13

Coding Reliability Studies

A substantial proportion of the PISA 2009 items were open-ended and required coding by trained personnel. It was important therefore that PISA implemented procedures which maximised the validity and consistency (both within and between countries) of this coding. Each country coded items on the basis of coding guides prepared by the Consortium (see Chapter 2) using the design described in Chapter 6. Training sessions to train coders from different countries on the use of the coding guides were held prior to both the field trial and the main survey.

This chapter describes the outcomes of three aspects of the coding reliability studies undertaken in conjunction with the field trial and the main survey. These are: *i)* the consistency analyses undertaken with the field trial data to assist the test developers in constructing valid, reliable scoring rubrics and to inform national centres about within-country coder reliability, *ii)* the consistency analyses undertaken with the main survey data to assess within-country coder reliability and *iii)* the international coder review undertaken to examine the between-country consistency in applying the coding guides. The objective of the international coder review was to estimate potential bias (either leniency or harshness) in the coding standards applied in each national centre, and to express this potential bias in PISA units.

CONSISTENCY ANALYSES

Both in the field trial and the main survey consistency analysis was used to estimate the level of agreement between coders of constructed-response items. In the field trial the primary purpose of the consistency analysis is to obtain data to inform the selection of items for the main survey – in the field trial, many more items were tried than were finally used in the main survey. An obvious goal of PISA is to ensure that coders largely agree in their categorisation of the answers.

The consistency analyses are based on data gained from having the same items coded by a number of different coders. For the PISA 2009 main survey only open-ended items from the first cluster in each booklet were multiple coded. This design also helped to ensure that the amount of missing data was minimised (the amount of missing data and non-responses increases towards the end of the booklet). For their main test language each country was required to randomly assign 100 booklets of each type that they were using for testing for multiple coding, and for minority languages the requirement was at least 50 booklets of each type. There were two groups of countries: those who did standard booklets only (booklets 1-13) and those who did some standard booklets and some non-standard easier booklets (booklets 8-13 and 21-27). There were 20 countries that chose this second option.[1]

All analysis was done by booklet. Each response was coded by four coders. Only students with four non-missing codes were used for analysis. The following notation is used in this chapter:

$i=1,....,I$ – items in the booklet

$c=1,...,C$ – country-by-language unit

$j=1,...,J_{i,c}$ – students in the country-by-language unit who attended to the booklet

$k=1,...,K_{ic}$ – coders in the country-by-language unit who coded items in the booklet during multiple coding exercise

$x_{ijk}=0, 1, 2, ...$ – code allocated by coder k to student j when coding item i.

To investigate the level of disagreement between coders, the data collected were used to first compute a coder-item disagreement index R_{ikc}. This index was computed for each coder k and each item i across all records j in the multiple coding exercise within a given country-by-language unit c. The index was computed as an average residual multiplied by 100 for readability purposes.

13.1

$$R_{ikc}= \frac{100}{J_{ic}} \sum_{j} \left| x_{ijk} - \frac{1}{4} \sum_{k} x_{ijk} \right|$$

R_{ikc} is then aggregated to compute other indices. A value of $R_{ikc}=0$ shows a perfect agreement among coders for all students responding to the item of a particular language in the country (e.g. shaded cells for item A in Table 13.1).

Each disagreement between coders contributes to an increase of the index. For example, if coder X disagrees by one score with three others, all of whom agree with each other, the residual for X would be 0.75 and the residual for each of three others would be 0.25. In the example in Table 13.1, coder 201 disagrees by one score with three

other coders 20% of the time when coding item B and there are no other cases of disagreement for this item (a fictitious situation). In this case R_{ikc}=15 for this coder and for the three other coders it is 5.

On the other hand, if two of the coders disagree with the two others in 20% of the cases and there are no other cases of disagreement (this is another fictitious situation with all residuals being 0.5), then R_{ikc}=10 for all coders (shaded cells for item C in Table 13.1).

In a real situation there is always a mix of different combinations of disagreement and the R_{ikc} would look more like shaded cells for items D and E in Table 13.1.

Table 13.1 Examples of various indices calculated on country-by-language level

	Item A	Item B	Item C	Item D	Item E	
Coder	Coder-item disagreement Rikc					Coder reliability index Dkc
201	0	15	10	9.88	11.82	9.34
202	0	5	10	4.45	10.91	6.07
203	0	5	10	5.14	10.45	6.12
204	0	5	10	5.14	10.45	6.12
Country-by-language item reliability index Sic	0	7.5	10	6.15	10.91	

The average across all coders was calculated as a country-by-language item reliability index S_{ic} for each item in each country-by-language unit (13.2) and the average across all items coded by a particular coder was calculated as a coder reliability index Q_{ic} (13.3). Examples of some possible S_{ic} values are shown in the bottom line in Table 13.1 and examples of some possible Q_{ic} values are shown in the last column in Table 13.1. In this example coder 201 appears less reliable than three other coders.

13.2

$$S_{ic} = \frac{1}{K_{ic}} \sum_{k} R_{ikc}$$

13.3

$$Q_{kc} = \frac{1}{I} \sum_{i} R_{ikc}$$

S_{ic} was further aggregated across all country-by-language units to the international item reliability index (T_i).

13.4

$$T_i = \frac{1}{C} \sum_{c} S_{ic}$$

The international item reliability index T_i for each item in the multiple-coding exercise is presented in Table 13.2. In this table we can see that on average mathematics items have fewer inconsistencies between coders than reading and science items. The ten items with the most discrepancies between coders across all domains are shown in bold. There are 8 (out of 57) of them in reading and 2 (out of 17) in science. There are no mathematics items in the top ten. The four highest on discrepancies items in reading were all link items from PISA 2000. The other four have much lower level of discrepancies. All new items improved slightly compared to the field trial.

Let C^ be a set of σ country-by-language units and δ be the number of items in the domain D ($D=r$ for reading, m for mathematics or s for science). The average for each country across all items in each of the three domains is then presented by national domain index N_{cD}.

13.5

$$N_{cD} = \frac{1}{\delta} \sum_{i \in D} \frac{1}{\sigma} \sum_{c \in C^{\wedge}} S_{ic}$$

The national domain index N_{cD} for three domains (reading, science and mathematics) is presented in Table 13.3. The countries' highest ten discrepancies across all domains are highlighted in dark blue and countries' lowest ten discrepancies are highlighted in dark grey. It should be noted that some countries that had a very high level of discrepancies during the field trial improved for the main survey. For example, Latvia had very high level of discrepancies in reading for the Field Trial, but is just outside one standard deviation from the mean for reading for the Main Survey. It can

be noted from the Table 13.3 that OECD countries have high level of discrepancies only for Science, the domain that they did not do during the Field Trial. Therefore, these discrepancies may be attributed to the lack of training.

An extremely low level of discrepancies (e.g. no discrepancies in Azerbaijan for mathematics) is also highlighted as a potential candidate for bias. To identify bias the international coder review is used. It is described in the next section.

[Part 1/2]

Table 13.2 International item reliability indices (Ti)

Mathematics		
ItemID	**Ti**	**Number of countries**
M155Q01	1.61	63
M155Q02D	4.03	64
M155Q03D	5.18	64
M406Q01	1.32	64
M406Q02	2.21	64
M442Q02	1.05	64
M446Q02	0.84	64
M462Q01D	1.80	64
M828Q01	4.41	64
M828Q02	1.89	64
M828Q03	1.09	64
Science		
ItemID	**Ti**	**Number of countries**
S131Q02D	3.35	64
S131Q04D	4.12	64
S269Q01	2.22	64
S269Q03D	2.82	64
S326Q01	4.35	64
S326Q02	3.77	64
S408Q03	5.04	64
S425Q03	7.22	64
S425Q04	3.51	64
S428Q05	3.61	64
S438Q03D	6.88	64
S465Q01	5.95	64
S498Q04	**7.86**	64
S514Q02	1.40	64
S514Q03	**4.39**	64
S519Q01	12.06	63
S519Q03	6.09	64

[Part 2/2]

Table 13.2 **International item reliability indices (Ti)**

Reading		
ItemID	**Ti**	**Number of countries**
R055Q02	6.60	64
R055Q03	3.38	64
R055Q05	2.77	64
R067Q04	**15.04**	64
R067Q05	**13.34**	64
R083Q02	0.37	44
R102Q04A	1.62	64
R104Q05	2.03	64
R111Q02B	**14.80**	64
R111Q06B	**14.53**	64
R219Q01E	2.99	64
R219Q02	4.65	64
R220Q01	4.98	64
R227Q03	3.76	64
R227Q06	1.17	64
R403Q03	1.06	20
R404Q10A	4.75	64
R404Q10B	6.18	64
R406Q01	2.47	64
R406Q02	**8.13**	64
R406Q05	2.99	64
R412Q08	5.56	64
R414Q06	4.65	44
R417Q03	4.44	20
R417Q04	4.44	20
R420Q02	0.94	64
R420Q06	6.42	64
R420Q10	4.98	64
R429Q08	1.28	20
R432Q05	4.69	64
R433Q05	4.58	20
R433Q07	1.08	20
R435Q05	4.57	20
R437Q07	6.68	64
R442Q02	2.48	44
R442Q03	1.70	44
R442Q05	4.89	44
R442Q06	6.87	44
R445Q01	3.61	20
R446Q06	2.47	64
R447Q06	6.71	44
R452Q03	0.71	44
R452Q06	5.21	44
R453Q04	**7.59**	63
R453Q06	4.46	64
R455Q02	6.19	64
R455Q03	0.76	64
R456Q02	3.80	64
R456Q06	1.72	64
R458Q07	**7.42**	44
R460Q01	2.08	64
R462Q02	2.18	20
R462Q05	5.07	20
R465Q02	1.56	20
R465Q05	5.05	20
R465Q06	**7.38**	20
R466Q02	2.23	64

Table 13.3 **National domain reliability indices**

		Mathematics	Reading	Science
OECD	Australia	2.47	6.23	**11.30**
	Austria	3.26	5.81	6.83
	Belgium	4.09	3.97	7.67
	Canada	6.09	7.10	**10.14**
	Chile	1.31	7.26	6.29
	Czech Republic	3.28	7.47	6.87
	Denmark	3.85	8.04	8.92
	Estonia	2.64	4.85	5.25
	Finland	1.81	4.41	4.85
	France	2.79	7.78	8.04
	Germany	4.34	6.05	6.85
	Greece	0.82	1.32	0.60
	Hungary	3.23	5.39	1.24
	Iceland	2.83	5.91	6.43
	Ireland	3.45	5.35	7.10
	Israel	4.37	7.48	**9.09**
	Italy	1.76	4.73	5.52
	Japan	1.37	2.85	1.77
	Korea	1.49	3.25	2.44
	Mexico	1.48	2.96	0.86
	Netherlands	2.84	6.72	5.44
	New Zealand	3.56	5.24	5.76
	Norway	3.34	4.88	8.17
	Poland	2.12	3.67	3.04
	Portugal	0.50	6.65	3.89
	Slovak Republic	1.73	4.27	4.00
	Slovenia	1.84	5.62	5.08
	Spain	4.09	6.19	7.98
	Sweden	3.74	6.00	6.08
	Switzerland	3.49	7.98	6.85
	Turkey	3.24	0.97	4.25
	United Kingdom	2.17	4.99	4.48
	United States	3.00	0.65	2.64
Partners	Albania	**0.28**	**0.44**	**0.34**
	Argentina	2.27	2.46	5.50
	Azerbaijan	**0.00**	0.55	**0.35**
	Brazil	**0.04**	1.40	1.02
	Bulgaria	1.28	8.53	5.08
	Colombia	2.70	**10.33**	7.02
	Croatia	0.83	1.85	2.74
	Dubai (UAE)	3.88	8.41	**10.67**
	Hong Kong-China	2.98	3.05	6.44
	Indonesia	1.44	6.72	5.71
	Jordan	**0.43**	1.52	1.48
	Kazakhstan	0.91	0.92	1.20
	Kyrgyzstan	1.45	1.88	1.37
	Latvia	4.92	7.92	**10.50**
	Lithuania	2.44	5.31	4.79
	Luxembourg	2.20	5.61	6.86
	Macao-China	0.86	0.83	1.13
	Montenegro	1.50	**9.65**	**9.55**
	Panama	1.29	7.60	5.60
	Peru	2.40	7.50	3.65
	Qatar	1.07	1.42	0.83
	Romania	1.18	6.73	0.83
	Russian Federation	0.49	0.93	1.11
	Serbia	3.26	3.90	5.51
	Shanghai-China	1.76	5.25	4.03
	Singapore	2.80	7.48	3.76
	Chinese Taipei	3.12	3.01	5.33
	Thailand	**0.15**	0.82	0.64
	Trinidad and Tobago	**0.16**	1.55	**0.46**
	Tunisia	3.30	8.65	**9.45**
	Uruguay	4.35	8.43	**9.65**
	International Average	2.31	4.89	4.97
	SD	1.35	2.68	3.05

Note: The countries' highest ten discrepancies across all domains are highlighted in dark blue and countries' lowest ten discrepancies are highlighted in dark grey.

INTERNATIONAL CODER REVIEW

For the PISA 2009 International Coding Review (ICR), the Consortium identified a set of items for inclusion in the study. Two booklets were chosen: booklet 8 (containing 8 manually coded reading items from cluster R2) and booklet 12 (containing 6 manually coded reading items from cluster R7). These items were also among those used previously in the multiple-coding study and had been coded four times by national coders as part of that study. The code assigned by the fourth national coder was entered into PISA data and is referred to as the reported code.

For each country-by-language unit from a national centre's data, up to 80 PISA records[2] (excluding those with a high number of missing responses for the multiple-coded items) were selected by the PISA Consortium from the data from booklets 8 and 12. The student IDs of the selected records were sent to the national centres.

In the PISA national centres, the corresponding booklets were located and scanned and these scanned images were sent to the PISA Consortium's linguistic verification expert. Where scanning was not possible, the original booklets were sent by post. The PISA Consortium's linguistic verification expert then erased the national coders' marks on all received copies of the booklets.

Coding of each student's response was then carried out a fifth time by a member of a team of independent reviewers who had been trained specifically for this task. These independent reviewers had previously been involved as part of the international translation verification team. The code assigned by the independent reviewer is referred to as the verifier code.

Reported scores and verifier scores were then calculated. These were obtained by scaling all the ICR students' data from all countries from cluster R2 in booklet 8 and cluster R7 in booklet 12 (including automatically scored and open-ended responses). Scaling using the reported code for the open-ended responses produced the reported score. Scaling using the verifier code for the open-ended responses produced the verifier score.

Each country's scores were then extracted and the reported scores and the verifier scores were compared. This comparison involved calculating the mean difference between the reported scores and the verified scores for each country for both booklets.[3] A 95% confidence interval was then calculated around the mean difference. If the confidence interval contained 0, the differences in score were considered as not statistically significant. Two hypothetical examples in Table 13.4 show that country A was initially found lenient (positive confidence interval: [5.93; 24.41]) and country B was found neither lenient nor harsh (confidence interval [-7.16; 4.641] contains 0).

Table 13.4 Examples of an initially lenient result and a neutral result

Country	Language	Mean difference between reported and verifier scores	N	Standard deviation	Confidence interval		Leniency/Harshness
					Low	High	
A	aaaa	15.17	80	41.53	5.93	24.41	Leniency
B	bbbb	-1.26	78	26.17	-7.16	4.641	

In addition, two types of inconsistencies between national codes and verifier codes were flagged:

- When the verifier code was compared with each of the four national codes in turn, fewer than two matches were observed.
- When the average raw score of the four national coders was at least 0.5 points higher or lower than the score based on the verifier code.

Cases are flagged if at least one of these conditions were met. Examples of flagged cases are given in Table 13.5.

Table 13.5 Examples of flagged cases

Country	StudentID	Question	Coder 1	Coder 2	Coder 3	Coder 4	Verifier	Flag (Y/N)
xxx	Xxxxx00001	R104Q05	0	1	1	1	1	N
xxx	Xxxxx00012	R104Q05	1	1	1	1	0	Y
xxx	Xxxxx00031	R104Q05	1	1	1	0	0	Y
xxx	Xxxxx00014	R104Q05	0	1	1	2	0	Y
xxx	Xxxxx00020	R104Q05	1	0	2	1	2	Y
xxx	Xxxxx00025	R104Q05	2	0	2	0	2	Y

The percentage of flagged cases was calculated for each item in each booklet. Table 13.6 shows that items R111Q02B and R111Q06B in booklet 8 had a high percentage of disagreement in nearly all countries (Table 13.7 shows the same information for booklet 12). These two items also showed a very high percentage of disagreement between national coders across all countries (Table 13.2). Therefore it was decided to exclude these items from calculations of leniency/harshness and to investigate these two items separately. They were adjudicated for English speaking countries. The Consortium adjudicator recoded, blind, all Australian, Irish and Qatar-English student responses in the ICR set for items R111Q02B and R111Q06B. Only 40% agreement with the verifier was obtained on the flagged cases, a result that supports the decision to exclude these items from the calculations of leniency/harshness and subsequently from PISA database.

After exclusion of items R111Q02B and R111Q06B, a country was selected for the adjudication process if it was found lenient or harsh for both booklets (see Table 13.8). This adjudication process involved additional coding by senior Consortium staff of a random sample of 30 student responses from each identified country. The following countries were initially found to be lenient and were adjudicated: Albania, Azerbaijan, Bulgaria, Indonesia, and Romania. The following country-by-language units were initially found to be harsh and were adjudicated: Israel (Arabic coders only), Kazakhstan (Kazakh coders only) and Sweden. It was decided to also adjudicate Brazil due to high number of items having a high percentage of flagged cases between verifier and national coders in both booklets and leniency in booklet 12.

The sampled student responses were back-translated into English, and the responses together with the four national codes and the verifier code for these selected cases were reviewed by the international adjudicator.

Systematic coder harshness or leniency on the national PISA score for each domain is confirmed if the percentage of agreement between verifier and adjudicator is above 50%.

[Part 1/2]

Table 13.6 Percentage of flagged records for Booklet 8 ICR items

	Language	R055Q02	R055Q03	R055Q05	R104Q05	R111Q02B	R111Q06B	R227Q03	R227Q06	Total	N
Albania	Albanian	11.25	8.75	18.75	3.75	42.50	25.00	12.50	6.25	10.21	80
Argentina	Spanish	15.94	1.45	5.80	0.00	17.39	14.49	10.14	1.45	5.80	69
Australia	English	3.75	2.50	2.50	0.00	33.75	11.25	5.00	0.00	2.29	80
Austria	German	1.25	6.25	0.00	0.00	27.50	17.50	1.25	0.00	1.46	80
Azerbaijan	Azerbaijani	38.75	5.00	2.50	3.75	22.50	30.00	8.75	2.50	10.21	80
Belgium	Dutch	20.00	8.75	0.00	2.50	36.25	41.25	3.75	0.00	5.83	80
Belgium	French	6.25	1.25	1.25	1.25	30.00	30.00	0.00	2.50	2.08	80
Brazil	Portuguese	17.65	3.92	27.45	0.00	39.22	13.73	13.73	0.00	10.46	51
Bulgaria	Bulgarian	8.75	6.25	6.25	2.50	31.25	32.50	5.00	16.25	7.50	80
Canada	English	8.75	2.50	0.00	0.00	35.00	15.00	11.25	2.50	4.17	80
Canada	French	2.50	1.25	5.00	1.25	22.50	23.75	10.00	0.00	3.33	80
Chile	Spanish	5.00	1.25	5.00	2.50	13.75	21.25	8.75	0.00	3.75	80
Colombia	Spanish	8.75	3.75	8.75	0.00	23.75	30.00	15.00	0.00	6.04	80
Croatia	Croatian	3.75	1.25	1.25	2.50	12.50	20.00	21.25	1.25	5.21	80
Czech Republic	Czech	3.75	0.00	1.25	1.25	38.75	15.00	6.25	0.00	2.08	80
Denmark	Danish	8.75	5.00	2.50	2.50	25.00	21.25	1.25	2.50	3.75	80
Dubai (UAE)	Arabic	8.82	8.82	26.47	5.88	23.53	26.47	2.94	2.94	9.31	34
Dubai (UAE)	English	19.64	1.79	3.57	1.79	21.43	14.29	1.79	0.00	4.76	56
Estonia	Estonian	3.13	0.00	0.00	1.56	17.19	6.25	6.25	1.56	2.08	64
Estonia	Russian	0.00	0.00	5.00	0.00	45.00	45.00	10.00	0.00	2.50	20
Finland	Finnish	3.75	0.00	0.00	5.00	26.25	18.75	2.50	1.25	2.08	80
France	French	3.75	1.25	3.75	2.50	21.25	17.50	6.25	0.00	2.92	80
Germany	German	7.14	0.00	0.00	3.57	25.00	25.00	0.00	0.00	1.79	28
Greece	Greek, Modern	11.25	3.75	3.75	1.25	33.75	15.00	8.75	1.25	5.00	80
Hong Kong-China	Chinese	5.00	3.75	1.25	0.00	25.00	36.25	7.50	0.00	2.92	80
Hungary	Hungarian	10.00	2.50	3.75	5.00	27.50	32.50	6.25	0.00	4.58	80
Iceland	Icelandic	8.86	5.06	6.33	3.80	83.54	30.38	8.86	1.27	5.70	79
Indonesia	Indonesian	8.75	0.00	7.50	6.25	31.25	17.50	10.00	3.75	6.04	80
Ireland	English	2.50	0.00	2.50	3.75	22.50	15.00	7.50	1.25	2.92	80
Israel	Arabic	5.00	2.50	2.50	0.00	27.50	7.50	40.00	0.00	8.33	40
Israel	Hebrew	18.75	1.25	5.00	0.00	31.25	22.50	1.25	1.25	4.58	80

[Part 2/2]

Table 13.6 **Percentage of flagged records for Booklet 8 ICR items**

	Language	R055Q02	R055Q03	R055Q05	R104Q05	R111Q02B	R111Q06B	R227Q03	R227Q06	Total	N
Italy	Italian	3.75	0.00	0.00	1.25	11.25	33.75	3.75	0.00	1.46	80
Japan	Japanese	21.25	3.75	8.75	7.50	33.75	35.00	3.75	1.25	7.71	80
Jordan	Arabic	16.25	2.50	11.25	5.00	50.00	20.00	7.50	0.00	7.08	80
Kazakhstan	Kazakh	25.00	10.00	7.50	7.50	37.50	55.00	17.50	0.00	11.25	40
Kazakhstan	Russian	7.50	2.50	5.00	5.00	17.50	17.50	5.00	0.00	4.17	40
Korea	Korean	8.75	0.00	2.50	1.25	55.00	23.75	5.00	0.00	2.92	80
Kyrgyzstan	Kyrgyz	12.50	4.69	12.50	4.69	14.06	10.94	10.94	0.00	7.55	64
Kyrgyzstan	Russian	3.57	0.00	7.14	0.00	10.71	3.57	7.14	3.57	3.57	28
Latvia	Latvian	7.94	6.35	3.17	7.94	30.16	26.98	0.00	1.59	4.50	63
Latvia	Russian	4.17	4.17	8.33	4.17	16.67	58.33	8.33	0.00	4.86	24
Lithuania	Lithuanian	2.50	2.50	2.50	1.25	7.50	13.75	10.00	0.00	3.13	80
Luxembourg	French	4.55	18.18	0.00	4.55	18.18	18.18	4.55	0.00	5.30	22
Luxembourg	German	7.81	1.56	0.00	1.56	15.63	28.13	3.13	0.00	2.34	64
Macao-China	Chinese	38.75	0.00	1.25	0.00	18.75	26.25	5.00	0.00	7.50	80
Mexico	Spanish	10.13	5.06	8.86	0.00	31.65	30.38	18.99	0.00	7.17	79
Montenegro	Serbian of a yekavian variant or Montenegrin	3.75	3.75	3.75	3.75	20.00	20.00	3.75	7.50	4.38	80
Netherlands	Dutch	20.00	2.50	0.00	0.00	42.50	18.75	6.25	1.25	5.00	80
New Zealand	English	6.25	2.50	5.00	2.50	31.25	15.00	1.25	0.00	2.92	80
Norway	Norwegian	3.75	0.00	1.25	0.00	18.75	15.00	1.25	0.00	1.04	80
Panama	Spanish	12.50	3.75	13.75	8.75	37.50	23.75	10.00	1.25	8.33	80
Peru	Spanish	10.00	7.50	11.25	1.25	12.50	16.25	23.75	0.00	8.96	80
Poland	Polish	6.25	11.25	0.00	2.50	28.75	16.25	3.75	1.25	4.17	80
Portugal	Portuguese	5.00	0.00	1.25	1.25	25.00	13.75	1.25	0.00	1.46	80
Qatar	Arabic	18.75	1.25	7.50	2.50	27.50	15.00	18.75	0.00	8.13	80
Qatar	English	7.50	5.00	5.00	2.50	22.50	12.50	5.00	0.00	4.17	40
Romania	Romanian	15.00	3.75	5.00	0.00	33.75	45.00	10.00	3.75	6.25	80
Russian Federation	Russian	7.50	0.00	6.25	2.50	27.50	12.50	13.75	2.50	5.42	80
Scotland	English	2.50	0.00	0.00	0.00	26.25	18.75	3.75	1.25	1.25	80
Serbia	Serbian	7.50	3.75	3.75	1.25	15.00	16.25	7.50	0.00	3.96	80
Shanghai-China	Chinese	1.25	1.25	3.75	0.00	32.50	31.25	6.25	0.00	2.08	80
Singapore	English	5.00	2.50	3.75	0.00	38.75	30.00	2.50	1.25	2.50	80
Slovak Republic	Slovak	6.25	2.50	2.50	0.00	28.75	16.25	5.00	0.00	2.71	80
Slovenia	Slovenian	5.88	2.94	2.94	0.00	19.12	10.29	10.29	0.00	3.68	68
Spain	Galician	7.50	2.50	2.50	7.50	32.50	17.50	7.50	2.50	5.00	40
Spain	Spanish	10.29	7.35	7.35	1.47	35.29	17.65	8.82	1.47	6.13	68
Sweden	Swedish	2.50	1.25	0.00	1.25	32.50	12.50	1.25	0.00	1.04	80
Switzerland	French	0.00	0.00	0.00	0.00	0.00	9.09	18.18	0.00	3.03	11
Switzerland	German	2.04	0.00	2.04	2.04	10.20	14.29	0.00	0.00	1.02	49
Chinese Taipei	Chinese	11.25	1.25	1.25	0.00	28.75	22.50	5.00	0.00	3.13	80
Thailand	Thai	13.75	1.25	10.00	0.00	20.00	15.00	17.50	1.25	7.29	80
Trinidad and Tobago	English	8.75	6.25	13.75	3.75	17.50	25.00	5.00	0.00	6.25	80
Tunisia	Arabic	12.50	3.75	10.00	3.75	25.00	36.25	3.75	0.00	5.63	80
Turkey	Turkish	8.75	8.75	2.50	0.00	41.25	17.50	13.75	0.00	5.63	80
United Kingdom (excl. Scotland)	English	2.50	0.00	1.25	0.00	20.00	18.75	2.50	1.25	1.25	80
United States	English	11.25	3.75	3.75	0.00	20.00	8.75	7.50	0.00	4.38	80
Uruguay	Spanish	3.80	3.80	10.13	1.27	8.86	18.99	8.86	0.00	4.64	79

[Part 1/2]

Table 13.7 **Percentage of flagged records for Booklet 12 ICR items**

	Language	R432Q05	R446Q06	R456Q02	R456Q06	R460Q01	R466Q02	Total	N
Albania	Albanian	26.25	8.75	15.00	11.25	17.50	2.50	13.54	80
Argentina	Spanish	5.13	11.54	10.26	1.28	7.69	1.28	6.20	78
Australia	English	1.25	2.50	3.75	0.00	3.75	1.25	2.08	80
Austria	German	5.00	0.00	2.50	1.25	3.75	1.25	2.29	80
Azerbaijan	Azerbaijani	26.25	45.00	6.25	3.75	1.25	20.00	17.08	80
Belgium	Dutch	0.00	5.00	11.25	0.00	0.00	7.50	3.96	80
Belgium	French	1.25	2.50	2.50	1.25	1.25	7.50	2.71	80
Brazil	Portuguese	10.20	2.04	22.45	10.20	10.20	12.24	11.22	49
Bulgaria	Bulgarian	10.00	1.25	6.25	5.00	17.50	1.25	6.88	80
Canada	English	1.25	1.25	2.50	1.25	1.25	0.00	1.25	80
Canada	French	0.00	1.25	2.50	3.75	0.00	6.25	2.29	80
Chile	Spanish	5.00	6.25	8.75	6.25	6.25	6.25	6.46	80
Colombia	Spanish	7.50	7.50	7.50	1.25	0.00	2.50	4.38	80
Croatia	Croatian	3.75	6.25	8.75	5.00	16.25	6.25	7.71	80
Czech Republic	Czech	2.50	16.25	0.00	0.00	0.00	16.25	5.83	80
Denmark	Danish	5.00	6.25	5.00	3.75	0.00	1.25	3.54	80
Dubai (UAE)	Arabic	23.53	0.00	26.47	0.00	8.82	5.88	10.78	34
Dubai (UAE)	English	1.79	3.57	7.14	3.57	0.00	5.36	3.57	56
Estonia	Estonian	4.69	0.00	4.69	4.69	1.56	0.00	2.60	64
Estonia	Russian	0.00	5.00	0.00	5.00	5.00	0.00	2.50	20
Finland	Finnish	0.00	0.00	2.50	1.25	5.00	1.25	1.67	80
France	French	2.50	0.00	1.25	0.00	2.50	3.75	1.67	80
Germany	German	3.33	0.00	3.33	0.00	0.00	10.00	2.78	30
Greece	Greek, Modern	7.50	1.25	2.50	0.00	2.50	0.00	2.29	80
Hong Kong-China	Chinese	3.75	1.25	10.00	1.25	1.25	0.00	2.92	80
Hungary	Hungarian	3.75	8.75	21.25	3.75	10.00	3.75	8.54	80
Iceland	Icelandic	3.85	16.67	8.97	3.85	2.56	7.69	7.26	78
Indonesia	Indonesian	35.00	8.75	15.00	2.50	5.00	5.00	11.88	80
Ireland	English	2.53	0.00	7.59	3.80	5.06	2.53	3.59	79
Israel	Arabic	12.50	10.00	22.50	7.50	0.00	2.50	9.17	40
Israel	Hebrew	2.50	3.75	7.50	3.75	5.00	1.25	3.96	80
Italy	Italian	5.00	5.00	3.75	1.25	1.25	6.25	3.75	80
Japan	Japanese	2.50	1.25	3.75	1.25	0.00	2.50	1.88	80
Jordan	Arabic	17.50	2.50	5.00	8.75	0.00	1.25	5.83	80
Kazakhstan	Kazakh	20.00	2.50	25.00	15.00	0.00	0.00	10.42	40
Kazakhstan	Russian	17.50	5.00	10.00	0.00	0.00	0.00	5.42	40
Korea	Korean	6.25	5.00	2.50	1.25	0.00	1.25	2.71	80
Kyrgyzstan	Kyrgyz	7.14	1.79	32.14	8.93	7.14	3.57	10.12	56
Kyrgyzstan	Russian	7.14	3.57	3.57	0.00	7.14	0.00	3.57	28
Latvia	Latvian	9.38	3.13	6.25	3.13	3.13	3.13	4.69	64
Latvia	Russian	0.00	0.00	0.00	4.35	4.35	4.35	2.17	23

[Part 2/2]

Table 13.7 **Percentage of flagged records for Booklet 12 ICR items**

	Language	R432Q05	R446Q06	R456Q02	R456Q06	R460Q01	R466Q02	Total	N
Lithuania	Lithuanian	7.50	5.00	7.50	3.75	2.50	1.25	4.58	80
Luxembourg	French	4.55	4.55	4.55	4.55	4.55	0.00	3.79	22
Luxembourg	German	0.00	1.56	0.00	0.00	6.25	1.56	1.56	64
Macao-China	Chinese	10.00	11.25	2.50	3.75	0.00	3.75	5.21	80
Mexico	Spanish	15.00	7.50	13.75	2.50	2.50	1.25	7.08	80
Montenegro	Serbian of a yekavian variant or Montenegrin	10.00	0.00	2.50	5.00	1.25	6.25	4.17	80
Netherlands	Dutch	16.25	1.25	6.25	0.00	1.25	1.25	4.38	80
New Zealand	English	2.50	0.00	0.00	0.00	1.25	0.00	0.63	80
Norway	Norwegian	1.25	1.25	8.75	0.00	1.25	7.50	3.33	80
Panama	Spanish	17.50	23.75	22.50	5.00	1.25	6.25	12.71	80
Peru	Spanish	11.25	5.00	5.00	1.25	0.00	2.50	4.17	80
Poland	Polish	5.00	1.25	5.00	2.50	0.00	2.50	2.71	80
Portugal	Portuguese	6.25	2.50	5.00	7.50	2.50	2.50	4.38	80
Qatar	Arabic	20.00	5.00	0.00	5.00	7.50	1.25	6.46	80
Qatar	English	10.00	10.00	27.50	2.50	0.00	5.00	9.17	40
Romania	Romanian	23.75	7.50	21.25	2.50	7.50	2.50	10.83	80
Russian Federation	Russian	13.75	2.50	8.75	3.75	2.50	0.00	5.21	80
Scotland	English	1.25	1.25	5.00	0.00	2.50	2.50	2.08	80
Serbia	Serbian	13.75	1.25	13.75	2.50	3.75	7.50	7.08	80
Shanghai-China	Chinese	6.25	10.00	2.50	0.00	3.75	0.00	3.75	80
Singapore	English	5.00	0.00	7.50	0.00	0.00	2.50	2.50	80
Slovak Republic	Slovak	3.75	1.25	2.50	1.25	13.75	8.75	5.21	80
Slovenia	Slovenian	8.57	5.71	10.00	2.86	4.29	12.86	7.38	70
Spain	Galician	2.50	2.50	5.00	2.50	0.00	15.00	4.58	40
Spain	Spanish	10.00	5.00	15.00	1.25	5.00	3.75	6.67	80
Sweden	Swedish	3.75	0.00	7.50	1.25	0.00	1.25	2.29	80
Switzerland	French	0.00	0.00	6.67	0.00	0.00	0.00	1.11	15
Switzerland	German	2.56	2.56	5.13	0.00	0.00	2.56	2.14	39
Chinese Taipei	Chinese	2.50	1.25	11.25	0.00	3.75	1.25	3.33	80
Thailand	Thai	7.59	2.53	6.33	7.59	3.80	3.80	5.27	79
Trinidad and Tobago	English	15.00	2.50	6.25	5.00	2.50	2.50	5.63	80
Tunisia	Arabic	12.50	3.75	17.50	1.25	1.25	1.25	6.25	80
Turkey	Turkish	2.50	1.25	23.75	1.25	0.00	2.50	5.21	80
United Kingdom (excl. Scotland)	English	1.25	1.25	13.75	1.25	5.00	2.50	4.17	80
United States	English	3.75	1.25	6.25	2.50	0.00	2.50	2.71	80
Uruguay	Spanish	5.06	2.53	8.86	6.33	2.53	0.00	4.22	79

[Part 1/2]

Table 13.8 **Leniency/Harshness analysis**

	Booklet 8 excluding R111Q02B and R111Q06B								Booklet 12							Overall
	Language	Mean	N	Std. deviation	CI_lo	CI_hi	t	Leniency/ Harshness	Mean	N	Std. deviation	CI_lo	CI_hi	t	Leniency/ Harshness	Leniency/ Harshness
Albania	Albanian	7.34	80	27.30	1.27	13.42	1.99	Lenient	15.17	80	41.53	5.93	24.41	1.99	Lenient	Lenient
Argentina	Spanish	6.35	69	25.19	0.30	12.40	2.00	Lenient	-1.26	78	26.17	-7.16	4.64	1.99		
Australia	English	1.82	80	22.04	-3.09	6.72	1.99		4.44	80	21.19	-0.28	9.15	1.99		
Austria	German	-0.94	80	20.91	-5.59	3.72	1.99		5.10	80	24.10	-0.27	10.46	1.99		
Azerbaijan	Azerbaijani	18.40	80	23.79	13.10	23.69	1.99	Lenient	10.96	80	33.04	3.61	18.31	1.99	Lenient	Lenient
Belgium	Dutch	0.48	80	32.87	-6.84	7.79	1.99		4.32	80	31.11	-2.60	11.24	1.99		
Belgium	French	-0.34	80	12.08	-3.03	2.35	1.99		-0.73	80	25.37	-6.38	4.91	1.99		
Brazil	Portuguese	6.86	51	30.13	-1.61	15.33	2.01		8.15	49	33.79	-1.56	17.85	2.01		
Bulgaria	Bulgarian	10.40	80	34.25	2.77	18.02	1.99	Lenient	13.34	80	46.01	3.10	23.58	1.99	Lenient	Lenient
Canada	English	1.41	80	24.54	-4.05	6.87	1.99		-3.07	80	24.48	-8.52	2.38	1.99		
Canada	French	3.38	80	22.61	-1.65	8.41	1.99		4.50	80	27.58	-1.64	10.64	1.99		
Chile	Spanish	0.00	80	23.23	-5.17	5.17	1.99		-0.63	80	29.50	-7.19	5.94	1.99		
Colombia	Spanish	-0.21	80	21.67	-5.03	4.61	1.99		1.63	80	30.55	-5.17	8.43	1.99		
Croatia	Croatian	2.62	80	19.21	-1.66	6.89	1.99		5.14	80	32.02	-1.98	12.27	1.99		
Czech Republic	Czech	-1.25	80	20.85	-5.89	3.39	1.99		2.58	80	29.33	-3.94	9.11	1.99		
Denmark	Danish	-4.16	80	21.60	-8.97	0.64	1.99		5.74	80	24.23	0.35	11.13	1.99	Lenient	
Dubai (UAE)	Arabic	1.06	34	25.27	-7.75	9.88	2.03		6.61	34	35.16	-5.66	18.88	2.03		
Dubai (UAE)	English	-2.89	56	21.38	-8.62	2.83	2.00		0.51	56	23.96	-5.91	6.92	2.00		
Estonia	Estonian	-2.68	64	20.33	-7.76	2.40	2.00		3.74	64	20.93	-1.49	8.96	2.00		
Estonia	Russian	-5.18	20	19.80	-14.44	4.09	2.09		-3.01	20	13.70	-9.42	3.40	2.09		
Finland	Finnish	-0.78	80	13.95	-3.88	2.33	1.99		3.30	80	23.61	-1.96	8.55	1.99		
France	French	4.39	80	18.20	0.33	8.44	1.99	Lenient	-2.14	80	35.08	-9.95	5.66	1.99		
Germany	German	-2.35	28	17.06	-8.97	4.26	2.05		0.85	30	32.41	-11.25	12.95	2.05		
Greece	Greek, Modern	-0.95	80	19.69	-5.33	3.44	1.99		-4.60	80	22.72	-9.66	0.46	1.99		
Hong Kong-China	Chinese	2.42	80	18.30	-1.65	6.49	1.99		8.17	80	32.69	0.89	15.44	1.99	Lenient	
Hungary	Hungarian	-2.04	80	20.68	-6.64	2.56	1.99		-11.70	80	48.35	-22.46	-0.94	1.99	Harsh	
Iceland	Icelandic	-3.44	79	28.17	-9.75	2.87	1.99		-20.33	78	34.18	-28.03	-12.62	1.99	Harsh	
Indonesia	Indonesian	16.35	80	52.59	4.64	28.05	1.99	Lenient	16.39	80	33.50	8.94	23.85	1.99	Lenient	Lenient
Ireland	English	-1.07	80	18.95	-5.28	3.15	1.99		1.95	79	24.84	-3.61	7.51	1.99		
Israel	Arabic	-13.47	40	26.68	-22.00	-4.93	2.02	Harsh	-15.65	40	32.91	-26.17	-5.13	2.02	Harsh	Harsh
Israel	Hebrew	-1.28	80	20.43	-5.82	3.27	1.99		-1.92	80	33.64	-9.41	5.56	1.99		
Italy	Italian	-1.37	80	18.06	-5.39	2.65	1.99		0.82	80	37.24	-7.46	9.11	1.99		
Japan	Japanese	5.05	80	25.02	-0.52	10.61	1.99		-3.19	80	24.43	-8.62	2.25	1.99		
Jordan	Arabic	-9.02	80	26.20	-14.85	-3.19	1.99	Harsh	-1.06	80	29.30	-7.58	5.46	1.99		
Kazakhstan	Kazakh	-9.02	40	25.29	-17.10	-0.93	2.02	Harsh	-9.72	40	24.18	-17.46	-1.99	2.02	Harsh	Harsh
Kazakhstan	Russian	-0.23	40	13.56	-4.57	4.11	2.02		3.83	40	22.64	-3.41	11.07	2.02		
Korea	Korean	-0.48	80	19.33	-4.79	3.82	1.99		-0.85	80	20.98	-5.52	3.82	1.99		
Kyrgyzstan	Kyrgyz	-1.96	64	23.87	-7.92	4.01	2.00		-13.33	56	24.12	-19.79	-6.87	2.00	Harsh	
Kyrgyzstan	Russian	-9.55	28	22.14	-18.14	-0.97	2.05	Harsh	7.22	28	20.38	-0.68	15.12	2.05		
Latvia	Latvian	5.50	63	29.43	-1.92	12.91	2.00		7.23	64	30.15	-0.30	14.76	2.00		
Latvia	Russian	1.47	24	17.87	-6.08	9.01	2.07		2.61	23	17.40	-4.91	10.14	2.07		

[Part 2/2]

Table 13.8 **Leniency/Harshness analysis**

	Booklet 8 excluding R111Q02B and R111Q06B								Booklet 12							Overall
	Language	Mean	N	Std. deviation	CI_lo	CI_hi	t	Leniency/ Harshness	Mean	N	Std. deviation	CI_lo	CI_hi	t	Leniency/ Harshness	Leniency/ Harshness
Lithuania	Lithuanian	1.28	80	14.85	-2.02	4.59	1.99		-9.69	80	23.21	-14.86	-4.53	1.99	Harsh	
Luxembourg	French	-5.83	22	23.48	-16.24	4.57	2.08		-3.38	22	17.75	-11.25	4.49	2.08		
Luxembourg	German	-0.87	64	15.15	-4.65	2.92	2.00		-1.00	64	21.77	-6.44	4.44	2.00		
Macao-China	Chinese	-12.99	80	22.79	-18.06	-7.92	1.99	Harsh	0.16	80	33.83	-7.37	7.69	1.99		
Mexico	Spanish	6.61	79	24.06	1.22	12.00	1.99	Lenient	-3.70	80	32.97	-11.04	3.64	1.99		
Montenegro	Serbian of a yekavian variant or Montenegrin	-5.38	80	20.31	-9.90	-0.86	1.99	Harsh	-0.62	80	25.81	-6.36	5.13	1.99		
Netherlands	Dutch	7.18	80	21.36	2.43	11.93	1.99	Lenient	4.31	80	29.34	-2.22	10.84	1.99		
New Zealand	English	-5.68	80	20.55	-10.26	-1.11	1.99	Harsh	-0.43	80	17.84	-4.40	3.53	1.99		
Norway	Norwegian	-0.70	80	14.42	-3.91	2.51	1.99		1.12	80	25.14	-4.47	6.72	1.99		
Panama	Spanish	6.00	80	36.78	-2.19	14.19	1.99		17.34	80	35.70	9.39	25.28	1.99	Lenient	
Peru	Spanish	13.40	80	26.52	7.50	19.30	1.99	Lenient	-3.34	80	21.83	-8.20	1.52	1.99		
Poland	Polish	2.89	80	24.82	-2.64	8.41	1.99		-1.38	80	27.91	-7.59	4.83	1.99		
Portugal	Portuguese	4.24	80	19.31	-0.06	8.54	1.99		-1.24	80	32.13	-8.39	5.91	1.99		
Qatar	Arabic	-9.49	80	28.46	-15.82	-3.15	1.99	Harsh	1.27	80	21.20	-3.44	5.99	1.99		
Qatar	English	0.31	40	19.34	-5.87	6.50	2.02		18.07	40	39.00	5.60	30.55	2.02	Lenient	
Romania	Romanian	5.92	80	19.81	1.51	10.32	1.99	Lenient	19.59	80	36.90	11.37	27.80	1.99	Lenient	Lenient
Russian Federation	Russian	-3.21	80	20.62	-7.80	1.38	1.99		-4.93	80	24.16	-10.31	0.44	1.99		
Scotland	English	-0.47	80	14.09	-3.61	2.66	1.99		-1.16	80	23.50	-6.39	4.07	1.99		
Serbia	Serbian	-2.12	80	18.52	-6.24	2.00	1.99		3.27	80	30.27	-3.47	10.01	1.99		
Shanghai-China	Chinese	9.15	80	27.28	3.08	15.22	1.99	Lenient	-5.93	80	36.00	-13.94	2.08	1.99		
Singapore	English	-9.27	80	29.46	-15.83	-2.72	1.99	Harsh	1.81	80	24.84	-3.71	7.34	1.99		
Slovak Republic	Slovak	3.22	80	16.12	-0.37	6.81	1.99		0.79	80	32.84	-6.52	8.10	1.99		
Slovenia	Slovenian	4.73	68	16.09	0.84	8.63	2.00	Lenient	-2.86	70	30.00	-10.01	4.30	1.99		
Spain	Galician	8.62	40	34.80	-2.50	19.75	2.02		11.87	40	27.14	3.19	20.55	2.02	Lenient	
Spain	Spanish	-3.21	68	21.98	-8.53	2.11	2.00		0.93	80	33.96	-6.62	8.49	1.99		
Sweden	Swedish	-3.31	80	14.86	-6.62	-0.01	1.99	Harsh	-21.88	80	39.08	-30.58	-13.18	1.99	Harsh	Harsh
Switzerland	French	-6.95	11	15.47	-17.34	3.44	2.23		-10.74	15	24.18	-24.13	2.65	2.14		
Switzerland	German	5.44	49	23.10	-1.19	12.08	2.01		4.33	39	17.63	-1.39	10.05	2.02		
Chinese Taipei	Chinese	-5.90	80	20.09	-10.37	-1.43	1.99	Harsh	4.00	80	27.37	-2.09	10.09	1.99		
Thailand	Thai	-2.20	80	22.50	-7.21	2.80	1.99		-4.01	79	20.79	-8.67	0.64	1.99		
Trinidad and Tobago	English	-3.24	80	30.36	-9.99	3.52	1.99		6.90	80	27.43	0.80	13.01	1.99	Lenient	
Tunisia	Arabic	7.02	80	25.87	1.26	12.78	1.99	Lenient	-10.11	80	35.83	-18.08	-2.13	1.99	Harsh	
Turkey	Turkish	0.25	80	18.08	-3.78	4.27	1.99		-11.51	80	26.19	-17.34	-5.69	1.99	Harsh	
United Kingdom (excl. Scotland)	English	1.85	80	14.34	-1.34	5.04	1.99		-2.99	80	26.76	-8.94	2.97	1.99		
United States	English	-0.66	80	18.38	-4.75	3.43	1.99		-0.05	80	20.08	-4.52	4.42	1.99		
Uruguay	Spanish	2.79	79	20.12	-1.72	7.30	1.99		5.18	79	21.36	0.40	9.97	1.99	Lenient	

The coder reliability studies formed part of the data adjudication process undertaken by the PISA Technical Advisory Group to ensure the quality of the data which was publicly released.

Notes

1. Albania, Argentina, Azerbaijan, Brazil, Bulgaria, Chile, Colombia, Dubai (UAE), Jordan, Kazakhstan, Kyrgyzstan, Mexico, Panama, Peru, Qatar, Romania, Serbia, Trinidad and Tobago, Tunisia, Uruguay.

2. For some adjudicated entities or certain languages all booklets were selected if, for a variety of reasons, there were fewer than 80 PISA records per booklet per country-by-language unit in the multiple coding exercise.

3. These results are further investigated by a Consortium adjudicator to confirm that the leniency or harshness was found to be on the national coder's side rather than a lenient or harsh international verifier.