
10

Chapter 10

Item Analysis and Review

Ina V.S. Mullis, Michael O. Martin, and Dana Diaconu

10.1 Overview

Before applying item response theory (IRT) scaling to the TIMSS 2003 achievement data to derive student mathematics and science achievement scores for analysis and reporting, the TIMSS & PIRLS International Study Center conducted a review of a range of diagnostic statistics to examine and evaluate the psychometric characteristics of each achievement item in the 49 countries and four Benchmarking participants in TIMSS 2003. This review played a crucial role in the quality assurance of the TIMSS 2003 data, enabling the detection of unusual item properties that could signal a problem or error for a particular country. For example, an item that was uncharacteristically easy or difficult, or had an unusually low discriminating power, could indicate a potential problem with either translation or printing. Similarly, a constructed-response item with unusually low scoring reliability could indicate a problem with a scoring rubric in a particular country. In the rare instances where such items were found, the country's translation verification documents and printed booklets were examined for flaws or inaccuracies and, if necessary, the item was removed from the international database for that country. This chapter describes the basic item statistics that were consulted and the review criteria that were applied, and provides examples from the assessment to illustrate the review process.

10.2 Statistics for Item Analysis

To begin the review process, the International Study Center computed item analysis statistics for each mathematics and science achievement item, showing the properties of the item in each of the 49 countries and four Benchmarking entities participating in TIMSS 2003. Exhibits 10.1 and 10.2 show examples of the statistics calculated for a multiple-choice and a con-

structed-response item, respectively. Statistics for each item were displayed alphabetically by country, with the international average for each statistic in the bottom row. For those countries that tested in more than one language, statistics were presented separately by language group. For all items, regardless of item format, statistics included the number of students that responded in each country, the difficulty level (the percentage of students that answered the item correctly), and the discrimination index (the point-biserial correlation between success on the item and a total score).[1] Also provided was an estimate of the item's difficulty using a Rasch one-parameter IRT model. The international means of the item difficulties and item discriminations served as guides to the overall statistical properties of the items.

Statistics displayed for multiple-choice items included the percentage of students that chose each option, as well as the percentage of students that omitted or did not reach the item, and the point-biserial correlation between the response to each option and the total score. Statistics displayed for constructed-response items (which could have one or two score levels) included the difficulty and discrimination of each score level. Constructed-response item displays also provided information about the reliability with which the item was scored in each country, with the total number of double-scored cases and the percent exact agreement between the scorers.

Detailed descriptions of the statistics provided in Exhibits 10.1 and 10.2 are listed below in order of appearance in the displays:

N: This is the number of students to whom the item was administered. If a student did not reach an item in the achievement booklet, the item was considered not administered for the purpose of the item analysis.[2]

Diff: Item difficulty is the percentage of students providing a fully correct response to the item. In the case of constructed-response items worth more than one point, this was the percentage of students receiving the maximum score. For the computation of this statistic, not reached items were treated as not administered.

Disc: Item discrimination was computed as the correlation between a correct response to the item and the total score on all of the items in the test booklet.[3] Items exhibiting good measurement properties should have a moderately positive correlation.

PCT_A, PCT_B, PCT_C, PCT_D, and PCT_E: Used for multiple-choice items only (see Exhibit 10.1), each column indicates the percentage of students choosing the particular response option for the item (A, B, C, D, or E). Not reached items were excluded from the denominator for these calculations.

1 For the purpose of computing the discrimination index, the total score was the percentage of items a student answered correctly.

2 In TIMSS, for the purposes of item analysis and item parameter estimation in scaling, items not reached by a student were treated as if they had not been administered. For purposes of estimating student proficiency, however, not reached items were treated as incorrectly answered.

3 For constructed-response items, the discrimination is the correlation between the number of score points and total score.

Exhibit 10.1 International Item Statistics for Item M012040

Trends in International Mathematics and Science Study - TIMSS 2003 Main Survey — July 4, 2004 — 10
International Item Statistics (Unweighted) - 8th Grade For Internal Review Only: DO NOT CITE OR CIRCULATE

Mathematics : Algebra (M012040 - M01_10)
Label: Find n from proportionality equation Type : MC Key: B

Country	N	Diff	Disc	Pct_A	Pct_B	Pct_C	Pct_D	Pct_E	Pct_In	Pct_OM	Pct_NR	PB_A	PB_B	PB_C	PB_D	PB_E	PB_In	PB_OM	RDIFF	Flags
Armenia	931	55.3	0.60	9.1	55.3	9.6	6.3	.	0.1	19.5	0.0	-0.21	0.60	-0.21	-0.17	.	0.00	-0.34	-1.28	____F.
Australia	800	74.5	0.44	8.6	74.4	8.9	6.4	.	0.0	1.8	0.2	-0.19	0.44	-0.29	-0.15	.	0.00	-0.14	-1.58	____F.
Bahrain	666	47.3	0.47	11.0	47.3	20.6	13.4	.	0.3	7.5	2.8	-0.12	0.47	-0.23	-0.17	.	0.00	-0.16	-1.23	_____.
Belgium (Flemish)	848	87.9	0.36	3.3	88.0	2.9	5.0	.	0.0	0.8	0.1	-0.21	0.36	-0.17	-0.19	.	0.00	-0.12	-1.95	__E_F.
Botswana	876	41.3	0.36	27.2	41.6	14.0	13.4	.	0.0	3.9	1.2	-0.15	0.36	-0.19	-0.13	.	0.00	-0.02	-1.34	_____.
Bulgaria	666	61.1	0.48	11.1	61.1	14.3	9.2	.	0.2	4.2	0.4	-0.21	0.48	-0.27	-0.15	.	-0.03	-0.13	-1.07	____F.
Chile	995	50.6	0.51	8.8	50.6	16.7	10.7	.	0.0	13.3	5.2	-0.11	0.51	-0.28	-0.09	.	0.00	-0.22	-1.30	____F.
Chinese Taipei	904	83.3	0.53	4.0	83.3	5.0	7.5	.	0.1	0.1	0.0	-0.30	0.53	-0.35	-0.23	.	-0.05	-0.03	-1.18	____F.
Cyprus	666	65.5	0.47	8.6	65.5	12.2	10.8	.	0.0	3.0	0.3	-0.20	0.47	-0.29	-0.21	.	0.00	-0.03	-1.57	____F.
Egypt	1167	68.7	0.51	6.7	68.7	15.7	6.2	.	0.0	2.7	0.5	-0.20	0.51	-0.36	-0.19	.	0.00	-0.07	-1.88	__E_F.
England	505	76.0	0.37	8.5	75.6	6.5	7.1	.	0.0	2.2	0.2	-0.15	0.37	-0.26	-0.14	.	0.00	-0.11	-1.54	____F.
Estonia	695	70.5	0.39	10.6	70.6	4.9	8.5	.	0.3	5.0	0.0	-0.27	0.39	-0.18	-0.09	.	-0.06	-0.12	-0.80	__H_F.
Ghana	845	29.0	0.27	26.3	29.0	24.1	15.1	.	0.2	5.2	2.3	-0.06	0.27	-0.12	-0.05	.	-0.06	-0.07	-1.22	_____.
Hong Kong, SAR	830	88.5	0.34	2.0	88.4	3.3	5.7	.	0.0	0.6	0.0	-0.19	0.34	-0.27	-0.13	.	0.00	-0.06	-1.46	____F.
Hungary	550	80.6	0.46	6.0	80.2	6.2	4.7	.	0.0	2.9	0.5	-0.25	0.46	-0.29	-0.13	.	0.00	-0.17	-1.48	____F.
Indonesia	957	61.0	0.45	17.1	61.0	11.0	7.8	.	0.0	3.0	1.4	-0.20	0.45	-0.26	-0.14	.	0.00	-0.10	-1.71	____F.
Iran, Islamic Rep. of	845	66.7	0.40	11.4	67.2	9.3	9.3	.	0.2	2.5	0.9	-0.20	0.40	-0.24	-0.17	.	0.01	-0.03	-2.10	__E_F.
Israel	717	71.7	0.48	7.8	71.5	8.5	7.0	.	0.0	5.2	0.4	-0.18	0.48	-0.30	-0.19	.	0.00	-0.16	-1.40	____F.
Italy	708	66.4	0.42	9.6	66.2	8.8	10.3	.	0.1	4.9	1.4	-0.19	0.42	-0.31	-0.08	.	-0.05	-0.11	-1.25	____F.
Japan	804	79.0	0.44	4.7	79.0	5.3	10.3	.	0.0	0.6	0.2	-0.21	0.44	-0.26	-0.21	.	0.00	-0.11	-0.96	__H_F.
Jordan	758	53.3	0.51	9.0	53.3	24.0	10.9	.	0.1	2.6	0.1	-0.13	0.51	-0.33	-0.17	.	-0.03	-0.13	-1.38	____F.
Korea, Rep. of	890	89.4	0.51	3.7	89.3	2.8	4.2	.	0.0	0.0	0.0	-0.27	0.51	-0.29	-0.30	.	0.00	0.00	-1.72	____F.
Latvia	603	73.7	0.48	9.1	73.8	9.8	4.8	.	0.0	2.5	0.2	-0.28	0.48	-0.27	-0.13	.	0.00	-0.14	-1.25	____F.
Lebanon	645	71.6	0.45	10.2	71.6	9.6	4.7	.	0.0	3.9	0.5	-0.24	0.45	-0.26	-0.12	.	0.00	-0.14	-2.13	__E_F.
Lithuania	848	66.0	0.48	13.0	66.0	8.8	7.4	.	0.1	4.6	0.1	-0.24	0.48	-0.25	-0.15	.	-0.04	-0.19	-0.95	__H_F.
Macedonia, Rep. of	650	49.5	0.54	12.3	49.5	17.5	9.2	.	0.6	10.8	0.3	-0.14	0.54	-0.29	-0.13	.	-0.05	-0.22	-0.95	__H_F.
Malaysia	873	83.3	0.36	7.0	83.3	4.5	4.5	.	0.0	0.8	0.5	-0.20	0.36	-0.22	-0.14	.	0.00	-0.09	-2.07	__E_F.
Moldova, Rep. of	662	62.4	0.47	9.5	62.4	16.3	6.5	.	0.0	5.3	0.5	-0.19	0.47	-0.26	-0.14	.	0.00	-0.15	-1.33	____F.
Morocco	521	55.5	0.43	9.6	55.1	18.8	9.4	.	0.2	6.9	1.9	-0.09	0.43	-0.29	-0.13	.	-0.05	-0.11	-1.77	____F.
Netherlands	517	86.1	0.32	7.9	86.1	2.1	2.9	.	0.0	1.0	0.2	-0.25	0.32	-0.14	-0.10	.	0.00	-0.05	-1.76	____F.
New Zealand	629	65.2	0.47	13.4	65.2	11.4	8.3	.	0.0	1.7	0.0	-0.18	0.47	-0.29	-0.18	.	0.00	-0.15	-1.17	____F.
Norway	677	60.2	0.46	11.1	60.3	16.2	6.8	.	0.1	5.5	1.0	-0.10	0.46	-0.32	-0.15	.	-0.02	-0.17	-1.28	____F.
Palestinian Nat'l Aut	911	53.1	0.50	11.6	52.4	23.3	9.4	.	0.3	3.0	0.5	-0.12	0.50	-0.34	-0.16	.	-0.07	-0.08	-1.53	____F.
Philippines	1264	52.9	0.44	17.7	53.5	17.3	10.1	.	0.1	1.3	0.7	-0.19	0.44	-0.27	-0.13	.	0.00	-0.02	-1.65	_____.
Romania	678	63.6	0.57	5.5	63.6	17.3	9.6	.	0.0	4.1	0.3	-0.20	0.56	-0.36	-0.18	.	0.00	-0.18	-1.20	____F.
Russian Federation	773	71.4	0.44	9.3	71.4	8.5	6.1	.	0.3	4.4	0.8	-0.23	0.44	-0.22	-0.14	.	-0.02	-0.18	-1.16	____F.
Saudi Arabia	699	33.6	0.42	10.4	33.6	31.8	16.2	.	0.0	8.0	1.7	-0.05	0.43	-0.25	-0.11	.	0.00	-0.11	-1.17	_____.
Scotland	587	78.9	0.38	10.1	78.9	4.9	4.8	.	0.0	1.4	0.0	-0.18	0.38	-0.23	-0.17	.	0.00	-0.13	-1.82	____F.
Serbia	718	64.3	0.53	12.1	64.3	10.0	6.7	.	0.0	6.8	0.6	-0.25	0.53	-0.24	-0.13	.	0.00	-0.26	-1.23	____F.
Singapore	993	93.2	0.30	1.7	93.2	1.1	3.7	.	0.0	0.3	0.0	-0.16	0.30	-0.18	-0.19	.	0.00	-0.01	-2.04	__E_F.
Slovak Republic	683	77.9	0.43	8.9	77.9	4.4	6.0	.	0.1	2.6	0.7	-0.24	0.43	-0.24	-0.14	.	-0.01	-0.14	-1.45	____F.
Slovenia	623	71.9	0.47	12.8	71.9	7.5	4.7	.	0.0	3.0	0.0	-0.26	0.47	-0.25	-0.14	.	0.00	-0.18	-1.55	____F.
South Africa	1415	27.0	0.47	14.6	27.0	33.1	19.3	.	0.7	5.3	4.4	-0.05	0.47	-0.20	-0.10	.	-0.09	-0.14	-1.01	____H__.
Sweden	707	65.7	0.40	15.4	65.3	8.2	6.5	.	0.0	4.5	1.1	-0.14	0.40	-0.23	-0.16	.	0.00	-0.15	-1.02	__H_F.
Syrian Arab Republic	811	47.7	0.41	13.3	47.7	23.2	11.7	.	0.0	4.1	0.7	-0.13	0.41	-0.24	-0.10	.	0.00	-0.10	-1.48	_____.
Tunisia	788	67.5	0.34	11.9	67.5	10.0	4.2	.	0.5	5.8	4.8	-0.14	0.34	-0.23	-0.01	.	-0.06	-0.14	-2.21	__E_F.
United States	1503	79.9	0.41	8.3	79.9	5.0	5.8	.	0.0	1.0	0.5	-0.23	0.41	-0.22	-0.18	.	0.00	-0.05	-1.86	__E_F.
International Avg.	.	65.7	0.44	10.3	65.7	11.8	8.1	.	0.1	4.0	0.9	-0.18	0.44	-0.26	-0.15	.	-0.01	-0.12	-1.46	__E_F.
Basque Country, Spain	420	76.0	0.42	4.3	76.0	8.8	8.1	.	0.0	2.9	0.7	-0.11	0.42	-0.31	-0.22	.	0.00	-0.04	-1.77	____F.
Ontario Province, Can	717	85.4	0.39	6.8	85.6	3.8	3.5	.	0.0	0.3	0.4	-0.26	0.39	-0.23	-0.13	.	0.00	-0.06	-2.07	__E_F.
Quebec Province, Can.	736	86.4	0.33	5.2	86.4	3.3	4.1	.	0.0	1.1	0.3	-0.18	0.33	-0.17	-0.18	.	0.00	-0.07	-1.83	____F.

Keys: Diff: Percent obtaining correct score; Disc: Discrimination; Pct_A...D: Percent choosing each option;
Pct_In, OM, NR: Percent Invalid, Omitted, and Not Reached Responses; PB_A...D: Point Biserial for each option;
PB_OM: Point Biserial for omitted. RDIFF= Difficulty (1-PL).
Flags: A= Ability not ordered/ Attractive distractor; C= Difficulty less than chance; D= Negative/low discrimination; E= Easier than average;
F= Distractor chosen by less than 10%; H= Harder than average; R= Scoring reliability < 80%; V= Difficulty greater than 95.

Exhibit 10.2 **International Item Statistics for Item S032680**

Trends in International Mathematics and Science Study - TIMSS 2003 Main Survey — August 12, 2004 — 95
International Item Statistics (Unweighted) - 8th Grade For Internal Review Only: DO NOT CITE OR CIRCULATE

Science : Chemistry (S032680 - S07_04)
Label: Identify iron, water and oxygen Type : CR Key: X

Country	N	Diff	Disc	Pct_0	Pct_1	Pct_2	Pct_3	Pct_OM	Pct_NR	PB_0	PB_1	PB_2	PB_3	PB_OM	RDIFF	Reliability Cases	Score	Code	Flags
Armenia	462	37.2	0.66	7.4	7.4	37.2	.	48.1	0.0	-0.16	0.06	0.63	.	-0.55	0.05	410	99.3	99.3	H_F.
Australia	407	52.3	0.61	16.0	25.6	52.3	.	6.1	0.0	-0.29	-0.25	0.60	.	-0.35	-0.33	193	100.0	100.0	_____.
Bahrain	337	25.5	0.52	23.7	35.9	25.5	.	14.8	0.0	-0.27	0.21	0.35	.	-0.40	-0.04	174	100.0	99.4	_____.
Belgium (Flemish)	434	58.5	0.60	11.3	27.0	58.5	.	3.2	0.0	-0.38	-0.26	0.56	.	-0.25	-0.57	227	100.0	99.6	_____.
Botswana	446	32.3	0.62	39.9	23.3	32.3	.	4.5	0.0	-0.48	0.03	0.56	.	-0.17	-0.74	219	95.4	90.4	_E__.
Bulgaria	335	40.3	0.58	16.7	19.7	40.3	.	23.3	0.0	-0.18	0.09	0.49	.	-0.50	0.02	175	92.6	91.4	_H__.
Chile	521	17.5	0.50	38.8	29.8	17.5	.	14.0	0.0	-0.19	0.06	0.46	.	-0.31	0.53	256	96.5	94.5	_H__.
Chinese Taipei	445	69.0	0.68	6.1	17.8	69.0	.	7.2	0.0	-0.30	-0.29	0.65	.	-0.46	-0.27	172	98.8	98.8	____F.
Cyprus	326	34.0	0.61	18.1	30.7	34.0	.	17.2	0.0	-0.24	-0.04	0.56	.	-0.40	-0.34	179	98.3	97.2	_____.
Egypt	591	42.1	0.70	27.6	22.7	42.1	.	7.6	0.0	-0.56	-0.01	0.63	.	-0.20	-0.45	66	100.0	100.0	_____.
England	251	68.9	0.62	5.6	21.1	68.9	.	4.4	0.0	-0.30	-0.30	0.58	.	-0.37	-0.61	59	98.3	98.3	____F.
Estonia	335	74.9	0.60	6.3	15.5	74.9	.	3.3	0.0	-0.37	-0.31	0.58	.	-0.28	-0.66	217	100.0	99.5	____F.
Ghana	427	11.0	0.58	52.5	15.2	11.0	.	21.3	0.0	-0.30	0.28	0.46	.	-0.23	-0.18	327	98.8	96.3	_____.
Hong Kong, SAR	424	67.9	0.50	9.2	19.8	67.9	.	3.1	0.0	-0.30	-0.14	0.44	.	-0.35	-0.32	205	100.0	99.0	____F.
Hungary	278	66.9	0.58	10.4	20.5	66.9	.	2.2	0.0	-0.36	-0.34	0.58	.	-0.18	-0.54	223	93.7	92.8	_____.
Indonesia	485	23.3	0.62	32.2	24.9	23.3	.	19.6	0.0	-0.36	0.13	0.53	.	-0.28	0.06	233	95.3	93.1	_H__.
Iran, Islamic Rep. of	431	40.6	0.54	16.9	31.3	40.6	.	11.1	0.0	-0.29	-0.16	0.53	.	-0.23	-0.53	193	95.9	93.8	_____.
Israel	358	51.7	0.59	8.9	31.8	51.7	.	7.5	0.0	-0.27	-0.26	0.57	.	-0.32	-0.71	183	92.9	92.3	_____.
Italy	363	36.6	0.49	15.7	35.3	36.6	.	12.4	0.0	-0.17	-0.16	0.48	.	-0.29	-0.07	206	99.0	99.0	_____.
Japan	399	70.2	0.59	9.0	17.3	70.2	.	3.5	0.0	-0.35	-0.29	0.57	.	-0.28	-0.55	203	98.5	98.0	____F.
Jordan	383	31.3	0.65	29.8	31.9	31.3	.	7.0	0.0	-0.48	0.05	0.55	.	-0.24	0.03	179	99.4	98.3	_H__.
Korea, Rep. of	444	66.9	0.66	10.8	18.2	66.9	.	4.1	0.0	-0.42	-0.23	0.61	.	-0.36	-0.35	205	100.0	99.5	_____.
Latvia	304	60.9	0.62	17.8	13.2	60.9	.	8.2	0.3	-0.42	-0.08	0.57	.	-0.32	-0.25	176	98.3	97.7	_____.
Lebanon	326	40.2	0.66	22.7	20.9	40.2	.	16.3	0.0	-0.37	-0.07	0.63	.	-0.34	-0.68	174	98.3	97.1	_____.
Lithuania	420	55.0	0.53	10.7	28.8	55.0	.	5.5	0.0	-0.27	-0.26	0.52	.	-0.25	-0.41	190	95.3	94.7	_____.
Macedonia, Rep. of	322	36.6	0.65	14.9	18.9	36.6	.	29.5	0.0	-0.26	0.07	0.58	.	-0.47	-0.09	142	100.0	100.0	_____.
Malaysia	432	68.1	0.55	9.3	19.2	68.1	.	3.5	0.0	-0.30	-0.28	0.54	.	-0.27	-0.90	235	99.6	99.1	_E_F.
Moldova, Rep. of	333	36.0	0.50	3.9	41.7	36.0	.	18.3	0.0	-0.06	-0.06	0.42	.	-0.42	-0.40	180	100.0	100.0	____F.
Morocco	254	14.2	0.44	28.7	35.0	14.2	.	22.0	0.0	-0.16	0.22	0.30	.	-0.33	0.18	104	90.4	87.5	_H__.
Netherlands	259	65.3	0.58	10.4	20.8	65.3	.	3.5	0.0	-0.33	-0.31	0.58	.	-0.26	-0.58	218	94.5	94.5	_____.
New Zealand	315	45.7	0.61	21.0	28.9	45.7	.	4.4	0.0	-0.35	-0.23	0.61	.	-0.26	-0.20	167	98.8	98.2	_____.
Norway	339	64.3	0.56	11.2	20.1	64.3	.	4.4	0.0	-0.36	-0.20	0.52	.	-0.28	-0.92	202	97.5	97.5	_E__.
Palestinian Nat'l Aut	449	26.7	0.60	34.1	33.9	26.7	.	5.3	0.0	-0.45	0.13	0.47	.	-0.25	-0.16	212	97.6	97.2	_____.
Philippines	642	17.4	0.62	39.9	21.0	17.4	.	21.7	0.0	-0.26	0.18	0.53	.	-0.36	0.02	352	96.3	93.8	_H__.
Romania	332	44.0	0.61	13.9	29.2	44.0	.	13.0	0.0	-0.25	-0.18	0.59	.	-0.37	-0.43	151	98.0	97.4	_____.
Russian Federation	395	62.0	0.56	13.2	18.0	62.0	.	6.8	0.0	-0.35	-0.24	0.56	.	-0.24	-0.40	194	100.0	100.0	_____.
Saudi Arabia	361	13.3	0.40	30.5	41.6	13.3	.	14.7	0.0	-0.16	0.03	0.38	.	-0.21	0.00	208	97.6	95.2	_____.
Scotland	291	66.7	0.53	9.6	20.3	66.7	.	3.4	0.0	-0.34	-0.25	0.51	.	-0.22	-0.82	76	98.7	96.1	_E_F.
Serbia	352	49.4	0.57	11.9	22.4	49.4	.	16.2	0.0	-0.21	-0.13	0.54	.	-0.40	-0.50	196	99.5	99.0	_____.
Singapore	497	71.0	0.60	6.2	20.9	71.0	.	1.8	0.0	-0.36	-0.33	0.56	.	-0.25	-0.82	200	100.0	100.0	_E_F.
Slovak Republic	342	61.7	0.56	12.6	19.3	61.7	.	6.4	0.0	-0.37	-0.11	0.50	.	-0.31	-0.28	196	99.0	99.0	_____.
Slovenia	312	58.7	0.56	11.2	22.8	58.7	.	7.4	0.0	-0.28	-0.24	0.55	.	-0.30	-0.44	99	98.0	98.0	_____.
South Africa	743	11.8	0.52	52.1	26.9	11.8	.	9.2	0.0	-0.38	0.22	0.41	.	-0.14	-0.34	396	99.0	97.2	_____.
Sweden	355	75.5	0.54	7.9	13.5	75.5	.	3.1	0.0	-0.34	-0.27	0.53	.	-0.24	-0.97	178	94.4	93.8	_E_F.
Syrian Arab Republic	404	26.5	0.57	30.2	30.7	26.5	.	12.6	0.0	-0.40	0.17	0.43	.	-0.25	-0.07	172	91.9	89.5	_____.
Tunisia	409	16.9	0.49	25.7	41.3	16.9	.	16.1	0.0	-0.22	0.21	0.34	.	-0.36	-0.19	196	100.0	100.0	_____.
United States	756	41.4	0.53	19.4	35.6	41.4	.	3.6	0.0	-0.32	-0.16	0.50	.	-0.23	0.00	370	97.8	97.0	_H__.
International Avg.	.	45.7	0.58	18.8	24.8	45.7	.	10.7	0.0	-0.31	-0.09	0.52	.	-0.30	-0.35	.	97.7	96.8	_E__.
Basque Country, Spain	216	42.1	0.61	16.7	32.9	42.1	.	8.3	0.0	-0.34	-0.24	0.61	.	-0.23	-0.34	229	97.8	97.4	_____.
Indiana State, US	195	41.0	0.46	17.4	39.5	41.0	.	2.1	0.0	-0.29	-0.23	0.46	.	-0.05	-0.12	90	97.8	97.8	_____.
Ontario Province, Can	361	56.8	0.60	15.8	24.7	56.8	.	2.8	0.0	-0.39	-0.26	0.59	.	-0.22	-0.42	175	92.6	91.4	_____.
Quebec Province, Can.	368	60.3	0.57	14.4	21.5	60.3	.	3.8	0.0	-0.36	-0.21	0.54	.	-0.26	-0.32	109	92.7	90.8	_____.

Keys: Diff: Percent obtaining maximum score; Disc: Discrimination; Pct_0...3: Percent obtaining score level; Pct_OM, NR: Percent Omitted, and Not Reached Responses;
PB_0...3: Point Biserial for each score level; PB_OM: Point Biserial for omitted; RDIFF= Difficulty (1-PL); Reliability (Cases): Responses Double Scored;
Reliability (Score): Percent Agreement on Score; Reliability (Code): Percent Agreement on Code.
Flags: A= Ability not ordered/ Attractive distractor; C= Difficulty less than chance; D= Negative/low discrimination; E= Easier than average;
F= Score obtained by less than 10%; H= Harder than average; R= Scoring reliability < 80%; V= Difficulty greater than 95.

PCT_0, PCT_1, PCT_2, and PCT_3: Used for constructed-response items only (see Exhibit 10.2), each column indicates the percentage of students scoring at the particular score level, up to and including the maximum score level for the item. Not reached items were excluded from the denominator for these calculations.

PCT_IN: Used for multiple-choice items only, this was the percentage of students that provided an invalid response to a multiple-choice item. Typically, invalid responses were the result of students selecting more than one response option for the same item.

PCT_OM: This is the percentage of students who, having reached the item, did not provide a response. Not reached items were excluded from the denominator when calculating this statistic.

PCT_NR: This is the percentage of student that did not reach the item. An item was coded as not reached when there was no evidence of a response to any subsequent items in the booklet and the response to the item preceding it was omitted.

PB_A, PB_B, PB_C, PB_D, and PB_E: Used for multiple-choice items only, these present the correlation between choosing each of the response options A, B, C, D, or E and the score on the test booklet. Items with good psychometric properties have near-zero or negative correlations for the distracter options (the incorrect options) and moderately positive correlations for the correct option.

PB_0, PB_1, PB_2, and PB_3: Used for constructed-response items only, these present the correlation between the score levels on the item (0, 1, 2, or 3) and the score on the test booklet. For items with good measurement properties the correlation coefficients should change from negative to positive as the score on the item increases.

PB_OM: This is the correlation between a binary variable - indicating an omitted response to the item - and the score on the test booklet. This correlation should be negative or near zero.

PB_IN: Used for multiple-choice items only, this presents the correlation between an invalid response to the item (usually caused by selecting more than one response option) and the score on the test booklet. This correlation also should be negative or near zero.

RDIFF: This is an estimate of the item's difficulty based on a Rasch one-parameter IRT model. The difficulty estimate is expressed in the logit metric (with a positive logit indicating a difficult item) and was scaled so that the average Rasch item difficulty was zero within each country.

Reliability - Cases: To provide a measure of the reliability of the scoring of the constructed-response items, those items in approximately 25 percent of the test booklets in each country were scored by two independent scorers. This column indicates the number of times each item was double-scored in a country.

Reliability - Score: This column contains the percentage of exact agreement between two independent scorers.

As an aid to reviewers, the item-analysis display includes a series of "flags" signaling the presence of one or more conditions that might indicate a problem with an item. The following conditions are flagged:

- Item difficulty exceeds 95 percent in the sample as a whole
- Item difficulty is less than 25 percent for 4-option multiple-choice items in the sample as a whole
- One or more of the distracter percentages is less than 10 percent
- One or more of the distracter percentages is greater than the percentage for the correct answer, or the point-biserial correlation for one or more of the distracters exceeds zero
- Item discrimination (i.e., the point-biserial for the correct answer) is less than 0.2
- Item discrimination does not increase with each score level (for constructed-response items with more than one score level)
- The Rasch difficulty estimate is harder than the average across all items
- The Rasch difficulty estimate is easier than the average across all items
- Difficulty levels on the item differ significantly for males and females
- Difference in item difficulty levels between males and females diverge significantly
- Scoring reliability is less than 70 percent (for constructed-response items only)

Although not all of these conditions necessarily indicate a problem, the flags are a useful way to draw attention to potential sources of concern.

In order to measure trends, TIMSS 2003 included items from TIMSS 1999 and TIMSS 1995 at the eighth grade and from TIMSS 1995 at the fourth grade.[4] For these trend items, the review included an examination of changes in item statistics between 1999 and 2003 at eighth grade and between 1995 and 2003 at fourth grade. An example item statistics display for an eighth-grade trend item is shown in Exhibit 10.3. Different from the item statistics presented in Exhibits 10.1 and 10.2, this display includes countries' statistics from both the TIMSS 1999 and TIMSS 2003 assessments. In review-

4 For more information on trend items, see Chapter 2.

ing these item statistics, the aim was to detect any unusual changes in item properties between assessments, which might indicate a problem in using the item to measure change.

10.2.1 Item-by-Country Interaction

Although countries are expected to exhibit some variation in performance across items, in general, countries with high average performance on the achievement test as a whole should perform relatively well on each of the items, and low-scoring countries should do less well on each of items. When this does not occur, i.e., when a high-scoring country has low performance on an item on which other countries are doing well, there is said to be an item-by-country interaction. When large, such item-by-country interactions may be a sign of an item that is flawed in some way and measures should be taken to address the problem.

To assist in detecting sizeable item-by-country interactions, the International Study Center produced a graphical display for each item showing the average probability across all countries of a correct response for a student of average international proficiency, compared with the probability of a correct response by a student of average proficiency in each country. Exhibit 10.4 provides an example of a TIMSS item-by-country interaction display. The probability for each country is presented as a 95 percent confidence interval, which includes a built-in Bonferroni correction for multiple comparisons. The limits for the confidence interval are computed as follows:

$$\text{Upper Limit} = \left(1 - \frac{e^{RDIFF_{ik} + SE_{RDIFF_{ik}} \times Z_b}}{1 + e^{RDIFF_{ik} + SE_{RDIFF_{ik}} \times Z_b}}\right)$$

$$\text{Lower Limit} = \left(1 - \frac{e^{RDIFF_{ik} - SE_{RDIFF_{ik}} \times Z_b}}{1 + e^{RDIFF_{ik} - SE_{RDIFF_{ik}} \times Z_b}}\right)$$

where $RDIFF_{ik}$ is the Rasch difficulty of item k within country i; $SE_{RDIFFik}$ is the standard error of the difficulty of item k in country i; and Z_b is the critical value from the Z distribution, corrected for multiple comparisons using the Bonferroni procedure.

The International Study Center also produced item-by-country interaction displays for each item in the trend study, showing for eighth grade the results from 1999 and 2003 separately in each display, and for fourth grade, the results from 1995 and 2003. An example of an item-by-country interaction display for a trend item is presented in Exhibit 10.5. Confidence intervals for 1999 and 2003 within a country appear side-by-side in the display

Exhibit 10.3 **International Item Statistics for Trend Item M012001**

Trends in International Mathematics and Science Study - TIMSS 2003 Main Survey 15:49 Wednesday, August 25, 2004
Percent of responses by Item Category - 8th Grade
For Internal Review Only: DO NOT CITE OR CIRCULATE

Mathematics: Number (M012001 - M01_01) Label: Number of squares in shaded fraction Item Type = MC Key = A

COUNTRY	Year	N	A	B	C	D	E	OTHER INCORRECT	DIFF	INVALID	NOT REACHED	OMIT	1.GIRL % Right	2.BOY % Right
Belgium (Flemish)	1999	5256	85.5	2.1	1.9	3.8	6.4	0.3	85.5	0.0	0.0	0.3	84.4	86.5
	2003	849	85.9	2.0	2.0	3.7	4.9	1.5	85.9	0.2	0.0	1.3	84.3	87.6
Bulgaria	1999	3270	66.0	5.4	6.4	9.0	10.4	2.8	66.0	0.2	0.1	2.6	64.7	67.3
	2003	669	46.6	8.8	11.4	13.3	13.0	6.9	46.6	0.0	0.0	6.9	45.2	48.1
Chile	1999	5862	18.3	3.8	7.0	23.8	42.0	5.1	18.3	0.0	0.0	5.0	16.3	20.3
	2003	1050	25.0	4.3	6.4	22.5	32.3	9.6	25.0	0.1	0.9	8.7	23.6	26.3
Chinese Taipei	1999	5772	78.4	2.4	5.8	4.9	8.5	0.1	78.4	0.0	0.0	0.0	78.9	77.8
	2003	904	78.7	2.2	6.0	4.6	8.5	0.0	78.7	0.0	0.0	0.0	77.4	79.9
Cyprus	1999	3109	60.2	7.9	7.5	10.5	13.0	0.9	60.2	0.2	0.0	0.8	58.5	61.8
	2003	668	56.1	8.1	11.1	8.2	12.0	4.5	56.1	0.0	0.0	4.5	55.5	56.8
England	1999	2946	58.7	2.7	4.7	12.7	20.7	0.4	58.7	0.1	0.0	0.3	54.3	62.9
	2003	506	66.8	2.4	6.5	8.7	15.2	0.4	66.8	0.0	0.0	0.4	64.5	69.6
Hong Kong, SAR	1999	5176	86.3	2.0	3.0	2.7	5.9	0.2	86.3	0.1	0.0	0.1	86.0	86.5
	2003	830	87.8	1.8	3.9	2.5	3.7	0.2	87.8	0.1	0.0	0.1	86.2	89.5
Hungary	1999	3178	67.2	2.8	4.2	10.9	12.7	2.3	67.2	0.2	0.0	2.1	65.9	68.4
	2003	548	66.8	3.6	8.6	8.6	10.2	2.2	66.8	0.0	0.0	2.2	68.0	65.7
Indonesia	1999	5847	24.6	7.8	20.8	17.9	27.4	1.5	24.6	0.1	0.0	1.4	24.3	24.9
	2003	970	22.6	10.7	22.9	15.8	22.8	5.3	22.6	0.0	0.1	5.2	19.9	25.5
Iran, Islamic Rep. o	1999	5291	47.2	4.4	5.5	15.3	25.5	2.1	47.2	0.0	0.1	2.0	41.3	51.3
	2003	853	42.0	7.0	7.7	15.2	25.9	2.1	42.0	0.1	0.1	1.9	41.8	42.1
Israel	1999	4191	49.1	8.5	7.2	13.5	17.8	3.9	49.1	0.1	0.4	3.4	44.7	53.6
	2003	720	65.1	6.1	5.8	7.9	13.1	1.9	65.1	0.1	0.0	1.8	61.6	69.3
Italy	1999	3328	48.5	5.0	6.7	12.5	25.2	2.0	48.5	0.0	0.0	2.0	44.4	52.8
	2003	718	50.6	4.6	5.0	10.6	24.2	5.0	50.6	0.7	0.3	4.0	46.5	54.9
Japan	1999	4735	79.6	2.5	4.8	5.3	7.7	0.2	79.6	0.0	0.0	0.1	80.4	78.9
	2003	806	78.9	3.1	4.2	6.1	6.7	1.0	78.9	0.0	0.1	0.9	79.5	78.4
Jordan	1999	5040	38.5	7.4	10.2	18.0	24.4	1.5	38.5	0.4	0.0	1.1	35.6	41.1
	2003	759	31.9	10.0	15.3	17.0	22.4	3.4	31.9	0.1	0.0	3.3	34.3	29.4
Korea, Rep. of	1999	6113	80.4	1.6	3.3	4.7	9.8	0.1	80.4	0.0	0.0	0.1	79.7	81.2
	2003	890	82.4	1.2	3.5	3.7	9.2	0.0	82.4	0.0	0.0	0.0	83.1	81.7
Latvia	1999	2870	56.2	4.6	7.0	12.8	17.4	2.0	56.2	0.3	0.0	1.7	52.0	60.7
	2003	604	61.6	4.5	7.0	9.1	15.1	2.8	61.6	0.0	0.0	2.8	56.3	66.8
Lithuania	1999	2359	40.4	8.4	8.0	16.9	21.5	4.8	40.4	0.0	0.0	4.8	39.3	41.7
	2003	849	48.3	8.0	7.3	12.8	20.3	3.3	48.3	0.4	0.0	2.9	44.3	52.0
Macedonia, Rep. of	1999	4022	42.2	3.7	6.8	19.8	24.0	3.5	42.2	0.3	0.0	3.2	41.2	43.2
	2003	652	37.0	5.4	7.8	19.6	18.6	11.7	37.0	1.1	0.0	10.6	36.2	37.6
Malaysia	1999	5577	72.5	3.6	4.1	8.1	11.3	0.5	72.5	0.0	0.0	0.5	73.6	71.1
	2003	879	67.8	4.4	4.9	10.2	11.3	1.4	67.8	0.2	0.0	1.1	72.5	61.4
Moldova, Rep. of	1999	3711	53.3	7.1	9.9	14.0	14.4	1.4	53.3	0.1	0.0	1.3	52.2	54.6
	2003	665	51.3	8.6	11.6	10.2	10.5	7.8	51.3	0.0	0.2	7.7	53.7	48.6
Morocco	1999	5384	19.7	17.3	20.2	19.1	16.9	6.8	19.7	1.1	0.0	5.7	20.2	19.2
	2003	516	21.1	11.8	16.7	17.4	19.4	13.6	21.1	0.2	0.2	13.2	22.3	19.6
Netherlands	1999	2957	75.1	2.6	3.3	5.8	12.6	0.7	75.1	0.0	0.0	0.7	72.0	78.4
	2003	518	81.7	2.5	2.1	4.4	8.7	0.6	81.7	0.0	0.0	0.6	81.3	81.6
New Zealand	1999	3603	57.3	2.9	6.3	12.5	20.4	0.5	57.3	0.2	0.0	0.3	57.4	57.2
	2003	629	57.4	4.5	8.3	11.0	17.8	1.1	57.4	0.0	0.0	1.1	59.3	55.1
Philippines	1999	6599	22.3	6.2	37.8	10.4	22.5	0.7	22.3	0.0	0.0	0.7	23.5	21.0
	2003	1273	23.0	8.5	35.3	10.4	21.8	1.0	23.0	0.1	0.0	0.9	22.7	23.5
Romania	1999	3425	53.8	6.0	8.8	13.4	16.3	1.6	53.8	0.5	0.0	1.1	53.0	54.6
	2003	680	45.3	8.1	12.1	14.1	15.1	5.3	45.3	0.1	0.1	5.0	44.7	45.8
Russian Federation	1999	4331	62.2	3.9	5.3	12.2	14.2	2.1	62.2	0.1	0.0	2.0	61.5	63.0
	2003	779	52.8	6.2	8.1	14.4	14.2	4.4	52.8	0.3	0.1	4.0	51.8	53.6
Singapore	1999	4966	87.5	1.4	2.1	3.2	5.5	0.3	87.5	0.0	0.0	0.3	87.7	87.4
	2003	993	86.9	2.2	2.8	3.9	3.9	0.2	86.9	0.0	0.0	0.2	86.7	87.1
Slovak Republic	1999	3490	59.1	4.4	6.7	12.3	15.5	2.0	59.1	0.0	0.0	2.0	57.5	60.8
	2003	688	56.7	5.4	8.0	12.1	13.2	4.7	56.7	0.1	0.1	4.4	52.8	60.6
South Africa	1999	8124	13.7	6.2	44.4	10.0	23.6	2.1	13.7	0.6	0.2	1.4	13.1	14.4
	2003	1480	15.5	9.3	37.2	9.8	19.4	8.9	15.5	2.2	0.8	5.9	14.7	15.9
Tunisia	1999	5037	35.1	7.6	10.8	17.8	23.1	5.6	35.1	1.3	0.0	4.2	30.9	39.5
	2003	827	26.5	10.3	10.6	14.5	21.6	16.4	26.5	1.9	0.2	14.3	22.9	30.5
United States	1999	8985	57.2	3.3	6.6	11.4	21.2	0.3	57.2	0.0	0.0	0.3	53.2	61.2
	2003	1510	60.6	3.7	7.0	10.3	17.9	0.5	60.6	0.0	0.1	0.5	55.2	66.6
International Avg.	1999	.	54.7	5.0	9.3	11.8	17.4	1.9	54.7	0.2	0.0	1.7	53.2	56.2

DIFF = Item Difficulty Other Incorrect = Sum of invalid, not reached and omitted
Because of missing gender information, some totals may appear inconsistent

to compare performance from one administration to the next. At the same time, the display can be used to detect item-by-country interactions across all countries.

10.3 Scoring Reliability

About one-third of the items in the TIMSS 2003 assessment were constructed-response items, comprising nearly half of the score points for the assessment.[5] An essential requirement for use of such items is that they be reliably scored by all participants. That is, a particular student response should receive the same score, regardless of the scorer. In conducting TIMSS 2003, measures taken to ensure that the constructed-response items were scored reliably in all countries included developing scoring guides for each constructed-response question (which provided descriptions of acceptable responses for each score point value),[6] and providing extensive training in the application of the scoring guides. Scoring procedures for organizing and monitoring the scoring sessions were outlined in the *TIMSS 2003 Survey Operations Manual* (TIMSS, 2002).

10.3.1 Within-Country Scoring Reliability

To gather and document information about the within-country agreement among scorers, a random sample of at least 200 student responses to each item was selected to be scored independently by two scorers.[7] The inter-rater agreement for each item in each country was examined as part of the item review process. The average and range of the within-country exact percent of agreement across all items is presented in Exhibit 10.6 for mathematics and Exhibit 10.7 for science at both grades. Agreement across items was high – on average across countries, exact percent agreement was 99 percent at both grades in mathematics and 97 percent at the eighth grade and 96 at the fourth grade in science. All countries had an average exact percent agreement above 96 percent at the eighth grade and 97 at the fourth grade in mathematics and above 90 percent at the eighth grade and 91 at the fourth grade in science.

10.3.2 Trend Item Scoring Reliability

The double scoring of a sample of the student test booklets provided a measure of the consistency within each country with which constructed-response questions were scored. TIMSS 2003 also took steps to show that those constructed-response items from 1999 that were used in 2003 were scored in the same way in both assessments. In anticipation of this, countries that participated in TIMSS 1999 sent samples of scored student booklets from the 1999

5 For details on the development of the TIMSS 2003 assessment items, see Chapter 2.

6 Discussion of the development of the scoring guides for constructed-response items is provided in Chapter 2.

7 Since individual items appear in at least two booklets, 100 of each of the 12 booklets were chosen randomly for double-scoring. For a sample of 4500, this amounts to almost 25% of the total sample.

Exhibit 10.4 **Example Item-by-Country Interaction Display for Item M012040**

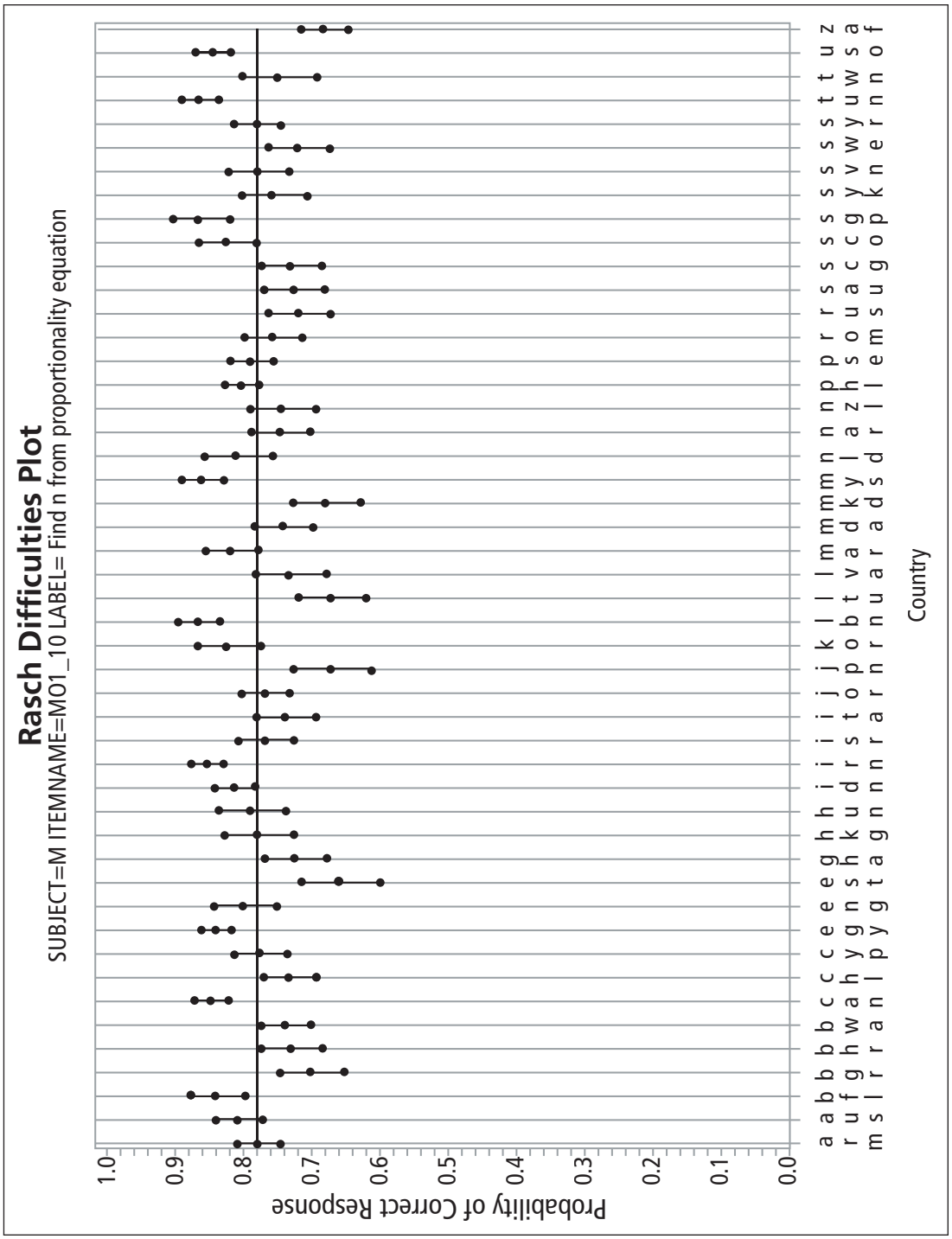
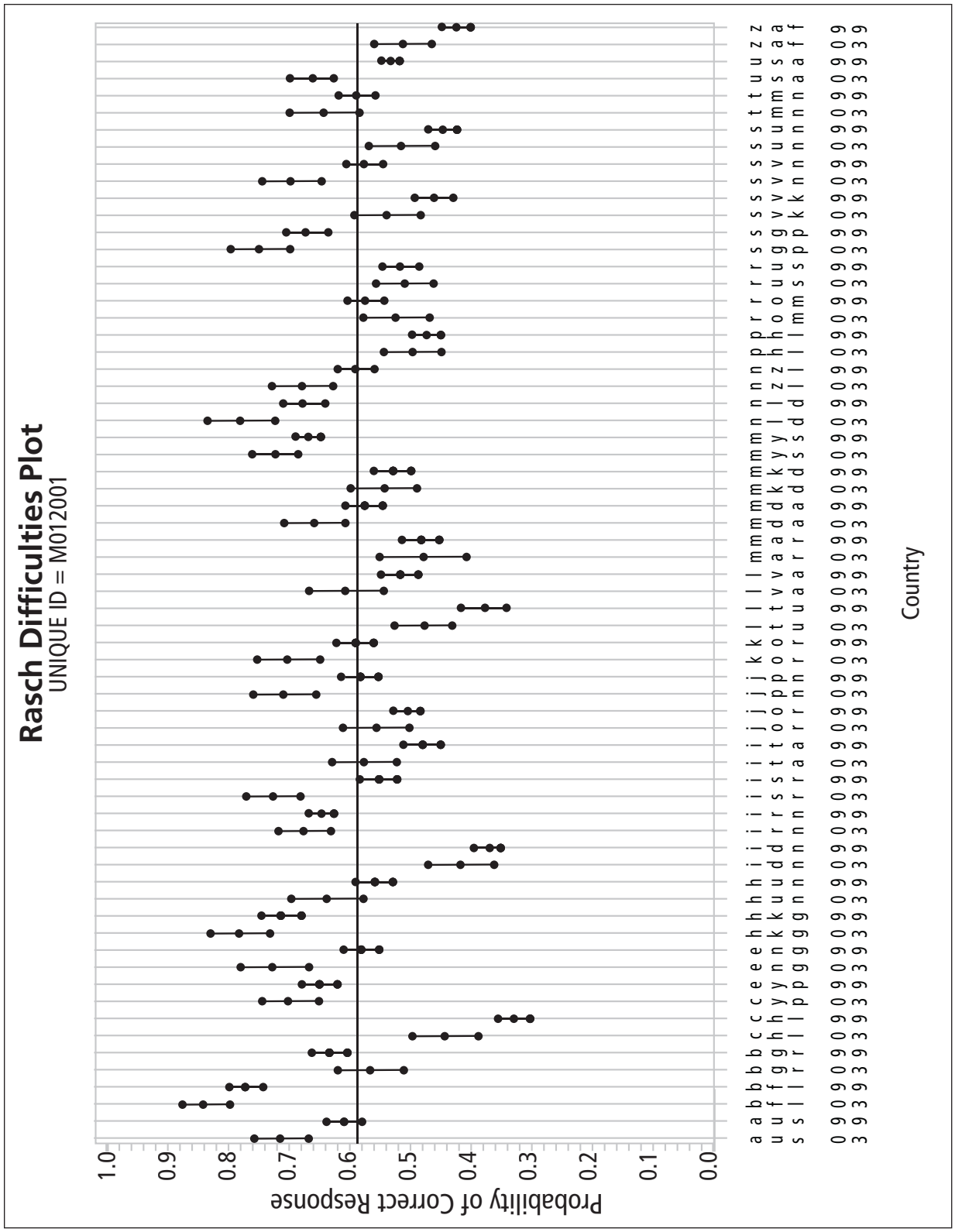


Exhibit 10.5 **Example Item-by-Country Interaction Display for Trend Item M012001**



eighth-grade data collection to the IEA Data Processing Center, where they were digitally scanned and stored in presentation software for later use. As a check on scoring consistency from 1999 to 2003, staff members working in each country on scoring the 2003 eighth-grade data were asked also to score these 1999 responses using the DPC software. The items from 1995 that were used in TIMSS 2003 all were in multiple-choice format, and therefore scoring reliability was not an issue. As shown in Exhibit 10.8 for mathematics and Exhibit 10.9 for science, there was a very high degree of scoring consistency, with 98 percent exact agreement in mathematics, on average, internationally, between the scores awarded in 1999 and those given by the 2003 scorers and 92 percent in science. There also was high agreement at the diagnostic score level, with 93 percent exact agreement, on average, in mathematics and somewhat less, 81 percent, in science.

10.3.3 Cross-Country Scoring Reliability Study

Although because of the many different languages in use in TIMSS, establishing the reliability of constructed-response scoring across all countries was not feasible, TIMSS 2003 did conduct a cross-country study of scoring reliability among northern-hemisphere countries whose scorers were proficient in English.[8] A sample of student responses to a subset of the eighth-grade mathematics and science constructed-response items was provided by the English-speaking southern hemisphere countries.

A sample of 150 student responses to each of 20 mathematics items and 21 science items (41 in total, representing about one-quarter of constructed-response items at eighth grade) was collected from Australia, Botswana, New Zealand, and Singapore. This set of 6,150 student responses in English was scored independently in each country that had at least one but preferably two scorers proficient in English. In all, 37 scorers from 20 countries participated in the study. Scoring for this study took place shortly after the within-country scoring reliability activities were completed. Making all possible comparisons among scorers gave 666 comparisons for each student response to each item, and 99,900 total comparisons when aggregated across all 150 student responses to that item. Agreement across countries was defined in terms of the percentage of these comparisons that were in exact agreement. Exhibits 10.10 and 10.11 show that scorer reliability across countries was high, with the percent exact agreement averaging 96 percent across the 20 mathematics items for the correctness score and 92 percent for the diagnostic score, and averaging 87 percent across the 21 science items for the correctness score and 76 percent for the diagnostic score.

8 See Chapter 6 for further details.

10.4 Item Review Procedures

The International Study Center thoroughly reviewed the item statistics for all participating countries to ensure that items were performing comparably across countries. In particular, items with the following problems were considered for possible deletion from the international database:

- An error was detected during TIMSS 2003 translation verification but was not corrected before test administration.
- Data checking revealed a multiple-choice item with more or fewer options than in the international version.
- The item analysis showed the item to have a negative biserial, or, for an item with more than one score point, a nonmonotonic relationship between score level and total score.
- The item-by-country interaction results showed a very large negative interaction for a particular country.
- For constructed-response items, the within-country scoring reliability data showed an agreement of less than 70 percent.
- For trend items, an item performed substantially differently in 1999 compared to 2003, or an item was not included in the 1999 assessment for a particular country.

When the item statistics indicated a problem with an item, the documentation from the translation verification[9] was used as an aid in checking the test booklets. If a question remained about potential translation or cultural issues, however, then the National Research Coordinator (NRC) was consulted before deciding how the item should be treated. If a problem could be detected by the International Study Center (such as a negative point-biserial for a correct answer or too few options for a multiple-choice item), the item was deleted from the international scaling.

The checking of the TIMSS 2003 achievement data involved 696 items for 49 countries and four Benchmarking participating at both grades (approximately 37,000 item-country combinations), and resulted in the detection of very few items that were inappropriate for international comparisons. Among the few items singled out in the review process were mostly items with differences attributable to either translation or printing problems. Appendix C provides a list of deleted items as well as a list of recodes made to constructed-response item codes.

9 See Chapter 4 for a description of the process for translating and verifying the TIMSS 2003 data-collection instruments.

Exhibit 10.6 **TIMSS 2003 Within-Country Constructed-Response Scoring Reliability Mathematics Items – Eighth Grade**

Countries	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Armenia	99	94	100	98	92	100
Australia	100	97	100	99	95	100
Bahrain	99	98	100	98	91	100
Belgium (Flemish)	99	96	100	98	91	100
Botswana	99	91	100	94	81	100
Bulgaria	96	70	100	92	64	99
Chile	99	95	100	97	91	100
Chinese Taipei	100	91	100	99	91	100
Cyprus	98	86	100	96	79	100
Egypt	100	97	100	99	97	100
England	99	93	100	98	91	100
Estonia	100	98	100	99	96	100
Ghana	99	97	100	95	90	99
Hong Kong, SAR	100	98	100	99	98	100
Hungary	98	90	100	96	80	100
Indonesia	98	90	100	94	82	100
Iran, Islamic Rep. of	99	94	100	96	90	100
Israel	98	93	100	93	83	99
Italy	99	95	100	98	92	100
Japan	99	94	100	98	91	100
Jordan	99	98	100	98	92	100
Korea, Rep. of	99	87	100	98	87	100
Latvia	98	90	100	96	79	100
Lebanon	100	94	100	99	91	100
Lithuania	97	71	100	95	62	100
Macedonia, Rep. of	100	97	100	99	95	100
Malaysia	100	98	100	99	97	100
Moldova, Rep. of	100	99	100	100	99	100
Morocco	97	89	100	92	82	99
Netherlands	97	84	100	95	78	100
New Zealand	99	96	100	97	88	100
Norway	98	91	100	96	86	100
Palestinian Nat'l Auth.	99	94	100	97	88	100

Exhibit 10.6 **TIMSS 2003 Within-Country Constructed-Response Scoring Reliability Mathematics Items – Eighth Grade** *(...Continued)*

Countries	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Philippines	99	97	100	97	92	100
Romania	100	98	100	99	94	100
Russian Federation	99	95	100	97	89	100
Saudi Arabia	99	94	100	95	81	99
Scotland	99	95	100	98	92	100
Serbia	99	96	100	98	94	100
Singapore	100	98	100	100	98	100
Slovak Republic	100	98	100	99	96	100
Slovenia	97	86	100	94	75	100
South Africa	99	95	100	97	90	99
Sweden	98	89	100	95	84	99
Tunisia	98	89	100	95	78	99
United States	97	86	100	94	75	99
International Avg.	**99**	**92**	**100**	**97**	**87**	**100**
Benchmarking Participants						
Basque Country, Spain	98	87	100	96	83	100
Indiana State, US	98	88	100	95	76	100
Ontario Province, Can.	97	80	100	93	72	100
Quebec Province, Can.	97	81	100	94	79	100

Exhibit 10.6 **TIMSS 2003 Within-Country Constructed-Response Scoring Reliability Mathematics Items – Fourth Grade**

Countries	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Armenia	99	98	100	98	95	100
Australia	100	98	100	99	97	100
Belgium (Flemish)	100	96	100	98	87	100
Chinese Taipei	99	83	100	97	76	100
Cyprus	98	91	100	95	82	100
England	99	91	100	98	90	100
Hong Kong, SAR	100	98	100	99	87	100
Hungary	98	91	100	95	78	100
Iran, Islamic Rep. of	100	98	100	99	96	100
Italy	98	92	100	96	81	100
Japan	99	95	100	98	94	100
Latvia	98	87	100	96	78	100
Lithuania	97	77	100	94	69	100
Moldova, Rep. of	100	100	100	100	100	100
Morocco	98	93	100	94	86	98
Netherlands	97	86	100	94	73	100
New Zealand	99	94	100	96	85	100
Norway	99	95	100	97	92	100
Philippines	99	96	100	97	91	100
Russian Federation	100	97	100	99	96	100
Scotland	99	98	100	98	93	100
Singapore	100	99	100	100	99	100
Slovenia	98	84	100	96	73	100
Tunisia	97	89	100	91	77	98
United States	97	88	100	95	82	100
International Avg.	**99**	**92**	**100**	**97**	**86**	**100**
Benchmarking Participants						
Indiana State, US	99	92	100	96	83	100
Ontario Province, Can.	98	87	100	96	84	100
Quebec Province, Can.	98	92	100	96	86	100

Exhibit 10.7 **TIMSS 2003 Within-Country Constructed-Response Scoring Reliability Science Items – Eighth Grade**

Countries	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Armenia	98	92	100	97	90	100
Australia	99	94	100	97	89	100
Bahrain	98	94	100	95	85	100
Belgium (Flemish)	97	89	100	93	83	100
Botswana	95	74	100	87	74	97
Bulgaria	91	72	99	84	64	99
Chile	97	91	100	94	89	99
Chinese Taipei	99	97	100	98	86	100
Cyprus	96	87	100	91	80	99
Egypt	100	98	100	100	97	100
England	98	92	100	96	85	100
Estonia	99	97	100	98	88	100
Ghana	98	93	100	93	83	99
Hong Kong, SAR	99	97	100	97	92	100
Hungary	96	87	100	92	83	100
Indonesia	96	87	100	86	68	99
Iran, Islamic Rep. of	98	87	100	95	84	100
Israel	95	89	100	84	66	98
Italy	98	91	100	96	90	100
Japan	97	81	100	93	80	100
Jordan	99	97	100	96	91	100
Korea, Rep. of	98	84	100	95	74	100
Latvia	94	78	100	87	50	100
Lebanon	100	98	100	99	95	100
Lithuania	90	69	100	82	58	100
Macedonia, Rep. of	99	96	100	97	92	100
Malaysia	99	98	100	99	97	100
Moldova, Rep. of	100	99	100	100	99	100
Morocco	94	86	100	86	69	95
Netherlands	90	70	100	84	61	100
New Zealand	98	92	100	93	84	100
Norway	95	83	100	91	80	100
Palestinian Nat'l Auth.	95	82	100	87	69	99
Philippines	98	89	100	94	83	99
Romania	99	96	100	98	94	100
Russian Federation	99	92	100	98	91	100

Exhibit 10.7 **TIMSS 2003 Within-Country Constructed-Response Scoring Reliability Science Items – Eighth Grade** *(...Continued)*

Countries	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Saudi Arabia	97	87	100	91	68	99
Scotland	97	89	100	94	85	100
Serbia	99	94	100	98	92	100
Singapore	100	99	100	99	98	100
Slovak Republic	99	95	100	97	89	100
Slovenia	90	70	100	81	61	100
South Africa	99	94	100	96	88	99
Sweden	92	76	100	85	68	99
Tunisia	98	90	100	94	73	100
United States	92	72	100	83	68	99
International Avg.	**97**	**88**	**100**	**92**	**80**	**99**
Benchmarking Participants						
Basque Country, Spain	96	87	100	92	79	100
Indiana State, US	94	82	100	87	67	100
Ontario Province, Can.	91	77	100	83	62	98
Quebec Province, Can.	92	80	100	84	66	100

Exhibit 10.7 **TIMSS 2003 Within-Country Constructed-Response Scoring Reliability Science Items – Fourth Grade**

Countries	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Armenia	99	97	100	97	91	100
Australia	99	94	100	98	91	100
Belgium (Flemish)	99	89	100	95	86	100
Chinese Taipei	98	89	100	96	89	100
Cyprus	94	76	100	89	75	99
England	98	87	100	96	86	100
Hong Kong, SAR	99	97	100	97	89	100
Hungary	95	80	100	91	78	100
Iran, Islamic Rep. of	96	85	100	93	83	99
Italy	94	77	100	90	77	100
Japan	97	86	100	94	83	100
Latvia	96	82	100	92	71	99
Lithuania	93	81	100	86	50	99
Moldova, Rep. of	100	100	100	100	100	100
Morocco	97	93	100	92	78	99
Netherlands	91	71	99	84	70	99
New Zealand	97	86	100	92	83	99
Norway	97	85	100	93	84	100
Philippines	97	89	100	91	77	99
Russian Federation	99	98	100	99	96	100
Scotland	98	90	100	96	85	100
Singapore	100	99	100	99	97	100
Slovenia	91	74	100	85	69	100
Tunisia	93	79	100	82	68	96
United States	93	70	100	86	68	99
International Avg.	**96**	**85**	**100**	**92**	**80**	**99**
Benchmarking Participants						
Indiana State, US	95	76	100	92	62	100
Ontario Province, Can.	95	80	100	90	75	100
Quebec Province, Can.	95	81	100	89	72	99

Exhibit 10.8 **TIMSS 2003 Trend Item Scoring Reliability Mathematics Items – Eighth Grade**

Countries	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Australia	98	88	100	94	73	100
Belgium (Flemish)	98	92	100	94	78	100
Bulgaria	99	82	100	94	71	100
Chile	99	97	100	92	73	100
Chinese Taipei	98	95	100	94	79	100
Cyprus	98	91	100	94	79	100
Hong Kong, SAR	98	91	100	96	84	100
Hungary	98	89	100	95	86	100
Indonesia	98	90	100	93	60	100
Iran, Islamic Rep.	98	83	100	89	24	99
Israel	98	91	100	92	74	100
Italy	99	91	100	97	86	100
Japan	98	87	100	96	76	100
Jordan	99	96	100	96	87	100
Korea, Rep. of	98	88	100	94	67	100
Latvia	90	34	100	78	32	100
Lithuania	98	93	100	94	74	100
Macedonia, Rep. of	99	85	100	96	70	100
Malaysia	99	91	100	95	84	100
New Zealand	99	96	100	94	85	100
Philippines	99	86	100	95	75	100
Romania	99	97	100	97	90	100
Russian Federation	98	94	100	92	62	100
Singapore	99	96	100	98	89	100
Slovak Republic	93	54	100	87	50	99
Slovenia	99	95	100	95	81	100
South Africa	99	92	100	93	47	100
United States	98	91	100	94	76	100
International Avg.	**98**	**88**	**100**	**93**	**72**	**100**
Benchmarking Participants						
Ontario Province, Can.	98	85	100	93	65	100
Quebec Province, Can.	98	85	100	93	65	100

Exhibit 10.9 **TIMSS 2003 Trend Item Scoring Reliability Science Items – Eighth Grade**

Countries	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Australia	93	75	100	81	56	100
Belgium (Flemish)	92	79	100	83	68	100
Bulgaria	96	87	100	83	45	100
Chile	91	80	100	77	47	100
Chinese Taipei	92	70	100	80	38	100
Cyprus	90	70	99	79	50	99
Hong Kong, SAR	89	74	100	80	58	100
Hungary	92	74	100	84	64	100
Indonesia	90	63	100	75	41	97
Iran, Islamic Rep.	92	68	100	82	55	99
Israel	93	80	100	81	46	100
Italy	94	86	100	88	73	100
Japan	92	72	100	84	62	100
Jordan	96	90	100	87	76	99
Korea, Rep. of	93	77	100	85	56	100
Latvia	79	36	100	65	21	98
Lithuania	86	66	100	74	40	100
Macedonia, Rep. of	99	89	100	98	80	100
Malaysia	92	80	100	74	35	100
New Zealand	94	87	99	79	52	98
Philippines	90	44	100	76	32	100
Romania	96	91	100	90	73	100
Russian Federation	93	80	100	79	55	99
Singapore	97	93	100	88	61	100
Slovak Republic	89	73	100	76	56	100
Slovenia	94	71	100	90	72	100
South Africa	93	71	100	79	19	100
United States	94	83	100	84	70	100
International Avg.	**92**	**75**	**100**	**82**	**54**	**100**
Benchmarking Participants						
Ontario Province, Can.	91	76	100	81	60	100
Quebec Province, Can.	91	76	100	81	60	100

Exhibit 10.10 Cross-Country Constructed-Response Scoring Reliability Data for Mathematics Items

Item Label	Total Valid Comparisons	Exact Percent Agreement	
		Correctness Score Agreement	Diagnostic Score Agreement
M022202	99900	99	98
M022156	99900	99	91
M022012	99900	94	86
M022261A	99900	99	98
M022261B	99900	99	98
M022261C	99900	90	84
M022227A	99900	99	99
M022227B	99900	97	90
M022227C	99900	94	86
M022234A	99900	95	88
M022234B	99900	91	87
M022110	99900	98	93
M032691	99900	98	94
M032640	99900	93	93
M032683	99900	92	85
M032681A	99900	99	99
M032681B	99900	99	98
M032681C	99900	97	97
M032233	99900	93	91
M032692	99900	95	95
Average		**96**	**92**

Exhibit 10.11 **Cross-Country Constructed-Response Scoring Reliability Data for Science Items**

Item Label	Total Valid Comparisons	Exact Percent Agreement	
		Correctness Score Agreement	Diagnostic Score Agreement
S032202	99900	83	73
S022283	99900	93	86
S022154	99900	83	70
S022191	99900	94	83
S022088A	99900	83	72
S022088B	99900	76	61
S022286	99900	91	77
S032625A	99900	97	94
S032625B	99900	92	72
S032120A	99900	78	61
S032120B	99900	87	69
S032063	99900	81	73
S032306	99900	88	83
S032640	99900	89	79
S032272	99900	95	88
S032650A	99900	90	84
S032650B	99900	87	80
S032056	99900	88	74
S032369	99900	80	71
S032565	99900	90	78
S032516	99900	84	74

10.5 Item Position in Booklet

As described in Chapter 2, TIMSS has a complicated student booklet design. Although each student completes just one booklet, there are 12 different student booklets at each grade level, with six blocks of mathematics and science items in each booklet. As illustrated in Exhibit 10.12, blocks of items appear in different positions in different booklets. For example, the items in block M1 appear as the first block in Booklet 1, as the second block in Booklet 6, and as the third block in Booklet 12. This allows the booklets to be linked together efficiently, but also to monitor and counterbalance any position effect.

An important step in the item review process, made possible by the counterbalanced booklet design, was to compare the characteristics of item blocks appearing in different booklet positions to detect any position effect. As the item statistics for each country were reviewed during this step, it became apparent that there was indeed an unexpectedly strong position effect in the data. As may be seen from Exhibit 10.13, this position effect occurred because some students in all countries did not reach all the items in the third block position, which was the end of the first half of each booklet before the break. The same effect was evident for the sixth block position, which was the last block in the booklets.

As described in Chapter 11, TIMSS addressed this problem using IRT scaling by treating items in the third and sixth block positions as if they were unique, even though they also appeared in other positions. For example, the mathematics items in block M1 from Booklet 1 (the first position) and from Booklet 6 (second position) were considered to be the same items for scaling and reporting purposes, but those in Booklet 12 (the third position) were scaled as items that were different and unique. This approach allowed all student responses to all items to be included in the calibration of the IRT scale and in estimating student achievement scores, while taking into account the booklet position effect. However, because items in blocks appearing in the third and sixth booklet positions were judged to have different properties to those same items when appearing in positions one, two, four, and five, student responses to items in positions three and six were not included when computing percent correct for individual example items, item statistics for use in scale anchoring, or average percent correct for measuring trends in mathematics or science content areas (see Chapter 12).

Exhibit 10.12 **TIMSS 2003 Booklet Design (Adapted from Exhibit 2.16)**

	Part 1			Part 2		
Booklet	**Position 1**	**Position 2**	**Position 3**	**Position 4**	**Position 5**	**Position 6**
1	M01	M02	S06	S07	M05	M07
2	M02	M03	S05	S08	M06	M08
3	M03	M04	S04	S09	M13	M11
4	M04	M05	S03	S10	M14	M12
5	M05	M06	S02	S11	M09	M13
6	M06	M01	S01	S12	M10	M14
7	S01	S02	M06	M07	S05	S07
8	S02	S03	M05	M08	S06	S08
9	S03	S04	M04	M09	S13	S11
10	S04	S05	M03	M10	S14	S12
11	S05	S06	M02	M11	S09	S13
12	S06	S01	M01	M12	S10	S14

Exhibit 10.13 **Average Percent Not Reached, by Booklet Position**

	Average Percent Not Reached					
	Position 1	**Position 2**	**Position 3**	**Position 4**	**Position 5**	**Position 6**
Grade 8						
Mathematics	0.4	2.5	8.4	0.3	0.9	7.7
Science	0.5	1.2	13.2	0.4	1.0	8.3
Grade 4						
Mathematics	1.1	5.4	10.9	0.7	3.1	13.2
Science	0.7	2.3	17.0	0.7	3.1	13.3

References

TIMSS (2002), *TIMSS 2003 Survey Operations Manual*, prepared by the International Study Center, Chestnut Hill, MA: Boston College.