions offered were for the new Mathematics items, all Creative Thinking items, and the four new Financial Literacy CR items.

Prior to the main survey, international coder trainings were held in January and February 2022, and were again conducted virtually for all domains. Full trainings were offered for all the new and all the trend Mathematics items, all the Creative Thinking items, and the new Financial Literacy items. Targeted

trainings were offered for the trend domains (Science, Reading, and the trend items in Financial Literacy); that is, the international experts reviewed analysis results from the field trial and considered items that have been historically challenging to code, and targeted items for which a refresher training would be most beneficial. Participants were also given the opportunity to ask questions about trend items if they were not already on the list prepared by the experts. Participants were also provided with the recorded trainings that were prepared for PISA 2018, which cover the items in the trend domains, and could be used to supplement the targeted virtual trainings.

Full trainings were provided virtually in April and May 2022 for PBA and New PBA participants for all domains. During these trainings, the coding guides were presented and explained, and participants had the opportunity to ask questions to have the coding rubrics clarified. Participants also practiced coding on sample responses and discussed any ambiguous or problematic situations as a group. When the discussion revealed areas where rubrics could be improved, those changes were noted and eventually implemented in an updated version of the coding guide that was made available after the meeting. The workshop-materials files were also updated as needed following the international trainings.

To support the national teams during the coding process, a coding query service was offered, which allowed national teams to submit coding questions and receive responses from the relevant domain experts. National teams were also able to see questions submitted by other countries/economies pertaining to the coding of new items, 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 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 from 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 source manuals and materials provided by the PISA international contractors. Coding teams were asked to facilitate discussion about any items that were challenging to code. Past experience has shown that when coders discuss items among themselves and with their lead coder, many issues can be resolved, and more consistent coding can be achieved.

The National Centres were responsible for organising training and coding. The recommended approach was to train at the item level. Under this approach, coders were fully trained on the coding rules for one item, and then proceeded with coding all responses for that one item. Once the item was fully coded, training was provided for the next item (blocked by unit), and so on. The approach of coding item by item has been shown to improve reliability by helping coders to apply the scoring rubric more consistently.

For PBA and new PBA participants, coder training was also recommended at the item level; however, training could be given at the unit level. Once the training was complete on the items within a single unit, coding could take place across booklet for all the items within that one unit.

Coding procedures

Since PISA 2015, coding CBA item responses has been facilitated through use of the Open-Ended Coding System (OECS), which allows coders to view student responses, defer responses for further review, and code responses directly in the system interface. The OECS supported coding teams in their work to code the CBA responses while ensuring that the coding design was appropriately implemented. Especially important during the COVID-19 pandemic, the OECS afforded coders the ability to work remotely. Detailed information about the system was included in the OECS manual provided to countries/economies.

Computer-based responses were coded on an item-by-item basis. For each item, coders receive a set of responses to be coded. Each set includes 1) student responses to be multiple coded as part of the within-country reliability monitoring process, and 2) student responses to be single coded. If the coder is one of the national team's two bilingual coders they also received anchor responses in English will also be included for across-country reliability monitoring. Because the generation of inter-rater agreement statistics were continuously being updated by the OECS as coders code (see Formula 13.1 in the *Reliability Studies* section), no pause in coding is required in the CBA to manually calculate these statistics, allowing coders to work at their own pace through all assigned responses.

When a coder logs into the system and selects an item to code, responses that require human coding appear on screen. Buttons at the top of the screen allow the coder to scroll through responses. In general, multiple-coded responses are populated first and then single-coded responses; for bilingual coders, anchor responses appear ahead of all student responses. For each response, the OECS displays the item stem or question, the individual response, and the available codes for the item, as well as a checkbox to *defer* the response to the lead coder and a checkbox to indicate that the response has been *recoded* from the originally applied code to a new code for some reason. It is expected that coders will 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 responses for all items were coded. Detailed information about the system was provided in the OECS manual.

For the PBA and New PBA, the coding designs were supported by the Data Management Expert (DME) system, and reliability was monitored through the Open-Ended Reporting System (OERS), an additional software that worked in conjunction with the DME to evaluate and report reliability for CR items. The coding process for paper-based participants involved using the actual paper booklets, with sections of some booklets single-coded and some sections coded multiple times. When a response is single coded, coders mark directly in the booklets. When a response is coded multiple times, only 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 an accurate evaluation of scoring reliability. Detailed information about the system was provided in the OERS manual to PBA countries/economies.

Unlike coding in the CBA, the process of coding in PBA and New PBA does require a pause between coding different sets of responses (anchor responses, multiple-coded responses, and single-coded responses), resulting in three distinct coding phases. In the first phase of coding, bilingual coders code the anchor responses, enter the data into the project database using the DME and evaluate the across-country scoring reliability using the OERS. In the second phase, at least 100 student responses for each item are multiple coded. Single-coded responses are addressed in the final phase. All anchor- and 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 distributions of single codes are also reviewed in the OERS, as a quality check.

National Centres used the output reports generated by the OECS and OERS to monitor irregularities and deviations in the coding process. The OECS and OERS generate the following reports of scoring reliability: i) percentage of first-digit code agreement on multiple and anchor coded responses and ii) coding category distribution across coders. NPMs were instructed to investigate whether a systematic pattern of irregularities existed and if the observed pattern was 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 to be retrained and all responses from a particular item or a particular coder were to be recoded, including those codes that showed agreement. Coding inconsistencies usually come from a misunderstanding of the general coding guidelines and/or a rubric for a particular item. Reliability studies conducted by the PISA contractors also made use of the OECS and OERS reports submitted by National Centres.

Reliability studies

Careful monitoring of scoring 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 for both items and individual coders. Through these processes of 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.

Within-country monitoring of scoring reliability

While coding was ongoing, score agreement and coding category distribution were the main indicators used by National Centres for monitoring coding.

- *Score agreement* refers to the proportion of scores (generally the first digit of assigned codes, denoting full, partial, and no credit) from one coder that exactly matched the scores of other coders on an identical set of multiple-coded responses for an item (including scores on partial credit item responses). Agreement can vary from 0 (0% agreement) to 1 (100% agreement). Each country/economy was expected to meet a scoring *standard* within-country and across-country proportion of at least 85% agreement on each item or coder in Mathematics, Reading, Science, and Financial Literacy; this standard was set to 70% agreement for Creative Thinking. Further, an average domain-level standard across all items in a domain of 92% was expected, except for Creative Thinking, which was also set to 70%. The design called for a minimum of one-hundred responses for each item to be multiple coded for the calculation of within-country score agreement; when fewer than 100 responses for an item in a particular country-by-language group were collected, as was the case of small samples, all responses were multiple coded. Additionally, ten (paper-based) or thirty (computer-based) English responses for each item were anchor 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.

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

Formula 15.1

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

where R_{ii} is the calculated agreement rate for coder C_j for item i, N is the total number of responses for item i, A is the number of automatically coded responses for item i, C is number of coders for the item, G_{ji} is the number of agreed codes for coder C_j for item i (max = (C-1)), and D_{ji} is the number of multiple-coded responses for item i coded by coder j so far (at the end of coding, this will equal 100 in a standard sample). The OERS reports calculated agreement similarly, with the exception that no responses were automatically coded (so, A = 0, simplifying the equation).

Score agreement across countries/economies, languages, and items

Scoring reliability was again reviewed by the PISA contractor following the completion of coding to check for scoring consistency of human-coded CR items within and across countries participating in PISA 2022. For comparability among country-by-language groups and between multiple-coded student responses and anchor-coded English responses, the proportion of automatically coded responses were disregarded, and only the scoring reliability of human coders was considered. The reliability studies included 78 CBA countries/economies, resulting in 124 country-by-language groups. One PBA country, and three new PBA countries, each with one language group. In total there were 128 country-by-language groups across modes of assessment.

In a review of country-level data, quality and consistency of score agreement within and across country-by-language groups was evaluated. High score agreement is generally reflective of quality coding: that national and international coder trainings were well-implemented, coding guides were reflective of the student responses, such that scores could be consistently applied, and the scores applied on human-coded CR items are reliably accurate. All country-by-language groups were reviewed to see if the score agreement standard was met on all items and domains. Table 15.4, Table 15.5 and Table 15.6 report the domain-level score agreement for all PISA 2022 participating countries and economies.

Overall, the majority of country-by-language groups administering the CBA met the domain-level within-country score agreement standard of 92% (or 70% in Creative Thinking):

- In Mathematics, 98.4% of country-by-language groups met the domain-level scoring standard; those below averaged 91.0% score agreement on multiple-coded responses.
- In Reading, 89.5% of country-by-language groups met the domain-level scoring standard; those below averaged 90.7% score agreement on multiple-coded responses.
- In Science, 73.4% of country-by-language groups met the domain-level scoring standard; those below averaged 90.7% score agreement on multiple-coded responses.
- In Financial Literacy, 86.7% of country-by-language groups met the domain-level scoring standard; those below averaged 91.4% score agreement on multiple-coded responses.
- In Creative Thinking, 97.0% of country-by-language groups met the domain-level scoring standard; those below averaged 69.6% score agreement on multiple-coded responses.

Note that in some cases, 100% score agreement was observed in certain country-by-language groups. This is more likely to occur when the number of responses being multiple coded is fewer than the recommended 100 student responses, usually due to a small sample size.

Quality in the coding of the English anchor responses is also important for ensuring that the coding guides have applied in the same way across countries/economies and language groups. Most country-by-language groups administering the CBA also met the relevant domain-level across-country score agreement standard of 85% (or 70% in Creative Thinking):

- In Mathematics, 92.7% of country-by-language groups met the domain-level scoring standard; those below averaged 86.0% score agreement on 30 anchor responses.
- In Reading, 87.1% of country-by-language groups met the domain-level scoring standard; those below averaged 86.0% score agreement on 30 anchor responses.
- In Science, 68.5% of country-by-language groups met the domain-level scoring standard; those below averaged 88.4% score agreement on 30 anchor responses.
- In Financial Literacy, 83.3% of country-by-language groups met the scoring standard; those below averaged 90.0% score agreement on 30 anchor responses.
- In Creative Thinking, 97.0% of country-by-language groups met the scoring standard; those below averaged 63.7% score agreement on 30 anchor responses.

Finally, all paper-based countries met the standard for across- and within-country score agreement all domains.

Table 15.7 summarizes Table 15.4, Table 15.5 and Table 15.6, providing an overall breakdown of score agreement of items by domain.

Across most domains and modes of assessment, across-country score agreement tended to be slightly lower than the within-country agreement by domain in the majority of country-by-language groups. This may be expected because, compared to multiple-coding, there are fewer bilingual coders (only two from the coding team) and fewer anchor-coded responses contributing to the calculation of agreement. However, the difference between multiple-coding and anchor-coding agreement by domain is generally minimal across country-by-language groups. In the domains of Mathematics, Reading, Science, and Financial Literacy, there was about 1-3% difference between the within-country agreement and the across-country agreement at the domain level in country-by-language groups, with only a few exceptions. In Creative Thinking, the domain level difference in agreement was closer to 7%, but with a lower threshold for standard of agreement, there is more room for fluctuation in agreement statistics, so this can also be expected.

Coder-level score agreement

Coder quality was also reviewed, particularly the percentage of coders in a country-by-language group that did not meet the standard level of agreement on across several items. Table 15.8 and Table 15.9 summarize overall coder quality and the impact of coder quality by item. In general, the coding standard indicates that all coders should agree with their fellow coders at least 85% of the time on each item, except in Creative Thinking, in which 70% score agreement was considered acceptable. Table 15.8 shows the percentage of coders who were unable to reach the 85% agreement threshold on 20% or more items assigned to them. In Mathematics, 2.5% of coders agreed with their fellow coders less than 85% of the time on at least 20% of new item responses selected for multiple coding, and 0.8% were below this standard for trend item responses; this was also true of 8.8% of coders in Reading, 3.1% of coders in Science, and 0.7% of coders in Financial Literacy. In Creative Thinking, 15.7% of coders agreed with their fellow coders less than 70% of the time on at least 20% of responses.

Because coder quality is reflected at the item level, the percentage of items in the domain over which two or more coders did not meet the standard level agreement on that item was also evaluated, and the results are presented in Table 15.9. Because there are a varying number of items in each domain, this table expresses the percentage of cases across all country-by-language groups. In other terms, however, about half of the CBA country-by-language groups had one new Mathematics item for which two coders did not meet the established 85% score agreement, and a fraction of that had this issue with a trend Mathematics item. Most country-by-language groups would have had about two reading items for which at least two coders did not meet the scoring standard, and in science, one item. About a third of all groups administering financial literacy would have had two coders below the standard on one item, and in Creative Thinking, all groups would have had about two items for which two or more coders did not reach 70% score agreement. There were no items across the paper-based domains for which there were two or more coders below the standard of score agreement on an item. These results overall suggest that any significant coding issues that may have arisen during coding were resolved at the National Centres.

Item-level agreement

The scales on which the PISA statistical framework is built are only as good as the scores used to establish them, so the overall agreement on student responses was also reviewed at the item-level, taking into account the proportion of responses that could be automatically coded. Here, the interest is to determine the proportion of items in a country-by-language group that did not meet the standard level of agreement.

Again, at the item level, the score agreement standard was set to 85% in all domains and modes of assessment, except for Creative Thinking, in which the standard was set to 70% score agreement. These standards were met for most items in each country-by-language group. Table 15.10 shows the number of country-by-language groups that had either no items in a domain ($n = 0$) below the standard, between one and five items ($1 \leq n \leq 5$), or up to ten items ($6 \leq n \leq 10$) in a domain below the score agreement standard. In the paper administration, all countries met the scoring standard on all items in all three domains. In the computer administration, all country-by-language groups met the standard on all items in Science and Financial Literacy. In Mathematics, one country-by-language group had items that failed to meet the standard, in Reading, four groups, and in Creative Thinking, five country-by-language groups had 1-5 items below the standard, and two had 6-10 items below the standard.

Machine-supported coding system

During the 2022 cycle, the CBA coding teams were able to benefit from the use of a machine-supported coding system (MSCS). The MSCS operates effectively due to a high response regularity among collected student data. Consider that, although an item's response field is open-ended, there is a commonality among students' raw responses, meaning that the same or similar correct or incorrect responses can be expected regularly throughout coding (Yamamoto et al., 2017[2]; 2018[3]). 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, greatly reducing the human coding burden and minimizing the error present in human-coded data, often associated with fatigue or carelessness.

Unlike commonly used automated scoring systems that generally involve algorithms, the MSCS relies entirely on text data that have already been human coded in past PISA cycles and during the field trial. These observed text responses and their associated verified codes from past administrations are stored in a *Coded Unique Response* (CUR) pool for each country-by-language group. In order for a text response to receive a verified code and be added to the CUR pool, the response must have appeared at least five times, and coders must have 100% agreement on the code to apply. The MSCS approach parallels automated scoring in the sense that a scoring model is first trained on existing historic data (2015 and 2018 PISA cycles and the 2022 field trial) and then applied to future data (2022 main survey). When raw student responses are received, and before they are distributed to human coders in the OECS, they are first checked to see if the MSCS can automatically apply a code. Raw responses fall into one of three categories: 1) nonresponse, 2) responses with verified coding in the CUR pool, and 3) infrequent or unseen responses that require human judgment. The MSCS can be applied to the first two categories. Human coding would only be required for unique, unseen responses (3). The MSCS is specific to each country-by-language group; responses that are identified for automatic coding are not shared among country-by-language groups. In brief, the MSCS identifies blank responses and the exact same responses that have been previously coded by humans and automatically applies the appropriate code, minimizing the need to score responses that have already been added to the database (Yamamoto et al., 2017[2]; 2018[3]; OECD, 2018[4]).

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

Table 15.11 and Table 15.12 summarize the efficiency of the MSCS with the reduction of human-coding burden in the PISA 2022 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, *CUR* and *nonresponse*, indicate the average and median percentage of responses across CBA items that were automatically coded by the MSCS as either a nonresponse or a verified response (correct – full and partial credit – 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 would have had to be coded by humans, including nonresponses. On average, across items and country-by-language groups, the coding burden for human coders was reduced for the 2022 field trial from a low of approximately 14% on new Mathematics items to a high of 31% on trend Mathematics items. For the 2022 main survey, the coding burden was reduced by a low of approximately 7% in Creative Thinking, for which only nonresponses were coded by the MSCS, to a high of 35% (about 15% CUR and 20% nonresponse) on trend Mathematics items.

For both field trial and main survey, approximately 7% to 20% of the total responses (on average) across all domains were empty responses and were automatically coded by the system. On new items, where no historic data were available, the MSCS reduced coding burden for human coders by 12% to 14% in Mathematics and 7% to 15% in Creative Thinking. For new Mathematics items that received modification following the field trial, only empty responses were automatically coded during the main survey, which may explain why only 5% of new Mathematics responses were automatically coded through the CUR pool. Because of the format and generally graphical nature of the Creative Thinking domain, only empty responses were automatically coded by the MSCS in both the field trial and the main survey. Overall, a similar or slightly higher percentage of responses were coded in each of the core domains in the 2022 PISA cycle than in the previous cycle.

References

OECD (2018), *PISA 2018 Technical Report*, PISA, OECD Publishing, Paris, https://www.oecd.org/pisa/data/pisa2018technicalreport/. [4]

Shin, H., M. von Davier and K. Yamamoto (2019), “Investigating Rater Effects in International Large-Scale Assessments”, in Veldkamp, B. and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement: Methodology of Educational Measurement and Assessment*, Springer, Cham. [1]

Yamamoto, K. et al. (2018), “Development and implementation of a machine-supported coding system for constructed-response items in PISA”, *Psychological Test and Assessment Modeling*, Vol. 60/2, pp. 145-164. [3]

Yamamoto, K. et al. (2017), “Developing a machine-supported coding system for constructed-response items in PISA”*, ETS Research Report*, No. RR-17-47, Educational Testing Service, Princeton, NJ, https://doi.org/10.1002/ets2.121. [2]

Chapter 15 tables

Table 15.1. Number of cognitive items by domain, item format, and coding method

			Mathematics (New)	Mathematics (Trend)	Reading	Science	Financial Literacy	Creative Thinking
CBA	Human Coded	Constructed Response	19	16	64	32	16	34
	Computer Scored	Simple Multiple Choice	80	18	104	33	12	0
		Complex Multiple Choice	35	14	27	47	14	2
		Constructed Response	26	26	2	3	4	0
		Total	**160**	**74**	**197**	**115**	**46**	**36**
PBA	Human Coded	Constructed Response		38	51	32		
	Computer Scored	Simple Multiple Choice		18	27	29		
		Complex Multiple Choice		12	9	24		
		Constructed Response		3	0	0		
		Total		**71**	**87**	**85**		
New PBA	Human Coded	Constructed Response		40	37	9		
	Computer Scored	Simple Multiple Choice		16	24	34		
		Complex Multiple Choice		8	5	23		
		Constructed Response		0	0	0		
		Total		**64**	**66**	**66**		

Table 15.2. CBA coding number of coders by domain

	Recommended Number of Coders by Number of Students Assessed				Example Workload*		
	< 4,500	4,501 – 8,000	8,001 – 13,000	> 13,000	Coders	Expected Coding Days	Responses per Coder
Mathematics	2 – 3	4 – 5	6 – 9	10 – 12	8	7.1	6,853
Reading	2 or 4	4 or 8	8 or 12	12 – 32	8	7.4	5,194
Science	2 – 3	4 – 5	6 – 9	10 – 12	8	6.3	6,107
Financial Literacy	2 – 3	4 – 5	6 – 9	10 – 12	8	4.8	4,691
Creative Thinking	2 – 3	4 – 5	6 – 9	10 – 24	8	8.9	5,974

Note: Example assumes a main sample size of 6 300 and a Financial Literacy sample size of 1 650.
Example assumes that coders in the core domains and Financial Literacy would be able to code approximately 1 000 responses per day, and coders in the Creative Thinking domain would be able to code approximately 700 responses per day.

Table 15.3. PBA and New PBA number of coders by domain

		Example Workload		
		Coders	Expected Coding Days	Responses per Coder
PBA	Mathematics	6	14	10,418
	Reading	6	15	7,733
	Science	6	9	3,328
New PBA	Mathematics	6	17	11,667
	Reading	6	16	10,500
	Science	6	4	2,625

Note: Example assumes that coders would be able to code approximately 1 000 responses per day.

Table 15.4. Summary of within- and across-country (%) scoring agreement for CBA participants for reading, mathematics and science

Please refer to Excel file Chapter_15_Tables.xlsx on line for this table.

Table 15.5. Summary of within- and across-country (%) agreement for Financial Literacy and Creative Thinking domains

		Country/Economy - Language	Within-country		Across-country	
			Financial Literacy	Creative Thinking	Financial Literacy	Creative Thinking
OECD		Australia - English		74.6%		88.7%
		Austria - German	93.0%		93.5%	
		Belgium - Dutch	96.6%	84.2%	96.6%	91.2%
		Belgium - French		76.8%		87.5%
		Belgium - German		76.8%		88.9%
		Canada - English	91.4%	76.2%	95.9%	91.0%
		Canada - French	92.6%	75.8%	94.6%	87.7%
		Chile - Spanish		76.7%		87.5%
		Colombia - Spanish		81.9%		86.7%
		Czech Republic - Czech	94.9%	89.7%	96.5%	91.7%
		Denmark - Danish	94.1%	86.8%	95.6%	90.6%
		Denmark - Faroese		98.1%		92.1%
		Estonia - Estonian		83.9%		96.9%
		Estonia - Russian		78.1%		86.6%

		Country/Economy - Language	Within-country		Across-country	
			Financial Literacy	Creative Thinking	Financial Literacy	Creative Thinking
		Finland - Finnish		83.0%		92.9%
		Finland - Swedish		96.2%		93.9%
		France - French		84.1%		90.7%
		Germany - German		86.0%		88.7%
		Greece - Greek		80.6%		87.4%
		Hungary - Hungarian	92.5%	80.7%	95.5%	92.7%
		Iceland - Icelandic		80.3%		93.0%
		Israel - Arabic		83.2%		92.9%
		Israel - Hebrew		80.6%		93.0%
		Italy - German	95.6%	82.2%	93.4%	94.2%
		Italy - Italian	92.8%	91.1%	95.0%	96.4%
		Korea - Korean		85.7%		87.2%
		Latvia - Latvian		86.1%		85.8%
		Latvia - Russian		86.4%		88.9%
		Lithuania - Lithuanian		90.4%		95.6%
		Lithuania - Polish		89.0%		93.0%
		Lithuania - Russian		89.7%		95.0%
		Mexico - Spanish		80.7%		83.6%
		Netherlands - Dutch	92.0%	77.0%	92.2%	89.4%
		New Zealand - English		80.3%		91.0%
		Norway - Bokmål	95.4%		97.1%	
		Norway - Nynorsk	96.3%		97.3%	
		Poland - Polish	93.3%	79.5%	95.3%	91.0%
		Portugal - Portuguese	90.9%	81.4%	93.7%	85.4%
		Slovak Republic - Hungarian		98.1%		90.4%
		Slovak Republic - Slovak		87.3%		88.9%
		Slovenia - Slovenian		85.1%		89.2%
		Spain - Basque	95.0%	69.9%	87.4%	79.8%
		Spain - Catalan	92.6%	75.6%	93.3%	82.3%
		Spain - Galician	92.4%	72.5%	92.4%	81.0%
		Spain - Spanish	92.1%	69.0%	91.1%	83.6%
	*	Spain - Valencian	100.0%	82.8%	93.9%	83.6%
		United States - English	95.4%		97.7%	
Partners		Albania - Albanian		93.2%		83.0%
		Baku (Azerbaijan) - Azeri		76.6%		68.8%
		Baku (Azerbaijan) - Russian		77.3%		74.8%
		Brazil - Portuguese	99.3%	86.8%	98.8%	95.7%
		Brunei Darussalam - English		76.3%		87.8%
		Bulgaria - Bulgarian	91.3%	81.0%	96.4%	89.2%
		Chinese Taipei - Chinese		79.1%		86.0%
		Costa Rica - Spanish	93.1%	80.0%	94.3%	82.6%
		Croatia - Croatian		94.8%		85.8%
		Cyprus - English		87.6%		90.4%
		Cyprus - Greek		78.0%		87.8%
		Dominican Republic - Spanish		90.7%		62.7%
		El Salvador - Spanish		84.8%		77.2%
		Hong Kong (China) - Chinese		94.0%		95.1%
		Hong Kong (China) - English		97.2%		94.1%
		Indonesia - Indonesian		83.4%		84.7%
		Jamaica - English		69.8%		86.0%
		Jordan - Arabic		83.3%		83.4%
		Kazakhstan - Kazakh		83.1%		89.2%
		Kazakhstan - Russian		87.0%		89.7%
		Macao (China) - Chinese		93.2%		93.6%

		Country/Economy - Language	Within-country		Across-country	
			Financial Literacy	Creative Thinking	Financial Literacy	Creative Thinking
		Macao (China) - English		91.8%		93.6%
	*	Macao (China) - Portuguese		100.0%		84.9%
		Malaysia - English	94.4%	76.2%	97.8%	86.6%
		Malaysia - Malay	92.9%	81.0%	98.0%	86.8%
		Malta - English		77.0%		87.6%
		Malta - Maltese		81.4%		88.1%
		Mongolia - Mongolian		81.8%		86.2%
		Morocco - Arabic		81.7%		88.6%
		Morocco - French		82.9%		87.4%
		North Macedonia - Macedonian		86.3%		89.3%
		Palestinian Authority - Arabic		95.9%		86.2%
	*	Palestinian Authority - English		98.5%		86.5%
	*	Panama - English		95.2%		77.8%
		Panama - Spanish		97.1%		59.5%
		Peru - Spanish	96.2%	87.5%	96.1%	91.4%
		Philippines - English		87.8%		90.9%
		Qatar - Arabic		93.2%		86.9%
		Qatar - English		100.0%		87.4%
		Republic of Moldova - Romanian		93.3%		100.0%
		Republic of Moldova - Russian		92.7%		99.7%
		Romania - Hungarian		79.4%		85.8%
		Romania - Romanian		81.0%		88.2%
		Saudi Arabia - Arabic	91.8%	83.7%	90.8%	88.5%
		Saudi Arabia - English	92.0%	97.2%	92.3%	90.0%
	*	Serbia - Hungarian		89.7%		90.7%
		Serbia - Serbian		85.7%		90.8%
		Singapore - English		86.7%		91.3%
		Thailand - Thai		92.0%		93.1%
	*	Ukraine - Russian		100.0%		87.7%
		Ukraine - Ukranian		82.0%		90.0%
		United Arab Emirates - Arabic	93.1%	79.1%	90.3%	79.5%
		United Arab Emirates - English	92.8%	78.3%	90.3%	81.0%
		Uruguay - Spanish		80.6%		92.5%
	*	Uzbekistan - Karakalpak		88.2%		84.7%
		Uzbekistan - Russian		91.8%		91.1%
		Uzbekistan - Uzbek		88.6%		87.0%

* Denotes a country-language group which assessed fewer than 200 students; therefore, there are fewer multiple coded responses contributing to the calculation of agreement in these groups.
Note: Originally assigned codes for Creative Thinking were rescored for some items during scaling; agreement in this table reflects the original human scoring.

Table 15.6. Summary of within- and across-country (%) scoring agreement for Paper-based countries

	Country/Economy - Language	Within-country Agreement			Across-country Agreement		
		Mathematics	Reading	Science	Mathematics	Reading	Science
Partners	Guatemala - Spanish	99.9%	99.8%	99.1%	98.5%	97.5%	96.5%
	Cambodia - Khmer	99.5%	99.5%	99.1%	99.6%	99.2%	99.3%
	Paraguay - Spanish	99.3%	97.5%	97.4%	99.0%	97.3%	97.2%
	Viet Nam - Vietnamese	99.9%	99.3%	99.7%	98.2%	94.0%	96.6%

Table 15.7. Average item-level score agreement (across country-language groups) by domain

		Mathematics (New)	Mathematics (Trend)	Reading	Science	Financial Literacy	Creative Thinking
CBA	Multiple-coded	95.4%	97.5%	95.4%	94.0%	93.9%	85.0%
	Anchor	93.8%	97.6%	94.8%	92.7%	94.4%	87.9%
PBA and New PBA	Multiple-coded		99.7%	99.0%	98.8%		
	Anchor		98.8%	97.0%	97.4%		

Table 15.8. Percentage of coders whose soring was below the standard inter-rater agreement on 20% or more of items, averaged across countries

	Mathematics (New)	Mathematics (Trend)	Reading	Science	Financial Literacy	Creative Thinking
CBA	2.5%	0.8%	8.8%	3.1%	0.7%	15.7%
PBA and New PBA		0.0%	0.0%	0.0%		

Note The standard is set to 85% agreement in mathematics, science, reading, and financial literacy; in Creative Thinking, it is set to 70% agreement.

Table 15.9. Percentage of items in a domain with at least two coders below the standard scoring agreement on the item (in the same country-by-language group)

	Mathematics (New)	Mathematics (Trend)	Reading	Science	Financial Literacy	Creative Thinking
CBA	3.0%	1.1%	4.2%	4.2%	2.1%	7.5%
PBA and N-PBA		0.0%	0.0%	0.0%		

Table 15.10. Number of country-language groups with score agreement below the domain standard

	N Items below the Standard	Mathematics (New)	Mathematics (Trend)	Science	Reading	Financial Literacy	Creative Thinking
CBA	$N = 0$	123	123	124	120	30	93
	$1 \leq N \leq 5$	1	1	0	2	0	5
	$6 \leq N \leq 10$	0	0	0	2	0	2
PBA and New PBA	$N = 0$		4	4	4		
	$1 \leq N \leq 5$		0	0	0		
	$6 \leq N \leq 10$		0	0	0		

Note: The standard is set to 85% agreement in Mathematics, Science, Reading, and Financial Literacy and 70% in Creative Thinking.

Table 15.11. Percentage of responses coded by the MSCS and by human coders across countries in the 2022 field trial

		Human Coded	Machine Coded		
			CUR	Nonresponse	Total
Mathematics (New)	Mean	85.81%	NA	14.19%	14.19%
	Median	93.85%	NA	6.15%	6.15%
Mathematics (Trend)	Mean	69.04%	11.94%	19.02%	30.96%
	Median	73.86%	1.59%	16.11%	26.14%
Reading	Mean	71.06%	10.70%	18.23%	28.94%
	Median	77.78%	0.00%	13.51%	22.22%
Science	Mean	78.44%	8.65%	12.90%	21.56%

		Human Coded	Machine Coded		
			CUR	Nonresponse	Total
	Median	81.48%	0.00%	8.44%	18.52%
Financial Literacy	Mean	84.42%	0.98%	14.60%	15.58%
	Median	87.09%	0.00%	12.27%	12.91%
Creative Thinking	Mean	85.23%	NA	14.77%	14.77%
	Median	89.91%	NA	10.09%	10.09%

Note: Mean values are the mean of the total percentage of responses coded by the MSCS within each domain across countries; median values are the median of those percentages across countries and, therefore, may not add up to 100%.
Note: CUR pool responses were not available for new items in Mathematics, Financial Literacy, or Creative Thinking.

Table 15.12. Percentage of responses coded by the MSCS and by human coders across countries in the 2022 main survey

		Human Coded	Machine Coded		
			CUR	Nonresponse	Total
Mathematics (New)	Mean	82.85%	4.76%	12.39%	17.15%
	Median	91.30%	0.00%	6.70%	8.70%
Mathematics (Trend)	Mean	65.22%	14.88%	19.90%	34.78%
	Median	69.53%	4.48%	16.67%	30.47%
Reading	Mean	71.32%	10.68%	18.00%	28.68%
	Median	78.20%	0.43%	13.67%	21.80%
Science	Mean	75.49%	11.42%	13.09%	24.51%
	Median	77.89%	1.70%	8.78%	22.11%
Financial Literacy	Mean	84.77%	2.99%	12.24%	15.23%
	Median	86.19%	0.00%	10.44%	13.81%
Creative Thinking	Mean	92.36%	NA	7.64%	7.64%
	Median	94.37%	NA	5.63%	5.63%

Note: Mean values are the mean of the total percentage of responses coded by the MSCS within each domain across countries; median values are the median of those percentages across countries and, therefore, may not add up to 100%.
Note: CUR pool responses were not available for some new items in Mathematics that had changes following the field trial; CUR pool responses were not applied in Creative Thinking.