Chapter 13: CODING DESIGN, CODING PROCESS, AND CODER RELIABILITY STUDIES

INTRODUCTION

The proficiencies of PISA for Development (PISA-D) respondents were estimated based on their performance on test items administered in the PISA-D assessment. All participating countries used paper-based assessments (PBAs). The majority of items were selected from previous cycles of PISA. The item pool was complemented with existing materials from other surveys, including PISA for Schools, PIAAC (Programme for the International Assessment of Adult Competencies), the STEP Skills Measurement Program, and LAMP (Literacy Assessment and Monitoring Programme).

The PISA-D tests consisted of both multiple-choice and constructed-response items. Multiple-choice items have predefined correct answers that can be automatically scored by machine, while constructed-response items require human coding. The breakdown of all PISA-D 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**

Coding Method	Item Format	Mathematics	Reading	Science	Reading Components
Human	Open Response	23	37	9	0
	Human Coded Simple	17	0	0	0
Automatic	Complex Multiple Choice	8	5	23	0
	Simple Multiple Choice	16	24	34	80
Total		64	66	66	80

Coding reliability is a critical component of PISA-D because it underpins the comparability of test results within and across countries. Coding reliability is established by having the same responses evaluated by different coders (multiple coding), followed by careful monitoring of the results of multiple codings through the use of the Data Management Expert (DME) and the Open-Ended Reporting System (OERS) coding software. These steps are essential quality-assurance procedures that provide evidence of the consistent application of coding rubrics by coders. Multiple coding of a subset of human-coded responses provided data for the calculation of item-level coder reliability (see section titled “Coder reliability studies”). The coding design for the PISA-D assessment included all human-coded items within each country. This chapter describes the coding design, procedures (preparation and training), and coder reliability studies.

CODING DESIGN[1]

PISA-D coding was based on the PISA-D Main Survey Integrated Design. In this design, each booklet contained items from two domains (see Table 13.2, where R=Reading, S=Science, M=Mathematics). The coding design for the PISA-D Main Survey aimed to assure the reliable coding process (within-country coder reliability) and to monitor the comparability of coding process across participating countries (across-country coder reliability). The coding design outlined how booklets should be organised and assigned to the team of coders within the National Centre for coding. National Data Managers (NDMs) needed to rely on aspects of this coding design for data entry procedures and data quality checks, which are described in detail in the Data Management Manual.

The design of multiple coding in the Main Survey integrated design is shown in Tables 13.2 and 13.3. For PISA-D participants, all booklets were organised into three coding groups. Group 1 consisted of booklets 1-4, Group 2 of booklets 5-8, and Group 3 of booklets 9-12. The booklets were then organised into one of 12 coding sets (CS) within each coding group, with 12 sets per coding group, as shown in Table 13.2. While the composition of all coding sets in a coding group was the same, the coding sequences were different. Only the first six coding sets in each group required multiple coding for a specified domain, while the last six coding sets in each group, as well as the clusters for other domains in the first six sets, only required single coding. Table 13.3 presents information about the design.

Table 13.2 **Main Survey integrated design**

	Booklet	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Coding Sets 1-12	1	R1	R2	S1	S2
	2	S2	S3	R2	R3
	3	R3	R4	S3	S4
	4	S4	S1	R4	R1
Coding Sets 13-24	5	S1	S2	M1	M2
	6	M2	M3	S2	S3
	7	S3	S4	M3	M4
	8	M4	M1	S4	S1
Coding Sets 25-36	9	M1	M2	R1	R2
	10	R2	R3	M2	M3
	11	M3	M4	R3	R4
	12	R4	R1	M4	M1

[1] For a better understanding of the PISA-D coding designs, we recommend reading the description of the PISA-D assessment design in Chapter 2 as important background information.

Table 13.3 **Coding set organisation**

Coding Sets											
1	2	3	4	5	6	7	8	9	10	11	12
Reading items multiple coded by four coders, other items single coded.						All items single coded.					
13	14	15	16	17	18	19	20	21	22	23	24
Science items multiple coded by four coders, other items single coded						All items single coded.					
25	26	27	28	29	30	31	32	33	34	35	36
Mathematics items multiple coded by four coders, other items single coded.						All items single coded.					

Also included in the coding design were anchor coding sets of 30 unique responses (CS00-R, CS00-M, and CS00-S) containing anchor booklets (Booklet 101, Science; Booklet 201, Reading; Booklet 301, Mathematics), which were used for coding reliability purposes.

CODING PROCEDURES

The coding designs for all responses for Math, Reading, and Science were supported by the DME system, and coding reliability was monitored through the OERS, a computer tool that works in conjunction with the DME software to evaluate and report reliability for paper-based constructed responses. Detailed information about the system was provided to participating countries in the OERS Manual.

When a booklet was single coded, coders marked directly in the booklet. When a booklet was multiple coded, the first three coders entered their codes into coding sheets (for data entry and monitoring reliability), and the final coder entered codes directly in the booklet. This final code was the actual response used for scoring and recorded in the public use data file.

The OERS worked in concert with the cognitive response database to generate two types of reliability reports: i) proportion agreement between coders, and ii) coding category distributions (the percentage of each code awarded to responses on an item). National centres utilised these OERS output reports to monitor irregularities and deviations in the coding process. Careful monitoring of coding reliability plays an important role in ensuring data quality control. Through coder reliability monitoring, coding inconsistencies or other coding problems within countries can be detected early, allowing action to be taken as soon as possible to correct these inconsistencies. Specifically, NPMs were instructed to investigate whether a systematic pattern of irregularities existed and if the pattern was attributable to a particular coder or item. In addition, they were instructed not to carry out resolution (e.g., changing the coding on individual responses to reach higher coding agreement). Instead, if systematic irregularities were identified, all responses on a particular item or from a particular coder needed to be recoded, regardless of

whether the coded responses showed inter-rater agreement or not. In this assessment round, general inconsistencies or problems were deemed due to a misunderstanding of rubrics for particular items, of scoring guidelines, or to the misuse of OERS. Coder reliability studies also made use of the OERS reports submitted by national centres (see the section on coder reliability studies for more detail).

Coding preparation

Prior to the assessment, national centers completed a number of key activities to prepare for human coding. NPMs were responsible for assembling a team of coders (i.e., separate coding teams for each domain). Their first task was to identify a lead coder who would be part of the coding team and who would additionally be responsible for the following tasks:

a) training coders within the country;

b) organising all materials and distributing them to coders;

c) monitoring the coding process;

d) monitoring the inter-rater reliability and taking action when the coding results were unacceptable and required further investigation;

e) retraining or replacing coders if necessary;

f) consulting with the international experts if item-specific issues arose; and

g) producing reliability reports.

The lead coder was required to be proficient in English (as international training and interactions with the contractors were in English only) and to attend the international coder trainings in Zambia in July 2016 (Field Trial) and in the United States in July 2017 (Main Survey), together with the NPM. In most cases, the lead coders for the Field Trial retained their 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:

a) All coders should have more than a secondary qualification (i.e., high school degree); university graduates were preferred.

b) All coders should have a good understanding of secondary level studies in their respective domains.

c) All coders should be available for the duration of the coding period, which was expected to last two to three weeks.

d) Due to normal attrition rates and unforeseen absences, it was strongly recommended that lead coders train a backup lead coder for their teams.

e) Two coders for each domain had to be bilingual in English and the language of the assessment. Both coded all responses in the anchor coding sets.

International coder training

All human-coded items in the PISA-D assessment had detailed coding guides, which included coding rubrics as well as examples of correct and incorrect responses. Since there was no item development done for PISA-D, the coding guides for the item pool were taken from the original study where the items were used.

Prior to the Field Trial, NPMs and lead coders were provided with a full item-by-item coder training, which covered all items across all domains. In preparation for the Main Survey, NPMs and lead coders were provided with a second round of full item-by-item coder training. During these trainings, the coding guides were presented and explained. Training participants practiced coding sample items and discussed any ambiguous or problematic situations as a group. By focusing on the sample responses that were most challenging to code, lead coders had the opportunity to ask questions and have the coding rubrics clarified as well as possible. When the discussion revealed areas where rubrics were unclear, additional sample responses were discussed and provided in an updated version of the coder training materials that were available after the meeting. All training workshop materials were made available for national training activities. This final set of materials provided to participating countries included presentations from the international trainings, as well as the set of sample responses used during the training, the official coding for each response, and a rationale for why each response was coded as shown.

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

National coder training provided by the National Centres

Each national center was required to develop a training package for its own coders to be used in national coder training. The training package consisted of an overview of PISA-D and its own adapted training manuals based on the manuals and materials used in the international coder training provided by the international PISA-D contractors. Coding teams for each domain were asked to work on the same schedule and in the same location in order to facilitate discussion about any items that proved challenging. Past experience has shown that if coders discuss items among themselves and with their lead coder in peer-to-peer learning, many issues can be resolved in a way that results in more consistent coding. However, each coder was responsible for independently coding the set of responses in the booklets assigned to him or her. Each coder was assigned a unique coder ID that was specific to each domain.

National centers were responsible for organising training and coding using one of the following two approaches (checking with contractors in the case of deviations):

a) Coder training 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 alone. Once coding that item was complete, training was provided for the next item, and so on.

b) Coder training at the "item set" level. In this alternative approach, coders were fully trained on a set of 13 to 18 items grouped by unit. Once the full training was complete, coding took place at the item level. However, to ensure that the coding rules were still fresh in coders' memories, a coding review was recommended before the coding of each item.

CODER RELIABILITY STUDIES

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

The purpose of evaluating **within-country coder reliability** was to ensure reliability among coders within a country and to identify any coding inconsistencies or problems in the scoring process so they could be addressed and resolved early in the coding process. The evaluation of within-country coder reliability was carried out by the multiple coding of sets of student responses—assigning identical student responses to different coders so those responses were coded multiple times within a country. Multiple coding of all student responses in an international large-scale assessment like PISA-D is not economical, so a coding design that combined multiple coding and single coding was utilised to reduce national costs and coder burden. A set of 60 cognitive booklets (120 unique responses) was randomly selected from each assessment booklet (1-12) to be multiple coded (see Table 13.3). The rest of the student cognitive booklets were divided evenly among coding sets to be single coded.

Accurate and consistent scoring within a country does not necessarily mean that coders from all countries are applying the coding rubrics in the same manner. Coding bias may be introduced if one country codes a certain response differently than other countries. Therefore, in addition to within-country coder reliability, it was also important to check the consistency of coders across countries. The evaluation of **across-country coder reliability** was made possible by the multiple coding of a set of **anchor** responses. For each constructed-response item, a set of 30 anchor responses in English was provided. The anchor responses were answers obtained from real students, and the authoritative coding—codes assigned by concensus by a team of content experts—for these responses was not released to countries. Two coders from each country coded the same sets of English anchor responses, in addition to the other student responses assigned to them. Because anchor responses are provided in the same mode as regular assessment responses, a country's coding results on anchor responses can be compared across countries to check for coding administration consistency.

Coder reliabilities were calculated in a form of exact agreements to evaluate coding consistency of human-coded constructed-response items within and across the countries participating in

PISA-D. During coding, responses to the PISA-D open-ended items were classified into four categories: full credit (1 for dichotomous items and 2 for polytomous items), partial credit (1 for polytomous items), no credit (0), not applicable (7), and no response (9). Where partial credit was not applicable, items had only full- or no-credit categories. Also, some items entailed different types of full- and/or partial-credit responses, which were captured using two-digit codes specified in the coding guides. Note that double-digit coding (e.g., 11 and 12 for two types of partial credit scores), not applicable, and no response codes were treated as separate coding categories in calculating the coder reliability. Thus, if the agreements within score levels are considered (i.e., combining two different double codes into one score, such as combining 11 and 12 into one "partial credit" code), the coder reliability becomes higher.

Three types of coder reliabilities were calculated:

1) domain-level proportion agreement

2) item-level proportion agreement

3) coding category distributions of coders on the same item

Proportion agreement at the domain and item levels as well as coding category distribution were the main indicators of coder reliability used in PISA-D.

- *Proportion agreement* refers to the percentage of each coder's coding that matched the other coders' codings on the identical set of multiple-coded responses for an item. It can vary from 0 (0% agreement) to 1 (100% agreement). Each country was expected to have an average within-country proportion agreement of at least 0.92 (92%) across all items, averaged across all coders, with a minimum 85% agreement for each item.
- *Coding category distribution* refers to the aggregation of the distributions of coding categories (such as "full credit," "partial credit," and "no credit") assigned by a coder to two sets of responses: a unique set of 120 responses for multiple coding and responses randomly allocated to the coder for single coding. Notwithstanding that negligible differences of coding categories among coders were tolerated, the coding category distributions between coders were expected to be statistically equivalent based on the standard chi-square distribution due to the random assignment of the single-coded responses.

Domain-level proportion agreement

Except for Senegal in Science, the average within-country coder reliability exceeded 92% in each domain across the seven countries (see Table 13.4). The Mathematics domain had higher average agreement (97.9%) than the other domains. The Reading domain had the second highest agreement (96.8%), while the Science domain was the lowest, with 96.5%.

Across-country coder reliability by domain in PISA-D has a trend similar to that of within-country agreement (see Table 13.5), with the exception of Senegal where average across-country agreement was below 92% for both Reading and Science. The Mathematics domain had higher average agreement (98.7%) than the other domains. The Reading domain had the second highest average agreement (96.5%), and Science had an average agreement of 95.7%.

Table 13.4: **Summary of within-country and across-country agreement (%) per domain for the participants**

Participants		Number of Coders	Within-Country Agreement			Across-Country Agreement		
Country	Language		Math	Reading	Science	Math	Reading	Science
Cambodia	Khmer	6 (Math) 6 (Reading) 6 (Science)	99.9	99.5	100.0	99.9	98.5	98.4
Ecuador	Spanish		98.9	96.7	96.6	99.7	97.9	98.4
Guatemala	Spanish		99.7	98.7	99.3	99.6	97.2	96.1
Honduras	Spanish		96.4	95.2	93.5	98.7	98.0	96.5
Paraguay	Spanish		97.7	96.8	96.7	99.0	96.7	98.2
Senegal	French		94.5	93.3	89.4	94.7	90.0	89.2
Zambia	English		98.0	97.3	99.6	99.1	97.5	93.3
Average			**97.9**	**96.8**	**96.5**	**98.7**	**96.5**	**95.7**
Median			**98.0**	**96.8**	**96.7**	**99.1**	**97.5**	**96.5**

Note: Senegal showed a procedural deviation in terms of the number of coders, which was discussed and accepted as the best alternative when two bilingual coders couldn't participate in reliability coding to fulfil the best practice coding design. Two anchor coders were unavailable for the entire duration of the coding session. As a result, these anchor coders were shadowed by two unilingual coders during training. These unilingual coders replaced the anchor coders in multiple and single coding as coders 7 and 8.

Item-level proportion agreement

At the item level, most countries showed proportion agreement above 85% on almost all items in all domains, for both within- and across-country agreement. There was no item that showed low coder reliability consistently across all countries. This indicates that most of the PISA-D participating countries showed an acceptable level of coder reliability in all domains within country and across countries (see Table 13.5). The only exception was Senegal for the across-country agreement: Senegal showed a large number of items with low coder reliability in Reading (10 of 37 items).

Table 13.5 **Numbers of items with proportion agreement < 85%**

Number of Items with Proportion Agreements < 85%	Within-Country Agreement			Across-Country Agreement		
	Mathematics (40 items)	Reading (37 items)	Science (9 items)	Mathematics (40 items)	Reading (37 items)	Science (9 items)
N = 0	7	6	6	5	6	5
1 ≤ N ≤ 5	0	1	1	2	0	2
6 ≤ N ≤ 10	0	0	0	0	1	0
N > 10	0	0	0	0	0	0

Note: "Items" in the table refers to human-coded constructed-response items.

Coding category distributions

In Mathematics, 11.9% of coders in all countries had significantly different coding category distributions from other coders on more than 20% of items (see Table 13.6). In Reading, it was 16.7%, while in Science, it was 40.5%. Although some of those percentages may appear high, all

participants reached an acceptable level of coder reliability, which is a minimum of 85% for an item and an average of 92% across all items, with the exception of Senegal within both Reading and Science domains for all items. This largely resulted from the different pools of responses upon which coding category distribution and proportion agreement were measured. As mentioned earlier, proportion agreement per item across coders was based only on the unique set of 120 responses for multiple coding, while coding category distributions per item across coders also took into account single coding.

Table 13.6 **Percentage of coders whose coding category distributions on more than 20% of coded items were significantly different from other coders, averaged across all countries**

Mathematics	Reading	Science
11.9%	16.7%	40.5%

Across all participating countries, the percentage of items over which more than two coders' coding category distributions were significantly different from other coders was 11.4% in Mathematics, 18.9% in Reading, and 22.2% in Science (see Table 13.7). Compared to Table 13.6, the percentages in the two tables are similar for both Mathematics and Reading, but much less for Science. This is partially due to the fact that there were only nine human-coded items in Science, which makes it much easier for coders to code more than 20% of items (only two items) with significantly different coding proportions.

Table 13.7 **Percentages of country × item pairs that have more than two coders' coding category distributions significantly different from other coders**

Mathematics	Reading	Science
11.4%	18.9%	22.2%

The scales on which the PISA-D statistical framework are built are only as good as the scores used to establish them. In sum, the results from the coder reliabilities revealed that the coding designs that were tailored to meet every PISA-D participant's specific survey needs and administration of coding procedures were executed as intended. The management of the coding process was reported to be smooth and efficient. While some countries showed less reliable human coding, on average, the participating countries achieved acceptable levels of coder reliability amid the operational challenges of manually managing the paper booklets bundles.