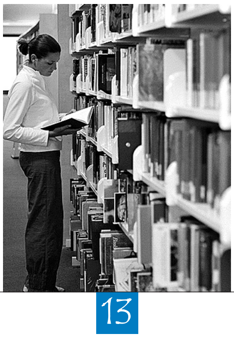


13

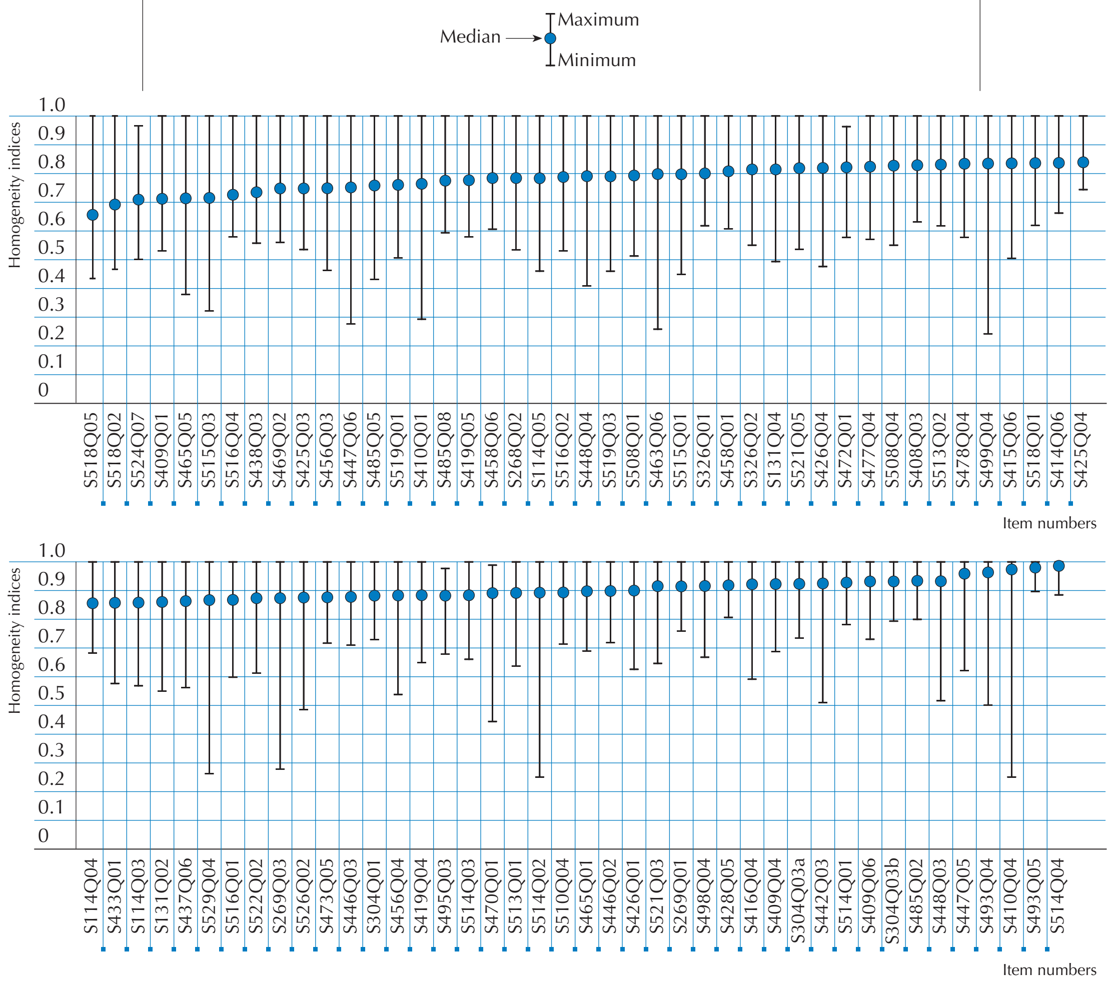
Coding and Marker Reliability Studies

As explained in the first section of this report, on test design (see Chapter 2), a substantial proportion of the PISA 2006 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 (see Chapter 2) using the design described in Chapter 6. Training sessions to train countries in the use of the coding guides were held prior to both the field trial and the main study.

This chapter describes the outcomes of three aspects of the coding and marking reliability studies undertaken in conjunction with the field trial and the main study. These are the homogeneity analyses undertaken with the field trial data to assist the test developers in constructing valid, reliable scoring rubrics; the variance component analyses undertaken with the main study data to examine within-country coder reliability; and an international coder review undertaken to examine the between-country consistency in applying the coding guides.

The methods used to compute the homogeneity indices and the variance components for PISA 2006 where the same as the methods used in PISA 2000 and PISA 2003. The methods for both homogeneity and variance components are fully discussed in Verhelst (2002).

Figure 13.1
Variability of the homogeneity indices for science items in field trial



HOMOGENEITY ANALYSES

Both in the field trial and the main study homogeneity analyses are used to estimate the level of agreement between coders of constructed-response items. In the field trial the primary purpose of the homogeneity analysis is to obtain data to inform the selection of items for the main study. In the field trial, many more items were tried than were used in the main study and one important purpose of the field trial was to select a subset of science items to be used in the main study. One obvious concern was to ensure that coders agreed to a reasonable degree in their categorisation of the answers.

For investigating the inter-coder agreement, the collected data were used to compute a homogeneity index by item and country. This coefficient theoretically can range from zero to one. A coefficient of one shows perfect agreement between coders. Figure 13.1 shows the distribution of the homogeneity indices for all science items in the field trial and for the selected science items for the main study.

If an item had a weak homogeneity index in the field trial, this was a signal to the Science Expert Group and to the test developers either that the item should not to be retained for the main study or that the coding guide required clarification.

Figure 13.2 shows the average of the homogeneity indices per science item for the items included in the main study. In general the chart shows a marked improvement in the level of agreement between coders in the main study compared to the field trial. Changes to coding schemes contributed to this improvement in a number of cases – for example: in *S425Q03*, double-digit coding was replaced by single-digit coding; in *S465Q01*, partial credit was eliminated; and, in *S519Q01*, partial credit was introduced. However, for most items there was no change to the coding scheme between the field trial and the main study. In these cases, much of the improvement can be attributed to improvements to the coding guides – for example, in *S485Q01*, the level descriptors were refined; examples were added for the descriptors in *S447Q05*; and, in *S514Q03*, the descriptors were revised and additional examples were included. The addition of more workshop examples, the expanded coder query database, and the extra experience gained by coders in the field trial also would have contributed significantly to the general tendency for improvement. The small decrease in the homogeneity index for *S493Q05* can be attributed to the change from partial credit to double-digit coding for the main study.

Figure 13.3, Figure 13.4, and Figure 13.5 show the distribution of the national homogeneity indices per item in the main study.

Figure 13.2

Average of the homogeneity indices for science items in field trial and main study

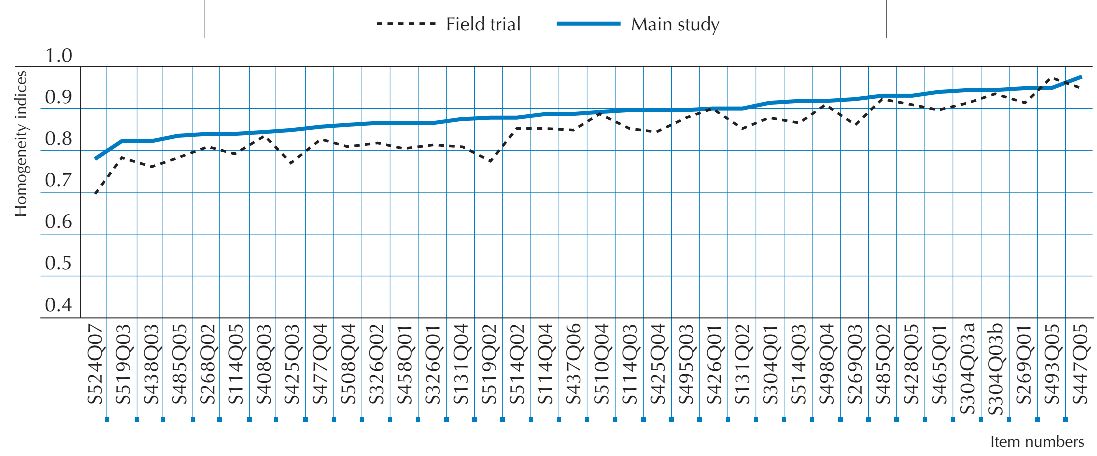


Figure 13.3
Variability of the homogeneity indices for each science item in the main study

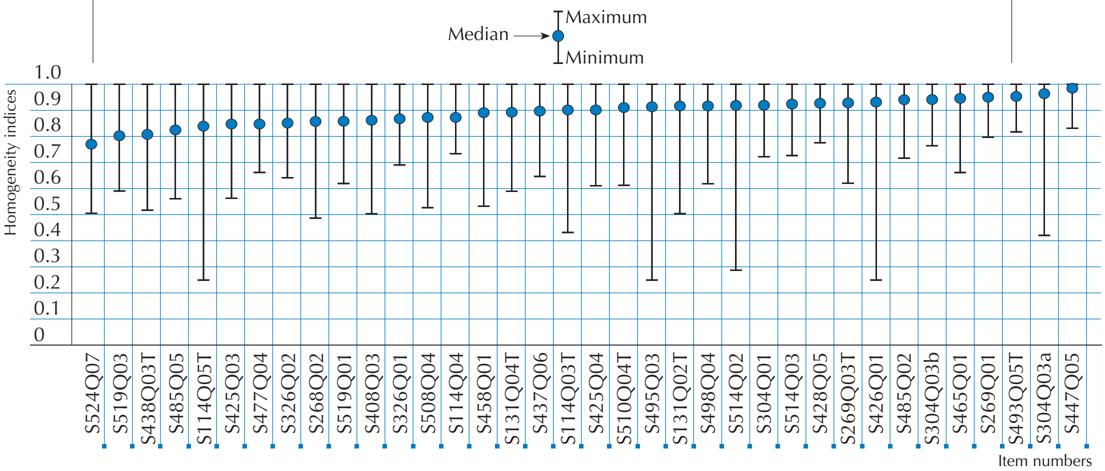


Figure 13.4
Variability of the homogeneity indices for each reading item in the main study

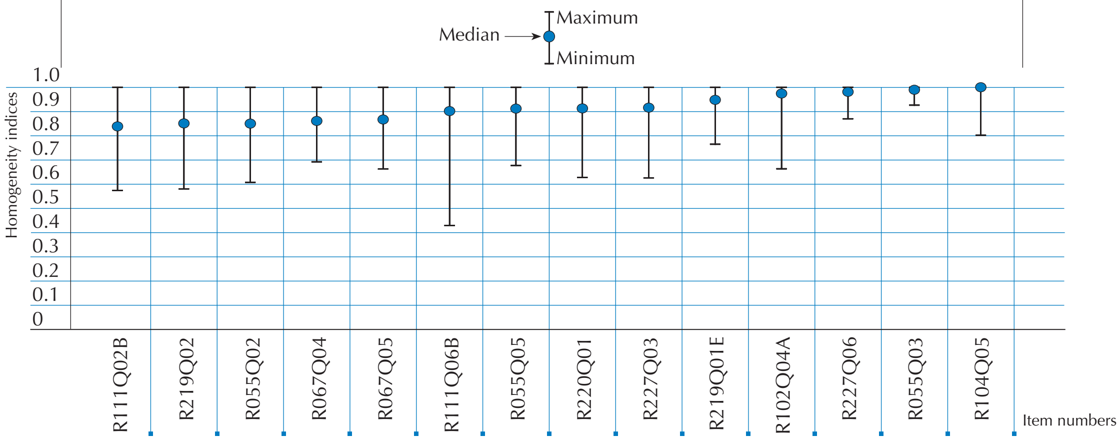


Figure 13.5
Variability of the homogeneity indices for each mathematics item in the main study

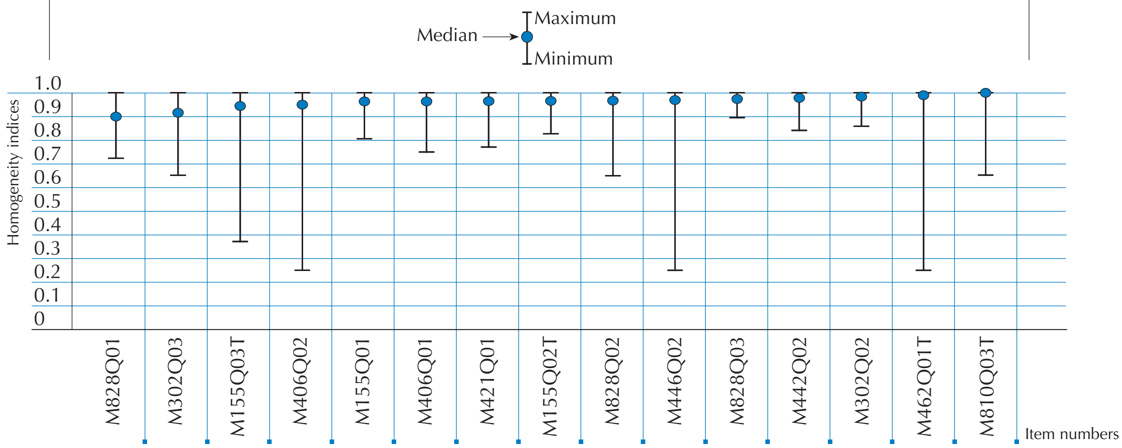
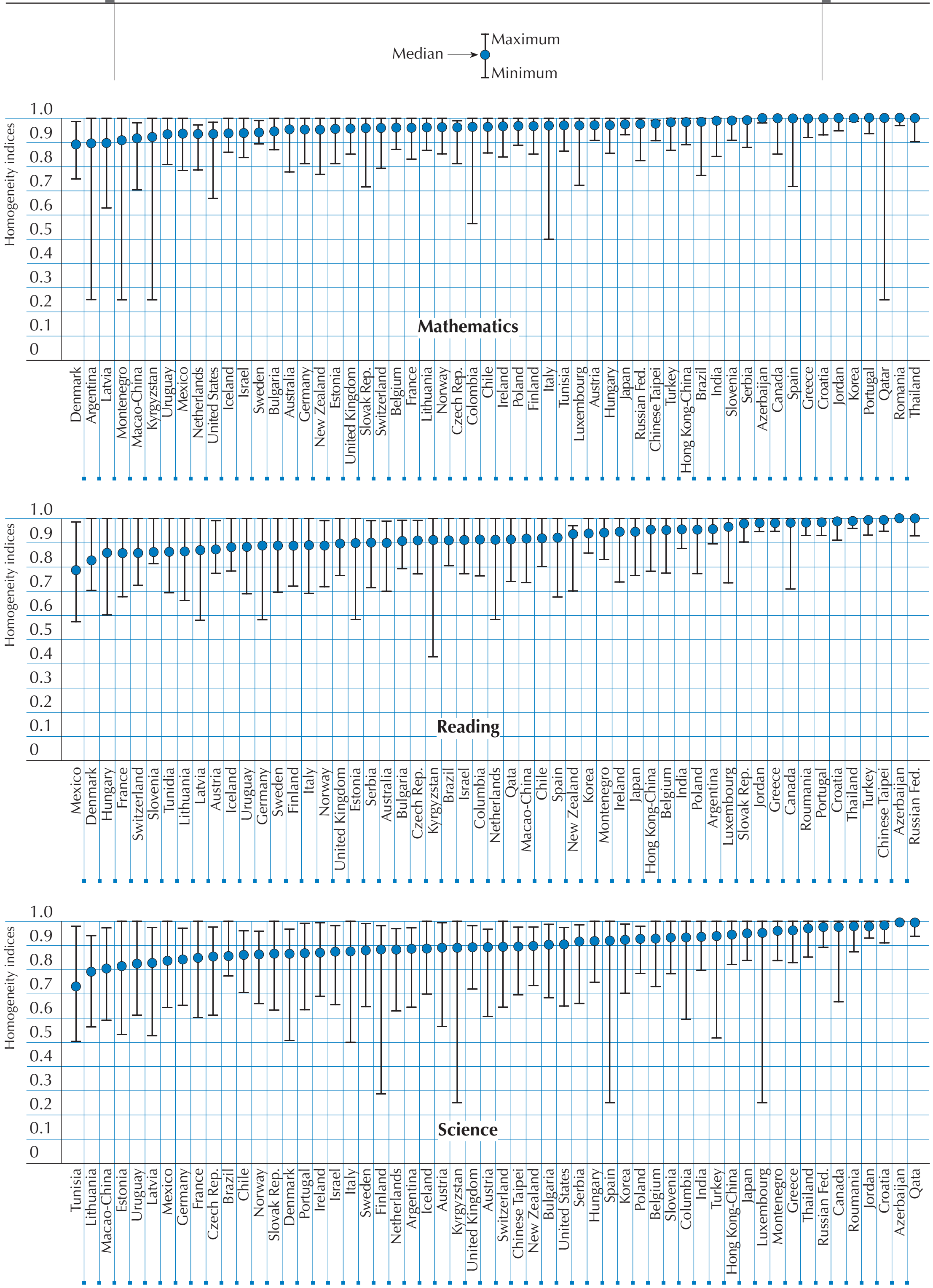


Figure 13.6
Variability of the homogeneity indices for the participating countries in the main study



For all items except one science item, *S524Q07,* the average index is greater than 0.80. Indices are higher for mathematics items which indicate that there is less disagreement between mathematics coders.

Figure 13.6 shows the distribution of homogeneity indices per domain and per country. There is more variability in the coding of reading and science than mathematics for most of the countries.

The results of the homogeneity analysis showed that the marking process of items is largely satisfactory and that on average countries are more or less reliable in the coding of the open-ended responses.

MULTIPLE MARKING STUDY OUTCOMES (VARIANCE COMPONENTS)

To obtain an estimate of the between-coder variability within each country, multiple coding was required for at least some student answers. Therefore, it was decided that multiple codings would be collected for open-ended items in both the field trial and the main study for a moderate number of students. In the main study, a selection of clusters from 600 students' booklets were multiply coded, with the full set of main study items requiring the judgement of a trained coder included in the exercise. The requirement was that the same four expert coders per domain (reading, mathematics and science) should code all items appearing together in the first two clusters of the test booklets 1, 3, 6, 8 and 10, and the first three clusters of booklet 5. A booklet 6 containing, for example, 14 reading items, would give a three-dimensional table for reading (100 students by 14 items by 4 markers), where each cell contains a single category. For each domain and each booklet, such a table was produced and processed in several analyses, which are described later. These data sets were required from each participating country.

Table 13.1 to Table 13.3 show the results of the variance components analysis for the multiply-marked items in mathematics, science, and reading, respectively. The variance components are each expressed as a percentage of their sum.

The tables show that those variance components associated with markers are small relative to the other components. This means that there are no significant systematic within-country marker effects.

Analyses of the type reported here can result in negative variance estimates. If the amount by which the component is negative is small, then this is a sign that the variance component is negligible (near zero). If the component is large and negative, then it is a sign that the analysis method is inappropriate for the data. In Table 13.1 to Table 13.3 countries with large inadmissible variance component estimates are indicated.

Generalisability coefficients

The generalisability coefficients are computed from the variance components using:

13.1

$$\rho_3\left(Y_{vg}, Y'_{vg}\right)=\frac{\sigma_A^2+\dfrac{\sigma^2_{AB+E^*}}{I}}{\sigma_A^2+\dfrac{\sigma^2_{AB+E^*}}{I}+\dfrac{\sigma^2_{ac}}{R}+\dfrac{\sigma^2_{abc+e^*}}{I\times R}}$$

and

13.2

$$\rho_3\left(Y_{vg}, Y'_{vg}\right)=\frac{\sigma_A^2+\dfrac{\sigma^2_{AB}}{I}}{\sigma_A^2+\dfrac{\sigma^2_{AB}+\sigma^2_{\varepsilon^*}}{I}+\dfrac{\sigma^2_{ac}}{R}+\dfrac{\sigma^2_{abc+e^*}}{I\times R}}$$

Table 13.1
Variance components for mathematics

	Student Component	Item Component	Marker Component	Student-item Interaction Component	Student-Marker Interaction Component	Item-Marker Interaction Component	Measurement Error component
Argentina	17.10	30.40	0.01	46.70	0.00	0.10	5.70
Australia	17.47	31.05	0.07	45.01	-0.02	0.10	6.32
Austria	25.21	19.34	0.00	51.76	0.02	0.07	3.60
Azerbaijan	9.53	27.04	0.00	63.31	0.00	0.00	0.11
Belgium (Dutch)	17.77	23.19	0.01	54.35	-0.09	0.03	4.73
Belgium (French)	23.82	17.55	0.03	54.09	0.17	0.02	4.32
Brazil	24.30	8.59	0.03	62.69	0.05	0.03	4.31
Bulgaria	16.86	17.02	0.00	59.11	-0.24	0.04	7.20
Canada (English)[1]	18.85	28.96	0.42	43.31	-20.00	-0.21	28.66
Canada (French)	11.73	30.86	0.01	52.52	0.03	0.12	4.72
Chile	17.58	21.00	0.02	55.57	-0.03	0.00	5.85
Colombia	14.69	21.93	0.00	59.18	-0.08	0.02	4.26
Croatia	13.84	23.03	0.00	62.20	0.01	0.01	0.91
Czech Republic	21.25	17.82	0.00	56.67	0.06	0.09	4.11
Denmark[1]	19.64	20.70	0.21	52.05	-4.87	0.15	12.13
Estonia (Estonian)[1]	10.71	30.09	0.01	52.19	-2.77	0.26	9.50
Estonia (Russian)	13.67	30.64	0.10	50.39	0.03	0.40	4.76
Finland	14.32	27.33	0.01	53.64	-0.06	0.08	4.69
France	23.78	17.25	0.02	53.40	0.05	0.09	5.42
Germany	18.72	21.24	0.00	53.14	-0.01	0.21	6.70
Greece	20.28	22.47	0.00	56.19	-0.01	0.00	1.06
Hong Kong-China	15.07	21.98	0.00	58.70	-0.06	0.10	4.21
Hungary	15.38	30.20	-0.01	51.08	0.04	0.03	3.28
Iceland	14.50	23.77	0.02	55.38	0.15	0.09	6.09
Indonesia	19.12	15.73	0.01	60.62	0.02	0.03	4.47
Ireland	16.38	29.41	0.01	48.39	-0.03	0.10	5.74
Israel	18.16	22.60	0.01	52.54	-0.04	0.10	6.63
Italy (German)	15.20	37.60	0.02	42.44	-0.06	0.09	4.71
Italy (Italian)	21.61	16.48	0.21	57.72	0.03	0.01	3.94
Japan	17.20	23.20	0.00	57.17	0.04	0.03	2.36
Jordan	13.09	18.00	0.00	67.75	0.00	0.01	1.15
Korea	20.66	18.43	0.00	60.36	-0.01	0.00	0.56
Kyrgyzstan (Kyrgyz)	6.12	6.31	-0.06	69.98	-0.42	0.64	17.44
Kyrgyzstan (Russian)	19.28	11.85	-0.02	63.56	-0.23	0.18	5.37
Latvia (Latvian)	16.37	19.34	0.30	49.87	0.08	1.21	12.83
Latvia (Russian)	13.47	26.62	0.33	46.52	0.42	0.60	12.04
Lithuania	18.69	21.88	0.01	54.41	-0.05	0.06	5.00
Luxembourg (French)	16.75	32.86	-0.02	44.01	-0.30	-0.04	6.74
Luxembourg (German)	23.12	19.92	0.00	54.45	-0.16	0.02	2.66
Macao-China	18.32	16.86	0.03	54.60	0.01	0.36	9.82
Mexico	14.35	19.35	0.04	56.47	0.07	0.13	9.58
Montenegro	17.89	11.30	0.06	58.26	-0.21	0.35	12.35
Netherlands	13.78	31.80	0.01	47.04	0.03	0.09	7.25
New Zealand	16.12	27.56	0.00	50.42	0.07	0.05	5.78
Norway	18.56	25.77	0.00	50.99	-0.06	0.02	4.72
Poland	24.57	13.30	0.00	57.94	0.05	0.04	4.10
Portugal	15.82	20.96	0.00	62.30	0.01	0.00	0.92
Qatar (Arabic)	14.44	9.16	0.00	74.83	-0.04	0.00	1.61
Qatar (English)	43.64	9.28	0.00	46.87	0.01	0.00	0.20
Romania	18.66	14.99	0.00	66.11	0.00	0.00	0.24
Russian Federation	20.30	25.91	0.02	50.33	0.00	0.08	3.37
Serbia	21.57	16.67	0.00	59.81	-0.03	0.00	1.99
Slovakia	22.10	21.58	0.00	50.22	0.00	0.07	6.03
Slovenia	15.72	18.08	0.00	64.36	0.43	0.01	1.41
Spain (Basque)	33.64	10.60	-0.01	53.17	0.00	-0.02	2.62
Spain (Catalan)	14.64	26.15	0.02	50.16	0.09	0.47	8.47
Spain (Galician)	14.83	30.01	0.06	48.90	-0.01	0.40	5.82
Spain (Spanish)	16.65	24.35	-0.05	54.24	0.05	0.30	4.44
Spain (Valencian)	5.70	36.88	0.14	46.23	-0.04	0.16	10.93
Sweden	16.05	27.62	-0.01	51.45	-0.03	0.04	4.87
Switzerland (French)	11.89	33.15	0.00	48.19	-0.02	0.08	6.71
Switzerland (German)	18.60	24.20	0.00	53.92	0.00	0.02	3.26
Chinese Taipei	20.13	15.33	0.00	61.05	-0.05	0.01	3.52
Thailand	20.52	18.17	0.00	60.05	0.05	0.01	1.21
Tunisia	16.04	10.82	0.01	68.03	-0.11	0.03	5.18
Turkey	27.17	9.63	0.00	60.26	0.00	0.02	2.93
United Kingdom (Scotland)	16.77	27.09	-0.01	51.35	-0.08	0.10	4.77
United Kingdom (The rest of)	17.02	32.82	0.01	44.69	-0.05	0.03	5.49
United States[1]	20.34	28.66	0.12	44.50	-5.78	0.03	12.13
Uruguay	16.42	20.70	0.01	56.24	-0.12	0.13	6.62

1. Countries with large inadmissible variance component estimates.

Table 13.2
Variance components for science

	Student Component	Item Component	Marker Component	Student-item Interaction Component	Student-Marker Interaction Component	Item-Marker Interaction Component	Measurement Error component
Argentina[1]	15.72	14.84	0.05	55.60	-3.30	0.20	16.89
Australia	17.26	23.19	0.00	47.53	0.02	0.43	11.56
Austria	17.37	20.17	0.00	50.23	-0.01	0.31	11.93
Azerbaijan	15.70	6.51	0.00	77.75	0.00	0.00	0.04
Belgium (Dutch)	13.78	28.44	0.02	49.48	0.00	0.17	8.12
Belgium (French)	17.39	22.53	0.02	54.44	0.04	0.04	5.54
Brazil	18.84	10.23	0.01	55.49	-0.08	0.65	14.86
Bulgaria	28.82	8.73	0.17	52.83	0.17	0.39	8.88
Canada (English)[1]	16.41	21.80	0.38	44.25	-10.49	0.46	27.19
Canada (French)	16.37	19.79	0.20	49.49	0.06	0.55	13.54
Chile	18.95	15.26	0.06	51.14	0.29	0.26	14.05
Colombia	15.28	13.22	0.01	61.50	0.01	0.07	9.91
Croatia	12.27	24.62	0.00	61.26	0.01	0.01	1.83
Czech Republic	16.80	21.08	0.02	48.07	-0.02	0.57	13.48
Denmark[1]	18.41	17.41	0.03	50.08	-1.98	0.27	15.78
Estonia (Estonian)[1]	16.41	26.43	0.10	42.93	-2.67	0.85	15.95
Estonia (Russian)	16.74	18.45	0.34	43.04	-0.14	1.37	20.20
Finland[1]	14.57	27.12	0.25	48.10	-1.58	0.36	11.18
France	16.37	24.24	0.05	46.27	0.05	0.43	12.58
Germany	16.08	18.59	0.09	50.13	0.15	0.80	14.15
Greece	18.55	19.32	0.00	59.00	0.02	0.02	3.07
Hong Kong-China	15.45	27.83	0.02	50.16	0.01	0.02	6.51
Hungary	16.06	15.43	0.01	59.70	0.13	0.12	8.56
Iceland	15.64	20.44	0.04	51.98	0.09	0.18	11.63
Indonesia	12.60	10.96	0.00	65.23	-0.93	0.56	11.57
Ireland	14.71	23.97	0.04	48.64	0.13	0.41	12.09
Israel	25.01	17.19	0.07	47.75	0.10	0.13	9.76
Italy (German)	16.11	21.08	-0.03	49.34	0.13	0.26	13.12
Italy (Italian)	16.19	15.99	0.63	56.47	0.00	0.14	10.57
Japan	19.37	22.93	0.01	54.02	0.03	0.03	3.61
Jordan	21.68	12.46	0.00	63.10	0.01	0.00	2.75
Korea	16.94	21.27	0.05	53.19	0.06	0.18	8.31
Kyrgyzstan (Kyrgyz)	10.79	7.64	0.01	65.64	0.28	0.35	15.30
Kyrgyzstan (Russian)	15.59	8.93	0.02	66.72	0.02	0.07	8.65
Latvia (Latvian)	13.92	19.55	0.10	48.34	0.12	1.10	16.87
Latvia (Russian)	16.15	22.47	-0.04	42.92	0.18	1.12	17.18
Lithuania	17.26	18.37	0.06	43.13	0.44	1.62	19.14
Luxembourg (French)	21.75	13.02	0.05	58.75	0.20	0.01	6.22
Luxembourg (German)	15.44	20.49	-0.02	56.92	0.10	0.27	6.80
Macao-China	12.76	23.01	0.44	44.02	0.07	1.39	18.31
Mexico	12.50	12.60	0.07	49.63	0.22	0.45	24.53
Montenegro	16.89	12.10	0.00	66.07	0.10	0.03	4.80
Netherlands	16.28	24.28	0.58	45.58	-0.31	0.73	12.87
New Zealand	18.50	19.56	0.08	50.95	0.06	0.12	10.73
Norway	17.80	14.33	0.09	52.65	0.04	0.50	14.59
Poland	14.72	23.42	0.01	54.92	0.02	0.03	6.87
Portugal	14.96	22.03	0.03	50.40	0.16	0.20	12.21
Qatar (Arabic)	17.95	14.35	0.00	66.09	0.03	0.00	1.59
Qatar (English)	21.19	15.59	0.00	61.83	-0.02	-0.01	1.41
Romania	18.44	10.98	0.00	68.08	-0.02	0.01	2.52
Russian Federation	15.99	16.22	0.00	65.18	0.00	0.00	2.60
Serbia	16.86	14.38	0.06	58.77	0.22	0.36	9.35
Slovakia	18.51	16.84	0.20	51.58	0.20	0.36	12.31
Slovenia	22.32	18.30	0.01	52.73	0.06	0.11	6.47
Spain (Basque)	13.59	21.27	0.04	57.83	-0.11	0.12	7.26
Spain (Catalan)	15.13	20.45	0.48	43.02	0.11	1.31	19.51
Spain (Galician)	11.88	23.02	0.13	50.36	0.14	0.47	13.99
Spain (Spanish)	14.73	21.99	0.43	52.56	0.02	0.27	10.00
Spain (Valencian)	17.16	6.92	0.55	49.05	-0.45	0.65	26.13
Sweden	17.52	19.97	0.00	51.49	0.07	0.20	10.76
Switzerland (French)	16.92	22.08	0.01	50.82	0.06	0.42	9.69
Switzerland (German)	20.69	19.54	0.05	50.05	0.09	0.23	9.36
Chinese Taipei	13.27	26.43	0.00	50.87	0.10	0.19	9.14
Thailand	15.72	17.45	0.01	62.73	-0.01	0.04	4.06
Tunisia	13.63	13.66	0.20	46.36	0.21	1.04	24.90
Turkey	17.33	11.62	0.25	59.48	0.17	0.26	10.89
United Kingdom (Scotland)	16.41	25.52	0.06	47.49	-0.04	0.20	10.35
United Kingdom (The rest of)	16.74	22.77	0.04	50.22	0.25	0.15	9.82
United States	20.67	17.06	0.01	51.45	0.06	0.15	10.60
Uruguay	15.82	15.23	0.04	53.34	0.09	0.75	14.73

1. Countries with large inadmissible variance component estimates.

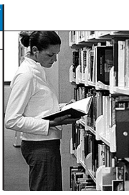


Table 13.3
Variance components for reading

	Student Component	Item Component	Marker Component	Student-item Interaction Component	Student-Marker Interaction Component	Item-Marker Interaction Component	Measurement Error component
Argentina	21.35	20.82	0.00	54.35	0.01	0.03	3.44
Australia	23.78	23.57	0.01	41.80	0.05	0.19	10.60
Austria	20.50	13.19	0.20	52.75	0.02	0.52	12.81
Azerbaijan	25.28	8.64	0.00	66.08	0.00	0.00	0.00
Belgium (Dutch)	11.44	26.77	0.05	49.91	-0.09	0.25	11.66
Belgium (French)	21.50	14.83	0.00	59.21	0.18	0.00	4.28
Brazil	13.94	19.18	0.08	56.03	0.11	0.27	10.39
Bulgaria	31.00	13.90	0.00	48.38	0.03	0.02	6.67
Canada (English)[1]	16.86	26.80	0.01	45.22	-10.00	-0.20	21.30
Canada (French)	18.56	21.19	0.03	46.47	0.11	0.76	12.89
Chile	15.01	31.49	0.01	44.11	0.13	0.15	9.10
Colombia	14.58	21.06	-0.01	52.57	0.20	0.19	11.42
Croatia	15.40	20.46	0.02	61.03	0.02	0.03	3.04
Czech Republic	27.10	14.40	0.00	48.17	0.13	0.39	9.81
Denmark[1]	19.07	12.83	-0.02	46.26	-2.34	1.61	22.58
Estonia (Estonian)[1]	10.76	27.07	-0.01	51.22	-2.28	0.18	13.06
Estonia (Russian)	17.53	22.53	-0.10	40.40	-0.26	2.11	17.79
Finland	14.55	19.31	0.10	53.07	0.04	0.17	12.76
France	19.76	24.01	0.26	39.17	-0.10	1.37	15.54
Germany	21.68	14.11	0.00	51.31	-0.01	0.09	12.83
Greece	22.47	23.43	0.01	52.00	-0.02	0.00	2.10
Hong Kong-China	14.07	28.02	0.03	49.10	0.00	0.35	8.43
Hungary	22.87	16.36	0.16	43.00	0.57	0.52	16.52
Iceland	19.31	10.33	0.01	54.22	0.04	0.62	15.48
Indonesia	11.82	18.34	0.01	64.22	0.02	0.09	5.51
Ireland	22.66	21.22	0.06	45.78	0.07	0.14	10.07
Israel	16.79	22.92	0.08	49.54	0.07	0.24	10.36
Italy (German)	20.24	19.88	0.12	44.21	-0.15	0.12	15.58
Italy (Italian)	20.56	22.60	-0.11	46.78	-0.06	0.22	10.01
Japan	20.64	11.12	0.01	62.33	0.10	0.10	5.70
Jordan	15.02	16.27	0.00	66.46	0.01	0.00	2.25
Korea	16.14	27.33	0.02	51.90	0.04	0.04	4.52
Kyrgyzstan (Kyrgyz)	5.79	6.91	-0.06	56.07	-0.35	0.48	31.15
Kyrgyzstan (Russian)	28.85	11.87	-0.02	51.91	0.06	0.18	7.16
Latvia (Latvian)	16.00	19.52	0.22	44.78	0.20	1.08	18.21
Latvia (Russian)	16.01	24.25	0.29	43.32	0.03	1.15	14.95
Lithuania	20.54	17.10	0.07	43.69	0.06	1.62	16.93
Luxembourg (French)	20.87	15.50	-0.01	57.46	0.17	0.00	6.01
Luxembourg (German)	25.32	14.35	0.00	53.28	0.27	0.02	6.76
Macao-China	10.09	29.36	0.13	45.75	0.08	0.77	13.82
Mexico	13.26	23.70	0.64	36.90	0.32	2.19	22.99
Montenegro	13.68	11.56	-0.01	67.32	0.98	0.01	6.45
Netherlands	16.50	17.90	0.01	53.33	-0.01	0.17	12.11
New Zealand	25.16	22.05	0.10	43.06	0.05	0.12	9.46
Norway	27.00	11.67	0.02	50.09	0.07	0.33	10.82
Poland	18.49	26.01	0.01	47.84	-0.02	0.07	7.60
Portugal	10.31	34.21	0.00	52.04	0.18	-0.01	3.27
Qatar (Arabic)	12.54	13.76	-0.01	64.69	0.07	0.08	8.86
Qatar (English)	21.17	19.44	-0.01	49.55	0.14	0.06	9.66
Romania	17.43	16.05	0.00	64.56	-0.03	0.01	1.97
Russian Federation	20.09	22.07	0.00	56.71	0.00	0.00	1.13
Serbia	18.94	14.08	0.04	53.45	0.11	0.24	13.14
Slovakia	15.95	25.65	0.00	54.64	0.00	0.08	3.69
Slovenia	19.16	22.90	0.00	45.59	0.01	0.25	12.09
Spain (Basque)	24.16	14.96	-0.01	44.31	0.00	0.25	16.33
Spain (Catalan)	16.20	24.84	0.82	37.18	0.04	1.79	19.12
Spain (Galician)	15.20	24.82	0.06	40.97	-0.02	0.56	18.41
Spain (Spanish)	19.28	23.30	0.26	42.92	0.21	0.33	13.69
Spain (Valencian)	29.85	18.79	1.20	28.88	0.29	1.44	19.55
Sweden	23.24	13.35	0.01	49.16	0.09	0.29	13.86
Switzerland (French)	14.60	23.53	-0.04	50.96	0.12	0.60	10.23
Switzerland (German)	18.70	15.67	0.05	52.11	-0.02	0.03	13.47
Chinese Taipei	13.21	37.15	0.00	48.09	-0.02	0.00	1.57
Thailand	14.89	20.25	0.00	63.23	0.00	0.01	1.62
Tunisia	16.24	16.85	-0.04	51.22	0.12	0.44	15.17
Turkey	14.57	19.68	0.00	63.89	0.01	0.00	1.84
United Kingdom (Scotland)	22.87	23.01	0.01	44.53	-0.01	0.10	9.49
United Kingdom (The rest of)	21.10	25.92	-0.01	44.14	0.02	0.05	8.77
United States[1]	26.42	22.04	-0.05	42.17	-2.10	-0.01	11.53
Uruguay	17.15	22.85	0.03	49.88	0.12	0.24	9.72

1. Countries with large inadmissible variance component estimates.

Table 13.4
Generalisability estimates for mathematics

	I=8 M=1		I=16 M=1		I=24 M=1	
	p3	p4	p3	p4	p3	p4
Argentina	0.97	0.72	0.98	0.84	0.99	0.89
Australia	0.97	0.73	0.98	0.85	0.99	0.89
Austria	0.99	0.78	0.99	0.88	0.99	0.92
Azerbaijan	1.00	0.55	1.00	0.71	1.00	0.78
Belgium (Dutch)	0.98	0.71	0.99	0.83	1.00	0.88
Belgium (French)	0.98	0.76	0.98	0.86	0.99	0.90
Brazil	0.98	0.74	0.99	0.85	0.99	0.90
Bulgaria	0.97	0.68	0.99	0.81	1.00	0.87
Canada (English)						
Canada (French)	0.97	0.62	0.98	0.77	0.98	0.83
Chile	0.97	0.70	0.98	0.82	0.99	0.87
Colombia	0.98	0.65	0.99	0.79	0.99	0.85
Croatia	0.99	0.64	1.00	0.78	1.00	0.84
Czech Republic	0.98	0.74	0.99	0.85	0.99	0.89
Denmark						
Estonia (Estonian)						
Estonia (Russian)	0.97	0.66	0.98	0.80	0.99	0.85
Finland	0.98	0.66	0.99	0.80	0.99	0.86
France	0.98	0.76	0.99	0.87	0.99	0.91
Germany	0.97	0.72	0.98	0.83	0.99	0.88
Greece	1.00	0.74	1.00	0.85	1.00	0.90
Hong Kong-China	0.98	0.66	0.99	0.80	0.99	0.86
Hungary	0.98	0.69	0.99	0.82	0.99	0.87
Iceland	0.96	0.65	0.97	0.78	0.98	0.84
Indonesia	0.98	0.70	0.99	0.82	0.99	0.88
Ireland	0.97	0.71	0.98	0.83	0.99	0.88
Israel	0.97	0.71	0.98	0.83	0.99	0.88
Italy (German)	0.98	0.72	0.99	0.84	0.99	0.89
Italy (Italian)	0.98	0.74	0.99	0.85	0.99	0.89
Japan	0.99	0.70	0.99	0.82	0.99	0.87
Jordan	0.99	0.60	1.00	0.75	1.00	0.82
Korea	1.00	0.73	1.00	0.85	1.00	0.89
Kyrgyzstan (Kyrgyz)	0.89	0.37	0.94	0.55	0.97	0.66
Kyrgyzstan (Russian)	0.98	0.70	1.00	0.83	1.00	0.88
Latvia (Latvian)	0.93	0.67	0.96	0.80	0.97	0.86
Latvia (Russian)	0.91	0.64	0.93	0.77	0.94	0.83
Lithuania	0.98	0.72	0.99	0.84	0.99	0.89
Luxembourg (French)	0.98	0.74	0.99	0.85	1.00	0.90
Luxembourg (German)	0.99	0.77	1.00	0.87	1.00	0.91
Macao-China	0.95	0.69	0.97	0.82	0.98	0.87
Mexico	0.94	0.63	0.96	0.77	0.97	0.84
Montenegro	0.95	0.68	0.98	0.81	0.99	0.87
Netherlands	0.96	0.67	0.97	0.80	0.98	0.86
New Zealand	0.97	0.69	0.98	0.82	0.98	0.87
Norway	0.98	0.73	0.99	0.84	0.99	0.89
Poland	0.98	0.76	0.99	0.86	0.99	0.90
Portugal	1.00	0.67	1.00	0.80	1.00	0.86
Qatar (Arabic)	0.99	0.60	1.00	0.75	1.00	0.82
Qatar (English)	1.00	0.88	1.00	0.94	1.00	0.96
Romania	1.00	0.69	1.00	0.82	1.00	0.87
Russian Federation	0.98	0.75	0.99	0.86	0.99	0.90
Serbia	0.99	0.74	1.00	0.85	1.00	0.89
Slovakia	0.97	0.76	0.99	0.86	0.99	0.90
Slovenia	0.98	0.65	0.97	0.78	0.97	0.83
Spain (Basque)	0.99	0.83	1.00	0.91	1.00	0.94
Spain (Catalan)	0.95	0.66	0.97	0.80	0.97	0.85
Spain (Galician)	0.97	0.69	0.98	0.81	0.99	0.87
Spain (Spanish)	0.98	0.69	0.98	0.82	0.99	0.87
Spain (Valencian)	0.90	0.45	0.93	0.62	0.95	0.71
Sweden	0.98	0.70	0.99	0.82	0.99	0.87
Switzerland (French)	0.96	0.64	0.97	0.78	0.98	0.84
Switzerland (German)	0.98	0.72	0.99	0.84	0.99	0.89
Chinese Taipei	0.99	0.72	0.99	0.84	1.00	0.88
Thailand	0.99	0.73	1.00	0.84	1.00	0.89
Tunisia	0.98	0.64	0.99	0.78	0.99	0.85
Turkey	0.99	0.78	0.99	0.87	1.00	0.91
United Kingdom (Scotland)	0.98	0.71	0.99	0.83	0.99	0.88
United Kingdom (The rest of)	0.97	0.73	0.99	0.85	0.99	0.89
United States						
Uruguay	0.97	0.68	0.99	0.81	0.99	0.87

Note: Countries with no value are displayed, because they fall outside the acceptable [0,1] range.

Table 13.5
Generalisability estimates for science

	I=8 M=1		I=16 M=1		I=24 M=1	
	p3	p4	p3	p4	p3	p4
Argentina						
Australia	0.94	0.70	0.97	0.82	0.98	0.87
Austria	0.94	0.69	0.97	0.82	0.98	0.87
Azerbaijan	1.00	0.62	1.00	0.76	1.00	0.83
Belgium (Dutch)	0.95	0.66	0.97	0.79	0.98	0.85
Belgium (French)	0.97	0.70	0.98	0.82	0.99	0.87
Brazil	0.94	0.68	0.96	0.81	0.98	0.87
Bulgaria	0.97	0.79	0.98	0.88	0.98	0.91
Canada (English)						
Canada (French)	0.93	0.67	0.96	0.80	0.97	0.86
Chile	0.93	0.69	0.95	0.81	0.96	0.86
Colombia	0.95	0.63	0.97	0.77	0.98	0.84
Croatia	0.99	0.61	0.99	0.76	0.99	0.82
Czech Republic	0.93	0.69	0.96	0.81	0.97	0.87
Denmark						
Estonia (Estonian)						
Estonia (Russian)	0.90	0.68	0.95	0.81	0.96	0.87
Finland						
France	0.93	0.69	0.96	0.82	0.97	0.87
Germany	0.92	0.66	0.95	0.79	0.96	0.85
Greece	0.99	0.71	0.99	0.83	0.99	0.88
Hong Kong-China	0.96	0.69	0.98	0.81	0.98	0.87
Hungary	0.95	0.65	0.97	0.79	0.97	0.84
Iceland	0.94	0.66	0.96	0.79	0.97	0.85
Indonesia	0.98	0.59	1.00	0.77	1.00	0.85
Ireland	0.93	0.66	0.95	0.79	0.96	0.85
Israel	0.96	0.77	0.98	0.87	0.98	0.91
Italy (German)	0.93	0.67	0.95	0.80	0.96	0.86
Italy (Italian)	0.95	0.66	0.97	0.79	0.98	0.85
Japan	0.98	0.73	0.99	0.84	0.99	0.89
Jordan	0.99	0.73	0.99	0.84	1.00	0.89
Korea	0.96	0.69	0.97	0.81	0.98	0.87
Kyrgyzstan (Kyrgyz)	0.90	0.51	0.92	0.67	0.94	0.75
Kyrgyzstan (Russian)	0.96	0.62	0.97	0.77	0.98	0.83
Latvia (Latvian)	0.90	0.63	0.94	0.77	0.95	0.83
Latvia (Russian)	0.90	0.68	0.94	0.80	0.95	0.86
Lithuania	0.89	0.68	0.92	0.80	0.94	0.85
Luxembourg (French)	0.97	0.72	0.98	0.84	0.98	0.88
Luxembourg (German)	0.96	0.66	0.97	0.79	0.98	0.85
Macao-China	0.89	0.62	0.93	0.76	0.95	0.83
Mexico	0.85	0.57	0.90	0.72	0.92	0.79
Montenegro	0.97	0.65	0.98	0.79	0.99	0.85
Netherlands	0.94	0.70	0.98	0.83	0.99	0.89
New Zealand	0.95	0.70	0.97	0.83	0.98	0.88
Norway	0.93	0.68	0.96	0.81	0.97	0.86
Poland	0.96	0.66	0.98	0.79	0.98	0.85
Portugal	0.93	0.65	0.95	0.79	0.96	0.84
Qatar (Arabic)	0.99	0.68	0.99	0.81	1.00	0.86
Qatar (English)	1.00	0.73	1.00	0.84	1.00	0.89
Romania	0.99	0.68	0.99	0.81	1.00	0.86
Russian Federation	0.99	0.65	0.99	0.79	0.99	0.85
Serbia	0.95	0.66	0.96	0.79	0.97	0.85
Slovakia	0.94	0.69	0.96	0.82	0.97	0.87
Slovenia	0.97	0.75	0.98	0.86	0.99	0.90
Spain (Basque)	0.96	0.63	0.98	0.77	0.99	0.84
Spain (Catalan)	0.89	0.66	0.93	0.79	0.95	0.85
Spain (Galician)	0.91	0.59	0.94	0.74	0.95	0.81
Spain (Spanish)	0.94	0.65	0.97	0.79	0.98	0.85
Spain (Valencian)	0.89	0.66	0.95	0.80	0.97	0.87
Sweden	0.94	0.69	0.97	0.82	0.97	0.87
Switzerland (French)	0.95	0.69	0.97	0.82	0.98	0.87
Switzerland (German)	0.96	0.73	0.97	0.85	0.98	0.89
Chinese Taipei	0.94	0.64	0.96	0.78	0.97	0.84
Thailand	0.98	0.65	0.99	0.79	0.99	0.85
Tunisia	0.85	0.60	0.90	0.75	0.93	0.81
Turkey	0.94	0.66	0.96	0.79	0.97	0.85
United Kingdom (Scotland)	0.95	0.70	0.97	0.82	0.98	0.87
United Kingdom (The rest of)	0.94	0.68	0.96	0.81	0.97	0.86
United States	0.95	0.73	0.97	0.84	0.98	0.89
Uruguay	0.92	0.65	0.95	0.79	0.96	0.84

Note: Countries with no value are displayed, because they fall outside the acceptable [0,1] range.

Table 13.6
Generalisability estimates for reading

	I=8 M=1		I=16 M=1		I=24 M=1	
	p3	p4	p3	p4	p3	p4
Argentina	0.99	0.75	0.99	0.86	0.99	0.90
Australia	0.96	0.78	0.97	0.88	0.98	0.91
Austria	0.94	0.71	0.97	0.83	0.98	0.88
Azerbaijan	1.00	0.75	1.00	0.86	1.00	0.90
Belgium (Dutch)	0.93	0.60	0.96	0.75	0.97	0.82
Belgium (French)	0.98	0.73	0.98	0.84	0.99	0.88
Brazil	0.94	0.62	0.96	0.77	0.97	0.83
Bulgaria	0.98	0.82	0.99	0.90	0.99	0.93
Canada (English)						
Canada (French)	0.93	0.71	0.96	0.83	0.97	0.88
Chile	0.94	0.69	0.96	0.81	0.97	0.87
Colombia	0.93	0.64	0.95	0.78	0.96	0.84
Croatia	0.98	0.66	0.99	0.79	0.99	0.85
Czech Republic	0.96	0.79	0.98	0.88	0.98	0.91
Denmark						
Estonia (Estonian)						
Estonia (Russian)	0.92	0.71	0.96	0.84	0.98	0.89
Finland	0.93	0.64	0.96	0.78	0.97	0.84
France	0.93	0.75	0.96	0.86	0.98	0.90
Germany	0.95	0.73	0.97	0.84	0.98	0.89
Greece	0.99	0.77	1.00	0.87	1.00	0.91
Hong Kong-China	0.95	0.66	0.97	0.80	0.98	0.85
Hungary	0.92	0.74	0.94	0.84	0.95	0.88
Iceland	0.93	0.69	0.96	0.82	0.97	0.87
Indonesia	0.97	0.58	0.98	0.73	0.98	0.80
Ireland	0.96	0.76	0.97	0.86	0.98	0.90
Israel	0.94	0.69	0.97	0.82	0.97	0.87
Italy (German)	0.94	0.73	0.97	0.85	0.98	0.90
Italy (Italian)	0.96	0.75	0.98	0.86	0.98	0.90
Japan	0.97	0.71	0.98	0.83	0.99	0.88
Jordan	0.99	0.64	0.99	0.78	0.99	0.84
Korea	0.97	0.70	0.98	0.82	0.99	0.87
Kyrgyzstan (Kyrgyz)	0.78	0.35	0.85	0.53	0.90	0.64
Kyrgyzstan (Russian)	0.97	0.80	0.98	0.89	0.99	0.92
Latvia (Latvian)	0.90	0.67	0.93	0.80	0.95	0.85
Latvia (Russian)	0.92	0.69	0.95	0.81	0.97	0.87
Lithuania	0.92	0.73	0.95	0.84	0.97	0.89
Luxembourg (French)	0.97	0.72	0.98	0.84	0.98	0.88
Luxembourg (German)	0.97	0.77	0.98	0.86	0.98	0.90
Macao-China	0.90	0.57	0.93	0.73	0.95	0.80
Mexico	0.85	0.63	0.90	0.77	0.92	0.83
Montenegro	0.93	0.57	0.93	0.71	0.93	0.77
Netherlands	0.94	0.67	0.96	0.80	0.97	0.86
New Zealand	0.96	0.79	0.98	0.88	0.98	0.92
Norway	0.96	0.78	0.98	0.87	0.98	0.91
Poland	0.96	0.73	0.98	0.84	0.99	0.89
Portugal	0.97	0.59	0.97	0.74	0.98	0.81
Qatar (Arabic)	0.95	0.58	0.96	0.73	0.97	0.80
Qatar (English)	0.95	0.74	0.97	0.85	0.98	0.89
Romania	0.99	0.68	1.00	0.81	1.00	0.86
Russian Federation	1.00	0.74	1.00	0.85	1.00	0.89
Serbia	0.94	0.69	0.96	0.82	0.97	0.87
Slovakia	0.98	0.69	0.99	0.81	0.99	0.87
Slovenia	0.94	0.73	0.97	0.84	0.98	0.89
Spain (Basque)	0.94	0.76	0.96	0.86	0.98	0.91
Spain (Catalan)	0.90	0.70	0.94	0.82	0.96	0.87
Spain (Galician)	0.90	0.67	0.94	0.81	0.96	0.86
Spain (Spanish)	0.93	0.73	0.95	0.84	0.96	0.88
Spain (Valencian)	0.92	0.83	0.95	0.90	0.97	0.93
Sweden	0.94	0.75	0.97	0.85	0.97	0.90
Switzerland (French)	0.94	0.65	0.96	0.79	0.97	0.85
Switzerland (German)	0.94	0.70	0.96	0.82	0.98	0.87
Chinese Taipei	0.99	0.68	1.00	0.81	1.00	0.87
Thailand	0.99	0.65	1.00	0.79	1.00	0.85
Tunisia	0.92	0.66	0.95	0.79	0.96	0.85
Turkey	0.99	0.64	0.99	0.78	1.00	0.84
United Kingdom (Scotland)	0.96	0.77	0.98	0.87	0.99	0.91
United Kingdom (The rest of)	0.96	0.76	0.98	0.86	0.98	0.91
United States						
Uruguay	0.95	0.69	0.97	0.82	0.97	0.87

Note: Countries with no value are displayed, because they fall outside the acceptable [0,1] range.

They provide an index of reliability for the multiple marking in each country. I denotes the number of items and M the number of markers. By using different values for I and M, one obtains a generalisation of the Spearman-Brown formula for test-lengthening. In Table 13.4 to Table 13.6 the formula is evaluated for the three combinations of $I = \{8, 16, 24\}$ and $M = 1$, using the variance component estimates from the corresponding tables presented above. For some countries, no values are displayed, because they fall outside the acceptable (0,1) range.

INTERNATIONAL CODING REVIEW

An international coding review (ICR) was conducted as one of the PISA 2006 quality control procedures in order to investigate the possibility of systematic differences among countries in the coding of open-ended items. The objective of this study was to estimate potential bias (either leniency or harshness) in each country's PISA results, and to express this potential bias in the same units as are used to report country performance on the PISA scales.

The need for the ICR arises because the manual coding of student responses to certain test items is performed by coders trained at the national level. This introduces the possibility of national-level bias in the resulting PISA scores. Coders in country A may interpret and apply the coding instructions more or less leniently than coders in country B.

The data used for the ICR were generated from the multiple coding study. That study, described above, had been implemented earlier to test consistency among coders within each country, and to compare that degree of consistency across countries. Some of the student responses and their multiple codes were selected from the multiple coding study for inclusion in the ICR. These responses, which had already been coded by four national coders, were coded a fifth time by an independent verifier (and in some cases were coded a sixth time by an international adjudicator) to enable estimation of a potential bias.

Background to changed procedures for PISA 2006

Similar ICR studies had been conducted as part of PISA 2000 and PISA 2003 surveys. However, during 2005 and 2006, a review of procedures that had been used previously suggested that improvements and efficiencies could be achieved. The main conclusions from the first two survey cycles were that on the basis of analyses using percentage of agreement among coders, verifiers and adjudicators, there was little evidence of any systematic problems with the application of coding standards; that the relatively small number of problems observed seemed to apply only to particular items (for example only some of the more difficult items) and to only one or two coders in particular national centres. The most useful outcomes of the process, therefore, had been in providing quite specific and detailed information to national centres that would assist them in their own review of coder training procedures, relating either to individual items or to individual coders.

The ICR review called for a simplification of procedures, and most importantly called for the addition of a new element – a way of quantifying the potential impact of any evidence of discrepant coding at the national level on a country's performance. Specifically, a potential bias (degree of harshness or leniency of the coding in each country) expressed in PISA score units, was seen as the most useful way of describing the outcomes of any future ICR.

ICR procedures

Revised procedures designed to estimate national-level bias in coding were developed during the latter part of 2006 and implemented during 2007, achieving simplification and improving effectiveness and efficiency in comparison with procedures used previously. Preliminary planning for the ICR saw the consortium identify a set of booklet types and a set of items for inclusion in the study. Three booklets were chosen: booklet 5 (from which 15 science items were selected, of the 42 science items in total requiring manual coding), booklet 6

(from which 14 of the available 17 manually coded reading items were selected), and booklet 8 (from which 9 mathematics items were selected, of the 20 mathematics items altogether requiring manual coding).

These booklets and items were also amongst those used previously in the multiple coding study. A random selection was made of 60 of these booklets for each domain from each distinct coding centre within all adjudicated PISA entities (and selecting a representative proportion of each language involved). This meant that 900 responses to science items, 840 responses to reading items, and 540 responses to mathematics items were available from each national coding centre for examination in the ICR. The codes that had been assigned to the student responses to these items by the four national coders involved previously in the multiple coder study were extracted. Coding of each student response a fifth time was then carried out by a member of a team of independent reviewers who had been trained specifically for this task. These independent reviewers had been involved as part of the international translation verification team. The code assigned by the independent reviewer was referred to as the verifier code.

The ICR analysis procedures were carried out in two related but independent parts. The first part was aimed at identifying countries in which evidence of coder bias exists, and estimating the magnitude of that bias. The second part was aimed at identifying particular items, student responses, and coders, that tended to generate coding discrepancies.

Part 1: Flagging countries

The main goal of the analysis of the ICR data was to express leniency or harshness of national coders as an effect on countries' mean performance in each PISA domain. For some countries, where national coding was performed by different teams each having responsibility for student responses in different languages, results were analysed separately for language-based subgroups. To perform this analysis, the domain-ability (using weighted likelihood estimates, or WLEs) of each of the 60 selected students was estimated twice: once using the original reported score on all items from that domain in the relevant booklet; and once with the verifier codes substituted for each item response from that booklet that had been included in the ICR. The scores for items not included in the ICR stayed unchanged in the two estimations. The reported scores for each student were derived from a mixture of about 25% of codes from each of the four national coders involved in the Multiple Coder Study. The abilities were transformed to the PISA scale. This resulted in a maximum of 60 pairs of ability estimates, from which 60 differences were calculated. The average of the differences in each country was an indication of the bias in country mean performance for that domain. In fact a 95% confidence interval was constructed around the mean difference, and if that interval did not contain the value zero then potential bias was indicated.

A *t*-test was then performed on the paired ability estimates to test for significance of the difference in country mean performance. If the country mean performance that was based on the verifier codes differed significantly from the mean performance based on the reported scores, the country was flagged as having a potential bias in their average score for that domain. Before confirming this potential bias, the consortium implemented one final quality check: a review to judge the quality of the verifier codes. This final review is referred to as adjudication.

Nineteen responses were randomly selected for each flagged country by domain (by language) combination for adjudication. Before selecting these responses, cases with perfect agreement amongst the five coders were excluded, because it is highly likely that the adjudicator would agree with the verifier in these cases. The 19 responses that were selected were sent to an international adjudicator, along with the five previously assigned codes. This review and adjudication was carried out by the consortium staff member responsible

for leading the relevant domain. The adjudicator provided a single definitive code to each of the sampled student responses, which had been back-translated into English for this purpose.

The overall percentage of agreement between verifier and adjudicator for one domain in one country was estimated based on their coding of the 19 responses. Two assumptions had to be made for this estimation: (1) that the percentage of agreement between verifier and adjudicator would have been 100% for the excluded responses that had perfect agreement among the first five coders, and (2) that the percentage of agreement on the 19 responses could be generalised to the responses that were randomly not selected for adjudication.

The percentage agreement, $\hat{P}$, between verifier and adjudicator was therefore estimated as follows:

13.3

$$\hat{P} = \frac{[n + (N - n)Z]100}{N}$$

where n is the number of responses for which there was perfect agreement among verifier and all four national coders, Z is the observed proportion of adjudicated responses for which the adjudicator and verifier agreed, and N is the total number of responses (usually 60).

The estimated percentage of agreement between verifier and adjudicator was used to assess the quality of the verifier codes. If the percentage was 90 or above, the coding from the verifier was deemed to be correct and the estimated national bias was reported. If the percentage was below 90, the verifier codes were deemed to be not sufficiently reliable to justify confirmation of the observed difference in country mean.

Part 2: Flagging responses

The second part of the ICR procedure for PISA 2006 aimed to give a more in-depth picture of differences between national coders and international verifiers by country, language, domain and item, in order to support evaluation and improvement processes within countries.

After international verifiers completed their coding of the 900 science, 840 reading and 540 mathematics responses for each country, their codes were compared to the four codes given by the national coders. 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;
- The average raw score of the 4 coders was at least 0.5 points higher or lower than the score based on the verifier code.

Examples of flagged cases are given in Table 13.6.

Table 13.7
Examples of flagged cases

CNT	Student ID	Question	Coder1	Coder2	Coder3	Coder4	Verifier	Flag (Y/N)
xxx	Xxxxx00001	R067Q04	0	1	1	1	1	N
xxx	Xxxxx00012	R067Q04	1	1	1	1	0	Y
xxx	Xxxxx00031	R067Q04	1	1	1	0	0	Y
xxx	Xxxxx00014	R067Q04	0	1	1	2	0	Y
xxx	Xxxxx00020	R067Q04	1	0	2	1	2	Y
xxx	Xxxxx00025	R067Q04	2	0	2	0	2	Y

In addition to flagging cases of discrepancy between national coders and verifier, the individual items figuring more frequently in these discrepancies were also identified for each country. The difference between the mean raw score from the four national codes and the raw score from the verifier code was calculated item by item. The 60 differences per item (in case of one test language) were averaged. A positive difference for a particular item was an indication of leniency of national coders for that item, a negative difference an indicator of harshness of national coders. The number and percentages of flagged responses and mean differences per item were reported back to national centres as described later in this chapter.

Outcomes

Sixty-seven units of analysis were involved in the ICR study for PISA 2006, each comprising a country or a language-based group within a country. Each unit was analysed for the three assessment domains of science, reading and mathematics. Of these 67 units, in the first stage of the analysis (Part 1: Flagging countries), 26 were flagged for adjudication in mathematics, 41 in reading and 29 in science. These are summarised in Table 13.8.

Table 13.8
Count of analysis groups showing potential bias, by domain

Potential difference indicated	Mathematics	Reading	Science	Total (%)
Harshness in national coding	9	13	14	36 (17.9%)
No significant difference	41	26	38	105 (52.2%
Leniency in national coding	17	28	15	60 (29.9%)
Total Analysis Groups	**67**	**67**	**67**	**201 (100%)**

In order to confirm the potential bias indicated by this flagging process, the overall consistency of the adjudicator and verifier codes was checked. Table 13.9 shows an overall summary of this comparison. In over 60% of the individual cases (across the three domains) the adjudicator agreed with the code assigned by the verifier.

Table 13.9
Comparison of codes assigned by verifier and adjudicator

Difference (Verifier-Adjudicator)	Number of Cases	Percent
–2	58	3.7
–1	293	18.6
0	952	60.5
1	241	15.3
2	30	1.9
Total Cases	**1574**	**100.0**

After adjudication, differences between mean performance for the 67 units of analysis using the reported codes and the verifier codes were judged to be significant in 22 units for mathematics, 20 for reading and 13 for science. The units are listed in Table 13.10. The '+' symbol indicates that the difference was positive, suggesting potential lenience in the national coding. The '–' symbol indicates that the difference was negative, suggesting potential harshness in the national coding. Blank cells indicate either no evidence of bias, or that evidence of bias was not confirmed by the adjudicator. Of the 55 units in which the difference was confirmed, 30 cases indicated positive bias (leniency in national coding) and 25 cases indicated negative bias (harshness in national coding).

In total, 25 cases of harshness in the standards applied in national coding centres were detected, alongside 30 cases of lenient coding at national level.

Table 13.10
Outcomes of ICR analysis part 1

	Reading	Mathematics	Science
Argentina		+	
Australia	+		
Austria			–
Azerbaijan	+	+	
Belgium (FLA)	–	+	
Belgium (FRA)		–	
Brazil			
Bulgaria			+
Canada (ENG)			
Canada (FRA)			
Chile		+	
Colombia			
Croatia	–		
Czech Republic			+
Denmark		–	
Estonia (EST)		+	+
Estonia (RUS)			
Finland			+
France		–	
Germany			
Greece	–		+
Hong Kong-China			–
Hungary		+	
Iceland		+	
Indonesia			
Ireland	+		
Israel			+
Italy			+
Japan			
Jordan			
Korea			–
Kyrgyzstan (KIR)	+		
Kyrgyzstan (RUS)	+		
Latvia (LVA)	–		
Latvia (RUS)	+	+	
Lithuania			
Luxembourg	+		
Macao-China	–	–	
Mexico			
Montenegro	–	–	
Netherlands			
New Zealand			
Norway			
Poland			
Portugal		–	
Qatar (ARA)	+	+	–
Qatar (ENG)	+	+	
Romania	–		
Russian Federation		–	
Serbia	–		
Slovak Republic			+
Slovenia			
Spain (BAQ)			
Spain (CAT)	–		
Spain (GLG)	–		
Spain (SPA)	–		
Sweden			
Switzerland (FRE)			
Switzerland (GER)	–		
Chinese Taipei		+	
Thailand			
Tunisia			
Turkey	–	+	
UK. England. Wales. N. Ireland			
UK. Scotland		–	+
Uruguay			
United States			
Count harsh (“-“)	**13**	**8**	**4**
Count lenient (“+“)	**9**	**12**	**9**
Count no difference	**45**	**47**	**54**

Table 13.11 [Part 1/3]
ICR outcomes by country and domain

		PISA score difference (reported-verifier)				PISA scores		
	Domain	Sign	CI_lo	CI_hi	Agree (%)	Ver	Rep	Adj
Argentina	Mathematics	ns	–2.72	4.55				
	Reading	+	5.06	13.84	97.40	15.90	17.10	15.90
	Science	+	1.16	6.54	94.80	21.40	21.90	21.50
Australia	Mathematics	+	0.88	8.58	97.40	17.50	17.60	17.50
	Reading	ns	–10.92	4.01				
	Science	ns	–3.49	1.97				
Austria	Mathematics	ns	–2.39	3.66				
	Reading	ns	–11.51	0.33				
	Science	–	–4.16	–0.02	95.80	38.30	37.80	38.20
Azerbaijan	Mathematics	+	7.28	13.46	98.10	10.20	11.60	10.60
	Reading	+	10.35	30.28	98.00	13.80	16.20	13.80
	Science	–	–5.88	–0.05	95.40	20.30	19.30	19.80
Belgium (FRE)	Mathematics	ns	–4.23	1.36				
	Reading	–	–20.67	–0.79	95.20	22.20	20.90	22.20
	Science	ns	–1.01	2.94				
Belgium (DUT)	Mathematics	–	–6.90	–0.06	95.20	18.00	17.60	17.70
	Reading	+	11.26	22.34	96.20	21.60	23.40	21.80
	Science	ns	–0.67	2.23				
Bulgaria	Mathematics	ns	–2.38	4.31				
	Reading	+	4.04	19.61	90.60	13.20	14.10	13.20
	Science	+	6.30	12.53	98.50	28.70	30.80	28.60
Brazil	Mathematics	ns	–5.76	1.20				
	Reading	ns	–3.30	10.81				
	Science	ns	–2.60	3.17				
Canada (ENG)	Mathematics	ns	0.99	11.33				
	Reading	ns	–3.78	5.76				
	Science	–	–5.92	1.46	90.40	37.70	37.60	37.80
Canada (FRE)	Mathematics	ns	–9.69	3.39				
	Reading	ns	–13.67	8.35				
	Science	–	–11.48	–0.61	87.50	31.30	30.00	31.00
Chile	Mathematics	–	–9.08	–1.69	94.20	11.30	10.80	10.80
	Reading	+	0.17	9.08	95.00	17.70	18.10	18.30
	Science	ns	–4.13	0.40				
Colombia	Mathematics	ns	–0.13	5.13				
	Reading	ns	–9.03	3.15				
	Science	ns	–2.27	1.94				
Croatia	Mathematics	–	–8.60	–1.16	96.50	13.40	12.70	12.90
	Reading	ns	–0.44	10.70				
	Science	ns	–2.26	1.63				
Czech Republic	Mathematics	ns	–2.27	3.05				
	Reading	+	3.75	15.54	91.10	19.90	20.50	20.20
	Science	+	0.94	7.42	93.30	43.90	44.60	44.30
Denmark	Mathematics	ns	–5.39	1.93				
	Reading	–	–15.45	–3.60	94.00	22.90	21.30	22.70
	Science	ns	–3.17	0.81				
Estonia	Mathematics	ns	–3.68	2.72				
	Reading	+	3.81	17.07	95.20	24.20	25.50	24.60
	Science	+	3.10	11.30	100.00	34.80	36.10	34.80
Estonia (RUS)	Mathematics	ns	–1.25	3.40				
	Reading	–	–11.99	7.61	81.50	21.00	20.80	21.50
	Science	–	–6.71	6.57	92.90	31.50	31.00	31.20
Finland	Mathematics	ns	–1.55	5.26				
	Reading	ns	–11.65	4.97				
	Science	+	1.43	5.71	95.30	42.60	42.90	42.80
France	Mathematics	ns	–6.67	0.55				
	Reading	–	–13.09	–0.43	92.30	23.80	23.40	23.70
	Science	ns	–1.21	3.92				
Germany	Mathematics	ns	–4.64	1.26				
	Reading	ns	–5.67	4.58				
	Science	ns	–4.93	0.72				
Greece	Mathematics	–	–5.58	–0.59	97.30	13.50	13.10	13.10
	Reading	ns	–8.82	0.42				
	Science	+	1.71	5.87	98.00	34.00	35.10	34.30
Hong Kong-China	Mathematics	ns	–2.79	3.36				
	Reading	ns	–5.32	6.82				
	Science	–	–5.64	–0.48	97.50	41.00	40.70	41.00
Hungary	Mathematics	ns	–0.16	6.67				
	Reading	+	3.86	18.76	93.10	21.60	22.50	21.90
	Science	+	1.69	6.19	93.00	37.10	37.50	37.40

Table 13.11 [Part 2/3]
ICR outcomes by country and domain

		PISA score difference (reported-verifier)				PISA scores		
	Domain	Sign	CI_lo	CI_hi	Agree (%)	Ver	Rep	Adj
Iceland	Mathematics	ns	–6.24	0.41				
	Reading	+	2.54	14.49	93.90	19.80	21.00	19.90
	Science	ns	–2.90	0.19				
Indonesia	Mathematics	ns	–0.60	11.63				
	Reading	–	–15.36	–2.25	88.40	12.10	11.60	11.80
	Science	+	2.15	7.16	86.40	21.70	22.50	21.30
Ireland	Mathematics	+	0.33	6.17	97.70	14.20	14.80	14.30
	Reading	+	1.67	11.91	90.70	22.80	23.20	23.00
	Science	ns	–1.98	3.04				
Israel	Mathematics	ns	–2.50	5.53				
	Reading	ns	–8.81	4.54				
	Science	+	0.31	5.17	94.50	35.10	35.60	35.40
Italy	Mathematics	ns	–6.87	0.14				
	Reading	ns	–4.37	3.66				
	Science	+	0.52	5.13	96.10	37.10	37.50	37.80
Jordan	Mathematics	ns	–7.84	2.71				
	Reading	ns	–0.28	9.94				
	Science	+	0.77	6.04	94.30	26.40	27.10	26.60
Japan	Mathematics	ns	–5.52	0.53				
	Reading	+	16.77	30.32	87.00	22.50	24.90	23.30
	Science	ns	–2.84	1.36				
Korea	Mathematics	ns	–3.85	3.27				
	Reading	+	16.33	27.12	90.50	24.00	26.20	24.90
	Science	–	–4.71	–0.78	94.70	37.90	37.70	38.00
Kyrgyzstan (KIR)	Mathematics	+	–1.10	7.76	99.50	4.90	5.10	4.80
	Reading	ns	–1.28	8.39				
	Science	–	–5.59	0.45	96.80	14.10	13.60	13.80
Kyrgyzstan (RUS))	Mathematics	+	–1.15	10.96	100.00	9.70	10.00	9.70
	Reading	ns	–11.09	19.11				
	Science	–	–7.79	2.70	92.90	17.70	17.60	18.00
Latvia (LVA)	Mathematics	–	–14.89	–5.63	94.00	14.40	14.00	14.40
	Reading	+	–5.23	7.51	89.10	23.00	22.70	23.50
	Science	ns	–6.02	0.01				
Latvia (RUS)	Mathematics	+	–3.44	14.04	95.70	15.20	14.60	15.40
	Reading	+	13.30	33.71	92.30	19.00	20.30	19.30
	Science	ns	–5.67	6.52				
Lithuania	Mathematics	ns	–4.13	1.67				
	Reading	–	–9.71	–1.43	92.40	19.20	19.20	19.90
	Science	ns	–5.01	1.04				
Luxembourg	Mathematics	+	1.93	7.85	96.60	13.60	14.30	13.90
	Reading	ns	–8.30	2.03				
	Science	ns	–2.36	1.48				
Macao-China	Mathematics	–	–7.50	–0.57	97.90	15.70	15.30	15.70
	Reading	–	–12.71	–0.22	94.60	20.10	19.60	20.10
	Science	ns	–4.64	1.11				
Mexico	Mathematics	–	–11.54	–3.57	93.20	11.40	10.60	11.10
	Reading	ns	–5.78	7.95				
	Science	–	–12.87	–8.45	87.90	26.90	25.60	26.60
Montenegro	Mathematics	–	–10.47	–1.37	98.70	11.10	10.60	10.90
	Reading	–	–17.56	–1.41	98.10	14.70	13.30	14.50
	Science	ns	–2.02	2.48				
Netherlands	Mathematics	ns	–2.72	6.15				
	Reading	+	0.79	15.65	79.60	21.40	22.20	22.10
	Science	+	1.36	8.22	80.60	38.10	39.20	38.60
New Zealand	Mathematics	ns	–1.45	4.86				
	Reading	ns	–0.01	11.38				
	Science	ns	–3.43	1.86				
Norway	Mathematics	ns	–0.72	4.59				
	Reading	+	17.46	30.65	92.50	19.50	21.20	20.00
	Science	ns	–3.80	0.41				
Poland	Mathematics	ns	–0.05	5.78				
	Reading	ns	–2.21	8.75				
	Science	ns	–3.48	0.91				
Portugal	Mathematics	–	–11.73	–3.65	90.90	15.30	14.30	14.70
	Reading	–	–28.44	–15.39	94.30	21.90	19.50	21.70
	Science	–	–14.93	–8.79	90.30	33.20	31.20	32.90

Table 13.11 [Part 3/3]
ICR outcomes by country and domain

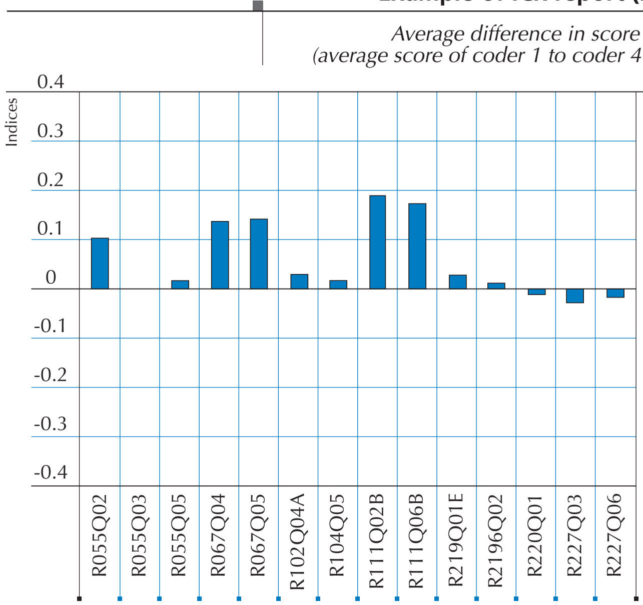
		PISA score difference (reported-verifier)				PISA scores		
	Domain	Sign	CI_lo	CI_hi	Agree (%)	Ver	Rep	Adj
Qatar (ARA)	Mathematics	+	0.54	15.16	98.80	6.00	6.90	6.60
	Reading	+	5.91	14.89	97.40	12.30	12.80	12.40
	Science	–	–5.32	–0.18	98.70	24.90	24.10	24.70
Qatar (ENG)	Mathematics	+	–0.95	15.25	99.20	11.90	12.90	13.20
	Reading	+	–1.83	26.91	97.40	23.60	24.60	23.80
	Science	–	–8.35	2.90	92.30	32.40	31.60	31.80
Romania	Mathematics	–	–6.30	–0.64	98.10	8.90	8.20	8.50
	Reading	ns	–13.55	0.38				
	Science	ns	–5.20	1.21				
Russian Federation	Mathematics	ns	–1.87	4.01				
	Reading	–	–26.37	–15.21	94.20	21.10	19.40	21.10
	Science	ns	–0.61	3.96				
Serbia	Mathematics	–	–10.13	–3.19	95.90	13.60	12.80	13.10
	Reading	ns	–8.39	1.48				
	Science	–	–5.70	–1.33	92.40	30.50	29.40	30.30
Scotland	Mathematics	ns	–3.91	2.76				
	Reading	–	–14.08	–2.32	92.90	22.80	22.40	23.00
	Science	+	0.96	6.87	95.70	39.60	40.30	42.60
Slovak Republic	Mathematics	ns	–4.58	2.25				
	Reading	+	6.02	15.37	91.90	18.50	19.90	19.30
	Science	+	1.49	5.78	94.40	38.60	39.10	38.90
Slovenia	Mathematics	ns	–3.66	3.17				
	Reading	+	2.62	14.24	91.50	20.90	21.10	21.10
	Science	+	2.24	7.16	93.40	39.30	39.90	39.50
Spain (BAQ)	Mathematics	ns	–13.48	3.21				
	Reading	ns	–6.97	19.02				
	Science	ns	–10.01	2.33				
Spain (CAT)	Mathematics	–	–10.18	–2.45	95.50	15.10	14.40	14.80
	Reading	ns	–12.06	0.17				
	Science	ns	–3.83	1.03				
Spain (GLG)	Mathematics	–	–10.45	–1.64	97.10	14.50	13.80	14.20
	Reading	+	–1.15	18.98	85.10	18.00	19.30	18.70
	Science	ns	–6.96	0.41				
Spain (SPA)	Mathematics	–	–4.46	–0.76	97.70	17.00	16.70	16.80
	Reading	ns	–5.50	3.88				
	Science	ns	–0.59	3.07				
Sweden	Mathematics	ns	–4.69	2.08				
	Reading	+	14.05	29.10	91.20	22.10	24.10	22.40
	Science	ns	–0.61	3.82				
Switzerland (FRE)	Mathematics	–	–11.10	6.86	92.20	16.60	16.20	16.80
	Reading	ns	–23.92	2.49		15.00	13.00	15.00
	Science	ns	–2.75	9.74				
Switzerland (GER)	Mathematics	–	–11.45	–2.55	95.90	18.30	17.90	18.00
	Reading	–	–22.28	–4.11	89.90	25.10	23.90	25.20
	Science	ns	–5.04	1.51				
Chinese Taipei	Mathematics	ns	–5.05	1.48				
	Reading	+	2.51	12.48	98.30	23.70	24.50	23.90
	Science	ns	–3.72	0.95				
Thailand	Mathematics	ns	–4.25	0.33				
	Reading	–	–16.39	–5.24	94.20	19.30	18.20	19.30
	Science	ns	–3.69	0.06				
Tunisia	Mathematics	ns	–4.77	0.95				
	Reading	+	5.94	19.20	91.20	12.20	13.00	12.30
	Science	ns	–2.85	1.82				
Turkey	Mathematics	–	–10.53	–2.66	96.10	13.40	12.60	12.80
	Reading	+	9.44	22.44	96.70	18.00	20.20	18.10
	Science	ns	–0.78	5.44				
United Kingdom	Mathematics	ns	–2.58	6.32				
	Reading	+	0.79	10.93	87.60	20.30	20.90	20.80
	Science	ns	–3.84	0.45				
Uruguay	Mathematics	ns	–6.10	1.34				
	Reading	+	2.18	13.74	88.00	19.10	19.80	19.50
	Science	ns	–2.66	2.13				
United States	Mathematics	ns	–0.06	5.79				
	Reading	+	1.33	10.42	89.30	23.90	24.30	24.50
	Science	ns	–0.71	5.55				

In Table 13.11 the outcomes of the ICR process are summarised for each country and by language group (where appropriate) and domain. In columns 3–5 of that table, information is reported about the estimated bias in the national score for the domain, in PISA score units, based on the difference observed when the score is calculated from national scores, and when calculated using the verifier score. The sign of any difference is reported, with the "+" symbol indicating leniency at the national level, "–" indicating harshness at the national level, and "ns" indicating no significant difference. The 95% confidence interval around the mean difference is reported in the next two columns. The column headed "Agree (%)" displays the estimated level of agreement between the adjudicator and the verifier, calculated according to the formula given earlier. And finally, three estimated PISA scores are given – those based on the codes given by the verifier, the country codes, and the adjudicator codes respectively.

At the conclusion of the ICR, a report was sent to each participant country summarising the outcomes of the international coding review for each test domain. The report contained several elements. One was a graph showing the discrepancies item by item within each domain between the average raw score based on codes given by the four national coders, and the raw score from the verifier's code, hence providing a fine-grained report at the item level of average discrepancies of national coders relative to an independent benchmark. The report also showed the number and the percentage of individual student responses that had been flagged in Part 2 of the ICR analysis. Finally, the report showed whether there was statistical evidence of bias in national coding, and the estimate of the extent of the bias in PISA score units. National centres were therefore given information that they could use to review their coding operation, and to inform planning for the recruitment and training of coders for future surveys.

An example of an ICR country report is provided in Figure 13.7. Looking at this example, the graph indicates a marked positive average difference between the mean of the four national coders' scores and the verifier score for five of the 14 reading items. Differences for the other nine reading items were much smaller, or non-existent. This provides evidence of leniency in the standards applied by coders in this country in the coding of five of the reading items. This information may be useful input to the coder training for the next PISA survey cycle.

Figure 13.7
Example of ICR report (reading)



Total number of flagged responses	**56**
Total per cent of flagged responses	**6.67%**
Confidence interval for estimated difference in PISA score (reported – verified)	**[5.06 , 13.84]**

To the right of this graph, the total number and percentage of flagged responses are given for this domain. In this example, 56 of the 840 reading item responses that were included in the ICR study from this country were flagged. That is, for about 6% of the student responses reviewed, differences were observed between the coding standards applied by the national coders and those applied by the international verifier.

The final element of the report is the estimated bias in the average reading score for this country expressed as a range of values, in PISA score units. The values are the 95% confidence interval about the mean estimate. This information is reported only in cases where the final adjudication process confirms the differences found by the international verifier.

The difference is calculated between the country's reported average reading score, and the score that would be calculated had the codes awarded by the international verifier been used in the scaling, but based only on the reading items in the test booklet used in the ICR. For this country, the degree of leniency estimated lies between about 5 and 14 points on the PISA reading scale.

Cautions

In interpreting the results of the international coder review, it should be borne in mind that the study gives only an indication of possible bias in national results.

First, only some of the manually coded items in each domain were included in the ICR, and the items selected for inclusion were not intended as a random sample of all manually coded items. The selection was made largely on practical and logistical grounds designed to minimise work for participating countries, namely, what was a selection of a small number of booklets that contained as many suitable items as possible. The behaviour of national coders on these items may not be an accurate representation of their behaviour in coding all items.

Related to this, the estimation of the magnitude of observed bias uses mean national ability estimates that are based only on one booklet for each domain, whereas reported PISA outcomes are based on a rotated design involving all 13 booklets. It is well known that positioning of items within test booklets has an impact on the calculation of item difficulty estimates, and therefore also student ability estimates. This further exacerbates the potential unreliability of the bias estimates.

Reader's Guide

Country codes – the following country codes are used in this report:

OECD countries

AUS	Australia
AUT	Austria
BEL	Belgium
BEF	Belgium (French Community)
BEN	Belgium (Flemish Community)
CAN	Canada
CAE	Canada (English Community)
CAF	Canada (French Community)
CZE	Czech Republic
DNK	Denmark
FIN	Finland
FRA	France
DEU	Germany
GRC	Greece
HUN	Hungary
ISL	Iceland
IRL	Ireland
ITA	Italy
JPN	Japan
KOR	Korea
LUX	Luxembourg
LXF	Luxembourg (French Community)
LXG	Luxembourg (German Community)
MEX	Mexico
NLD	Netherlands
NZL	New Zealand
NOR	Norway
POL	Poland
PRT	Portugal
SVK	Slovak Republic
ESP	Spain
ESB	Spain (Basque Community)
ESC	Spain (Catalonian Community)
ESS	Spain (Castillian Community)
SWE	Sweden
CHE	Switzerland
CHF	Switzerland (French Community)
CHG	Switzerland (German Community)
CHI	Switzerland (Italian Community)
TUR	Turkey
GBR	United Kingdom
IRL	Ireland
SCO	Scotland
USA	United States

Partner countries and economies

ARG	Argentina
AZE	Azerbaijan
BGR	Bulgaria
BRA	Brazil
CHL	Chile
COL	Colombia
EST	Estonia
HKG	Hong Kong-China
HRV	Croatia
IDN	Indonesia
JOR	Jordan
KGZ	Kyrgyztan
LIE	Liechtenstein
LTU	Lithuania
LVA	Latvia
LVL	Latvia (Latvian Community)
LVR	Latvia (Russian Community)
MAC	Macao-China
MNE	Montenegro
QAT	Qatar
ROU	Romania
RUS	Russian Federation
SRB	Serbia
SVN	Slovenia
TAP	Chinese Taipei
THA	Thailand
TUN	Tunisia
URY	Uruguay

References

Adams, R.J., Wilson, M. & **Wang, W.C.** (1997), The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement,* No. 21, pp. 1-23.

Adams, R.J., Wilson, M. R. & **Wu, M.L.** (1997), Multilevel item response models: An approach to errors in variables regression, *Journal of Educational and Behavioural Statistics,* No. 22 (1), pp. 46-75.

Adams, R.J. & **Wu, M.L.** (2002), *PISA 2000 Technical Report,* OECD, Paris.

Bollen, K.A. & **Long, S.J.** (1993) (eds.), *Testing Structural Equation Models*, Newbury Park: London.

Beaton, A.E. (1987), Implementing the new design: The NAEP 1983-84 technical report (Rep. No. 15-TR-20), Princeton, NJ: Educational Testing Service.

Buchmann, C. (2000), Family structure, parental perceptions and child labor in Kenya: What factors determine who is enrolled in school? *Soc. Forces,* No. 78, pp. 1349-79.

Buchmann, C. (2002), Measuring Family Background in International Studies of Education: Conceptual Issues and Methodological Challenges, in Porter, A.C. and Gamoran, A. (eds.). *Methodological Advances in Cross-National Surveys of Educational Achievement* (pp. 150-97), Washington, DC: National Academy Press.

Creemers, B.P.M. (1994), *The Effective Classroom*, London: Cassell.

Cochran, W.G. (1977), *Sampling techniques,* third edition, New York, NY: John Wiley and Sons.

Ganzeboom, H.B.G., de Graaf, P.M. & **Treiman, D.J.** (1992), A standard international socio-economic index of occupational status, *Social Science Research*, No. 21, pp. 1-56.

Ganzeboom H.B. & **Treiman, D.J.** (1996), Internationally comparable measures of occupational status for the 1988 international standard classification of occupations, *Social Science Research*, No. 25, pp. 201-239.

Grisay, A. (2003), Translation procedures in OECD/PISA 2000 international assessment, *Language Testing,* No. 20 (2), pp. 225-240.

Hambleton, R.K., Swaminathan, H. & **Rogers, H.J.** (1991), *Fundamentals of item response theory*, Newbury Park, London, New Delhi: SAGE Publications.

Hambleton, R.K., Merenda, P.F. & **Spielberger, C.D.** (2005), *Adapting Educational and Psychological Tests for Cross-Cultural Assessment,* IEA Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey.

Harkness, J.A., Van de Vijver, F.J.R. & **Mohler, P.Ph** (2003), *Cross-Cultural Survey Methods,* Wiley-Interscience, John Wiley & Sons, Inc., Hoboken, New Jersey.

Harvey-Beavis, A. (2002), Student and School Questionnaire Development, in R.J. Adams and M.L. Wu (eds.), *PISA 2000 Technical Report,* (pp. 33-38), OECD, Paris.

International Labour Organisation (ILO) (1990), *International Standard Classification of Occupations: ISCO-88.* Geneva: International Labour Office.

Jöreskog, K.G. & **Sörbom, Dag** (1993), *LISREL 8 User's Reference Guide*, Chicago: SSI.

Judkins, D.R. (1990), Fay's Method of Variance Estimation, *Journal of Official Statistics*, No. 6 (3), pp. 223-239.

Kaplan, D. (2000), *Structural equation modeling: Foundation and extensions*, Thousand Oaks: SAGE Publications.

Keyfitz, N. (1951), Sampling with probabilities proportionate to science: Adjustment for changes in probabilities, *Journal of the American Statistical Association,* No. 46, American Statistical Association, Alexandria, pp. 105-109.

Kish, L. (1992), Weighting for Unequal, *Pi. Journal of Official Statistics*, No. 8 (2), pp. 183-200.

LISREL (1993), K.G. Jöreskog & D. Sörbom, [computer software], Lincolnwood, IL: Scientific Software International, Inc.

Lohr, S.L. (1999), *Sampling: Design and Analysis*, Duxberry: Pacific Grove.

Macaskill, G., Adams, R.J. & **Wu, M.L.** (1998), Scaling methodology and procedures for the mathematics and science literacy, advanced mathematics and physics scale, in M. Martin and D.L. Kelly, Editors, *Third International Mathematics and Science Study, technical report Volume 3: Implementation and analysis*, Boston College, Chestnut Hill, MA.

Masters, G.N. & **Wright, B.D.** (1997), The Partial Credit Model, in W.J. van der Linden, & R.K. Hambleton (eds.), *Handbook of Modern Item Response Theory* (pp. 101-122), New York/Berlin/Heidelberg: Springer.

Mislevy, R.J. (1991), Randomization-based inference about latent variables from complex samples, *Psychometrika,* No. 56, pp. 177-196.

Mislevy, R.J., Beaton, A., Kaplan, B.A. & **Sheehan, K.** (1992), Estimating population characteristics from sparse matrix samples of item responses, *Journal of Educational Measurement,* No. 29 (2), pp. 133-161.

Mislevy, R.J. & **Sheehan, K.M.** (1987), Marginal estimation procedures, in Beaton, A.E., Editor, 1987. *The NAEP 1983-84 technical report*, National Assessment of Educational Progress, Educational Testing Service, Princeton, pp. 293-360.

Mislevy, R.J. & **Sheehan, K.M.** (1989), Information matrices in latent-variable models, *Journal of Educational Statistics*, No. 14, pp. 335-350.

Mislevy, R.J. & **Sheehan, K.M.** (1989), The role of collateral information about examinees in item parameter estimation, *Psychometrika,* No. 54, pp. 661-679.

Monseur, C. & **Berezner, A.** (2007), The Computation of Equating Errors in International Surveys in Education, *Journal of Applied Measurement,* No. 8 (3), 2007, pp. 323-335.

Monseur, C. (2005), An exploratory alternative approach for student non response weight adjustment, *Studies in Educational Evaluation*, No. 31 (2-3), pp. 129-144.

Muthen, B. & **L. Muthen** (1998), [computer software], *Mplus* Los Angeles, CA: Muthen & Muthen.

Muthen, B., du Toit, S.H.C. & **Spisic, D.** (1997), *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*, unpublished manuscript.

OECD (1999), *Classifying Educational Programmes. Manual for ISCED-97 Implementation in OECD Countries,* OECD, Paris.

OECD (2003), *Literacy Skills for the World of Tomorrow: Further results from PISA 2000*, OECD, Paris.

OECD (2004), *Learning for Tomorrow's World – First Results from PISA 2003*, OECD, Paris.

OECD (2005), *Technical Report for the OECD Programme for International Student Assessment 2003*, OECD, Paris.

OECD (2006), *Assessing Scientific, Reading and Mathematical Literacy: A framework for PISA 2006,* OECD, Paris.

OECD (2007), *PISA 2006: Science Competencies for Tomorrow's World*, OECD, Paris.

PISA Consortium (2006), *PISA 2006 Main Study Data Management Manual*, *https://mypisa.acer.edu.au/images/mypisadoc/opmanual/pisa2006_data_management_manual.pdf*

Rasch, G. (1960), Probabilistic models for some intelligence and attainment tests, Copenhagen: Nielsen & Lydiche.

Routitski A. & **Berezner, A.** (2006), Issues influencing the validity of cross-national comparisons of student performance. Data Entry Quality and Parameter Estimation. Paper presented at the Annual Meeting of the American Educational Research Association (AERA) in San Francisco, 7-11 April, *https://mypisa.acer.edu.au/images/mypisadoc/aera06routitsky_berezner.pdf*

Rust, K. (1985), Variance Estimation for Complex Estimators in Sample Surveys, *Journal of Official Statistics*, No. 1, pp. 381-397.

Rust, K.F. & **Rao, J.N.K.** (1996), Variance Estimation for Complex Surveys Using Replication Techniques, *Survey Methods in Medical Research*, No. 5, pp. 283-310.

Shao, J. (1996), Resampling Methods in Sample Surveys (with Discussion), *Statistics*, No. 27, pp. 203-254.

Särndal, C.-E., Swensson, B. & **Wretman, J.** (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.

SAS® CALIS (1992), W. Hartmann [computer software], Cary, NC: SAS Institute Inc.

Scheerens, J. (1990), School effectiveness and the development of process indicators of school functioning, *School effectiveness and school improvement,* No. 1, pp. 61-80.

Scheerens, J. & **Bosker, R.J.** (1997), *The Foundations of School Effectiveness*, Oxford: Pergamon.

Schulz, W. (2002), Constructing and Validating the Questionnaire composites, in R.J. Adams and M.L. Wu (eds.), *PISA 2000 Technical Report*, OECD, Paris.

Schulz, W. (2004), Mapping Student Scores to Item Responses, in W. Schulz and H. Sibberns (eds.), *IEA Civic Education Study, Technical Report* (pp. 127-132)*,* Amsterdam: IEA.

Schulz, W. (2006a), *Testing Parameter Invariance for Questionnaire Indices using Confirmatory Factor Analysis and Item Response Theory,* Paper presented at the Annual Meetings of the American Educational Research Association (AERA) in San Francisco, 7-11 April.

Schulz, W. (2006b), *Measuring the socio-economic background of students and its effect on achievement in PISA 2000 and PISA 2003*, Paper presented at the Annual Meetings of the American Educational Research Association (AERA) in San Francisco, 7-11 April.

Thorndike, R.L. (1973), *Reading comprehension in fifteen countries,* New York, Wiley: and Stockholm: Almqvist & Wiksell.

Travers, K.J. & **Westbury, I.** (1989), *The IEA Study of Mathematics I: Analysis of Mathematics Curricula*, Oxford: Pergamon Press.

Travers, K.J., Garden R.A. & **Rosier, M.** (1989), Introduction to the Study, in Robitaille, D. A. and Garden, R. A. (eds), *The IEA Study of Mathematics II: Contexts and Outcomes of School Mathematics Curricula,* Oxford: Pergamon Press.

Verhelst, N. (2002), Coder and Marker Reliabilaity Studies, in R.J. Adams & M.L. Wu (eds.), *PISA 2000 Technical Report*. OECD, Paris.

Walberg, H.J. (1984), Improving the productivity of American schools, *Educational Leadership,* No. 41, pp. 19-27.

Walberg, H. (1986), Synthesis of research on teaching, in M. Wittrock (ed.), *Handbook of research on teaching* (pp. 214-229), New York: Macmillan.

Walker, M. (2006), *The choice of Likert or dichotomous items to measure attitudes across culturally distinct countries in international comparative educational research*. Paper presented at the Annual Meetings of the American Educational Research Association (AERA) in San Francisco, 7-11 April.

Walker, M. (2007), Ameliorating Culturally-Based Extreme Response Tendencies To Attitude items, *Journal of Applied Measurement,* No. 8, pp. 267-278.

Warm, T.A. (1989), Weighted Likelihood Estimation of Ability in Item Response Theory, *Psychometrika*, No. 54 (3), pp. 427-450.

Westat (2007), *WesVar® 5.1* Computer software and manual, Rockville, MD: Author (also see *http://www.westat.com/wesvar/*).

Wilson, M. (1994), Comparing Attitude Across Different Cultures: Two Quantitative Approaches to Construct Validity, in M. Wilson (ed.), *Objective measurement II: Theory into practice* (pp. 271-292), Norwood, NJ: Ablex.

Wolter, K.M. (2007), *Introduction to Variance Estimation*. Second edition, Springer: New York.

Wu, M.L., Adams, R.J. & **Wilson, M.R.** (1997), *ConQuest®: Multi-Aspect Test Software* [computer program manual], Camberwell, Vic.: Australian Council for Educational Research.

List of abbreviations – the following abbreviations are used in this report:

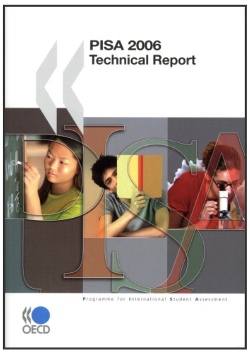
ACER	Australian Council for Educational Research
AGFI	Adjusted Goodness-of-Fit Index
BRR	Balanced Repeated Replication
CBAS	Computer Based Assessment of Science
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CITO	National Institute for Educational Measurement, The Netherlands
CIVED	Civic Education Study
DIF	Differential Item Functioning
ENR	Enrolment of 15-year-olds
ESCS	PISA Index of Economic, Social and Cultural Status
ETS	Educational Testing Service
IAEP	International Assessment of Educational Progress
I	Sampling Interval
ICR	Inter-Country Coder Reliability Study
ICT	Information Communication Technology
IEA	International Association for the Evaluation of Educational Achievement
INES	OECD Indicators of Education Systems
IRT	Item Response Theory
ISCED	International Standard Classification of Education
ISCO	International Standard Classification of Occupations
ISEI	International Socio-Economic Index
MENR	Enrolment for moderately small school
MOS	Measure of size
NCQM	National Centre Quality Monitor
NDP	National Desired Population
NEP	National Enrolled Population
NFI	Normed Fit Index
NIER	National Institute for Educational Research, Japan
NNFI	Non-Normed Fit Index
NPM	National Project Manager
OECD	Organisation for Economic Cooperation and Development
PISA	Programme for International Student Assessment
PPS	Probability Proportional to Size
PGB	PISA Governing Board
PQM	PISA Quality Monitor
PSU	Primary Sampling Units
QAS	Questionnaire Adaptations Spreadsheet
RMSEA	Root Mean Square Error of Approximation
RN	Random Number
SC	School Co-ordinator
SE	Standard Error
SD	Standard Deviation
SEM	Structural Equation Modelling
SMEG	Subject Matter Expert Group
SPT	Study Programme Table
TA	Test Administrator
TAG	Technical Advisory Group
TCS	Target Cluster Size
TIMSS	Third International Mathematics and Science Study
TIMSS-R	Third International Mathematics and Science Study – Repeat
VENR	Enrolment for very small schools
WLE	Weighted Likelihood Estimates

Table of contents

LIST OF BOXES

LIST OF FIGURES

LIST OF TABLES



From:

PISA 2006 Technical Report

Access the complete publication at:

https://doi.org/10.1787/9789264048096-en

Please cite this chapter as:

OECD (2009), “Coding and Marker Reliability Studies”, in *PISA 2006 Technical Report*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/9789264048096-14-en