13

Scaling Outcomes

INTERNATIONAL CHARACTERISTICS OF THE ITEM POOL

When main study data were received from each participating country, they were first verified and cleaned using the procedures outlined in Chapter 11. Files containing the achievement data were prepared and national-level Rasch and traditional test analyses were undertaken. The results of these analyses were included in the reports that were returned to each participant.

Table 13.1 ■ **Number of sampled students by country and booklet**

	Booklet														
	1	2	3	4	5	6	7	8	9	10	11	12	13	UH	Total
Australia	992	977	961	970	954	975	964	961	954	969	946	950	978		12 551
Austria	356	352	350	348	355	355	348	361	341	350	353	350	349	29	4 597
Belgium	660	658	649	644	645	647	660	661	656	665	657	667	680	247	8 796
Brazil	351	346	347	335	346	335	340	360	318	348	328	345	353		4 452
Canada	2 184	2 143	2 117	2 154	2 143	2 142	2 137	2 157	2 140	2 162	2 127	2 163	2 184		27 953
Czech Republic	466	464	468	472	483	482	497	478	487	469	473	472	481	128	6 320
Denmark	321	323	329	335	311	343	327	317	322	325	301	330	334		4 218
Finland	430	427	430	443	435	451	461	457	463	457	456	436	450		5 796
France	321	339	342	326	326	331	332	330	330	330	323	334	336		4 300
Germany	348	343	350	348	356	357	348	348	355	356	349	342	352	108	4 660
Greece	371	361	375	347	343	344	338	352	359	351	358	362	366		4 627
Hong Kong-China	349	348	339	344	349	348	345	346	331	351	339	344	345		4 478
Hungary	347	344	329	332	344	339	330	332	341	324	327	338	344	394	4 765
Iceland	254	258	256	261	263	259	259	265	261	251	255	254	254		3 350
Indonesia	817	817	826	830	815	820	810	831	842	848	846	834	825		10 761
Ireland	283	302	295	292	292	303	311	302	299	306	302	302	291		3 880
Italy	887	885	889	902	881	869	889	909	919	898	913	901	897		11 639
Japan	355	371	362	361	362	356	364	358	369	362	361	371	355		4 707
Korea	419	416	412	417	416	409	423	426	425	413	430	417	421		5 444
Latvia	359	363	357	356	360	358	357	358	354	348	353	358	346		4 627
Liechtenstein	28	26	23	27	25	26	23	25	23	27	25	27	27		332
Luxembourg	317	318	311	308	306	304	303	299	296	289	295	289	288		3 923
Macao-China	93	91	98	99	99	98	99	97	100	96	94	94	92		1 250
Mexico	2 321	2 304	2 327	2 319	2 318	2 330	2 294	2 308	2 296	2 293	2 298	2 272	2 303		29 983
Netherlands	299	315	308	288	299	306	296	309	296	297	298	299	298	84	3 992
New Zealand	339	352	347	335	338	338	342	342	343	347	353	376	359		4 511
Norway	310	309	314	320	314	322	310	308	299	310	321	316	311		4 064
Poland	349	342	339	330	318	325	340	342	331	345	331	356	335		4 383
Portugal	355	365	362	353	366	347	350	355	346	339	361	360	349		4 608
Russian Federation	461	469	472	473	470	449	455	455	439	461	456	462	452		5 974
Serbia	339	359	341	347	338	328	341	325	347	320	346	332	342		4 405
Slovak Republic	563	558	567	565	564	559	547	567	563	563	560	551	561	58	7 346
Spain	838	827	835	817	828	827	798	837	847	836	846	824	831		10 791
Sweden	368	368	372	369	364	362	350	354	333	342	344	344	354		4 624
Switzerland	634	665	652	646	648	649	649	663	646	639	647	629	653		8 420
Thailand	393	404	412	408	396	390	408	414	410	410	403	389	399		5 236
Tunisia	363	366	363	361	369	356	364	361	361	360	367	365	365		4 721
Turkey	383	391	379	364	378	381	380	375	370	365	366	363	360		4 855
United Kingdom	756	741	742	712	747	738	729	743	730	725	714	716	742		9 535
United States	418	425	420	408	431	427	407	418	421	437	424	404	416		5 456
Uruguay	462	467	455	457	439	450	455	447	446	442	435	441	439		5 835
Total	**21 259**	**21 299**	**21 222**	**21 123**	**21 134**	**21 135**	**21 080**	**21 253**	**21 109**	**21 126**	**21 081**	**21 079**	**21 217**	**1 048**	**276 165**

After processing at the national level, a set of international-level analyses was undertaken. Some involved summarising national analyses, while others required an analysis of the international data set.

The final international cognitive data set (that is, the data set of coded achievement booklet responses) (available as *intcogn.txt*) consisted of 276 165 students from 42 participating countries. Table 13.1 shows the total number of sampled students, broken down by participating country and test booklet.

Test targeting

Each of the domains was separately scaled to examine the targeting of the tests. Figures 13.1, 13.2, 13.3 and 13.4 show the match between the international item difficulty distribution and the international

Figure 13.1 ■ **Item plot for mathematics items**

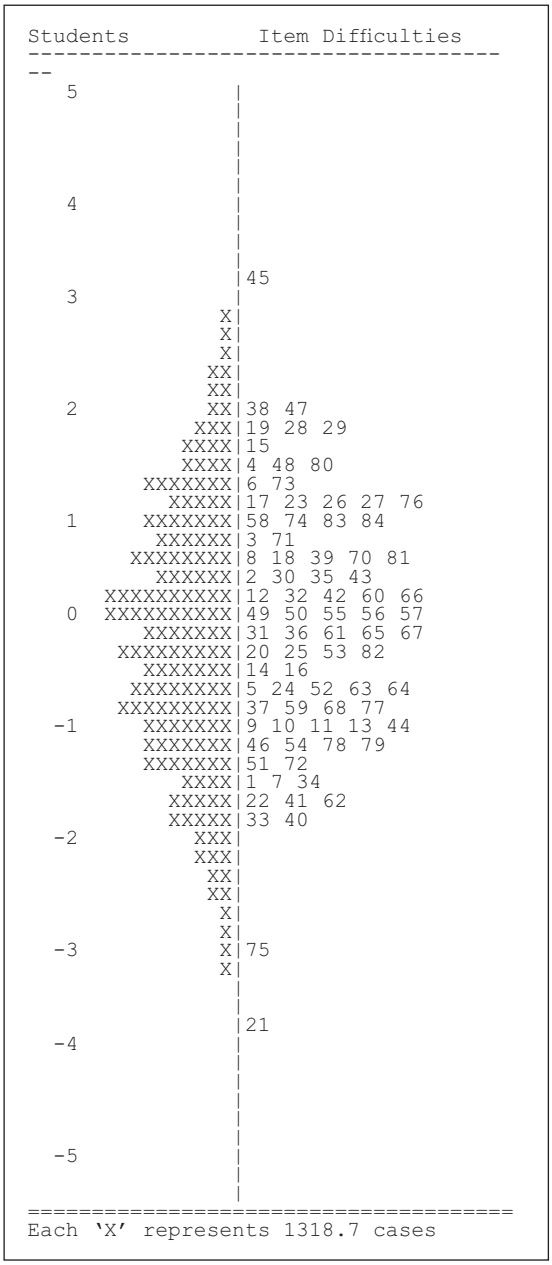
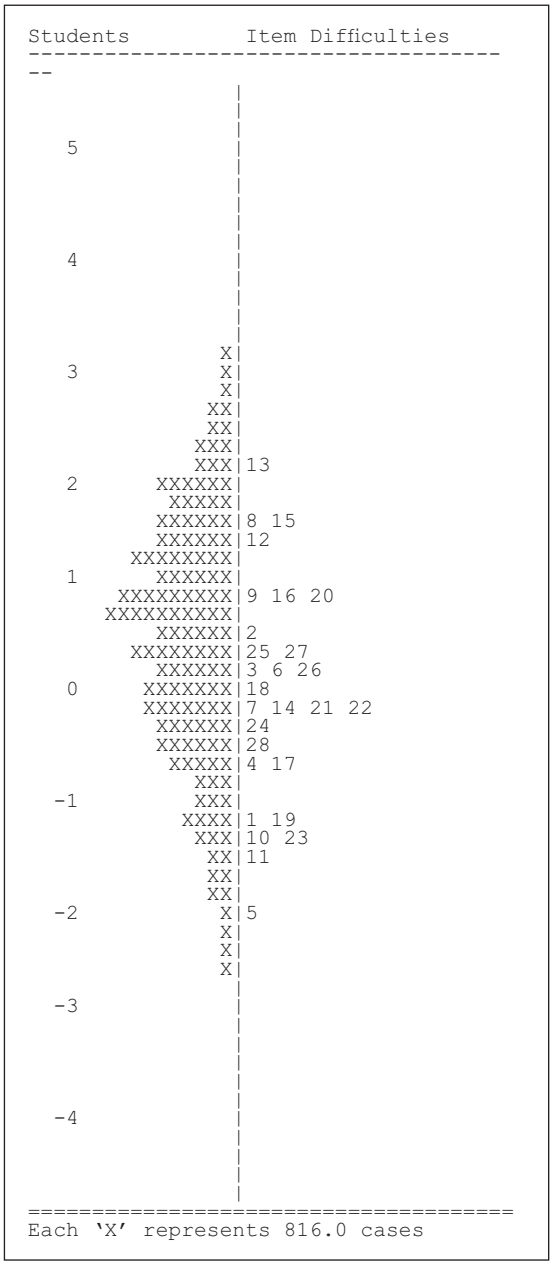


Figure 13.2 ■ **Item plot for reading items**



distribution of student achievement for each of mathematics, reading, science and problem solving, respectively. The figures consist of two panels. The left panel, students, shows the distribution of students' Rasch-scaled achievement estimates. Students at the top end of this distribution have higher achievement estimates than students at the lower end of the distribution. The right panel, item difficulties, shows the distribution of Rasch-estimated item difficulties.

In each of the figures, the student achievement distribution, shown by 'X', is well matched to the item difficulty distribution. The figures are constructed so that when a student and an item are located at the same height on the scale then the student has a 50 per cent chance of responding correctly to the item.

Figure 13.3 ■ **Item plot for science items**

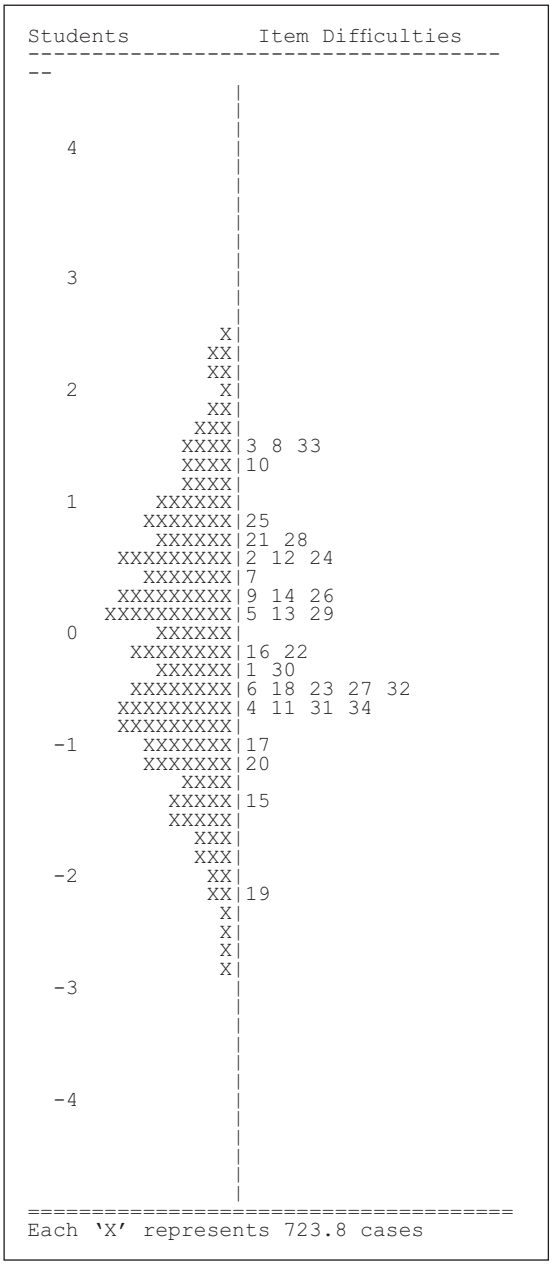
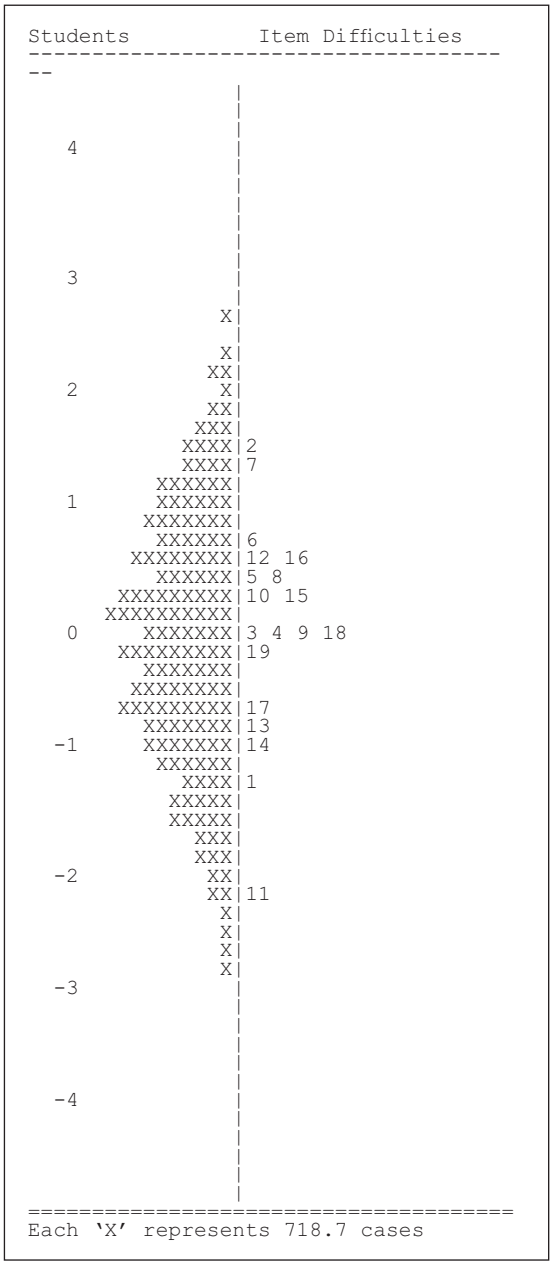


Figure 13.4 ■ **Item plot for problem-solving items**



Test reliability

A second test characteristic that is of importance is the test reliability. Table 13.2 shows the reliability for each of the four overall scales (mathematics, reading, science and problem solving) before conditioning and based upon four separate scalings. The international reliability for each domain after conditioning is reported later in Table 13.6. Appendix 11 shows the reliabilities for each country.

Table 13.2 ■ **Reliabilities of each of the four overall scales when scaled separately**

Domain	Reliability
Mathematics	0.845
Reading	0.799
Science	0.789
Problem solving	0.761

Domain intercorrelations

Correlations between the ability estimates for individual students in each of the four domains, the so-called latent correlations, as estimated by *ConQuest* (Wu *et al.,* 1997) are given in Table 13.3. It is important to note that these latent correlations are unbiased estimates of the true correlation between the underlying latent variables. As such they are not attenuated by the unreliability of the measures, and will generally be higher than the typical product moment correlations that have not been disattenuated for unreliability. The results in the table are reported for both OECD countries and for all participating countries.[1]

Table 13.3 ■ **Latent correlation between the four domains**

	Reading		Science		Problem solving	
	r	SE	r	SE	r	SE
Mathematics						
OECD countries	0.77	0.003	0.82	0.002	0.89	0.001
All participating countries	0.77	0.002	0.82	0.002	0.89	0.001
Reading						
OECD countries			0.83	0.002	0.82	0.002
All participating countries			0.82	0.001	0.82	0.002
Science						
OECD countries					0.79	0.002
All participating countries					0.78	0.002

Mathematics subscales

A seven-dimensional scaling was performed on the achievement data, consisting of:

- Scale 1: mathematics items – space and shape (M1)
- Scale 2: mathematics items – change and relationships (M2)
- Scale 3: mathematics items – uncertainty (M3)
- Scale 4: mathematics items – quantity (M4)
- Scale 5: problem solving items (PS)
- Scale 6: reading items (R)
- Scale 7: science items (S)

Table 13.4 ■ **Correlation between scales**

	M1		M2		M3		M4	
	r	SE	r	SE	r	SE	r	SE
M1								
OECD countries			0.89	0.001	0.88	0.001	0.89	0.001
All participating countries			0.90	0.001	0.89	0.001	0.90	0.001
M2								
OECD countries					0.92	0.001	0.92	0.001
All participating countries					0.92	0.001	0.93	0.001
M3								
OECD countries							0.90	0.001
All participating countries							0.90	0.001
Problem solving								
OECD countries	0.79	0.002	0.83	0.002	0.81	0.002	0.82	0.002
All participating countries	0.80	0.002	0.83	0.001	0.82	0.001	0.83	0.001
Reading								
OECD countries	0.67	0.003	0.73	0.002	0.73	0.002	0.73	0.002
All participating countries	0.68	0.003	0.74	0.002	0.74	0.002	0.73	0.002
Science								
OECD countries	0.73	0.002	0.77	0.002	0.77	0.002	0.76	0.002
All participating countries	0.74	0.002	0.77	0.002	0.78	0.002	0.76	0.002

SCALING OUTCOMES

The procedures for the national and international scaling are outlined in Chapter 9.

Item deletions

The items were first scaled by country and their fit was considered at the national level, as was the consistency of the item parameter estimates across countries. consortium staff then adjudicated items, considering the items' functioning both within and across countries in detail. Those items considered to be "dodgy" (see Chapter 9) were then reviewed in consultation with NPMs. The consultations resulted in the deletion of a few items at the national level and two items at the international level.

At the international level, the two deleted items were S327Q02 and M434Q01T. The nationally deleted items are listed in Table 13.5. All deleted items were recoded as not applicable and were not included in either the international scaling nor in generating plausible values.

Table 13.5 ■ **Items deleted at the national level**

Item	Country
M144Q03	Iceland (booklet 4 only)
M155Q01	Korea
M179Q01T	Italy (Italian version only)
M273Q01	Denmark (booklet 7 only)
M402Q02	Hungary
M442Q02	Uruguay
M603Q02	Canada
M704Q01T	Switzerland (Italian version only)
M800Q01	Uruguay
R055Q03	Austria, Luxembourg (German version only), Germany, Switzerland (German version only), Belgium (German version only), Italy (German version only), Liechtenstein
R102Q04a	Korea
R111Q6B	Tunisia
R219Q01E	Tunisia
R219Q01T	Tunisia
R227Q01	Spain (Catalonian and Castilian versions),
S131Q02T	Russia
S252Q02	Spain (Castilian, Galician, and Valencian versions)
S268Q02T	Norway
S326Q01	Portugal
X414Q01	Russia
X603Q02T	Italy (Italian version only)
X603Q03	Italy (Italian version only)

International scaling

The international scaling was performed on the calibration data set of 15 000 students (500 randomly selected students from each of the 30 OECD

countries). The item parameter estimates from this scaling are reported in Appendices 12, 13, 14 and 15. The item parameters were estimated using four separate one-dimensional models. As discussed later, a booklet facet was used in the item response model.

Generating student scale scores

Applying the conditioning approach described in Chapter 9 and anchoring all of the item parameters at the values obtained from the international scaling, plausible values were generated for all sampled students. Table 13.6 gives the reliabilities at the international level for the generated scale scores. The increase in reliability of the results reported in Table 13.6 over those presented in Table 13.2 is due to the use of multidimensional scaling and conditioning.

Table 13.6 ■ **Final reliability of the PISA scales**

Domain	Reliability
Mathematics (overall)	0.918
Space and shape	0.865
Change and relationships	0.905
Uncertainty	0.905
Quantity	0.895
Reading	0.848
Science	0.843
Problem solving	0.874

TEST LENGTH ANALYSIS

Table 13.7 shows the number of missing responses and the number of missing responses recoded as not reached,[2] by booklet. Table 13.9 shows this information by country.

The average number of not reached items differs from one country to another. It is worth noting that countries with higher averages of not-reached items also have higher averages of missing data. Table 13.8 provides the percentage distribution of not-reached items per booklet. The percentage of students who reached the last item ranges from 77 to 89 per cent (*i.e.* the percentages of students with zero not-reached items).

Table 13.7 ■ **Average number of not-reached items and missing items by booklet**

Booklet	Missing	Not reached
1	4.15	1.34
2	5.43	1.67
3	4.09	0.72
4	5.20	1.19
5	5.93	1.58
6	6.40	1.58
7	5.93	1.52
8	5.52	1.83
9	6.16	2.07
10	5.63	2.07
11	4.80	1.94
12	4.73	2.04
13	5.52	2.19
14	5.10	1.77
Total	**5.34**	**1.67**

Table 13.8 ■ **Percentage distribution of not-reached items by booklet**

Number of not-reached items	Booklet													
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
0	84.3	79.0	89.3	81.5	83.9	85.3	84.2	81.4	78.2	77.1	78.2	82.1	77.9	79.3
1	0.6	1.7	2.0	1.2	1.7	0.9	2.0	1.4	3.2	3.2	1.4	0.5	1.7	2.4
2	0.5	3.4	0.9	2.3	1.3	0.8	0.7	2.1	1.9	0.6	1.5	0.9	5.0	0.9
3	0.9	1.6	0.7	2.8	1.0	1.4	1.0	1.5	1.6	1.6	1.0	0.9	0.5	0.9
4	4.5	0.9	0.8	2.1	1.5	0.4	1.0	0.5	0.4	2.6	1.3	2.5	0.6	2.5
5	0.2	2.8	0.4	1.8	1.1	0.7	1.2	1.6	1.0	0.4	4.4	0.4	0.8	1.3
6	0.3	1.1	1.6	1.4	0.7	1.7	1.7	1.0	0.7	1.0	1.1	0.8	0.6	0.3
7	2.1	0.8	0.7	1.6	1.3	0.9	1.6	1.5	0.6	0.8	2.0	1.5	1.6	0.6
8	1.1	1.2	0.8	1.4	0.5	0.7	0.5	1.0	1.3	2.2	1.1	0.3	0.7	2.3
> 8	5.6	7.6	2.9	3.9	7.0	7.2	6.1	8.0	11.1	10.3	8.0	10.3	10.6	9.6

TIMING ISSUES

Timing issues are important for any testing sessions. A test that is too long (ie, contains too many items) will not only frustrate students, but also threaten the validity of the test because many students may rush to complete the test. A test that is too short (ie, too contains too few items) will result in disruptions at the end of the testing session because students will finish the test well before the end of the testing time.

The field trial incorporated procedures to collect some timing information with the consideration that such a collection should not disrupt the testing session, and should not cause undue burden for the students, test administrators and data entry staff.

As a result, five time points were included in each test booklet, requiring students to record the time as they reached these five points. The first time point was at the start of the test questions. The subsequent time points were at the end of each of the four blocks in the test booklet. An example of a timing point is shown below in Figure 13.2.

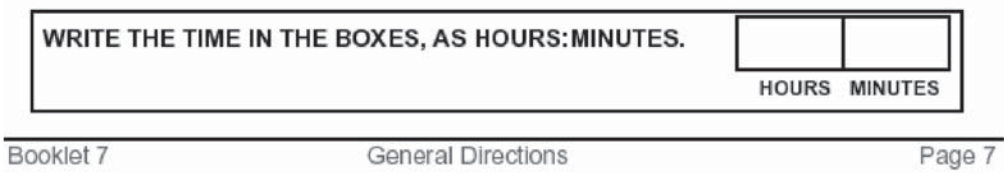
In subsequent sections of this document will denote the five time points as t_1, t_2, t_3, t_4, t_5.

Data collected as described above require a considerable amount of cleaning. Students may move around the test booklet in a somewhat random manner. A simple recording system of time points would not be able to capture such movements, and it was not the intention of this data collection to capture such movements. Consequently, records with any missing time points and any non-increasing time-points were removed from the data analyses. Only timing records with $t_1 < t_2 < t_3 < t_4 < t_5$ were used in the analyses.

Table 13.9. ■ **Average number of missing items and not-reached items by country**

Country	Missing	Not reached
Australia	3.79	0.91
Austria	5.10	0.33
Belgium	4.09	0.99
Brazil	9.32	5.23
Canada	3.04	0.76
Czech Republic	4.62	0.64
Denmark	6.11	1.36
Finland	2.85	0.76
France	5.10	1.32
Germany	5.31	0.62
Greece	8.53	2.45
Hong Kong-China	2.66	0.50
Hungary	5.32	1.43
Iceland	4.11	1.14
Indonesia	7.88	3.43
Ireland	3.30	0.52
Italy	6.07	1.41
Japan	5.24	1.08
Korea	3.46	0.42
Latvia	5.45	1.70
Liechtenstein	3.71	0.40
Luxembourg	6.16	0.89
Macao-China	3.50	1.30
Mexico	6.12	3.73
Netherlands	1.57	0.15
New Zealand	3.55	0.49
Norway	6.65	1.29
Poland	5.99	0.97
Portugal	6.05	1.57
Russian Federation	6.50	3.24
Serbia	11.24	1.47
Slovak Republic	6.10	1.11
Spain	5.76	1.54
Sweden	5.07	1.46
Switzerland	5.11	0.90
Thailand	5.48	2.19
Tunisia	8.81	4.21
Turkey	7.01	1.29
United Kingdom	3.78	0.40
United States	3.23	0.50
Uruguay	10.28	5.56

Figure 13.2 ■ **Example of a timing point**



Using these timing markers it is possible to define time intervals $D_{ij} = t_j - t_i$, so that D_{12}, D_{23}, D_{34}, D_{45} are the times taken to complete blocks, 1, 2, 3 and 4 respectively. D_{13} is the time taken to complete the first two blocks and D_{14} is the time taken to complete the first three blocks. Each of the D_{ij} was expressed in hours. In analysing the data, records were deleted for cases in which any of the individual block lengths had a value outside the range of 0 to 1, D_{13} was outside the range 0.5 to 2 and D_{14} was outside the range 0.8 to 2.2. While somewhat arbitrary this choice was made because each block had an expected length of approximately 30 minutes.

There were 28 770 students in the data file, and around 20 000 students had valid timing data. Table 13.10 gives the means and standard deviations of D_{12}, D_{23}, D_{34}, D_{45}, D_{13} and D_{14}, for each booklet.

Table 13.10 ■ **Means, the number of cases and standard deviations of duration time (in hours) by block and by field trial booklet**

Booklet		D12	D23	D34	D45	D13	D14
	Mean	0.5894	0.5469	0.4215	0.3520	1.171	1.553
1	N	1 969	1 958	1 783	1 385	1 958	1 722
	S.D.	0.1736	0.2034	0.1430	0.1313	0.3417	0.3093
	Mean	0.5351	0.5438	0.4719	0.4063	1.105	1.537
2	N	2 039	1 975	1 791	1 211	1 969	1 738
	S.D.	0.1671	0.1976	0.1577	0.1368	0.3340	0.3255
	Mean	0.6006	0.5493	0.4570	0.3496	1.184	1.587
3	N	1 805	1 775	1 553	1 094	1 772	1 480
	S.D.	0.1763	0.2012	0.1473	01392	0.3450	0.3089
	Mean	0.5734	0.5326	0.4819	0.3726	1.37	1.537
4	N	2 140	2 120	1 862	1 298	2 123	1 781
	S.D.	0.1755	0.2040	0.1514	0.1431	0.3494	0.3214
	Mean	0.5965	0.6007	0.4155	0.3122	1.239	1.604
5	N	2 051	2 009	1 819	1 301	2 039	1 753
	S.D.	0.1754	0.2000	0.1495	0.1314	0.3356	0.3022
	Mean	0.5205	0.4149	0.5766	0.3812	0.9631	1.513
6	N	2 125	2 104	1 956	1 428	2 057	1 879
	S.D.	0.1659	0.1685	0.1811	0.1422	0.2935	0.3032
	Mean	0.5308	0.5195	0.3874	0.3914	1.080	1.480
7	N	2 038	2 026	1 961	1 524	2 030	1 873
	S.D.	0.1734	0.11908	0.1385	0.1533	0.3278	0.3128
	Mean	0.5162	0.5813	0.4532	0.3675	1.143	1.560
8	N	1 972	1 895	1 779	1 259	1 918	1 710
	S.D.	0.1824	0.1984	0.1456	0.1493	0.3375	0.3120
	Mean	0.5687	0.4671	0.5076	0.3663	1.070	1.537
9	N	2 050	2 087	1 858	1 327	2 050	1 782
	S.D.	0.1707	0.1852	0.1596	0.1319	0.3302	0.3134
	Mean	0.5481	0.5105	0.4139	0.4201	1.093	1.485
10	N	1 976	1 972	1 858	1 373	1 958	1 787
	S.D.	0.1680	0.1927	0.1451	0.1472	0.3287	0.3162
	Mean	0.5574	0.5253	0.4571	0.3726	1.117	1.535
11	N	20 165	19 921	18 220	13 200	19 874	17 505
	S.D.	0.1754	0.2009	0.1615	0.1439	0.3402	0.3152

Table 13.10 needs to be matched to the test design so that the timing information can be related to the actual test clusters of the assessment material. Table 13.11 shows the PISA 2003 field trial test design.

Table 13.11 ■ **PISA 2003 field trial test design**

Booklet	Block 1 30 minutes	Block 2 30 minutes	Block 3 30 minutes	Block 4 30 minutes
1	M1	M11	S2	M2
2	M2	M12	M11	M3
3	M3	M13	M12	M4
4	M4	M14	M13	M5
5	M5	P1	M14	M6
6	M6	P2	P1	M7
7	M7	P3	P2	M8
8	M8	P4	P3	M9
9	M9	S1	P4	M10
10	M10	S2	S1	M1

For example, the column headed D_{12} in Table 13.10 gives the timing information for mathematics clusters 1 to 10 (M1 to M10), corresponding to the booklets 1 to 10. It can be seen that M8 is the shortest cluster, while M3 is the longest cluster among the first 10 mathematics clusters.

Students were given a break after one hour of testing. The duration of the break varied from country to country. The break was usually just a few minutes in length. The timing information in Table 13.10 includes this break in the computation, as no information is available about the length of the break in this data set. Consequently, the time duration D_{23} is likely to be a slight over-estimate of the actual time taken to complete this block. The over-estimate is probably about 0.05. The fact that the means for the first block are all greater than 0.5 suggests that D_{23} is likely to include this break time, and D_{34} is less likely to include this break. The means for D_{13} are mostly over one hour. That is, the majority of students took more than one hour to complete the first two blocks.

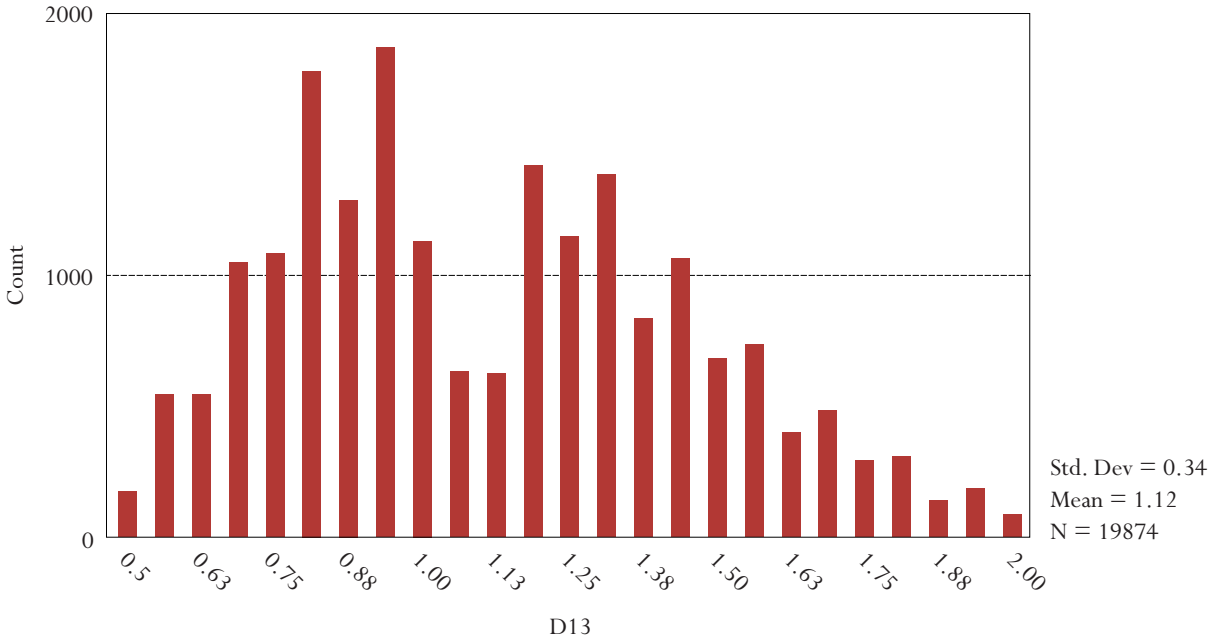
Despite the inclusion of the break time in the computation, there is a trend that the time duration taken to complete a block decreases as testing goes on. For example, the block 3 durations (D_{34}) are all much shorter than block 1 (D_{12}) durations. This observation is consistent with earlier findings that as testing goes on, students' motivation wanes and there are far more missing responses as well as guessing in the latter part of a test than in the earlier part of a test. This observation suggests that the most reliable timing information is the block 1 (D_{12}) information. So these should be used to compute average time taken per item, and not the information from blocks 2, 3 and 4. Certainly, block 4 information is the least reliable. As students run out of time at the end of the test, the block 4 timing cannot be regarded as time taken to complete this block. Nevertheless, blocks 2, 3 and 4 timing information is still useful for comparing the relative lengths of the clusters within these blocks.

D_{13} provides information about the time taken to complete half of the test. A histogram of D_{13} is shown in Figure 13.3.

The dip in the middle of Figure 13.3 is likely to be caused by the break after one hour of testing. More than half of the students needed more than one hour to complete the first half of the test. Ninety per cent of the students completed the first half of the test after 95 minutes from the start of the test. This suggests that the field trial blocks were, on average, too long (ie, contained too much material) even though 43 per cent completed the first half of the test in less than one hour.

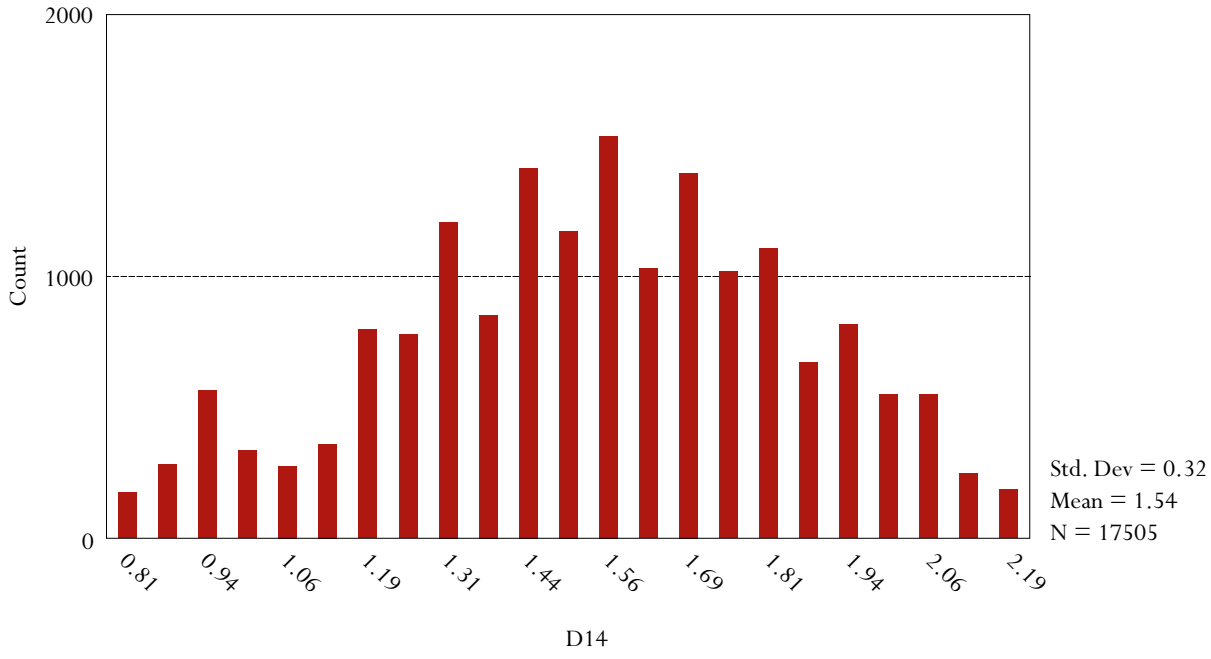
D_{14}, in Table 13.10, provides information about the time taken to complete the first three quarters of the test. A histogram of D_{14} is shown in Figure 13.4.

Figure 13.3 ■ **Histogram of time taken to complete half of the test**



About half of the students completed the first three quarters of the test one-and-a-half hours after the testing started. And around 92% of students completed the first three-quarters of the test 120 minutes

Figure 13.4 ■ **Histogram of time taken to complete the first three-quarters of the test**



after the testing started. While there is some evidence that the students caught up during the third block, the test still appears to be a little long. Besides, students might have caught up because they knew that time was running out, and started to skip questions or spent less time on some questions.

Average time per item

For the first block of the test, Table 13.12 shows the relationship between the number of items, number of words and the average time required to complete the cluster. The average amount of time per item is also reported.

Table 13.12 ■ **Time required in relation to the number of items and number of words**

	Number of items	Number of words	Time (minutes)	
			Per cluster	Per item
M1	16	1338	35.40	2.21
M2	15	1230	32.10	2.14
M3	18	1377	36.00	2.00
M4	17	1277	34.40	2.02
M5	14	1217	35.80	2.56
M6	15	1181	31.20	2.08
M7	13	1061	31.80	2.45
M8	14	1336	31.00	2.21
M9	17	1420	34.10	2.01
M10	17	1294	32.90	1.94
Average			**33.47**	**2.15**

Figure 13.5 shows the relationship between the time required and the number of items in the cluster.

Figure 13.5 ■ **Time required versus the number of items in the cluster**

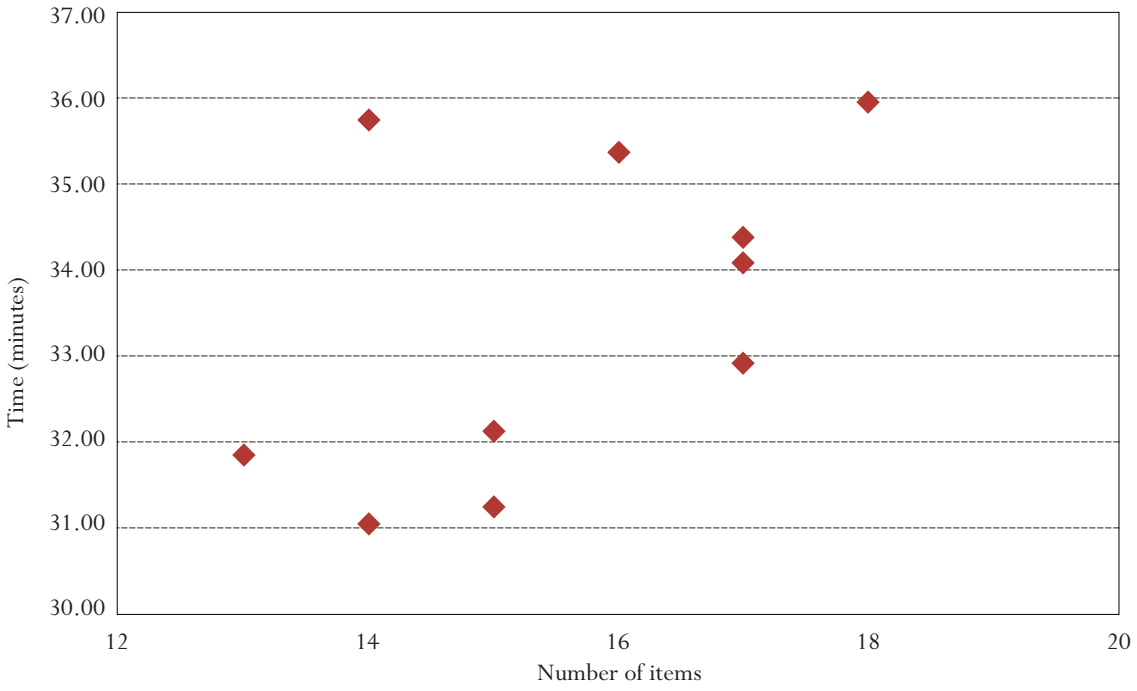


Figure 13.6 ■ **Time required versus the number of words in the cluster**

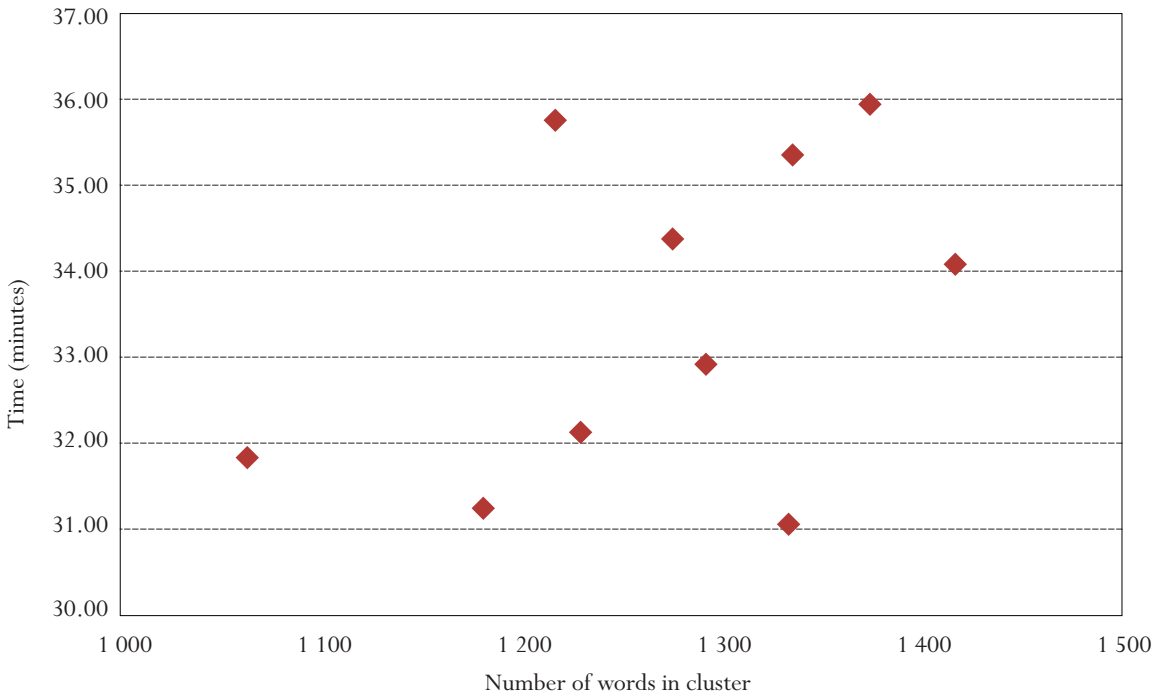


Figure 13.6 shows the relationship between the time required and the number of words in the cluster.

While there is a positive correlation between the time required and the number of items in a cluster, there are also some cases where a cluster with relatively few items took an above average amount of time. For example, a closer examination is required for cluster M5, where the average time per item is the highest. That is, there appear to be other factors (such as item format or item difficulty) that affect the length of time required to complete a cluster. Figure 13.6 shows that there is also a positive correlation between the time required and the number of words in a cluster.

While the average time per item is 2.15 minutes, it should be noted that using this estimate to fill a two-hour test would result in a test that approximately 50 per cent of the students would fail to finish. That is, the time required per item so that around 90 per cent of the students can finish the test within the two-hour time needs to be worked out. For block 1, 90 per cent of the students finished the cluster within 48 minutes. That is an average of 3.08 minutes per item. This is an estimate of the time taken per item when students are considered to be focused on the task. It does not take into account that students will work through the later clusters faster (skipping more items) because of fatigue and loss of motivation.

Number of not-reached items by booklet

At the booklet level, it is also important to monitor the number of not-reached items. The following gives the frequencies of not-reached items for each field trial booklet.

Table 13.13 ■ **Frequencies of not-reached items by field trial booklet**

Booklet	Mean	N	SD
1	3.39	2 879	7.357
2	4.79	2 872	8.609
3	4.94	2 889	8.647
4	3.93	2 891	6.493
5	4.22	2 886	7.616
6	2.57	2 834	4.823
7	1.78	2 860	4.414
8	3.56	2 869	6.600
9	3.58	2 886	6.661
10	2.81	2 874	6.575
97	0.00	11	0.000
99	0.00	19	0.000
Total	356	28 770	6.971

Table 13.13 shows that the average number of not-reached items was between 2 to 5 per booklet. Table 13.13 also suggests that booklets with all mathematics items have higher number of not-reached items. It could be the case that the mathematics clusters are longer on average than problem-solving and science clusters, or that there are more omissions for problem-solving and science items.

Based on the timing distribution for cluster 1, the expected number of not-reached items given a total number of items for a booklet could be computed. Table 13.14 gives the results. Based upon the data in this table a target cluster size of 12 items was adopted for the PISA 2003 main study.

Table 13.14 ■ **The expected number of not-reached items as a function of the total number of items in a booklet**[1]

Assumed time per item (minutes)	Number of items per 30-minute cluster	Number of items in a two-hour booklet	Expected number of not-reached items
1.93	15.6	62	4.4
2.0	15.0	60	3.7
2.1	14.3	57	2.9
2.2	13.6	55	2.3
2.3	13.0	52	1.8
2.4	12.5	50	1.4
2.5	12.0	48	1.1
2.6	11.5	46	0.8
2.7	11.1	44	0.6
2.8	10.7	43	0.5

1. This was the field trial number of items. There were, on average, 15.6 items for the first ten mathematics clusters in the field trial.

BOOKLET EFFECTS

Because the PISA 2003 test design was balanced, the item parameter estimates that are obtained from the scaling are not influenced by a booklet effect, as was the case in PISA 2000. But, due to the different location of domains within each of the booklets it was expected that there would still be booklet influences on the estimated proficiency distributions.

After scaling the PISA 2003 data for each country separately, achievement scores for mathematics, reading, problem solving and science could be compared across countries and across booklets. Tables 13.15, 13.16, 13.17 and 13.18 present student scale scores for the four domains, standardised to have a mean of 10 and

a standard deviation of 2 for each domain and country combination. The table rows represent countries (or sub-regions within countries) and the columns represent booklets. The purpose of these analyses and tables is to examine the nature of any booklet effects, therefore the countries are not named.

If Tables 13.15, 13.16, 13.17 and 13.18 are examined in conjunction with the test design (see Table 2.1 in Chapter 2) the explanation for the patterns in the booklet means is quite clear. From Table 13.15, it can be seen that the mathematics scores are systematically lower on booklets 8, 9, 10 and 11, those being the booklets that only have mathematics at the end. In Table 13.16, the reading scores are systematically lower on booklets 1, 2, 7 and 8. In Table 13.17, the science scores are systematically lower on booklets 5, 6, 12 and 13, and in Table 13.18, the problem-solving scores are systematically lower on booklets 3, 4, 9 and 10.

Table 13.15 ■ **Mathematics means for each country by booklet**

Booklet													
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**UH**
10.40	10.21	10.23	10.16	10.10	10.22	9.94	9.65	9.35	9.38	9.82	10.32	10.19	-
10.28	10.20	10.23	10.31	10.10	9.93	10.07	9.82	9.57	9.61	9.78	9.89	10.20	6.53
10.28	10.38	10.38	10.36	10.18	9.87	10.09	9.40	9.65	9.25	9.92	9.93	10.33	5.68
10.65	10.65	10.78	10.69	10.37	10.17	9.91	9.01	8.46	8.75	9.46	10.49	10.52	-
10.31	10.30	10.22	10.23	10.18	10.13	9.97	9.39	9.51	9.52	9.73	10.26	10.24	-
10.36	10.35	10.31	10.41	10.15	9.95	10.00	9.51	9.57	9.05	9.74	10.24	10.35	-
10.42	10.40	10.25	10.29	10.03	9.90	9.99	9.67	9.50	9.16	9.67	10.30	10.40	7.10
10.21	10.27	10.12	10.16	10.14	10.13	10.17	9.46	9.68	9.31	9.93	10.05	10.39	6.22
10.54	10.36	10.32	10.46	10.14	9.94	9.86	9.22	9.26	9.13	9.94	10.47	10.30	-
10.26	10.50	10.33	10.22	10.01	10.03	10.21	9.07	9.45	9.11	9.99	10.36	10.53	-
10.35	10.57	10.53	10.23	9.98	10.08	10.32	9.37	9.16	8.89	9.64	10.39	10.31	-
10.56	10.40	10.15	10.33	10.16	10.16	10.34	9.27	9.50	9.27	9.54	10.19	10.22	-
10.42	10.69	10.35	10.39	9.99	9.96	10.14	9.21	9.30	9.13	9.59	10.40	10.49	-
10.47	10.32	10.18	10.35	10.12	10.08	9.88	9.51	9.51	9.26	9.88	10.25	10.29	-
10.38	10.56	10.31	10.54	10.32	10.17	10.28	9.02	9.56	9.26	9.53	9.73	10.32	-
10.35	10.18	10.17	10.20	10.14	10.16	9.87	9.54	9.50	9.59	9.93	10.28	10.09	-
10.51	10.53	9.89	10.07	10.33	9.83	10.02	9.17	9.43	9.55	9.72	10.35	10.56	-
10.62	10.71	10.70	10.60	10.22	10.16	10.21	9.02	8.94	8.66	9.24	10.35	10.49	-
10.16	10.02	10.46	10.27	10.00	9.96	10.19	10.16	9.40	9.48	9.76	9.84	10.29	-
10.45	10.40	10.38	10.39	10.27	10.08	10.00	9.37	9.44	9.01	9.66	10.22	10.26	7.97
10.46	10.60	10.84	10.83	10.23	9.59	10.35	9.12	9.18	8.78	9.67	9.62	10.81	-
10.23	10.25	10.32	10.06	10.25	10.30	9.92	9.67	9.25	9.82	9.88	10.02	10.06	-
10.39	10.46	10.44	10.63	10.10	10.21	10.03	9.53	9.37	8.90	9.64	10.13	10.16	-
10.35	10.52	10.23	10.65	10.49	10.18	10.29	8.90	9.31	9.05	9.53	10.53	10.04	-
10.18	10.55	10.50	10.65	10.21	10.10	10.06	9.53	9.11	9.03	9.30	10.27	10.51	-
10.55	10.79	10.25	10.19	10.05	10.24	10.11	9.26	8.95	8.94	9.73	10.49	10.52	-
10.47	10.45	10.51	10.43	10.34	10.10	10.20	9.05	9.08	9.44	9.50	10.14	10.39	-
10.44	10.51	10.17	10.28	10.23	10.05	10.53	9.17	9.50	9.03	9.51	10.22	10.35	-
10.35	10.62	10.29	10.48	10.18	10.16	9.86	9.09	9.59	9.43	10.04	9.74	10.36	-
10.57	10.54	10.63	10.47	10.32	10.18	10.18	8.85	8.95	8.87	9.63	10.47	10.42	-
10.36	10.28	10.39	10.27	10.19	10.10	10.01	10.00	9.27	9.15	9.60	10.11	10.29	-
10.19	10.23	10.28	10.37	10.03	10.00	9.97	10.11	9.57	9.55	9.62	10.08	10.02	-
9.94	10.23	10.39	10.48	10.49	9.96	10.22	9.98	9.36	9.13	9.63	10.04	10.16	-
10.47	10.26	10.46	10.42	10.15	10.13	10.05	9.42	9.25	9.42	9.76	9.94	10.16	-
10.48	10.56	10.60	10.53	10.15	10.06	10.11	9.24	9.28	8.94	9.50	9.87	10.66	-
10.40	10.52	10.88	10.62	9.86	10.00	10.34	9.55	8.80	8.76	9.85	10.15	10.35	-
10.83	10.61	10.76	10.58	10.24	10.11	10.06	8.97	8.47	8.77	9.63	10.26	10.69	-
10.09	10.31	10.12	10.25	10.11	10.02	9.89	9.69	9.67	9.61	9.82	10.09	10.32	5.74
10.30	10.41	10.67	10.27	10.19	10.13	10.08	9.31	9.05	9.23	9.61	10.28	10.42	-
10.28	10.17	10.17	10.09	10.17	10.07	10.12	9.54	9.35	9.45	9.82	10.54	10.19	-
10.49	10.52	10.59	10.53	10.20	9.93	10.10	8.77	9.20	8.92	9.84	10.20	10.73	-
10.49	10.50	10.41	10.34	9.87	10.10	10.27	9.32	9.10	9.33	9.59	10.24	10.39	-
10.49	10.56	10.59	10.71	10.30	10.16	10.17	8.85	8.94	8.81	9.62	9.97	10.72	-
10.32	10.41	10.58	10.40	10.20	9.90	10.15	9.49	9.29	9.09	9.59	10.16	10.43	7.20
10.29	10.46	10.13	10.25	10.09	10.10	10.04	9.08	9.30	9.13	9.94	10.53	10.55	-
10.36	10.57	10.67	10.51	10.21	10.05	10.68	9.02	8.94	9.02	9.70	9.79	10.54	-
10.64	10.83	10.91	10.93	10.25	9.56	10.44	9.03	8.89	8.25	9.72	9.66	10.84	-
10.49	10.44	10.43	10.33	9.96	10.00	10.29	9.37	9.01	9.31	9.88	10.25	10.19	-
10.75	10.89	10.78	10.68	10.23	10.15	10.19	8.55	8.49	8.43	9.61	10.52	10.64	-
10.41	9.59	10.08	9.99	10.34	10.27	10.08	9.44	9.58	9.92	10.05	9.79	10.45	-
10.55	10.55	10.56	10.21	10.38	10.00	10.04	8.96	9.17	8.98	9.60	10.39	10.49	-

Table 13.16 ■ **Reading means for each country by booklet**

Booklet						
1	**2**	**7**	**8**	**9**	**10**	**11**
9.73	9.61	9.65	9.96	10.26	10.31	10.50
9.78	9.85	9.87	10.03	9.97	10.22	10.28
9.89	9.57	9.67	9.82	10.12	10.38	10.54
9.53	8.71	10.07	9.26	10.87	10.92	10.78
9.80	9.86	9.50	9.74	10.16	10.38	10.56
9.76	9.73	9.63	10.01	10.00	10.24	10.63
9.81	9.71	9.82	9.79	10.16	10.26	10.45
10.04	9.85	9.81	9.81	9.98	10.26	10.25
9.93	9.70	9.37	9.80	9.90	10.44	10.93
9.36	9.40	9.62	9.86	10.30	10.52	10.86
9.68	9.59	9.64	9.90	10.40	10.36	10.51
9.25	9.34	9.60	9.85	10.17	10.99	10.73
9.61	9.64	9.51	9.69	10.37	10.32	10.82
9.84	9.55	9.75	9.72	10.38	10.29	10.44
9.69	9.69	9.82	9.86	10.05	10.32	10.58
9.99	9.85	9.64	9.70	10.11	10.20	10.53
9.90	10.12	9.44	9.64	9.80	10.46	10.69
9.32	9.10	9.65	9.69	10.56	10.68	11.01
9.77	10.03	9.50	9.92	9.85	10.25	10.70
9.67	9.46	9.67	9.92	10.32	10.33	10.68
9.41	9.12	9.71	9.63	10.44	10.64	10.98
9.92	10.01	9.64	9.91	9.93	10.15	10.44
9.76	9.65	9.69	9.99	10.13	10.40	10.39
10.01	9.67	9.54	9.64	10.27	10.46	10.35
10.03	9.61	9.89	9.75	10.05	10.41	10.23
9.67	9.27	9.63	9.89	10.32	10.55	10.67
9.76	9.41	9.67	9.68	10.16	10.71	10.55
9.56	9.44	9.77	10.08	10.24	10.49	10.42
9.51	9.38	9.76	10.00	10.40	10.16	10.67
9.54	9.36	9.51	9.65	10.42	10.63	10.88
9.65	9.93	9.52	10.00	9.79	10.34	10.77
9.62	9.92	9.22	9.96	9.90	10.58	10.78
9.85	9.76	9.79	10.44	9.82	10.19	10.13
9.84	9.58	9.75	9.73	10.29	10.29	10.58
9.49	9.52	9.66	10.05	10.39	10.35	10.57
9.50	9.60	9.58	9.85	10.04	10.29	11.13
9.50	8.97	9.73	9.68	10.47	10.75	10.90
9.61	9.91	9.52	10.06	9.95	10.26	10.69
9.50	9.80	9.71	9.57	10.25	10.49	10.67
9.93	9.84	9.77	9.75	10.01	10.17	10.52
9.95	9.29	9.90	9.75	10.28	10.33	10.53
9.95	9.16	9.98	9.75	10.46	10.41	10.34
9.20	9.21	9.77	9.94	10.38	10.72	10.83
9.64	9.65	9.64	9.99	10.11	10.33	10.64
9.48	9.75	9.63	9.85	10.27	10.46	10.64
9.48	9.33	9.59	9.63	10.32	10.64	11.00
9.32	9.00	9.61	9.56	10.60	10.68	11.22
9.68	9.45	9.63	9.76	10.20	10.40	10.95
9.17	9.12	9.65	9.67	10.70	10.68	11.12
9.94	9.63	9.90	9.53	10.16	10.31	10.52
9.75	9.47	9.70	9.50	10.25	10.54	10.81

Table 13.17 ■ **Science means for each country by booklet**

Booklet						
5	**6**	**7**	**8**	**9**	**12**	**13**
9.87	9.96	10.26	10.30	10.41	9.50	9.70
9.77	10.18	9.94	10.25	10.54	9.38	9.95
10.11	9.93	10.26	10.31	10.41	9.29	9.71
9.95	10.05	10.55	10.73	10.82	8.77	9.18
9.95	10.07	10.16	10.28	10.26	9.41	9.88
9.92	10.08	10.09	10.35	10.41	9.34	9.81
9.80	10.04	10.04	10.22	10.58	9.32	9.99
9.89	10.14	9.93	10.22	10.40	9.34	10.06
10.13	9.67	10.33	10.31	10.20	9.73	9.67
9.68	10.33	10.23	10.47	10.83	8.85	9.57
9.88	10.30	10.24	10.30	10.71	8.83	9.78
10.02	9.84	10.09	10.39	11.23	8.94	9.51
9.99	9.94	10.31	10.45	10.71	9.24	9.33
9.81	10.14	10.19	10.29	10.52	9.21	9.78
10.27	9.80	10.48	10.37	10.32	9.47	9.31
9.96	9.96	10.06	10.21	10.48	9.60	9.73
10.38	9.64	10.19	9.98	10.08	9.86	9.87
9.59	10.25	10.23	10.79	11.13	8.72	9.34
9.77	10.16	9.89	10.11	10.44	9.55	10.10
9.76	10.52	9.86	10.41	10.89	8.91	9.65
10.15	9.77	10.80	10.63	10.69	9.11	8.87
9.85	10.05	10.03	10.41	10.43	9.61	9.61
9.83	10.20	10.16	10.50	10.63	8.91	9.73
9.79	10.26	10.21	10.43	10.72	8.98	9.60
9.47	10.61	10.31	9.91	10.77	9.18	9.77
9.53	10.13	10.18	10.63	10.82	8.93	9.77
9.81	10.30	10.07	10.47	10.71	9.00	9.73
9.90	10.02	10.19	10.61	10.71	8.96	9.58
9.82	10.25	9.97	10.39	10.66	8.85	9.95
9.58	10.35	10.27	10.49	10.95	8.95	9.38
9.69	10.45	9.71	10.32	10.93	8.94	9.97
9.84	10.09	10.06	10.33	10.41	9.56	9.69
9.82	9.92	10.35	10.78	10.24	9.26	9.75
9.90	10.12	10.03	10.28	10.64	9.27	9.73
9.71	10.28	10.10	10.60	10.73	8.98	9.60
9.67	10.47	10.17	10.47	10.78	9.05	9.31
9.61	10.52	10.16	10.69	11.25	8.44	9.30
10.11	9.84	10.08	10.22	10.34	9.66	9.74
9.94	10.10	10.31	10.19	10.58	9.33	9.59
9.93	9.90	10.28	10.07	10.41	9.60	9.86
9.89	10.22	9.93	10.35	10.72	9.31	9.62
9.85	10.00	10.33	10.45	10.49	9.28	9.63
9.79	10.33	10.29	10.62	10.83	8.77	9.41
9.76	10.16	10.04	10.46	10.66	9.08	9.82
9.98	9.99	10.27	10.33	10.47	9.41	9.56
9.98	9.80	10.50	10.40	10.86	9.20	9.19
9.97	9.67	10.73	10.81	10.87	8.73	9.23
9.97	10.00	10.28	10.38	10.60	9.21	9.52
9.91	10.04	10.57	10.79	11.02	8.72	8.91
9.89	10.07	10.16	10.30	10.53	9.51	9.53
9.90	10.29	10.12	10.37	10.80	8.96	9.54

Table 13.18 ■ **Problem-solving means for each country by booklet**

Booklet						
3	**4**	**9**	**10**	**11**	**12**	**13**
9.68	9.65	9.86	9.98	10.15	10.17	10.50
9.92	9.96	9.81	9.99	10.09	10.07	10.15
9.77	9.59	9.88	10.02	10.30	10.11	10.31
9.74	8.81	10.06	9.61	10.73	10.53	10.51
9.66	9.75	9.81	10.05	9.68	10.37	10.66
9.68	9.75	9.68	10.04	10.02	10.25	10.57
9.67	9.73	9.83	9.90	10.16	10.25	10.44
9.72	9.91	9.76	9.93	10.16	10.21	10.31
9.73	9.60	9.86	9.90	10.23	10.37	10.34
9.52	9.37	9.88	9.99	10.32	10.45	10.47
9.79	9.27	10.20	9.86	9.96	10.32	10.60
9.38	9.34	9.92	10.35	10.09	10.20	10.60
9.55	9.21	10.06	10.00	10.33	10.38	10.51
9.77	9.35	9.94	10.11	10.24	10.14	10.43
9.73	9.60	9.89	9.88	10.06	10.29	10.54
9.69	9.92	9.77	10.11	10.06	10.13	10.32
9.32	9.67	9.65	10.16	10.14	10.25	10.84
9.50	9.00	9.78	9.68	10.45	10.74	10.80
10.24	9.92	9.87	9.60	10.26	10.01	10.11
9.74	9.22	10.01	9.89	10.37	10.51	10.25
9.94	8.78	10.08	9.49	10.74	10.72	10.25
9.96	9.29	10.01	9.93	10.30	10.26	10.23
9.62	9.44	10.00	9.88	10.08	10.47	10.52
9.32	9.47	9.72	10.05	10.35	10.67	10.44
9.13	9.42	9.73	10.27	10.12	10.37	10.95
9.34	9.09	9.90	9.91	10.41	10.72	10.62
9.59	9.31	9.89	9.89	10.35	10.34	10.59
9.32	9.22	10.10	10.35	10.23	10.38	10.45
9.38	10.09	10.03	10.06	10.49	9.85	10.08
9.77	9.07	9.84	9.68	10.52	10.45	10.63
10.08	9.67	9.90	9.49	10.25	10.24	10.38
10.11	9.65	10.05	9.76	10.07	10.19	10.16
10.12	9.97	9.43	10.05	9.90	10.02	10.44
9.80	9.70	9.74	9.92	10.14	10.17	10.56
9.79	9.17	10.01	9.60	10.41	10.41	10.62
10.29	9.24	9.82	9.34	10.53	10.54	10.30
9.53	8.80	9.99	9.54	10.90	10.63	10.64
9.53	9.75	9.65	10.10	10.15	10.23	10.58
10.01	9.30	9.84	9.72	10.24	10.34	10.55
9.61	9.85	9.72	9.95	9.97	10.11	10.75
9.70	9.56	9.95	9.75	10.16	10.37	10.50
9.59	9.17	9.94	9.92	10.26	10.37	10.74
9.33	9.00	9.97	9.88	10.53	10.50	10.86
9.97	9.38	9.92	9.78	10.27	10.34	10.36
9.53	9.32	9.89	9.97	10.36	10.48	10.52
10.00	9.09	10.13	9.27	10.83	10.68	10.05
9.41	9.02	9.67	9.54	10.49	10.86	10.98
9.90	9.43	9.65	9.90	10.41	10.30	10.43
9.43	8.67	9.81	9.67	10.65	10.87	10.99
9.67	9.69	9.77	10.21	10.04	10.17	10.44
9.83	9.27	9.82	9.69	10.31	10.57	10.52

Correction of the booklet effect

Modelling the effect

Modelling the order effects in terms of item positions in a booklet or at least in terms of cluster positions in a booklet would result in a very complex model. For the sake of simplicity in the international scaling, the effect, as in PISA 2000, was modelled at the booklet level, separately for each domain.

When estimating the item parameters, booklet effects were included in the measurement model to prevent confounding item difficulties and booklet effects. For the *ConQuest* model statement, the calibration model was: item + item*step + booklet.

The booklet parameter, formally defined in the same way as item parameters, reflects booklet difficulty.[3]

Estimating the parameters

The calibration model given above was used to estimate the international item parameters. It was estimated using the international calibration sample of 15 000 students, and not-reached items in the estimation were treated as not administered.

The booklet parameters obtained from this analysis were not used to correct for the booklet effect. Instead, a set of booklet parameters was obtained by scaling the entire data set of OECD countries using booklet as a conditioning variable and a senate weight. The students who responded to the UH booklet were excluded from the estimation. The booklet parameter estimates obtained are reported in Table 13.19. The booklet effects are the amount that must be added to the proficiencies of student who responded to each booklet. That is a positive value indicates a booklet that was harder than the average while a negative value indicates a booklet that was easier than the average. Since the booklet effects are deviations from an average they sum to zero for each domain.

Table 13.20 shows the booklet effects after transformation to the PISA scales.

Table 13.19 ■ **Estimated booklet effects in logits**

	Domain			
Booklet	**Mathematics**	**Reading**	**Science**	**Problem solving**
1	–0.24	0.18		
2	-0.22	0.24		
3	-0.21			0.16
4	-0.20			0.27
5	-0.09		0.07	
6	-0.05		-0.06	
7	-0.04	0.20	-0.09	
8	0.36	0.11	-0.20	
9	0.41	-0.12	-0.33	0.07
10	0.46	-0.23		0.06
11	0.15	-0.38		-0.13
12	-0.13		0.41	-0.17
13	-0.21		0.19	-0.26

Table 13.20 ■ **Estimated booklet effects on the PISA scale**

	Domain			
Booklet	**Mathematics**	**Reading**	**Science**	**Problem solving**
1	-18.5	14.0		
2	-17.1	19.3		
3	-16.4			13.5
4	-15.5			23.2
5	-6.8		6.4	
6	-3.7		-5.5	
7	-2.8	16.1	-7.8	
8	27.9	8.7	-18.0	
9	31.5	-9.4	-29.7	6.0
10	35.7	-18.1		4.9
11	12.0	-30.6		-11.2
12	-10.2		37.2	-14.5
13	-16.1		17.3	-21.9

Applying the correction

To correct the student scores for the booklet effects, two alternatives were considered:

- Correcting all students' scores using one set of the internationally estimated booklet parameters; or
- Correcting the students' scores using nationally estimated booklet parameters for each country.

When choosing between these two alternatives a number of issues were considered. First, it is important to recognise that the sum of the booklet correction values is zero for each domain, so the application of either of the above corrections does not change the country means or rankings. Second, if a national correction was applied then the mean, within a country, will be the same for each booklet. As such, this approach would incorrectly remove a component of expected sampling and measurement error variation. Third, the booklet corrections are essentially an additional set of item parameters that capture the effect of the item locations in the booklets.

In PISA all item parameters are treated as international values so that all countries are therefore treated in exactly the same way. Perhaps the following scenario best illustrates the justification for this. Suppose students in a particular country found the reading items on a particular booklet surprisingly difficult, even though those items have been deemed as central to the PISA definition of PISA reading literacy and have no technical flaws, such as a translation or coding error. If a national correction were used, then an adjustment would be made to compensate for the greater difficulty of these items in that particular country. The outcome would be that two students from two different countries who responded in the same way to these items would be given different proficiency estimates. This differential treatment of students based upon their country has not been deemed as suitable in PISA. Moreover, this form of adjustment would have the effect of masking real underlying differences in literacy between students in those two countries, as indicated by those items.

Applying an international correction was therefore deemed the most desirable option from the perspective of cross-national consistency.

Table 13.21 ■ Standard deviations of mean scores across booklets

	Mathematics		Reading		Science		Problem solving	
	SD of booklet means	SE of PISA mean	SD of booklet means	SE of PISA mean	SD of booklet means	SE of PISA mean	SD of booklet means	SE of PISA mean
Australia	6.24	2.15	4.71	2.13	7.09	2.10	5.59	1.98
Austria	10.85	3.27	11.31	3.76	9.05	3.44	12.38	3.18
Belgium	6.52	2.29	4.98	2.58	8.24	2.48	5.44	2.20
Brazil	23.00	4.83	36.29	4.58	17.08	4.35	19.67	4.84
Canada	6.22	1.82	7.03	1.75	9.04	2.02	11.95	1.74
Czech Republic	5.63	3.55	5.24	3.46	7.56	3.38	5.97	3.42
Denmark	7.22	2.74	10.02	2.82	16.49	2.97	4.27	2.54
Finland	5.67	1.87	9.30	1.64	5.18	1.92	5.82	1.86
France	11.00	2.50	5.14	2.68	18.17	2.99	4.13	2.67
Germany	7.93	3.32	11.49	3.39	10.34	3.64	9.14	3.24
Greece	17.64	3.90	22.50	4.10	22.20	3.82	19.45	3.97
Hong Kong-China	13.59	4.54	12.26	3.69	14.21	4.26	18.25	4.18
Hungary	4.32	2.84	3.85	2.47	13.68	2.77	8.90	2.86
Iceland	7.31	1.42	6.00	1.56	7.87	1.47	4.82	1.38
Indonesia	15.21	3.91	8.90	3.38	15.02	3.21	15.59	3.29
Ireland	12.32	2.45	11.81	2.63	9.01	2.69	8.48	2.34
Italy	10.63	3.08	9.09	3.04	13.96	3.13	11.20	3.10
Japan	9.18	4.02	12.63	3.92	20.30	4.14	12.43	4.05
Korea	12.54	3.24	11.94	3.09	7.96	3.54	11.55	3.06
Latvia	11.37	3.69	6.60	3.67	5.27	3.89	11.54	3.90
Liechtenstein	12.87	4.12	18.31	3.58	12.89	4.33	17.16	3.95
Luxembourg	5.80	0.97	4.34	1.48	3.64	1.50	6.14	1.37
Macao-China	13.87	2.89	6.11	2.16	12.85	3.03	16.46	2.53
Mexico	15.43	3.64	13.21	4.09	19.03	3.49	19.65	4.30
Netherlands	10.34	3.13	9.78	2.85	12.42	3.15	8.48	2.95
New Zealand	7.75	2.26	8.25	2.46	11.55	2.35	11.84	2.17
Norway	7.04	2.38	9.75	2.78	6.05	2.87	9.12	2.60
Poland	12.32	2.50	9.24	2.88	7.50	2.86	3.60	2.78
Portugal	6.73	3.40	13.42	3.73	5.04	3.46	9.37	3.87
Russian Federation	15.63	4.20	17.15	3.94	16.61	4.14	19.88	4.59
Serbia	9.22	3.75	6.61	3.56	5.24	3.50	7.59	3.32
Slovak Republic	5.89	3.35	5.24	3.12	6.48	3.71	5.52	3.38
Spain	6.01	2.41	7.95	2.60	11.08	2.61	5.74	2.73
Sweden	9.18	2.56	6.25	2.42	7.33	2.72	6.28	2.44
Switzerland	4.68	3.38	8.30	3.28	7.64	3.69	7.53	3.05
Thailand	12.70	3.00	8.62	2.81	11.62	2.70	21.10	2.72
Tunisia	21.83	2.54	23.19	2.81	20.05	2.56	16.78	2.11
Turkey	10.13	6.74	5.99	5.79	5.46	5.89	8.38	6.03
United Kingdom	9.36	2.43	8.11	2.46	10.58	2.52	11.63	2.38
United States	17.58	2.95	8.65	3.22	7.58	3.08	8.07	3.13
Uruguay	31.35	3.29	34.75	3.43	33.51	2.90	35.52	3.68

Remaining booklet effects

The choice of a common correction does, however, leave deviations from equal booklet means in the data and these deviations vary over countries. These deviations occur because of sampling error, measurement error and any remaining item- or booklet-by-country interactions in the data. The results in Appendix 3 show the mean for each country by booklet after the international correction has been implemented. The annexes also show the country ranks that would have resulted using each booklet.

In Table 3.21, the results in the appendix are summarised by showing the standard deviation of the means across booklets. As a point of comparison the standard error of the PISA mean is also shown.

Under the assumption that the scaling model is correct, all of the variation between the booklet means should be explainable through sampling and measurement error. While there is variation across countries and booklet in the standard errors of the booklet means, they are typically about two to three times the size of the standard error of the PISA mean. It follows that where the standard deviations of the booklet means exceed the standard error of the PISA means by a factor of about three, there are remaining item- or booklet-by-country interactions in the data. The observation of these booklet variations is an important outcome of PISA that should not be neglected when analysing, reporting and interpreting PISA results.

Imputing data for students who did not respond to a domain

The PISA conditioning variables are prepared using procedures based on those used in the United States National Assessment of Educational Progress (Beaton, 1987) and in TIMSS (Macaskill *et al.,* 1998). The steps involved in this process are as follows:

- *Step 1.* Five variables (booklet ID, gender, mother's occupation, father's occupations and school mean mathematics score) were prepared to be directly used as conditioning variables. The booklet ID was dummy coded so that booklet 9 was used as the reference booklet. Booklet 9 had to be chosen as the reference booklet because it is the only booklet that contains items from all four assessment domains. For mother's and father's occupation the international socio-economic index of occupational status (ISEI) was used. For each student the mean mathematics achievement for that student's school was estimated using the mean of the weighted likelihood estimates for mathematics for each of the students who also attended that student's school.
- *Step 2.* Each variable in the Student Questionnaire was dummy coded. The details of this dummy coding are provided in Appendix 10.
- *Step 3.* For each country, a principal components analysis of the dummy-coded variables was performed, and component scores were produced for each student (a sufficient number of components to account for 95 per cent of the variance in the original variables).
- *Step 4.* The item-response model was fitted to each national data set and the national population parameters were estimated using item parameters anchored at their international location, and conditioning variables derived from the national principal components analysis and from *Step 1*.
- *Step 5.* Five vectors of plausible values were drawn using the method described in Chapter 9. The vectors were of length seven, one for each of the PISA 2003 reporting scales.

In PISA 2000 the plausible values for those students who did not respond to any items from a domain were removed from the database and a set of weight adjustments was provided for dealing with the smaller data

set. The assumption under this approach is that the students who did not get domain scores were missing at random. For PISA 2003 the plausible values for all domains have been retained for all students. This approach has a number of advantages. First, the database structure is simpler and analysis is simpler because the use of a weight adjustment is not necessary. Second, the missing at random assumption is loosened somewhat. The plausible value generation assumes that the relationships between the domain for which no items are observed and all other variables (both conditioning variables and the other domains) is the same for both the students who did respond to items from a domain and those who did not. Using all of this relationship information, and all available information about the student, an imputation is made. Because of the amount of data that is available to make the imputation, the analysis of the full data set will produce more accurate results than will analysis of the data set that omits students who did not respond to a domain. Additionally, it can be expected that, due to sampling variation, the characteristics of the students who did not respond to a domain will be slightly different from the characteristics of those who did, these differences will be appropriately adjusted for in the imputation and the estimated characteristics of, for example, the reading proficiency distribution for all students will be slightly different from the estimated characteristics of the reading proficiency distribution for the subset of students who responded to reading items.

The one disadvantage of this approach is that the average performances on the reference booklet (booklet 9) will influence the imputations for students who did not respond to items from a domain. As noted above, booklet- and item-by-country interactions do result in variations across booklets in the country means. If a country has an unusually low or high performance on the reference booklet, for a particular domain, then this unusual performance will influence the imputations for all students that did not respond to that domain. The consequential effect is that the reference booklet will be given more weight than the other booklets in the assessment of national means.

Tables 13.22, 13.23 and 13.24 show the mean and standard errors of the mean for each country using all students in the database, and using the subset of students who responded to items in each domain for reading, science and problem solving. The tables also show the difference between the mean of all students and the mean of the assessed students and the ratio of the error variances for the two estimates of the mean.

For the majority of the cases the variance ratio is less than one. This indicates that the error variances associated with the estimate of the mean for all students is less than that for the assessed students. It is important to realise that this is not an artificial result that is merely due to an increase in sample size, but is a genuine reduction in the error caused by the increase in the total available information about the proficiency distribution.

For a number of countries the difference between the means is reasonably large. In the case of reading, amongst OECD countries the difference is significant for Denmark. For science the differences are significant for the following OECD countries: Canada, Denmark, Greece and Mexico. For problem solving, none of the differences are significant for OECD countries.

Table 13.22 ■ Comparison of reading means for all students and reading-assessed students

	All students		Reading-assessed students only		Difference	Ratio of error variance
	Mean	SE of mean	Mean	SE of mean	(All - Assessed)	(All/Assessed)
Australia	525	2.1	526	2.1	-0.8	1.01
Austria	491	3.8	499	4.0	-7.8	0.89
Belgium	507	2.6	508	2.9	-1.3	0.78
Brazil	403	4.6	383	5.3	19.6	0.76
Canada	528	1.7	529	1.9	-1.3	0.87
Czech Republic	489	3.5	490	3.6	-1.6	0.93
Denmark	492	2.8	502	3.2	-9.3	0.79
Finland	543	1.6	541	2.1	2.6	0.62
France	496	2.7	499	3.1	-2.5	0.73
Germany	491	3.4	493	3.7	-2.1	0.86
Greece	472	4.1	460	4.4	12.4	0.87
Hong Kong-China	510	3.7	517	4.4	-7.1	0.72
Hungary	482	2.5	480	2.9	1.5	0.73
Iceland	492	1.6	492	2.1	-0.3	0.53
Indonesia	382	3.4	379	3.5	3.1	0.94
Ireland	515	2.6	521	3.1	-5.6	0.72
Italy	476	3.0	471	3.4	4.2	0.82
Japan	498	3.9	507	3.9	-9.0	0.99
Korea	534	3.1	540	3.3	-5.8	0.89
Latvia	491	3.7	486	3.9	4.5	0.89
Liechtenstein	525	3.6	528	5.8	-3.3	0.38
Luxembourg	479	1.5	479	1.9	0.5	0.61
Macao-China	498	2.2	500	3.3	-2.2	0.43
Mexico	400	4.1	394	4.5	5.5	0.82
Netherlands	513	2.9	517	3.0	-3.6	0.91
New Zealand	522	2.5	524	2.8	-2.0	0.75
Norway	500	2.8	495	3.1	4.3	0.81
Poland	497	2.9	494	3.1	2.8	0.87
Portugal	478	3.7	473	3.9	4.7	0.93
Russian Federation	442	3.9	439	4.4	3.4	0.79
Serbia	412	3.6	412	3.8	0.0	0.88
Slovak Republic	469	3.1	470	3.3	-0.9	0.89
Spain	481	2.6	478	3.1	2.3	0.71
Sweden	514	2.4	515	2.9	-0.7	0.70
Switzerland	499	3.3	504	3.8	-5.1	0.75
Thailand	420	2.8	419	2.8	1.0	1.02
Tunisia	375	2.8	367	2.9	7.6	0.93
Turkey	441	5.8	439	6.0	2.0	0.92
United Kingdom	507	2.5	508	2.9	-1.2	0.74
United States	495	3.2	495	3.7	-0.3	0.75
Uruguay	434	3.4	417	3.9	17.2	0.77

Table 13.23 ■ Comparison of science means for all students and science assessed students

	All students		Science assessed students only		Difference	Ratio of error variance
	Mean	SE of mean	Mean	SE of mean	(All – Assessed)	(Assessed/All)
Australia	525	2.1	531	2.3	-5.8	0.80
Austria	491	3.4	495	3.4	-3.8	1.00
Belgium	509	2.5	512	2.6	-3.5	0.90
Brazil	390	4.3	386	4.4	3.2	0.98
Canada	519	2.0	528	2.3	-9.3	0.76
Czech Republic	523	3.4	523	3.9	0.0	0.74
Denmark	475	3.0	486	3.1	-11.1	0.94
Finland	548	1.9	551	2.3	-2.5	0.72
France	511	3.0	516	2.9	-4.8	1.06
Germany	502	3.6	506	3.7	-3.4	0.95
Greece	481	3.8	465	3.9	16.0	0.94
Hong Kong-China	539	4.3	545	4.5	-5.1	0.90
Hungary	503	2.8	498	3.0	5.0	0.85
Iceland	495	1.5	493	2.1	2.0	0.49
Indonesia	395	3.2	398	3.5	-2.7	0.83
Ireland	505	2.7	511	3.0	-5.3	0.79
Italy	486	3.1	480	3.4	6.9	0.83
Japan	548	4.1	536	4.6	11.2	0.82
Korea	538	3.5	541	3.8	-2.4	0.87
Latvia	489	3.9	487	4.5	1.7	0.74
Liechtenstein	525	4.3	532	6.9	-6.5	0.39
Luxembourg	483	1.5	482	2.0	1.1	0.56
Macao-China	525	3.0	517	4.1	7.7	0.56
Mexico	405	3.5	393	3.9	12.2	0.80
Netherlands	524	3.1	529	3.5	-4.2	0.81
New Zealand	521	2.4	525	2.5	-3.7	0.86
Norway	484	2.9	483	3.6	1.3	0.62
Poland	498	2.9	493	3.5	4.3	0.68
Portugal	468	3.5	472	3.8	-4.4	0.85
Russian Federation	489	4.1	485	4.4	4.3	0.87
Serbia	436	3.5	434	3.5	2.0	0.97
Slovak Republic	495	3.7	493	4.1	2.2	0.83
Spain	487	2.6	480	2.9	6.8	0.79
Sweden	506	2.7	510	2.8	-3.8	0.97
Switzerland	513	3.7	517	3.9	-3.9	0.90
Thailand	429	2.7	425	3.0	4.2	0.79
Tunisia	385	2.6	380	2.7	4.5	0.89
Turkey	434	5.9	433	6.0	1.5	0.98
United Kingdom	518	2.5	523	2.8	-4.9	0.80
United States	491	3.1	494	3.5	-3.0	0.78
Uruguay	438	2.9	422	3.1	16.6	0.88

Table 13.24 ■ Comparison of problem solving means for all students and problem solving assessed students

	All students		Problem solving assessed students only		Difference	Ratio of error variance
	Mean	SE of mean	Mean	SE of mean	(All - Assessed)	(Assessed/All)
Australia	530	2.0	533	2.2	-3.0	0.84
Austria	506	3.2	508	3.4	-1.8	0.86
Belgium	525	2.2	525	2.3	-0.1	0.94
Brazil	371	4.8	369	5.0	2.2	0.94
Canada	529	1.7	530	2.0	-0.8	0.78
Czech Republic	516	3.4	518	3.6	-1.8	0.91
Denmark	517	2.5	520	2.6	-3.1	0.97
Finland	548	1.9	546	2.2	2.0	0.71
France	519	2.7	517	2.8	1.8	0.90
Germany	513	3.2	514	3.5	-0.9	0.85
Greece	448	4.0	448	4.0	0.1	0.96
Hong Kong	548	4.2	548	4.6	0.2	0.84
Hungary	501	2.9	498	3.0	3.1	0.91
Iceland	505	1.4	502	2.0	2.9	0.49
Indonesia	361	3.3	357	3.5	4.4	0.90
Ireland	498	2.3	496	2.9	2.9	0.63
Italy	469	3.1	471	3.3	-2.0	0.91
Japan	547	4.1	545	4.3	1.8	0.90
Korea	550	3.1	547	3.1	3.4	0.96
Latvia	483	3.9	480	3.8	2.1	1.04
Liechtenstein	529	3.9	533	6.1	-3.9	0.42
Luxembourg	494	1.4	498	1.5	-4.0	0.82
Macao-China	532	2.5	530	4.0	2.4	0.40
Mexico	384	4.3	382	4.6	1.9	0.88
Netherlands	520	3.0	524	3.1	-3.3	0.92
New Zealand	533	2.2	533	2.8	-0.5	0.60
Norway	490	2.6	490	3.0	-0.5	0.76
Poland	487	2.8	486	3.0	0.4	0.87
Portugal	470	3.9	468	4.2	1.4	0.87
Russian Federation	479	4.6	479	4.7	-0.2	0.97
Serbia	420	3.3	421	3.5	-0.9	0.90
Slovak Republic	492	3.4	490	3.5	1.7	0.91
Spain	482	2.7	481	2.9	1.5	0.90
Sweden	509	2.4	512	2.7	-3.6	0.81
Switzerland	521	3.0	526	2.8	-4.4	1.21
Thailand	425	2.7	419	2.8	5.7	0.95
Tunisia	345	2.1	349	2.6	-4.4	0.68
Turkey	408	6.0	412	6.2	-4.3	0.93
United Kingdom	510	2.4	512	2.6	-2.2	0.82
United States	477	3.1	480	3.2	-2.8	0.97
Uruguay	411	3.7	413	3.6	-2.2	1.04

In each case these differences can be explained by characteristics of the students who did not respond to items from the respective domain. In Denmark, students performed surprisingly poorly on booklet 9 when responding to both the science and the reading items. In contrast they performed quite well (relative to other booklets) on problem solving. In addition, it has been noted that the non-responding students (for each domain) have a lower value in the index of economic, social and cultural status (ESCS) than students who did respond to items on each domain. Given the positive correlation between ESCS and achievement, the lower values of ESCS for the students who were not assessed in a domain, and the lower than expected scores on booklet 9, it can be expected that the imputations for the non-assessed students will lead to a reduction in the mean scores in reading and science for Denmark.

In the case of Canada, the mean on science of all students is nine points lower than the mean of the assessed students. This is because Canadian students have not performed well on booklet 9. Interestingly, it appears that the fatigue effect that normally results in PISA booklet differences is less pronounced in Canada than in other countries.

For each of Greece, Hungary and Mexico, a higher than expected performance on the reference booklet has resulted in the mean science scores for all students being higher than the mean science scores for the assessed students.

COMPUTATION OF THE LINK ERROR

Link errors (as discussed in Chapter 9) were obtained by estimating the item parameters for PISA 2000 and PISA 2003 using the international calibration samples. Tables 13.25, 13.26, 13.27 and 13.28 show the item parameter estimates for the items that were common to the two studies for reading, science and the two common mathematics scales (space and shape, and change and relationships) respectively.

The column headed "Difference" in each of these tables shows the amount by which the difference between the estimated item parameters differs from the average difference. The standard deviation of these differences divided by the square root of the number of link items gives the standard errors of the differences under the assumption that the link items are a random sample from some universe of possible link items between 2000 and 2003.

The link standard errors in logits, and on the PISA scale, are given in Table 13.29.

Table 13.25 ■ **Comparison of reading item parameters for PISA 2000 and PISA 2003**

Item name	Difficulty estimate 2003	Centred difficulty estimate 2003	Difficulty estimate 2000	Centred difficulty estimate 2000	Difference	Difference squared
R055Q01	-1.28	-1.28	-1.377	-1.347	-0.072	0.005
R055Q02	0.63	0.63	0.496	0.526	-0.101	0.010
R055Q03	0.27	0.27	0.067	0.097	-0.175	0.031
R055Q05	-0.69	-0.69	-0.877	-0.847	-0.154	0.024
R067Q01	-2.08	-2.08	-1.726	-1.696	0.388	0.151
R067Q04	0.25	0.25	0.516	0.546	0.292	0.085
R067Q05	-0.18	-0.18	0.182	0.212	0.394	0.155
R102Q04A	1.53	1.53	1.206	1.236	-0.290	0.084
R102Q05	0.87	0.87	0.905	0.935	0.067	0.005
R102Q07	-1.42	-1.42	-1.566	-1.536	-0.116	0.013
R104Q01	-1.47	-1.47	-1.235	-1.205	0.268	0.072
R104Q02	1.44	1.44	1.105	1.135	-0.306	0.094
R104Q05	2.17	2.17	1.875	1.905	-0.267	0.071
R111Q01	-0.19	-0.19	-0.053	-0.023	0.164	0.027
R111Q02B	1.54	1.54	1.365	1.395	-0.147	0.022
R111Q06B	0.89	0.89	0.808	0.838	-0.051	0.003
R219Q01T	-0.59	-0.59	-0.550	-0.520	0.069	0.005
R219Q01E	0.10	0.10	0.278	0.308	0.210	0.044
R219Q02	-1.13	-1.13	-0.917	-0.887	0.243	0.059
R220Q01	0.86	0.86	0.785	0.815	-0.041	0.002
R220Q02B	-0.14	-0.14	-0.144	-0.114	0.027	0.001
R220Q04	-0.10	-0.10	0.163	0.193	0.297	0.088
R220Q05	-1.39	-1.39	-1.599	-1.569	-0.184	0.034
R220Q06	-0.34	-0.34	-0.172	-0.142	0.196	0.038
R227Q01	0.40	0.40	0.196	0.226	-0.170	0.029
R227Q02T	0.16	0.16	0.045	0.075	-0.086	0.007
R227Q03	0.46	0.46	0.295	0.325	-0.132	0.017
R227Q06	-0.56	-0.56	-0.916	-0.886	-0.327	0.107

Table 13.26 ■ **Comparison of science item parameters for PISA 2000 and PISA 2003**

Item name	Difficulty estimate 2003	Centred difficulty estimate 2003	Difficulty estimate 2000	Centred difficulty estimate 2000	Difference	Difference squared
S114Q03T	-0.29	-0.30	-0.373	-0.346	0.049	0.002
S114Q04T	0.54	0.54	0.377	0.404	0.133	0.018
S114Q05T	1.48	1.47	1.307	1.334	0.139	0.019
S128Q01	-0.66	-0.67	-0.557	-0.530	-0.138	0.019
S128Q02	0.20	0.20	0.284	0.311	-0.116	0.013
S128Q03T	-0.52	-0.53	-0.527	-0.500	-0.030	0.001
S129Q01	0.42	0.42	0.620	0.647	-0.231	0.053
S129Q02T	1.53	1.53	1.497	1.524	0.004	0.000
S131Q02T	0.26	0.26	0.028	0.055	0.201	0.041
S131Q04T	1.41	1.40	1.438	1.465	-0.063	0.004
S133Q01	-0.60	-0.60	-0.356	-0.329	-0.274	0.075
S133Q03	0.64	0.64	0.313	0.340	0.295	0.087
S133Q04T	0.13	0.13	0.250	0.277	-0.151	0.023
S213Q01T	0.36	0.35	0.419	0.446	-0.094	0.009
S213Q02	-1.46	-1.46	-1.484	-1.457	-0.005	0.000
S252Q01	-0.18	-0.19	0.026	0.053	-0.241	0.058
S252Q02	-0.97	-0.97	-1.123	-1.096	0.124	0.015
S252Q03T	-0.46	-0.47	-0.176	-0.149	-0.323	0.104
S256Q01	-2.21	-2.22	-2.491	-2.464	0.245	0.060
S268Q01	-1.10	-1.11	-1.250	-1.223	0.117	0.014
S268Q02T	0.80	0.79	0.578	0.605	0.188	0.035
S268Q06	-0.17	-0.17	-0.236	-0.209	0.034	0.001
S269Q01	-0.46	-0.46	-0.460	-0.433	-0.030	0.001
S269Q03T	0.56	0.55	0.497	0.524	0.026	0.001
S269Q04T	0.89	0.88	0.712	0.739	0.141	0.020

Table 13.27 ■ **Comparison of space and shape item parameters for PISA 2000 and PISA 2003**

Item name	Difficulty estimate 2003	Centred difficulty estimate 2003	Difficulty estimate 2000	Centred difficulty estimate 2000	Difference	Difference squared
M033Q01	-1.52048	-1.496	-1.38728	-1.410	0.022	0.000506
M034Q01T	0.45924	0.432	0.592436	0.518	0.074	0.005508
M144Q01T	-1.01169	-0.666	-0.87849	-0.580	-0.299	0.089232
M144Q02T	1.08967	1.235	1.222866	1.321	-0.098	0.009674
M144Q03	-1.81466	-1.491	-1.68146	-1.405	-0.277	0.076556
M144Q04T	0.43081	0.641	0.564006	0.727	-0.163	0.02664
M145Q01T	-0.5594	-0.906	-0.4262	-0.820	0.394	0.1549
M266Q01T	1.85779	1.782	1.990986	1.868	0.123	0.015071
M273Q01T	-0.13004	-0.307	0.003156	-0.221	0.224	0.050146

Table 13.28 ■ **Comparison of change and relationships item parameters for PISA 2000 and PISA 2003**

Item name	Difficulty estimate 2003	Centred difficulty estimate 2003	Difficulty estimate 2000	Centred difficulty estimate 2000	Difference	Difference squared
M124Q01	0.53645	0.797	0.478116	0.682	-0.204	0.041691
M124Q03T	1.27627	1.488	1.217936	1.373	-0.155	0.024138
M150Q01	-0.68604	-0.913	-0.74437	-1.028	0.283	0.080274
M150Q02T	-1.12923	-0.979	-1.18756	-1.094	-0.094	0.00881
M150Q03T	0.00896	0.322	-0.04937	0.207	-0.257	0.065882
M155Q01	-0.74461	-0.891	-0.80294	-1.006	0.203	0.04111
M155Q02T	-0.64269	-0.480	-0.70102	-0.595	-0.106	0.011305
M155Q03T	1.71785	1.616	1.659516	1.501	0.158	0.025032
M155Q04T	-0.23021	-0.391	-0.28854	-0.506	0.217	0.047157
M192Q01T	0.47659	0.578	0.418256	0.463	-0.045	0.002029

Table 13.29 ■ **Standard errors for the PISA 2000 to PISA 2003 links**

Scale	Standard error on logits	Standard error on PISA scale
Reading	0.041	3.744
Science	0.033	2.959
Space and shape	0.077	6.008
Change and relationships	0.062	4.84

TRANSFORMING THE PLAUSIBLE VALUES TO PISA SCALES

As described in Chapter 9 the PISA 2003 reporting scales for reading and science are the same as those used in PISA 2000. For mathematics and problem solving new scales were prepared for PISA 2003. The transformations for mapping the PISA 2003 logits to the PISA reporting scales are given below for each domain.

Reading

After computing the plausible values, on the logit metric, and following the procedures described in Chapter 9, it was noted that there were substantial differences between the optimal linking transformations for male and female students. The resulting transformations were as follows:

For male students:

$$P_{2000} = \left\{ \frac{\left[\left(0.8823 I_{2003} + 0.0204 \right) - 0.5076 \right]}{1.1002} \times 100 + 500 \right\} \qquad (13.1)$$

For female students:

$$P_{2000} = \left\{ \frac{\left[\left(0.8739 I_{2003} + 0.0970 \right) - 0.5076 \right]}{1.1002} \times 100 + 500 \right\} \qquad (13.2)$$

For students with missing gender code:

$$P_{2000} = \left\{ \frac{\left[\left(0.8830 l_{2003} + 0.0552 \right) - 0.5076 \right]}{1.1002} \times 100 + 500 \right\} \tag{13.3}$$

The coefficients 0.5076, 1.1002, 100 and 500 are required to transform the PISA 2000 logits to the PISA 2000 scale. The scale factors of 0.8823, 0,8739 and 0.8830 and shifts of 0.0204, 0.0970 and 0.0552 transform the PISA 2003 logit scale to the PISA 2000 logit scale for males, females, and missing gender code students respectively.

Science

For science the transformation is given by:

$$P_{2000} = \left\{ \frac{\left[\left(1.0063 l_{2003} + 0.0155 \right) + 0.0933 \right]}{1.1086} \times 100 + 500 \right\} \tag{13.4}$$

The multiplication by 1.0063 and addition of 0.0155 transforms the 2003 logits to the 2000 logit scale, and then the 2000 logit is transformed to the PISA 2000 scale.

Problem solving

For problem solving the transformations are simpler because they do not involve the transformation of the PISA 2003 logits to the PISA 2000 scale.

$$P_{2003} = \left\{ \frac{\left[l_{2003} + 0.0973 \right]}{1.1751} \times 100 + 500 \right\} \tag{13.5}$$

Mathematics

Similarly for mathematics the transformations are simpler because they do not involve the transformation of the PISA 2003 logits to the PISA 2000 scale.

$$P_{2003} = \left\{ \frac{\left[l_{2003} + 0.1344 \right]}{1.2838} \times 100 + 500 \right\} \tag{13.6}$$

Space and shape

$$P_{2003} = \left\{ \frac{\left[\left(0.996 l_{2000} + 0.008 \right) + 0.1342 \right]}{1.2837} \times 100 + 500 \right\} \tag{13.7}$$

Change and relationships

$$P_{2003} = \left\{ \frac{\left[\left(0.985 l_{2000} + 0.059 \right) + 0.1342 \right]}{1.2837} \times 100 + 500 \right\} \qquad (13.8)$$

Notes

1 Note that both Luxembourg and the United Kingdom have been excluded from these calculations.

2 For the definition of "not reached" see Chapter 18.

3 Note that because the design was balanced the inclusion of the booklet term in the item response model did not have an appreciable effect on the item parameter 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

BRA Brazil
HKG Hong Kong-China
IND Indonesia
LVA Latvia
LVL Latvia (Latvian Community)
LVR Latvia (Russian Community)
LIE Liechtenstein
MAC Macao-China
RUS Russian Federation
YUG Serbia and Montenegro (Serbia)
THA Thailand
TUN Tunisia
URY Uruguay

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
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CITO	National Institute for Educational Measurement, The Netherlands
CIVED	Civic Education Study
DIF	Differential Item Functioning
ESCS	Economic, Social and Cultural Status
ENR	Enrolment of 15-year-olds
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
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

References

Adams, R.J., Wilson, M.R. and **W. Wang** (1997), "The multidimensional random coefficients multinomial logit model", *Applied Psychological Measurement 21,* pp. 1-24.

Aiken, L. R. (1974), "Two scales of attitudes toward mathematics," *Journal for Research in Mathematics Education 5,* National Council of Teachers of Mathematics, Reston, pp. 67-71.

Andersen, Erling B. (1997), "The Rating Scale Model", in van der Linden, W. J. and R.K. Hambleton (eds.), *Handbook of Modern Item Response Theory,* Springer, New York/Berlin/Heidelberg.

Bandura, A. (1986), *Social Foundations of Thought and Action: A Social Cognitive Theory,* Prentice Hall, Englewood Cliffs, N.J.

Baumert, J. and **O. Köller** (1998), "Interest Research in Secondary Level I : An Overview", in L. Hoffmann, A. Krapp, K.A. Renninger & J. Baumert (eds.), *Interest and Learning*, IPN, Kiel.

Beaton, A.E. (1987), *Implementing the New Design: The NAEP 1983-84 Technical Report* (Report No. 15-TR-20), Educational Testing Service, Princeton, N.J.

Bryk, A. S. and **S.W. Raudenbush** (1992), *Hierarchical Linear Models: Applications and Data Analysis Methods,* SAGE Publications, Newbury Park.

Bollen, K.A. and **S.J. Long** (eds.) (1993), *Testing Structural Equation Models*, SAGE publications, Newbury Park.

Branden, N. (1994), *Six Pillars of Self-Esteem*. Bantam, New York.

Brennan, R.L. (1992), *Elements of Generalizability Theory*, American College Testing Program, Iowa City.

Buchmann, C. (2000), *Measuring Family Background in International Studies of Educational Achievement: Conceptual Issues and Methodological Challenges,* paper presented at a symposium convened by the Board on International Comparative Studies in Education of the National Academy of Sciences/National Research Council on 1 November, in Washington, D.C.

Cochran, W.G. (1977), *Sampling Techniques* (3rd edition), Wiley, New York.

Cronbach, L.J., G.C. Gleser, H. Nanda and **N. Rajaratnam** (1972), *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*, Wiley and Sons, New York.

Eccles, J.S. (1994), "Understanding Women's Educational and Occupational choice: Applying the Eccles *et al.* Model of Achievement-Related Choices", *Psychology of Women Quarterly 18*, Society for the Psychology of Women, Washington, D.C., pp. 585-609.

Eccles, J.S. and **A. Wigfield** (1995), "In the mind of the achiever: The structure of adolescents' academic achievement-related beliefs and self-perceptions", *Personality and Social Psychology Bulletin 21*, Sage Publications, Thousand Oaks, pp. 215-225.

Ganzeboom, H.B.G., P.M. de Graaf and **D.J. Treiman** (1992), "A standard international socio-economic index of occupational status", *Social Science Research 21*, Elsevier, pp. 1-56.

Gifi, A. (1990), *Nonlinear Multivariate Analysis*, Wiley, New York.

Greenacre, M.J. (1984), *Theory and Applications of Correspondence Analysis*, Academic Press, London.

Grisay, A. (2003), "Translation procedures in OECD/PISA 2000 international assessment", *Language Testing 20,* Holder Arnold Journals, pp. 225-240.

Gustafsson, J.E and **P.A. Stahl** (2000), *STREAMS User's Guide, Version 2.5 for Windows,* MultivariateWare, Mölndal, Sweden.

Hacket, G. and **N. Betz.** (1989), "An Exploration of the mathematics Efficacy/mathematics Performance Correspondence", *Journal of Research in Mathematics Education 20,* National Council of Teachers of Mathematics, Reston, pp. 261-273.

Harvey-Beavis, A. (2002), "Student and Questionnaire Development" in OECD, *PISA 2000 Technical Report*, OECD, Paris.

Hatcher, L. (1994), *A Step-by-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling,* SAS Institute Inc., Cary.

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

Jöreskog, K.G. and **Dag Sörbom** (1993), *LISREL 8 User's Reference Guide,* Scientific Software International, Chicago.

Judkins, D.R. (1990), "Fay's Method for Variance Estimation", *Journal of Official Statistics 6,* Statistics Sweden, Stockholm, pp. 223-239.

Kaplan, D. (2000), *Structural Equation Modeling: Foundation and Extensions,* SAGE Publications, Thousand Oaks.

Keyfitz, N. (1951), "Sampling with probabilities proportionate to science: Adjustment for changes in probabilities", *Journal of the American Statistical Association 46,* American Statistical Association, Alexandria, pp. 105-109.

Lepper, M. R. (1988), "Motivational considerations in the study of instruction", *Cognition and Instruction 5,* Lawrence Erlbaum Associates, Mahwah, pp. 289-309.

Ma, X. (1999), "A Meta-Analysis of the Relationship Between Anxiety Toward mathematics and Achievement in mathematics", *Journal for Research in Mathematics Education 30,* National Council of Teachers of Mathematics, Reston, pp. 520-540.

Macaskill, G., R.J. Adams and **M.L. Wu** (1998), "Scaling methodology and procedures for the mathematics and science literacy, advanced mathematics and physics scales", in M. Martin and D.L. Kelly (eds.) *Third International Mathematics and Science Study, Technical Report Volume 3: Implementation and Analysis,* Center for the Study of Testing, Evaluation and Educational Policy, Boston College, Chestnut Hill.

Marsh, H. W. (1990), *Self-Description Questionnaire (SDQ) II: A theoretical and Empirical Basis for the Measurement Of Multiple Dimensions of Adolescent Self-Concept: An Interim Test Manual and a Research Monograph,* The Psychological Corporation, San Antonio.

Marsh, H. W. (1994), "Confirmatory factor analysis models of factorial invariance: A multifaceted approach" *Structural Equation Modeling 1,* Lawrence Erlbaum Associates, Mahwah, pp. 5-34.

Marsh, H. W. (1999), *Evaluation of the Big-Two-Factor Theory of Motivation Orientation: Higher-order Factor Models and Age-related Changes*, paper presented at the 31.62 Symposium, Multiple Dimensions of Academic Self-Concept, Frames of Reference, Transitions, and International Perspectives: Studies From the SELF Research Centre. Sydney: University of Western Sydney.

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

Meece, J., A. Wigfield and **J. Eccles** (1990), "Predictors of Maths Anxiety and its Influence on Young Adolescents' Course Enrolment and Performance in Mathematics", *Journal of Educational Psychology 82,* American Psychological Association, Washington, D.C., pp. 60-70.

Middleton, J.A. and **P.A. Spanias** (1999), "Findings, Generalizations, and Criticisms of the Research", *Journal for Research in Mathematics Education 30*, National Council of Teachers of Mathematics, Reston, pp. 65-88.

Mislevy, R.J. (1991), "Randomization-based inference about latent variable from complex samples", *Psychometrika 56*, Psychometric Society, Greensboro, pp. 177-196.

Mislevy, R.J. and **K.M. Sheehan** (1987), "Marginal estimation procedures", in A.E. Beaton (ed.), *The NAEP 1983-1984 Technical Report* (Report No. 15-TR-20), Educational Testing Service, Princeton, N.J.

Mislevy, R.J. and **K.M. Sheehan** (1980), "Information matrices in latent-variable models", *Journal of Educational Statistics 14.4,* American Educational Research Association and American Statistical Association, Washington, D.C., and Alexandria, pp. 335-350.

Mislevy, R.J., A.E. Beaton, B. Kaplan and **K.M. Sheehan.** (1992), "Estimating population characteristics form sparse matrix samples of item responses", *Journal of Educational Measurement 29*, National Council on Measurement in Education, Washington, D.C., pp. 133-161.

Multon, K. D., S. D. Brown and **R.W. Lent** (1991), "Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation", *Journal of Counselling Psychology 38,* American Psychological Association, Washington, D.C., pp. 30-38.

Muthén, B. O., S. H. C. du Toit and **D. Spisic** (1997), "Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical outcomes", *Psychometrika*, Psychometric Society, Greensboro.

Muthen, L. and **B. Muthen** (2003), *Mplus User's Guide Version 3.1,* Muthen & Muthen, Los Angeles.

Nishisato, S. (1980), *Analysis of Categorical Data: Dual Scaling and its Applications,* University of Toronto Press, Toronto.

OECD (Organisation for Economic Co-Operation and Development) (1999), *Classifying Educational Programmes: Manual for ISCED-97 Implementation in OECD Countries,* OECD, Paris.

OECD (2001), *Knowledge and Skills for Life: First Results from PISA 2000*, OECD, Paris.

OECD (2002), *PISA 2000 Technical Report,* OECD, Paris.

OECD (2003), *Student Engagement at School: A Sense of Belonging and Participation: Results from PISA 2000,* OECD, Paris.

OECD (2004a), *The PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills*, OCED, Paris.

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

OECD (2004c), *Problem Solving for Tomorrow's World – First Measures of Cross-Curricular Competencies from PISA 2003,* OECD, Paris.

OECD (2005a), *PISA 2003 Data Analysis Manual: SAS® Users,* OECD, Paris.

OECD (2005b), *PISA 2003 Data Analysis Manual: SPSS® Users,* OECD, Paris.

Owens L. and **J. Barnes** (1992), *Learning Preference Scales*, Australian Council for Educational Research, Hawthorn.

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

Rust, K. (1985), "Variance estimation for complex estimators in sample surveys", *Journal of Official Statistics 1,* Statistics Sweden, Stockholm, pp. 381-397.

Rust, K. and **J.N.K. Rao** (1996), "Variance estimation for complex surveys using replication techniques", *Statistical Methods in Medical Research 5,* Holder Arnold Journals, pp. 283-310.

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

Schaffer, E. C., P. S. Nesselrodt and **S. Stringfield** (1994), "The Contribution of Classroom Observation to School Effectiveness Research" in Reynolds *et. al.* (eds.), *Advances in School Effectiveness Research and Practice,* Pergamon, Oxford/New York/Tokyo.

Schulz, W. (2003), *Validating Questionnaire Constructs in International Studies. Two Examples from PISA 2000,* paper presented at the Annual Meeting of the American Educational Research Association (AERA) in Chicago, 21-25 April.

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

Sirotnik, K. (1970), "An analysis of variance framework for matrix sampling", *Educational and Psychological Measurement 30,* SAGE Publications, pp. 891-908.

Slavin, R. E. (1983), "When does cooperative learning increase student achievement?" *Psychological Bulletin 94,* American Psychological Association, Washington, D.C., pp. 429-445.

Statistical Solutions (1992), *BMDP Statistical Software*, Statistical Solutions, Los Angeles.

Teddlie, C. and **D. Reynolds** (2000) (eds.), *The International Handbook of School Effectiveness Research,* Falmer Press, London/New York.

Thorndike, R.L. (1973), *Reading Comprehension Education in Fifteen Countries: An Empirical Study*, Almquist & Wiksell, Stockholm.

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

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

Verhelst, N. (2004), "Generalizability Theory", in Council of Europe, *Reference Supplement to the Preliminary Pilot version of the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, (Section E), Council of Europe (DGIV/EDU/LANG (2004) 13), Strasbourg.

Warm, T. A. (1989), "Weighted Likelihood Estimation of Ability in Item Response Theory", *Psychometrika* 54, Psychometric Society, Greensboro, pp. 427-45.

Wigfield, A., J. S. Eccles and **D. Rodriguez** (1998), "The development of children's motivation in school contexts", in P. D. Pearson. and A. Iran-Nejad (eds.), *Review of Research in Education 23,* American Educational Research Association, Washington D.C., pp. 73-118.

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

Wolter, K.M. (1985), *Introduction to Variance Estimation*, Springer-Verlag, New York.

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

Zimmerman, B.J. and **D.H. Schunk** (eds.) (1989), *Self-Regulated Learning and Academic Achievement. Theory, Research and Practice,* Springer, New York.

Table of Contents