13

Coding Reliability Studies

A substantial proportion of the PISA 2012 items were open-ended and required coding by trained personnel. It was important therefore that PISA implemented procedures that 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 using the design described in Chapter 2. 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 Main Survey. These are *i)* the consistency analyses undertaken to inform the Technical Advisory Group about levels of coding agreement for each of the items that require coder judgement, *ii)* the consistency analyses to assess within-country coder agreement and *iii)* the international coder review undertaken to examine possibilities of countries' coding bias. The consistency analyses are described in the next section and the analyses undertaken for international coder review are described in subsequent sections.

CONSISTENCY ANALYSES

Similar to previous cycles, the consistency analyses were undertaken in relation to a subset of constructed-response items. In PISA 2012 all constructed-response items were classified into two sets. The majority of constructed-response items were classified as *constructed-response expert* items, indicating that they would need some judgement from the coders and, therefore, would need to be included in the multiple-coding exercise and the subsequent analyses. A small number of constructed-response items was classified as *constructed-response manual*, which required coding by coders but did not require multiple-coding due to fairly simple, straightforward coding instructions for the item in the coding guide. Constructed-response manual items are the ones that on the one hand cannot be automatically coded due to limitations of the data management software *KeyQuest*, but on the other hand do not require an expert judgment. An example of such instruction can have code '1' for π or 3.14 or any other approximation of π, and 0 for any other response. The symbol π cannot be entered in *KeyQuest* and such item would be coded manually. More details about item classification can be found in Annex A, Tables A1.1 to A1.7.

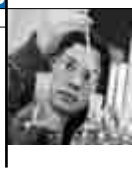
The number of constructed-response expert items varied between domains and also depended on the set of booklets administered by the country (standard or easier). The size of the data available for analysis for each domain depended on the number of constructed-response expert items and whether the test was administered in the country in the major or minor language. The way in which items were allocated to coders for multiple coding depended on whether an item was coded by the country on line or on paper.

PISA 2012 offered seven domains in total. There were four paper-based domains: mathematics, reading, science and financial literacy and three computer-based domains: problem solving, mathematics and reading. Participating countries and economies that have more than one language of instruction administered the test in more than one language, however, if the Consortium expected fewer than 50 students per booklet type for a minor language for a particular domain the locale (country-by-language unit) was exempted from the multiple coding of this domain because the amount of data would be insufficient for analysis. In the Main Survey 76 locales participated in the multiple-coding exercise. Table 13.1 shows which locales participated in multiple coding for which domains and with which options.

In the paper-based assessment there were two groups of countries: those that did standard booklets only (booklets 1-13) and those that did some standard booklets and some non-standard easier booklets (booklets 8-13 and 21-27). There were 17 participants that chose this second option. Both easier and standard booklets contained new and link mathematics items as well as science and reading link items. In addition, there were 18 participants from both groups that administered the financial literacy test (see Chapter 2 for details on the PISA 2012 test design)

In the computer-based assessment there were also two groups of participants. Twelve of them assessed their students in only one computer-based domain, problem solving. In addition, there were 32 participants that assessed their students in three computer-based domains: problem solving, computer-based mathematics and digital reading.

In PISA 2012, eleven participants opted to code constructed-response paper-based items using an online coding system. This system was primarily designed for the coding of the constructed-response computer-based items and was used to code constructed-response computer-based items by all participants administering the PISA 2012 computer-based assessment. Coding of the paper-based items in the online coding system was not compulsory and most of the participants coded constructed-response paper-based items in the paper test booklets and in the specially designed multiple-coding sheets and then entered data into the data management software *KeyQuest*.



[Part 1/2]

Table 13.1 Participation in multiple coding by domain, locale, option

	Locale (country-by-language unit)	Paper-based domains/options				Computer-based domains	
		Mathematics, reading and science	Easier booklet	Financial literacy	On-line coding	Problem solving	Mathematics and digital reading
OECD	Australia-English	Y		Y	Y	Y	Y
	Austria-German	Y			Y	Y	Y
	Belgium-Flemish	Y		Y		Y	Y
	Belgium-French	Y				Y	Y
	Canada-English	Y				Y	Y
	Canada-French	Y				Y	Y
	Chile-Spanish	Y	Y			Y	Y
	Czech Republic-Czech	Y		Y		Y	
	Denmark-Danish	Y				Y	Y
	Estonia-Estonian	Y		Y		Y	Y
	Finland-Finnish	Y				Y	
	France-French	Y		Y		Y	Y
	Germany-German	Y				Y	Y
	Greece-Greek	Y					
	Hungary-Hungarian	Y				Y	Y
	Iceland-Icelandic	Y			Y		
	Ireland-English	Y				Y	Y
	Israel-Arabic	Y		Y	Y	Y	Y
	Israel-Hebrew	Y		Y	Y	Y	Y
	Italy-Italian	Y		Y		Y	Y
	Japan-Japanese	Y				Y	Y
	Korea-Korean	Y			Y	Y	Y
	Luxembourg-French	Y					
	Luxembourg-German	Y					
	Mexico-Spanish	Y	Y				
	Netherlands-Dutch	Y			Y	Y	
	New Zealand-English	Y		Y			
	Norway-Norwegian	Y				Y	Y
	Poland-Polish	Y		Y		Y	Y
	Portugal-Portuguese	Y				Y	Y
	Slovak Republic-Slovak	Y		Y		Y	Y
	Slovenia-Slovenian	Y		Y		Y	Y
	Spain-Basque	Y				Y	Y
	Spain-Catalan	Y		Y		Y	Y
	Spain-Spanish	Y		Y		Y	Y
	Sweden-Swedish	Y			Y	Y	Y
	Switzerland-French	Y			Y		
	Switzerland-German	Y			Y		
	Turkey-Turkish	Y				Y	
	United Kingdom-English	Y				Y[3]	
	United States-English	Y		Y		Y	Y
Partners	Albania-Albanian	Y					
	Argentina-Spanish	Y	Y				
	Brazil-Portuguese	Y	Y			Y	Y
	Bulgaria-Bulgarian	Y	Y			Y	
	Colombia-Spanish	Y	Y	Y	Y	Y	Y
	Costa Rica-Spanish	Y	Y				
	Croatia-Croatian	Y		Y		Y	
	Cyprus-English[1,2]	Y	Y			Y	
	Cyprus-Greek[1,2]	Y	Y			Y	
	Hong Kong-China-Chinese	Y				Y	Y
	Indonesia-Indonesian	Y					
	Jordan-Arabic	Y	Y				
	Kazakhstan-Kazakh	Y	Y				
	Kazakhstan-Russian	Y	Y				
	Latvia-Latvian	Y		Y			
	Lithuania-Lithuanian	Y					
	Macao-China-Chinese	Y				Y	Y
	Malaysia-English	Y				Y	
	Malaysia-Malay	Y				Y	
	Montenegro-Montenegrin	Y				Y	
	Peru-Spanish	Y	Y				
	Qatar-Arabic	Y					
	Qatar-English	Y					
	Romania-Romanian	Y	Y				
	Russian Federation-Russian	Y		Y		Y	Y

[Part 2/2]

Table 13.1 Participation in multiple coding by domain, locale, option

Locale (country-by-language unit)	Paper-based domains/options				Computer-based domains	
	Mathematics, reading and science	Easier booklet	Financial literacy	On-line coding	Problem solving	Mathematics and digital reading
Serbia-Serbian	Y	Y			Y	
Shanghai-China-Chinese	Y		Y		Y	Y
Singapore-English	Y				Y	Y
Chinese Taipei-Chinese	Y			Y	Y	Y
Thailand-Thai	Y					
Tunisia-Arabic	Y	Y				
United Arab Emirates-Arabic	Y	Y			Y	Y
United Arab Emirates-English	Y	Y			Y	Y
Uruguay-Spanish	Y	Y		Y	Y	
Viet Nam-Vietnamese	Y	Y				

1. Footnote by Turkey: The information in this document with reference to « Cyprus » relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

2. Footnote by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

3. England only.

As was the case in the previous cycles, for the PISA 2012 Main Survey a subset of constructed-response expert items from the first cluster in each booklet was multiple coded. Given that each item appeared in each cluster, this design provided around a hundred students per item for major languages and, at the same time, ensured that the amount of missing data was minimised (the amount of missing data and non-responses increases towards the end of the booklet). For the paper-based multiple coding for their main test language each National Centre was required to randomly assign 100 booklets of each type that they were using for testing, and for minority languages the requirement was at least 50 booklets of each type. Four coders participated in the multiple coding exercise.

For the computer-based coding for their main test language in each participant the online coding system randomly assigned at least 100 records of each constructed-response expert item for multiple coding, and for minority languages it assigned at least 50 records of such items for multiple coding. The actual number of responses assigned for multiple coding depended on the number of coders involved in the coding of the item and the number of records available for coding. For example, if four coders coded an item in the main test language and there is a sufficient number of records for single and multiple coding then 100 records of this item would be randomly chosen by the system for multiple coding. If there were five coders, then the number of responses allocated for multiple coding would increase to 125 to ensure that each of these responses are coded 4 times and each coder coded 100 responses from the pool, and so on.

All analysis was done by item. Each response was coded by four coders. Only students with four non-missing codes were used for analysis. The statistics were first aggregated by locale-domain and then for each item internationally.

The following notation is used for consistency analysis:

$i=1,....,I$ – items in the domain

$c=1,...,C_i$ – locale that retained the item[1]

$j=1,...,J_{i,c}$ – students in each locale who attended to the item i

$k=1,...,K_{ic}$ – coders in each locale who coded item i during multiple coding exercise in the locale c

$x_{ijk}=0, 1, 2, ...$ – 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 of the absolute value of the 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} was 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 13.2).

Each disagreement between coders contributes to an increase of the index. For example, if one coder disagrees by one score with three others, all of whom agree with each other, the residual for this one would be 0.75 and the residual for each of three others would be 0.25. In the example in Table 13.2, 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 13.2).

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 13.2.

Table 13.2 Fictitious examples of various indices calculated on locale-domain level

Coder ID	Item A	Item B	Item C	Item D	Item E	Q_{kc}
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**
S_{ic}	**0**	**7.5**	**10**	**6.15**	**10.91**	

For each item in each locale, a locale item reliability index S_{ic} was computed as follows:

13.2

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

and the average across all items coded by a particular coder, Q_{kc} was calculated as:

13.3

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

Examples of some S_{ic} values are shown in the bottom line in Table 13.2 and examples of some Q_{kc} values are shown in the last column in Table 13.2. In this example coder 201 appears less reliable than the three other coders. Coder reliability indices were reported to the countries in the national reports to inform countries of the quality of their coders. This index was not aggregated further.

S_{ic} was further aggregated across all OECD locales that retained the item *i* to form the OECD item reliability index (T_i) for all items except financial literacy and easier mathematics paper-based items. The financial literacy items and the easier mathematics items were aggregated across all locales that retained item *i*.

13.4

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

The OECD/international item reliability index T_i for each item in the multiple-coding exercise is presented in Table 13.3. As was the case in the previous PISA administrations, the items with T_i >7.5 were considered to have high inconsistency of coding and highlighted in grey. The threshold of 7.5 is a rule of thumb which is based on two cycles of experience of analysing variability of coding data for the Field Trial and the Main Survey. As explained previously it can be interpreted as equivalent to the case when one of the coders disagree with three others 20% of the time while three others agree between themselves. Or two coders disagree with two others 15% of the time. The threshold was accepted as high because it does not appear often in the paper-based domains.

Table 13.3 **OECD/International item reliability indices (Ti)**

Item	Number of locales	Ti	S.E.	Ti_SD	Item	Number of locales	Ti	S.E.	Ti_SD
Computer-based mathematics					**Paper-based mathematics**				
CM015Q03	28	5.10	(0.571)	3.019	PM00FQ01	39	3.46	(0.338)	2.113
CM028Q03	28	1.08	(0.186)	0.986	PM00KQ02	41	0.69	(0.096)	0.614
CM038Q05	28	1.87	(0.231)	1.220	PM155Q01	41	1.17	(0.180)	1.149
CM038Q06	28	4.41	(0.476)	2.517	PM155Q02	41	3.50	(0.332)	2.126
Problem solving					PM155Q03	41	5.65	(0.555)	3.554
CP002Q06	33	4.93	(0.394)	2.264	PM406Q01	41	1.08	(0.165)	1.053
CP018Q05	32	2.16	(0.252)	1.424	PM406Q02	41	2.44	(0.333)	2.130
CP034Q05	33	1.15	(0.145)	0.832	PM446Q02	41	1.24	(0.572)	3.664
CP036Q02	33	1.64	(0.352)	2.021	PM462Q01	41	1.65	(0.213)	1.367
CP036Q03	33	1.30	(0.128)	0.737	PM828Q01	41	6.24	(0.475)	3.044
CP041Q02	33	4.62	(0.470)	2.702	PM903Q01	39	3.71	(0.455)	2.839
Digital reading					PM905Q02	39	1.95	(0.190)	1.185
CR002Q05	28	6.23	(0.589)	3.115	PM906Q02	41	8.28	(0.621)	3.976
CR013Q07	28	4.46	(0.512)	2.710	PM909Q03	41	0.51	(0.096)	0.618
CR014Q01	28	5.73	(0.585)	3.096	PM923Q04	39	1.00	(0.160)	1.001
CR017Q07	28	7.26	(0.740)	3.915	PM936Q02	20	0.84	(0.185)	0.827
CR021Q08	23	9.68	(1.194)	5.726	PM942Q02	20	0.70	(0.285)	1.274
Paper-based reading					PM943Q02	39	0.24	(0.063)	0.394
PR220Q01	41	3.87	(0.350)	2.243	PM948Q03	20	0.31	(0.077)	0.345
PR404Q10A	41	4.13	(0.345)	2.209	PM949Q03	39	2.61	(0.322)	2.009
PR404Q10B	41	5.99	(0.494)	3.166	PM953Q02	39	3.78	(0.322)	2.008
PR406Q01	40	2.41	(0.266)	1.684	PM953Q04	39	2.45	(0.225)	1.403
PR406Q02	41	8.05	(0.725)	4.641	PM954Q02	39	1.19	(0.157)	0.983
PR406Q05	41	2.64	(0.321)	2.055	PM954Q04	39	0.89	(0.162)	1.011
PR412Q08	41	5.53	(0.385)	2.467	PM955Q02	41	2.36	(0.240)	1.539
PR420Q06	41	5.37	(0.473)	3.028	PM955Q03	41	1.51	(0.195)	1.246
PR432Q05	41	2.67	(0.220)	1.406	PM961Q02	20	0.62	(0.165)	0.740
PR437Q07	41	6.27	(0.490)	3.140	PM961Q05	20	9.57	(1.739)	7.778
PR446Q06	41	1.79	(0.298)	1.907	PM991Q02	20	0.93	(0.233)	1.042
PR453Q04	41	7.57	(0.666)	4.265	PM992Q03	41	0.82	(0.141)	0.905
PR453Q06	40	3.63	(0.380)	2.405	PM995Q02	39	0.98	(0.468)	2.921
PR455Q02	41	6.04	(0.504)	3.228	**Science**				
PR456Q02	41	3.11	(0.400)	2.563	PS131Q02	40	3.15	(0.373)	2.357
PR456Q06	41	1.34	(0.140)	0.896	PS131Q04D	39	3.77	(0.335)	2.093
PR466Q02	41	1.99	(0.226)	1.449	PS269Q01	41	2.15	(0.268)	1.716
Financial literacy					PS269Q03D	41	2.62	(0.323)	2.067
PF004Q03	19	1.59	(0.322)	1.405	PS326Q01	41	4.42	(0.408)	2.614
PF024Q02	20	7.05	(1.118)	4.998	PS326Q02	41	4.18	(0.449)	2.872
PF028Q02	20	6.96	(0.760)	3.401	PS408Q03	41	5.99	(0.516)	3.305
PF036Q01	20	4.85	(0.622)	2.781	PS425Q03	41	6.95	(0.676)	4.329
PF051Q01	20	2.29	(0.394)	1.760	PS425Q04	41	3.31	(0.433)	2.775
PF051Q02	20	7.60	(1.110)	4.962	PS428Q05	40	3.01	(0.311)	1.965
PF054Q01	20	3.72	(0.510)	2.282	PS438Q03	41	7.68	(0.624)	3.998
PF058Q01	20	3.14	(0.550)	2.458	PS465Q01	40	6.08	(0.495)	3.128
PF068Q01	20	3.07	(0.519)	2.321	PS514Q02	41	1.45	(0.241)	1.541
PF082Q01	20	3.90	(0.577)	2.580	PS514Q03	41	4.61	(0.408)	2.611
PF102Q02	20	7.21	(1.033)	4.621	PS519Q01	41	12.12	(0.972)	6.226
PF103Q01	20	3.38	(0.358)	1.603	PS519Q03	40	6.51	(0.773)	4.888
PF106Q01	20	3.42	(0.516)	2.309					

There were no such items in the computer-based mathematics and problem solving assessments. There was one item with T_i >7.5 in the financial literacy and digital reading and there were two such items in the paper based domains of reading, mathematics and science. Most of the items in paper-based mathematics, computer-based mathematics and problem solving have a satisfactory T_i <3 (highlighted in blue) which means that in these domains most of the items on average were coded consistently across all coders in all locales. Computer-based reading and paper-based domains of reading, science, and financial literacy were more difficult to code and as a result most of the items in these domains have T_i>3.

Table 13.4 compares the international item reliability indices for link items between 2009 and 2012 cycles of PISA. It shows that the index is a stable measure. The change between cycles is statistically significant only for three items: the coding of the reading items PR220Q01 and PR432Q05 improved in 2012 and the coding of the mathematics item PM828Q01 become less consistent.

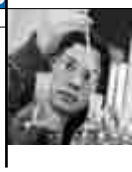


Table 13.4 Comparison of OECD/International item reliability indices (Ti) for link items between PISA 2009 and PISA 2012 cycles

Item	Ti_2012	S.E.	Ti_2009	S.E.	Z-value
Mathematics					
PM155Q01	1.17	(0.180)	1.61	(0.167)	-1.779
PM155Q02	3.50	(0.332)	4.03	(0.402)	-1.017
PM155Q03	5.65	(0.555)	5.18	(0.484)	0.639
PM406Q01	1.08	(0.165)	1.32	(0.123)	-1.170
PM406Q02	2.44	(0.333)	2.21	(0.255)	0.539
PM446Q02	1.24	(0.572)	0.84	(0.114)	0.697
PM462Q01	1.65	(0.213)	1.80	(0.172)	-0.538
PM828Q01	6.24	(0.475)	4.41	(0.377)	3.016
Reading					
PR220Q01	3.87	(0.350)	4.98	(0.428)	-2.001
PR404Q10A	4.13	(0.345)	4.75	(0.392)	-1.186
PR404Q10B	5.99	(0.494)	6.18	(0.525)	-0.254
PR406Q01	2.41	(0.266)	2.47	(0.209)	-0.163
PR406Q02	8.05	(0.725)	8.13	(0.699)	-0.080
PR406Q05	2.64	(0.321)	2.99	(0.296)	-0.797
PR412Q08	5.53	(0.385)	5.56	(0.500)	-0.045
PR420Q06	5.37	(0.473)	6.42	(0.570)	-1.411
PR432Q05	2.67	(0.220)	4.69	(0.487)	-3.784
PR437Q07	6.27	(0.490)	6.68	(0.583)	-0.532
PR446Q06	1.79	(0.298)	2.47	(0.312)	-1.590
PR453Q04	7.57	(0.666)	7.59	(0.626)	-0.028
PR453Q06	3.63	(0.380)	4.46	(0.382)	-1.527
PR455Q02	6.04	(0.504)	6.19	(0.498)	-0.214
PR456Q02	3.11	(0.400)	3.80	(0.356)	-1.286
PR456Q06	1.34	(0.140)	1.72	(0.182)	-1.651
PR466Q02	1.99	(0.226)	2.23	(0.227)	-0.747
Science					
PS131Q02	3.15	(0.373)	3.35	(0.334)	-0.399
PS131Q04D	3.77	(0.335)	4.12	(0.444)	-0.627
PS269Q01	2.15	(0.268)	2.22	(0.188)	-0.214
PS269Q03D	2.62	(0.323)	2.82	(0.368)	-0.416
PS326Q01	4.42	(0.408)	4.35	(0.413)	0.111
PS326Q02	4.18	(0.449)	3.77	(0.371)	0.713
PS408Q03	5.99	(0.516)	5.04	(0.515)	1.298
PS425Q03	6.95	(0.676)	7.22	(0.630)	-0.285
PS425Q04	3.31	(0.433)	3.51	(0.295)	-0.364
PS428Q05	3.01	(0.311)	3.61	(0.399)	-1.184
PS438Q03	7.68	(0.624)	6.88	(0.588)	0.928
PS465Q01	6.08	(0.495)	5.95	(0.546)	0.165
PS514Q02	1.45	(0.241)	1.40	(0.160)	0.169
PS514Q03	4.61	(0.408)	4.39	(0.402)	0.490
PS519Q01	12.12	(0.972)	12.06	(1.028)	0.063
PS519Q03	6.51	(0.773)	6.09	(0.600)	0.325

Let C be a set of σ test languages within the economy participating in the reliability exercise for the domain, D, and δ be the number of items in the domain D retained in the locale (see the list of all items deleted at the national level in Table 12.10, Chapter 12). The average disagreement for each participant across all items in each of the domains is then presented by national domain reliability indices N_{cD}.

13.5

$$N_{cD} = \sum_{c \epsilon C, \delta \epsilon D} \frac{S_{ic}}{\sigma \delta}$$

The national domain indices N_{cD} are presented in Table 13.5 for paper-based domains and in Table 13.6 for computer-based domains. N_{cD} >7.5 are highlighted in grey as unusually high and N_{cD} < 0.5 are highlighted in blue as unusually low. These tables confirm the observation from the previous table that some domains were easier to code consistently than others. The most consistent were the mathematics domains (both paper-based and computer-based) and problem solving with average N_{cD} across all participants less than 3. Paper-based domains of reading, science and financial literacy were coded less consistently with average N_{cD} across all participants around 4.5 (for paper-based reading N_{cD} = 4.37, for paper-based science N_{cD} = 4.66 and for financial literacy N_{cD} = 4.51). The most difficult domain to code was digital reading with N_{cD} = 6.03. This was based on the existence of only four expert-coded items in the digital reading component and should be treated with caution. The online coding software provided a highly

Table 13.5 National domain reliability indices (N_{cD}) for paper-based domains

		Mathematics		Reading		Science		Financial literacy	
		N_{cD}	S.E.	N_{cD}	S.E.	N_{cD}	S.E.	N_{cD}	S.E.
OECD	Australia	2.39	(0.402)	5.22	(0.740)	5.79	(0.802)	5.75	(0.698)
	Austria	1.80	(0.433)	4.77	(0.543)	5.25	(0.881)		
	Belgium[1]	2.70	(0.399)	2.55	(0.311)	4.62	(0.576)	3.65	(0.673)
	Canada	4.43	(0.645)	5.99	(0.659)	10.30	(1.113)		
	Chile	3.34	(0.839)	6.03	(1.040)	8.47	(1.306)		
	Czech Republic	3.04	(0.757)	5.96	(1.109)	3.44	(0.455)	4.16	(0.816)
	Denmark	2.55	(0.414)	5.53	(0.807)	5.93	(0.942)		
	Estonia	2.71	(0.580)	3.46	(0.877)	4.89	(0.875)	5.94	(1.407)
	Finland	1.49	(0.336)	4.33	(0.654)	5.70	(0.865)		
	France	3.17	(0.764)	6.70	(0.909)	7.41	(1.476)	7.21	(1.167)
	Germany	3.24	(0.682)	5.86	(0.662)	5.60	(0.754)		
	Greece	1.07	(0.235)	1.68	(0.309)	2.81	(0.328)		
	Hungary	1.49	(0.357)	3.47	(0.554)	1.98	(0.449)		
	Iceland	2.26	(0.508)	4.87	(0.802)	4.16	(0.841)		
	Ireland	3.48	(0.742)	4.94	(0.878)	6.41	(0.969)		
	Israel	2.66	(0.345)	4.25	(0.654)	5.02	(0.664)	4.20	(0.549)
	Italy	2.48	(0.746)	3.17	(0.446)	6.09	(1.061)	2.56	(0.529)
	Japan	1.02	(0.240)	1.48	(0.235)	2.89	(0.398)		
	Korea	0.73	(0.211)	1.52	(0.310)	1.31	(0.287)		
	Luxembourg	1.84	(0.312)	2.60	(0.337)	3.79	(0.514)		
	Mexico	4.30	(1.430)	7.46	(1.187)	6.71	(1.066)		
	Netherlands	2.81	(0.462)	4.58	(0.801)	6.23	(1.091)		
	New Zealand	2.70	(0.495)	4.57	(0.678)	4.34	(0.635)	4.76	(0.691)
	Norway	3.17	(0.677)	3.81	(0.451)	6.98	(1.446)		
	Poland	2.74	(0.467)	2.41	(0.396)	2.81	(0.419)	2.56	(0.357)
	Portugal	1.41	(0.316)	4.92	(0.808)	1.39	(0.275)		
	Slovak Republic	1.12	(0.271)	5.07	(0.894)	3.55	(0.630)	1.30	(0.259)
	Slovenia	1.81	(0.436)	3.05	(0.400)	4.10	(0.750)	3.89	(0.827)
	Spain	2.48	(0.320)	5.87	(0.558)	5.53	(0.742)	4.02	(0.653)
	Sweden	3.57	(0.576)	4.26	(0.536)	4.85	(0.969)		
	Switzerland	1.39	(0.238)	2.23	(0.316)	2.51	(0.451)		
	Turkey	1.30	(0.301)	3.40	(0.752)	0.96	(0.237)		
	United Kingdom	2.01	(0.306)	5.17	(0.599)	5.34	(0.544)		
	United States	2.12	(0.447)	4.25	(0.683)	6.03	(0.814)	5.34	(1.189)
Partners	Albania	0.07	(0.070)	0.00	(0.000)	0.52	(0.354)		
	Argentina	0.23	(0.066)	0.19	(0.063)	0.59	(0.094)		
	Brazil	3.76	(1.160)	7.99	(1.096)	8.40	(1.439)		
	Bulgaria	3.73	(0.838)	6.92	(1.201)	5.78	(0.971)		
	Colombia	2.66	(0.640)	2.87	(0.385)	4.40	(0.664)	4.28	(0.953)
	Costa Rica	0.35	(0.150)	8.04	(1.458)	12.98	(2.240)		
	Croatia	1.67	(0.347)	5.59	(1.066)	6.52	(1.277)	6.44	(1.036)
	Cyprus[2, 3]	0.85	(0.184)	0.27	(0.077)	0.60	(0.102)		
	Hong Kong-China	2.61	(0.524)	4.28	(0.863)	7.70	(1.346)		
	Indonesia	4.86	(1.120)	17.47	(1.763)	11.57	(1.512)		
	Jordan	0.17	(0.051)	0.27	(0.090)	0.50	(0.120)		
	Kazakhstan	0.61	(0.090)	0.76	(0.118)	1.99	(1.018)		
	Latvia	4.04	(0.903)	11.75	(1.512)	10.45	(1.605)	9.29	(1.560)
	Lithuania	1.95	(0.435)	3.26	(0.512)	3.16	(0.591)		
	Macao-China	1.44	(0.260)	3.90	(0.524)	1.38	(0.169)		
	Malaysia	5.50	(0.897)	9.09	(0.932)	8.58	(0.974)		
	Montenegro	1.37	(0.373)	6.77	(1.172)	9.26	(1.182)		
	Peru	1.38	(0.673)	2.65	(0.393)	3.69	(0.715)		
	Qatar	0.67	(0.139)	1.56	(0.240)	0.48	(0.172)		
	Romania	0.32	(0.091)	6.31	(0.882)	0.75	(0.175)		
	Russian Federation	0.45	(0.155)	1.60	(0.309)	1.01	(0.180)	4.57	(0.737)
	Serbia	3.52	(0.750)	5.59	(0.769)	3.70	(0.502)		
	Shanghai-China	1.25	(0.375)	2.23	(0.482)	0.80	(0.177)	1.32	(0.448)
	Singapore	0.30	(0.084)	0.08	(0.046)	0.46	(0.191)		
	Chinese Taipei	1.28	(0.285)	2.15	(0.433)	2.99	(0.712)		
	Thailand	0.00	(0.000)	0.12	(0.056)	0.13	(0.073)		
	Tunisia	2.81	(0.737)	6.59	(1.022)	12.38	(1.875)		
	United Arab Emirates	1.93	(0.402)	2.60	(0.325)	3.25	(0.473)		
	Uruguay	2.44	(0.805)	6.19	(0.751)	3.32	(0.461)		
	Viet Nam	2.83	(1.091)	7.17	(2.461)	7.76	(1.689)		
	Mean (all participants)	**2.12**		**4.37**		**4.66**		**4.51**	
	SD (all participants)	**1.26**		**2.93**		**3.14**		**2.00**	

1. Only the Flemish community of Belgium participated in the financial literacy assessment.

2. Footnote by Turkey: The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

3. Footnote by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

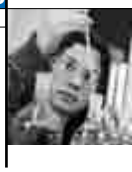


Table 13.6 **National domain reliability indices (N_{cD}) for computer-based domains**

	Participant	Problem solving		Mathematics		Digital reading	
		N_{cD}	S.E.	N_{cD}	S.E.	N_{cD}	S.E.
OECD	Australia	3.56	(1.00)	3.73	(1.25)	11.08	(1.08)
	Austria	2.73	(1.06)	2.06	(1.00)	5.58	(1.06)
	Belgium	2.07	(0.63)	3.92	(1.03)	5.76	(0.87)
	Canada	3.07	(0.61)	3.19	(0.70)	9.63	(2.30)
	Chile	3.63	(1.25)	4.17	(1.22)	8.08	(0.80)
	Czech Republic	2.60	(0.57)				
	Denmark	2.52	(0.66)	4.47	(1.87)	10.05	(2.21)
	Estonia	2.42	(0.60)	3.13	(0.96)	5.79	(1.26)
	Finland	1.03	(0.46)				
	France	4.51	(2.34)	5.47	(2.21)	13.20	(1.89)
	Germany	3.10	(0.88)	3.88	(1.82)	8.23	(0.96)
	Hungary	2.27	(0.71)	1.97	(0.56)	4.14	(0.67)
	Ireland	2.99	(0.97)	2.61	(0.48)	8.60	(1.38)
	Israel	2.30	(0.45)	4.00	(1.02)	6.09	(1.40)
	Italy	3.42	(0.87)	4.81	(1.73)	3.19	(0.57)
	Japan	2.14	(0.47)	2.26	(1.10)	4.04	(0.97)
	Korea	0.65	(0.36)	0.73	(0.18)	2.33	(0.79)
	Netherlands	3.57	(0.77)				
	Norway	2.67	(0.89)	3.62	(1.20)	4.40	(1.27)
	Poland	1.09	(0.27)	1.13	(0.92)	5.00	(0.94)
	Portugal	4.18	(1.31)	1.55	(0.31)	6.85	(2.59)
	Slovak Republic	3.77	(1.64)	1.97	(1.03)	6.37	(0.97)
	Slovenia	0.96	(0.37)	1.01	(0.20)	2.14	(0.61)
	Spain	1.33	(0.26)	1.62	(0.48)	5.26	(0.68)
	Sweden	2.68	(0.70)	6.08	(1.69)	6.51	(1.37)
	Turkey	3.05	(0.98)				
	United Kingdom[1]	5.00	(1.90)				
	United States	3.24	(1.04)	4.67	(1.82)	6.20	(1.13)
Partners	Brazil	2.41	(1.04)	1.99	(0.84)	5.32	(0.81)
	Bulgaria	4.40	(1.28)				
	Colombia	3.12	(0.85)	4.51	(2.11)	6.71	(0.56)
	Croatia	5.28	(2.51)				
	Cyprus[2, 3]	1.67	(0.48)				
	Hong Kong-China	2.55	(1.19)	4.50	(1.86)	5.43	(2.95)
	Macao-China	0.06	(0.06)	0.45	(0.30)	2.47	(0.62)
	Malaysia	2.49	(0.53)				
	Montenegro	7.08	(2.49)				
	Russian Federation	1.38	(0.20)	2.35	(0.93)	4.86	(0.94)
	Serbia	2.21	(0.88)				
	Shanghai-China	1.97	(0.46)	3.29	(1.42)	5.30	(0.78)
	Singapore	1.60	(0.80)	2.71	(1.64)	6.48	(3.16)
	Chinese Taipei	2.15	(1.17)	1.56	(0.65)	4.08	(0.67)
	United Arab Emirates	1.68	(0.44)	2.11	(0.50)	3.91	(0.68)
	Uruguay	3.27	(0.73)				
	Mean	**2.72**		**2.99**		**6.03**	
	SD	**1.31**		**1.44**		**2.52**	

1. England only.

2. Footnote by Turkey: The information in this document with reference to « Cyprus » relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

3. Footnote by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

sophisticated means of coding student responses, which accommodate all but four of the reading items, and these were the items requiring the most complex judgements. Historically, reading items have always been more difficult to code than mathematics items.

INTERNATIONAL CODER REVIEW

Control scripts

With the introduction of the online coding system the opportunity was provided in the PISA 2012 administration to develop an objective alternative for the international coder review. The item developers provided responses for constructed response expert items for each domain and correct coding for each response. These responses are referred to as *control scripts* in this chapter. National Centres translated control scripts and scanned translations into the online coding system where they were presented to coders as student responses, indistinguishable from other student responses.

This was done for all domains that were coded on line (all computer-based domains and for some countries for paper-based domains). Control scripts were provided to allow for international bias analysis by comparison of codes given to the same response by coders from different National Centres on the one hand and by item developers on the other hand. Table 13.7 shows participation in the control-script exercise by domain. Fifty-five locales coded control scripts for at least one domain. The use of control scripts enabled National Centres to monitor the quality of their coding in real time, since the online coding system allowed National Centres to re-train coders when observed discrepancies between coders and provided control scripts were high.

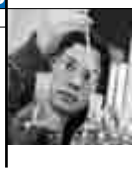
Table 13.7 Participation in control scripts bias analysis by domain

		Paper-based domains		Computer-based domains	
	Locale	Mathematics, reading, science	Financial literacy	Mathematics, digital reading	Problem solving
OECD	Australia-English	Y	Y	Y	Y
	Austria-German	Y		Y	Y
	Belgium-Flemish			Y	Y
	Belgium-French			Y	Y
	Canada-English			Y	Y
	Canada-French			Y	Y
	Chile-Spanish			Y	Y
	Czech Republic-Czech				Y
	Denmark-Danish	Y		Y	Y
	Estonia-Estonian			Y	Y
	Estonia-Russian			Y	Y
	Finland-Finnish				Y
	France-French			Y	Y
	Germany-German			Y	Y
	Hungary-Hungarian			Y	Y
	Iceland-Icelandic	Y			
	Ireland-English			Y	Y
	Israel-Arabic	Y	Y	Y	Y
	Israel-Hebrew	Y	Y	Y	Y
	Italy-Italian			Y	Y
	Japan-Japanese			Y	Y
	Korea-Korean	Y		Y	Y
	Netherlands-Dutch				Y
	Norway-Norwegian			Y	Y
	Poland-Polish			Y	Y
	Portugal-Portuguese			Y	Y
	Slovak Republic-Slovak			Y	Y
	Slovenia-Slovenian			Y	Y
	Spain-Basque			Y	Y
	Spain-Catalan			Y	Y
	Spain-Spanish			Y	Y
	Sweden-Swedish	Y		Y	Y
	Switzerland-French	Y			
	Switzerland-German	Y			
	Turkey-Turkish				Y
	United Kingdom-English				Y[3]
	United States-English	Y	Y	Y	Y
Partners	Brazil-Portuguese			Y	Y
	Bulgaria-Bulgarian				Y
	Colombia-Spanish	Y	Y	Y	Y
	Croatia-Croatian				Y
	Cyprus-English[1, 2]				Y
	Cyprus-Greek[1, 2]				Y
	Hong Kong-China-Chinese			Y	Y
	Macao-China-Chinese			Y	Y
	Malaysia-English				Y
	Malaysia-Malay				Y
	Montenegro-Montenegrin				Y
	Russian Federation-Russian			Y	Y
	Shanghai-China-Chinese			Y	Y
	Singapore-English			Y	Y
	Chinese Taipei-Chinese	Y		Y	Y
	United Arab Emirates-Arabic	Y		Y	Y
	United Arab Emirates-English	Y		Y	Y
	Uruguay-Spanish	Y			Y

1. Footnote by Turkey: The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

2. Footnote by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

3. England only.



The items and the number of control scripts for each domain are listed in Table 13.8. The number of control scripts per item was determined by item developers. To avoid dependency between scripts each script represented a different type of answer. This approach, however, often provided only single digit number of scripts per item, which is essentially equivalent to the single digit number of student responses per item per locale. Because the number of scripts available was relatively small, and the number of countries participating in this new approach to international coder review was limited, the volume of material generated through the use of control scripts was not sufficient to perform a robust analysis of the outcomes of the procedures used. Nevertheless, it is expected on the basis of the experience gained that higher levels of participation in future would lead to better data volumes and this would permit analysis of outcomes to be carried out. In future control scripts can be used if there are more constructed response items in each of the computer-based domains or if all participants in the paper-based domains use online coding or both. For the Field Trial, control scripts still can be used as an effective tool to improve coding guides and to identify items that are difficult to code.

Comparison of student achievement in constructed response and all other items

Since the use of control scripts did not provide data of sufficient volume for identification of bias (Table 13.8) a different statistical procedure was employed. In summary, the procedure compared two differences between student achievements in each of the 100 achievement categories. The difference I_j (j=1,…100) between achievement in constructed response and all other items internationally was used as a benchmark. This statistic was based on the plausible values for all PISA students who participated in the domain. It was compared to the differences L_{kj} between student achievement in constructed response and all other items in each participant k. This statistics was based on the plausible values for the

Table 13.8 The list of items for which control scripts were provided

Item ID	Number of control scripts	Item ID	Number of control scripts
Computer-based mathematics		**Paper-based mathematics**	
CM015Q03	10	PM00FQ01	8
CM028Q03	8	PM00KQ02	8
CM038Q05	8	PM155Q01	3
CM038Q06	9	PM155Q02	5
Problem solving		PM155Q03	4
CP002Q06	14	PM406Q01	4
CP018Q05	8	PM406Q02	4
CP034Q05	5	PM462Q01	6
CP036Q02	5	PM828Q01	2
CP036Q03	6	PM903Q01	8
CP041Q02	11	PM905Q02	8
Digital reading		PM906Q02	7
CR002Q05	16	PM949Q03	8
CR013Q07	14	PM953Q04	8
CR014Q01	17	PM955Q03	7
CR017Q07	18	PM961Q05	7
CR021Q08	19	PM991Q02	7
Science		**Paper-based reading**	
PS131Q02	9	PR404Q10A	5
PS131Q04	9	PR404Q10B	4
PS269Q01	10	PR406Q01	6
PS269Q03	9	PR406Q02	7
PS326Q01	8	PR406Q05	7
PS326Q02	7	PR412Q08	3
PS408Q03	8	PR420Q06	5
PS425Q03	8	PR420Q10	4
PS425Q04	9	PR432Q05	4
PS428Q05	9	PR437Q07	5
PS438Q03	10	PR446Q06	2
PS465Q01	10	PR453Q04	3
PS498Q04	10	PR453Q06	4
PS514Q02	10	PR455Q02	5
PS514Q03	10	PR456Q02	4
PS519Q01	10	PR456Q06	3
PS519Q03	8	PR466Q02	4
Financial literacy			
PF004Q03	12		
PF024Q02	13		
PF028Q02	13		
PF036Q01	19		
PF051Q01	15		
PF082Q01	14		
PF102Q02	17		

subset of students from this participant. The assumption was that if L_{kj} behaves statistically differently from $_{Ij}$ persistently across j, it may partially be attributed to coding bias.

We know from previous research (Routitsky and Turner, 2003) that there can be differences in performance on items of different format (e.g. multiple choice and constructed response items) and that the magnitude of this difference varies for students of different abilities. Therefore, L_{kj} were expected to vary across achievement categories (j=1,…,100) within participants (as well as between them) and were compared to the corresponding I_j which was used as a benchmark.

In detail, the procedure was implemented as follows.

International item parameters were used for all domains. For paper-based mathematics, only common items were used; the items that were unique to standard booklets and items that were unique to easier booklets were excluded to facilitate comparison between countries that used easier booklets and countries that used standard booklets.

For each domain, the items in the item parameter file were divided into two groups. One group contained constructed response items ("CR" item group), and the other group contained the rest of the items ("Rest" item group). Item parameters of each group were adjusted to a parameter mean of zero nationally: if an item was deleted from participant data, a separate item parameter file was created by excluding this item and re-adjusting all item parameters to the mean of zero.

The ACER *ConQuest* (Adams, Wu and Wilson, 2012) programme file was created to estimate plausible values for student achievement based on each item group of each domain within each participant using a 2-dimensional model. For each domain the plausible values estimated by ACER *ConQuest* were read into SPSS© (2010) and processed as described below.

Let W_D be the weighted number of students for the domain D across all participants.

Let $\{RP_{s,i}\}$ (s=1,…,W_D; i=1,…,5), be a set of plausible values derived for the "Rest" item group of the domain D.

For each i=1,…,5 RP_{si} was sorted in ascending order and divided into 100 equally weighted sets A_{ji} (j=1,…,100; i=1,…,5) of the W_D=W_D/100 size. For each i=1,…,5 the new variable S_i was constructed. All students from A_{ji} were assigned $S_i = j$, meaning that according to the plausible value i the student belongs to the achievement group j. Note that for the same student the value of j could be different for different plausible values.

Let $\{CP_{s,i}\}$ (s=1,…,W_D; i=1,…,5), be a set of plausible values derived for the "CR" (constructed response) item group of the domain D.

For each set A_{ji} (j=1,…,100; i=1,…,5) the mean difference was calculated as follows

13.6

$$MI_{ji} = \frac{\sum_{s \in A_{ji}} v_s \left(CP_{s,i} - RP_{s,i}\right)}{w_D},$$

where v_s is a total student weight for students (see Chapter 8 for details about weight estimation).

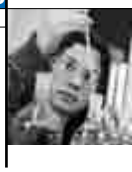
The difference I_j (j=1,…,100) between achievement in constructed response and all other items internationally was calculated as the average between 5 differences MI_{ji}:

13.7

$$I_j = \frac{\sum_{i=1}^{5} MI_{ji}}{5}$$

I_j can be interpreted as achievement in constructed response items relative to the achievement in all other items and will be called in the rest of this chapter *relative international achievement*.

The differences L_{kj} between student achievement in constructed response and all other items in each participant k were calculated as follows.



Let $B_{k\,ji}$ be a subset of A_{ji} from a participant k: $B_{k\,ji} \subset A_{ji}$ and m_{kiD} the weighted number of students in this set. Then

13.8

$$ML_{kji} = \frac{\sum_{s \in B_{kji}} v_s \left(CP_{s,i} - RP_{s,i}\right)}{m_{kiD}}$$

13.9

$$L_{kj} = \frac{\sum_{i=1}^{5} ML_{kji}}{5}$$

L_{kj} can be interpreted as achievement in constructed response items relative to the achievement in all other items within the locale and will be called in the rest of this chapter *relative locale achievement*.

Standard errors for I_j and L_{kj} were calculated using the balanced repeated replication method. Standard errors were used to run z-tests with α=0.05 to find whether L_{kj} is significantly different from I_j. Z-test showed that the difference L_{kj} - I_j was statistically significantly different from zero for some j within some participant k. However, the differences were not systematic across different achievement groups j. Therefore, the next step was to identify the size of this difference for each participating country and economy. To identify the size of the difference between L_{kj} - I_j within a particular participant the following approach was employed.

Let CIL_{kj} be a lower boundary of the confidence interval of the difference L_{kj} - I_j. Then,

if $CIL_{kj} > 0$, the adjusted plausible values for constructed response items $RCP_{s,i}$ were computed for all plausible values $s \in B_{k\,ji}$ as

13.10

$$RCP_{s,i} = CP_{s,i} - CIL_{kj}$$

Let CIU_{kj} be an upper boundary of the confidence interval of the difference L_{kj} - I_j. Then,

if $CIL_{kj} < 0$, corrected plausible values for constructed response items $RCP_{s,i}$ were computed for all plausible values $s \in B_{k\,ji}$ as

13.11

$$RCP_{s,i} = CP_{s,i} - CIU_{kj}$$

Finally, if $CIL_{kj} < 0 < CIU_{kj}$,

13.12

$$RCP_{s,i} = CP_{s,i}$$

The adjusted plausible values for constructed response items $RCP_{s,i}$ were then compared to the initial plausible values $CP_{s,i}$ within each participating country/economy by calculating the average difference G_{kj} [13.14] and its standard error as well as standard deviation SD(G_{kj}) using the balanced repeated replication method.

13.13

$$MG_{kji} = \frac{\sum_{s \in B_{kji}} v_s \left(RCP_{s,i} - CP_{s,i}\right)}{m_{kiD}}$$

13.14

$$G_{kj} = \frac{\sum_{i=1}^{5} MG_{kji}}{5}$$

Table 13.9 Percent of students in the lowest level of proficiency and amount of difference between national and international relative achievement in constructed response items (G_{kj}) by domain

		Paper-based domains								Computer-based domains			
		Mathematics		Reading		Science		Financial literacy		Problem solving		Digital reading	
	Participant	% below Level 1	G_{kj}	% below Level 1a	G_{kj}	% below Level 1	G_{kj}	% below Level 1	G_{kj}	% below Level 1	G_{kj}	% below Level 2	G_{kj}
OECD	Australia	6.1	-0.13	4.01	-0.08	3.4	-0.14	3.39	-0.01	5.03	-0.02	12.46	-0.01
	Austria	5.7	-0.07	5.66	-0.03	3.6	-0.06			6.49	0.00	20.23	0.01
	Belgium[1]	7.0	-0.10	5.74	-0.05	5.8	-0.05	2.72	-0.03	9.08	0.00	17.19	0.00
	Canada	3.6	-0.16	2.86	-0.07	2.4	-0.14			5.10	-0.02	8.46	-0.01
	Chile	22.0	0.06	9.08	0.03	8.1	0.01			15.15	0.03	29.30	0.03
	Czech Republic	6.8	-0.06	4.12	-0.04	3.3	-0.16	3.09	-0.01	6.53	-0.02		
	Denmark	4.4	-0.11	3.90	-0.03	4.7	-0.05			7.30	0.00	14.23	0.01
	Estonia	2.0	-0.13	1.46	-0.08	0.5	-0.24	0.79	-0.03	4.01	-0.01	11.43	0.00
	Finland	3.3	-0.14	3.12	-0.09	1.8	-0.20			4.46	0.00		
	France	8.7	-0.06	6.99	-0.04	6.1	-0.06	8.68	0.00	6.63	-0.01	13.77	0.00
	Germany	5.5	-0.11	3.79	-0.07	2.9	-0.17			7.48	0.00	19.14	0.00
	Greece	14.5	0.01	8.47	-0.02	7.4	-0.01						
	Hungary	9.9	-0.03	5.95	-0.02	4.1	-0.06			17.22	0.01	32.48	0.03
	Iceland	7.5	-0.04	7.66	-0.02	8.0	-0.01						
	Ireland	4.8	-0.09	2.12	-0.10	2.6	-0.15			7.02	0.00	9.41	-0.01
	Israel	15.9	0.00	10.69	-0.03	11.2	-0.01	11.65	0.03	21.86	0.05	31.03	0.09
	Italy	8.5	-0.06	6.77	-0.07	4.9	-0.15	7.93	0.01	5.18	0.00	15.68	0.00
	Japan	**3.2**	**-0.23**	**3.06**	**-0.17**	**2.0**	**-0.32**			1.79	-0.07	4.92	-0.04
	Korea	**2.7**	-0.26	**2.15**	-0.15	**1.2**	-0.27			2.14	-0.05	3.95	-0.03
	Luxembourg	8.8	-0.03	8.33	-0.02	7.2	-0.03						
	Mexico	22.8	0.14	13.58	0.11	12.6	0.24						
	Netherlands	3.8	-0.11	3.72	-0.04	3.1	-0.11			7.36	0.00		
	New Zealand	7.5	-0.07	5.29	-0.05	4.7	-0.10	7.26	-0.01				
	Norway	7.2	-0.05	5.41	-0.05	6.0	-0.05			8.12	0.00	16.65	0.00
	Poland	3.3	-0.10	2.47	-0.09	1.3	-0.16	1.88	-0.01	10.04	0.00	22.39	0.00
	Portugal	8.9	-0.05	6.47	-0.02	4.7	-0.04			6.48	0.00	19.16	0.01
	Slovak Republic	11.1	-0.02	12.00	0.00	9.2	0.00	10.75	0.01	10.72	0.01	22.56	0.01
	Slovenia	5.1	-0.07	6.17	0.00	2.4	-0.03	5.32	0.00	11.39	0.02	25.12	0.04
	Spain	7.8	-0.05	5.75	-0.06	3.7	-0.12	4.94	0.00	13.14	0.00	26.16	0.01
	Sweden	9.5	-0.03	8.84	-0.03	7.3	-0.04			8.82	0.00	16.72	0.00
	Switzerland	3.6	-0.13	3.43	-0.06	3.0	-0.11						
	Turkey	15.5	0.00	5.06	-0.01	4.4	0.01			10.98	0.06		
	United Kingdom[2]	7.8	-0.06	5.44	-0.06	4.3	-0.11			5.55	-0.01		
	United States	8.0	-0.04	4.32	-0.02	4.2	-0.02	6.03	0.00	5.66	-0.01	12.61	0.00
Partners	Albania	32.5	0.12	27.98	0.11	23.5	0.16						
	Argentina	34.9	0.14	25.86	0.11	19.8	0.15						
	Brazil	35.2	0.20	18.77	0.16	18.6	0.28			21.89	0.06	37.16	0.09
	Bulgaria	20.0	0.02	20.80	0.01	14.4	0.01			33.33	0.24		
	Colombia	41.6	0.24	20.41	0.09	19.8	0.18	32.64	0.09	33.16	0.12	54.85	0.12
	Costa Rica	23.6	0.11	8.12	0.04	8.6	0.11						
	Croatia	9.5	-0.02	4.75	-0.03	3.2	-0.05	5.28	0.00	12.05	0.01		
	Cyprus[3, 4]	19.0	0.02	15.81	0.09	14.4	0.05			19.55	0.11		
	Hong Kong-China	**2.6**	**-0.29**	**1.51**	**-0.22**	**1.2**	**-0.37**			3.33	-0.03	7.57	-0.03
	Indonesia	42.3	0.22	20.41	0.16	24.7	0.36						
	Jordan	36.5	0.20	22.40	0.09	18.2	0.15						
	Kazakhstan	14.5	0.05	21.50	0.20	11.3	0.15						
	Latvia	4.8	-0.05	4.38	-0.04	1.8	-0.07	1.97	-0.02				
	Lithuania	8.7	-0.03	5.55	-0.01	3.4	-0.05						
	Macao-China	**3.2**	**-0.17**	2.46	-0.08	1.4	-0.17			1.55	-0.04	6.96	0.00
	Malaysia	23.0	0.08	22.23	0.10	14.5	0.12			22.66	0.14		
	Montenegro	27.5	0.08	17.55	0.05	18.7	0.14			30.00	0.17		
	Peru	47.0	0.27	30.43	0.18	31.5	0.39						
	Qatar	47.0	0.26	32.55	0.23	34.6	0.36						
	Romania	14.0	0.02	12.88	0.02	8.7	0.05						
	Russian Federation	7.5	-0.05	6.29	0.00	3.6	-0.03	5.53	0.00	6.76	0.00	24.61	0.03
	Serbia	15.5	0.01	11.86	0.02	10.3	0.03			10.27	0.03		
	Shanghai-China	**0.8**	**-0.52**	**0.39**	**-0.32**	**0.3**	**-0.52**	**0.32**	**-0.29**	3.09	-0.01	7.88	-0.01
	Singapore	**2.2**	**-0.34**	**2.41**	-0.17	**2.2**	-0.27			2.01	-0.05	4.35	-0.08
	Chinese Taipei	**4.5**	**-0.27**	3.05	-0.11	1.6	-0.19			3.44	-0.01	11.08	0.00
	Thailand	19.1	0.05	8.87	0.05	7.0	0.05						
	Tunisia	36.5	0.17	21.68	0.07	21.3	0.19						
	United Arab Emirates	20.5	0.06	13.71	0.06	11.3	0.07			30.28	0.17	50.48	0.20
	Uruguay	29.2	0.09	21.13	0.06	19.7	0.10			32.39	0.22		
	Viet Nam	3.6	-0.13	1.60	-0.11	0.9	-0.22						
	R^2	**0.77**		**0.74**		**0.75**		**0.29**		**0.87**		**0.82**	
	R^2 adjusted for outliers	**0.94**		**0.82**		**0.83**		**0.91**		**N/A**		**N/A**	

1. Only the Flemish community of Belgium took part in the assessment of financial literacy.

2. Only England took part in the assessment of problem solving.

3. Footnote by Turkey: The information in this document with reference to « Cyprus » relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

4. Footnote by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

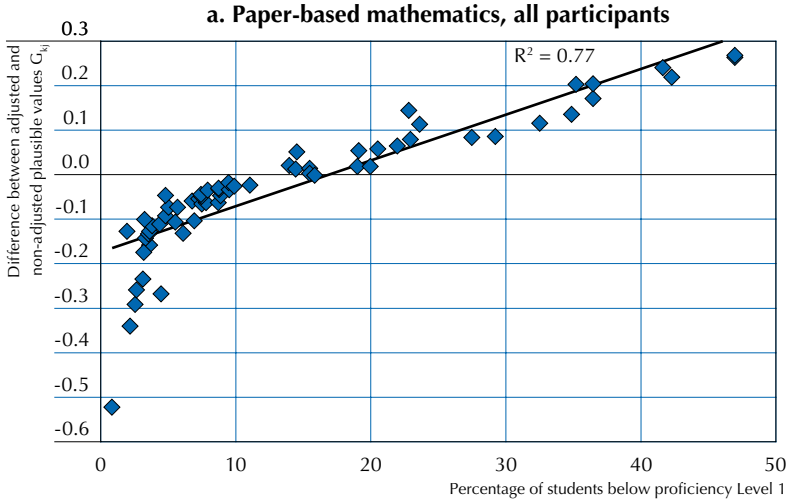
Due to the way adjusted plausible values were calculated, G_{kj} can be interpreted as the difference between relative national and international constructed response achievement (that is achievement in constructed response items relative to the achievement in all other items). Index G_{kj} was calculated for all domains except computer-based mathematics because computer-based mathematics had only 4 constructed response items; and for all participants except Liechtenstein. Data from only four items were deemed to be insufficient to calculate plausible values for all students. The number of students in data for Liechtenstein (293) was insufficient to estimate results separately for each type of item in each of 100 sets B_{kji}.

As mentioned earlier, we know from previous research that differential behaviour of various item formats depends on the level of a student's achievement (Routitsky and Turner, 2003). Thus, we would like to see how much of the variation in difference between national and international relative constructed response achievement G_{kj} can be explained by the percent of students in the lowest level of proficiency for each domain before we ascribe responsibility for any of this variation to country specific coding bias. The levels of proficiency are described in Volume I of the *PISA 2012 Results* (OECD, 2014). The lowest level of proficiency was chosen because students at this level are most likely to skip constructed response items and so would be least affected by coding bias and, therefore, correlation between the percentage of students in the lowest level of achievement and G_{kj} will be least confounded by coding bias. Table 13.9 shows side-by-side for each participant (except Liechtenstein) the percentage of students in the lowest level of proficiency for each domain except computer-based mathematics.

Figure 13.1a illustrates the relationship between G_{kj} and the percentage of students below proficiency level 1 for paper-based mathematics. It shows that 77% in G_{kj} variation is explained by the proportion of low achieving students in the country. G_{kj} shows that students from low achieving countries are achieving relatively better on the constructed response items than the students from the high achieving countries (relative to their achievement on all other items). One possible explanation of this could be that low achieving countries have some positive bias towards their students or high achieving countries have some negative bias towards their students or both. Eliminating this, were it the case, may only increase the distance between countries not the general ranking. However an alternative explanation is that G_{kj} is higher in the low achieving countries due to the fact that their achievement in all other items is so much lower.

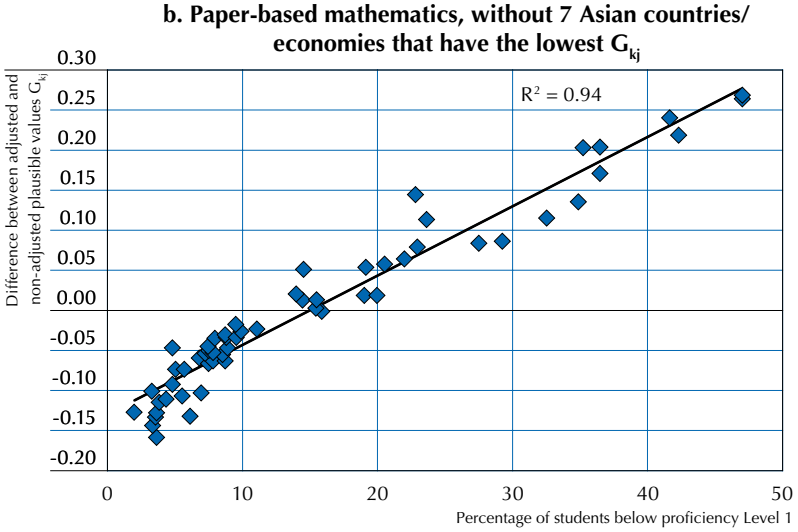
■ Figure 13.1 [Part 1/2] ■

Relationship between G_{kj} and percentage of students below proficiency Level 1 for paper-based mathematics



■ Figure 13.1 [Part 2/2] ■

Relationship between G_{kj} and percentage of students below proficiency Level 1 for paper-based mathematics

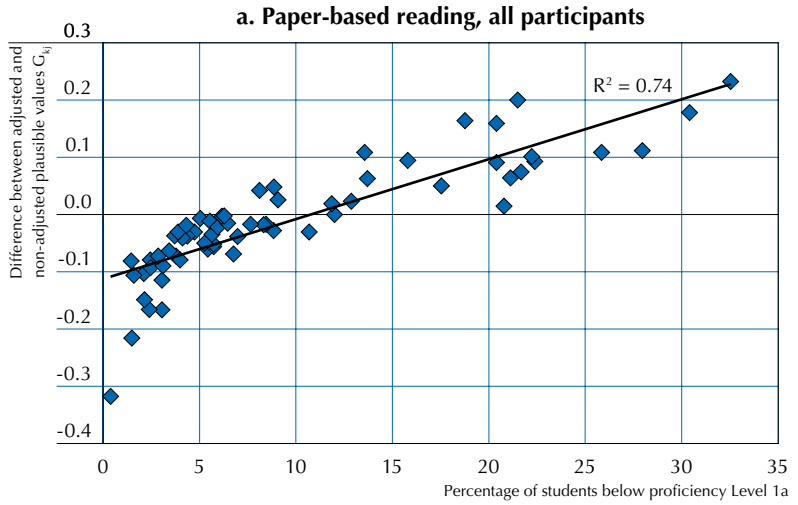


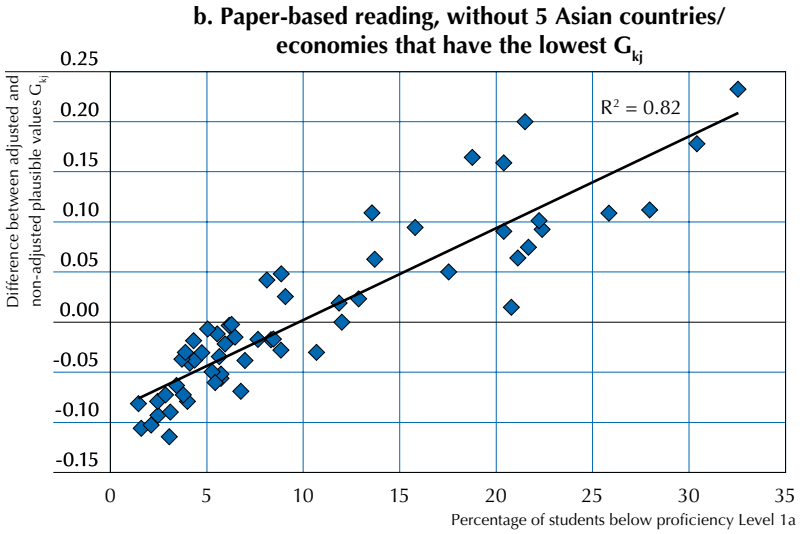
This is indeed the case and this would explain why the correlation with the proportion of students from the lowest level of proficiency is so high. We can also see some outliers in the left bottom part of the graph. These outliers belong to seven Asian participants that have G_{kj} ranging from -0.52 to -0.17 and a percentage of students below proficiency Level 1 ranging from 0.8% to 4.5% (Figure 13.1a and Table 13.9). These participants are Shanghai-China, Singapore, Hong Kong-China, Chinese Taipei, Korea, Japan and Macao-China, and are highlighted in bold in Table 13.9. Numerous researches comparing education in eastern and western countries (Leung et al., 2006) noticed that curriculum, teaching methods and assessment practices in these participants are different from those in other regions and have some similarities with each other. One possibility is that it is these factors that contribute to the variation in G_{kj} above and beyond the variation explained by the percentage of students in the lower level of achievement. The mechanism for this, however, is unclear and other reasons should be explored in the future. If we calculate R^2 without the above seven Asian participants, we can see that for the rest of PISA participants the proportion of low achieving students explains 94 % of variation in G_{kj} (Figure 13.1b).

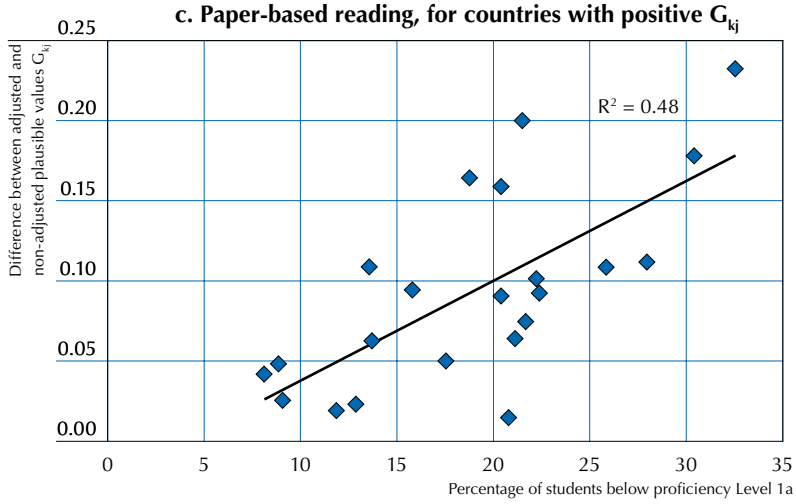
There are similar results for paper-based reading and science. Figure 13.2 shows that 74 % in G_{kj} variation for reading is explained by the proportion of low achieving students in the country and Figure 13.3 shows that 75% in G_{kj} variation for science is explained by the proportion of low achieving students in the country.

■ Figure 13.2 ■

Relationship between G_{kj} and percentage of students below proficiency Level 1a for paper-based reading

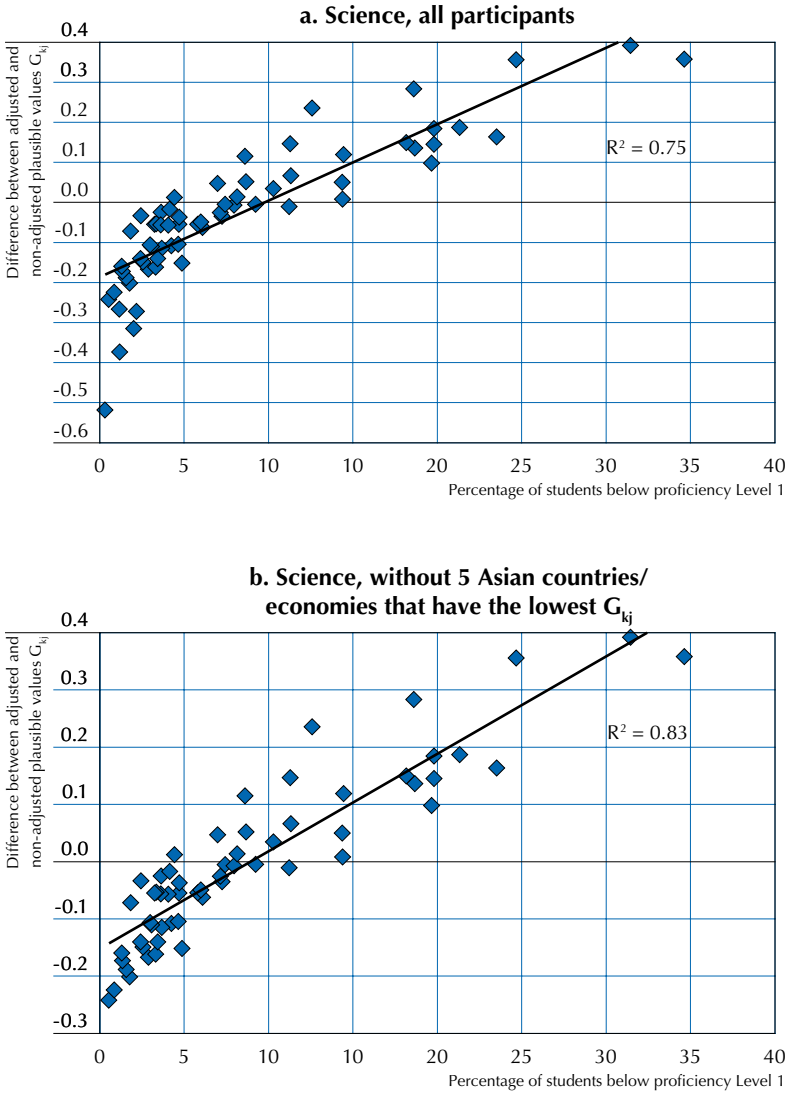






■ Figure 13.3 ■

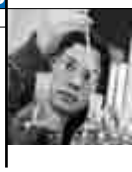
Relationship between G_{kj} and percentage of students below proficiency Level 1 for science



Results for financial literacy (Figure 13.4) seemed to be different but if Shanghai-China – which is the only and very clear outlier – is not taken into account, for the rest of participating countries R^2=91%, which is comparable to the paper-based mathematics result.

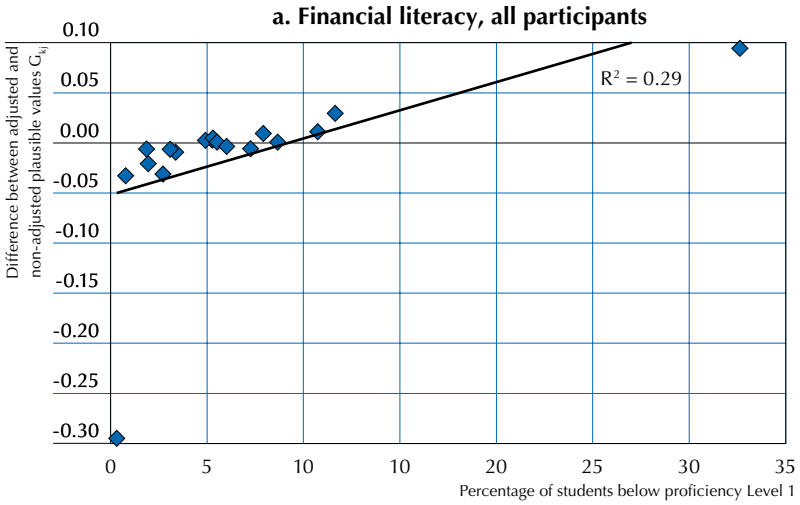
For the computer-based domains of problem solving and digital reading (Figures 13.5 and 13.6) there are no clear outliers and R^2 is higher than non-adjusted R^2 for paper-based domains. For problem solving 87% in G_{kj} variation is explained by the proportion of low achieving students in the country and for digital reading 82% in G_{kj} variation is explained by the proportion of low achieving students in the country.

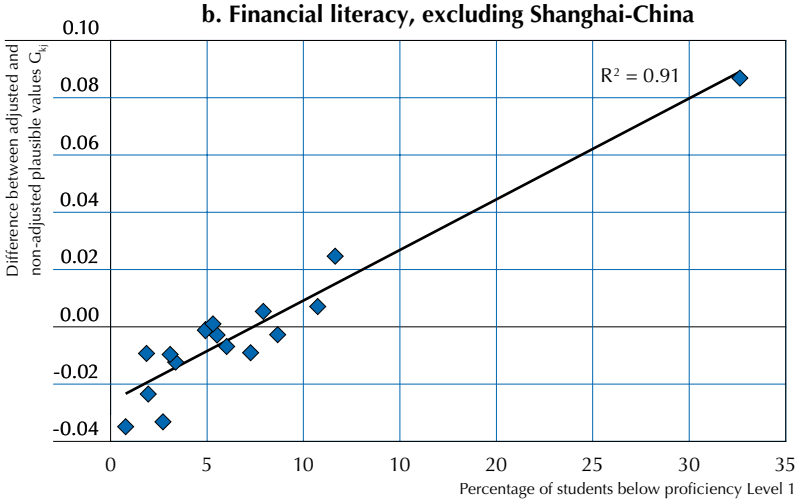
Given that in addition to the differences between the percentage of students in different proficiency levels, there are some curriculum, teaching methods and assessment practices differences between PISA participants that can contribute to the variation in G_{kj} beyond and above the variation that is attributed to the percent of students in the lowest proficiency level, we can't conclude that there is a bias in coding of constructed response items in any particular PISA economy.



■ Figure 13.4 ■

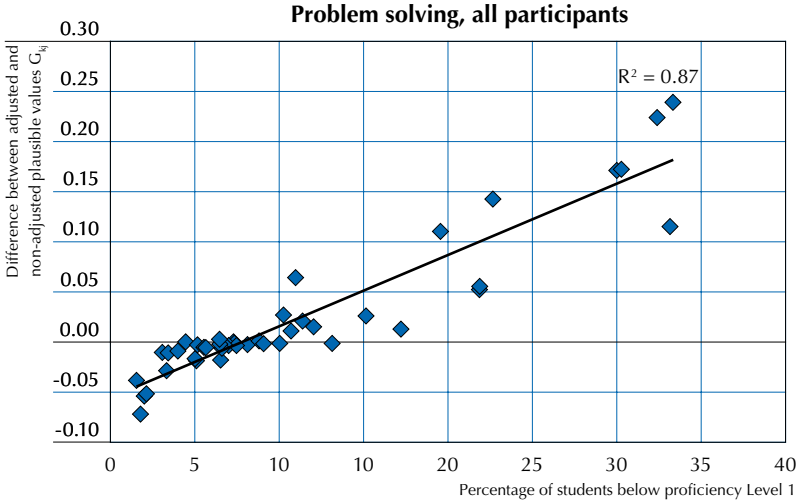
Relationship between G_{kj} and percentage of students below proficiency Level 1 for financial literacy





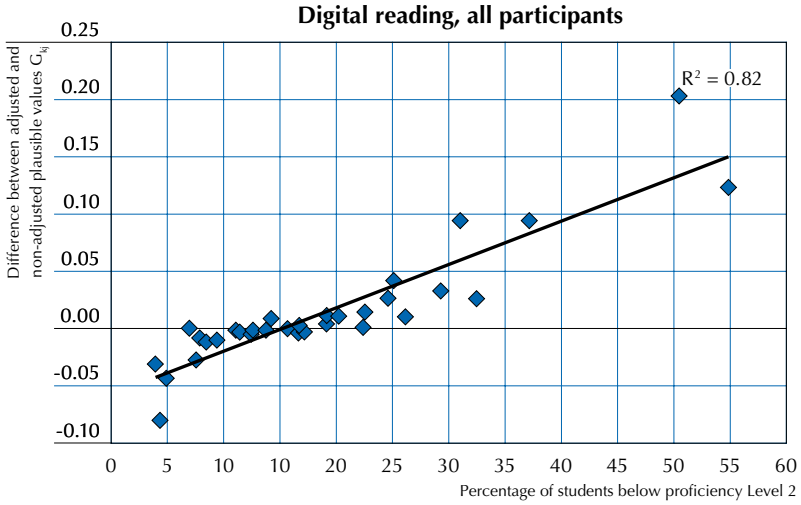
■ Figure 13.5 ■

Relationship between G_{kj} and percent of students below proficiency Level 1 for problem solving



■ Figure 13.6 ■

Relationship between G_{kj} and percentage of students below proficiency Level 2 for digital reading



Note

1. Some items have been removed from analysis from some locales during adjudication process due to printing, translation and other errors (see Table 12.10, in Chapter 12, for the complete list of such items).

References

Adams, R., M. Wu, and **M. Wilson** (2012), *ACER ConQuest 3.1*, ACER, Melbourne.

Leung, F. K. S., K.D. Graf and **F. J. Lopez-Real** (ed.) (2006), "Mathematics Education in Different Cultural Traditions: A Comparative Study of East Asia and the West: The 13th ICMI Study", New ICMI Study Series, Volume 9, Springer, New York.

OECD (2014), *PISA 2012 Results: What Students Know and Can Do (Volume I, Revised edition, February 2014): Student Performance in Mathematics, Reading and Science),* PISA, OECD Publishing, Paris.
http://dx.doi.org/10.1787/9789264208780-en

Routitsky, A. and **R. Turner** (2003), "Item Format Types and their Influence on Cross-national Comparisons of Student Performance", Presentation given to the Annual Meeting of the American Educational Research Association (AERA) in Chicago, USA. Retrieved from *http://works.bepress.com/cgi/viewcontent.cgi?article=1013&context=alla_routitsky*

SPSS, IBM (2010), SPSS for Windows® (version 19), SPSS. Inc., Chicago, Illinois.