Reviewing the TIMSS and PIRLS 2011 Achievement Item Statistics

Pierre Foy
Michael O. Martin
Ina V.S. Mullis
Gabrielle Stanco

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 across the countries that participated in the 2011 TIMSS and PIRLS assessments. This review of item statistics is essential to the successful application of item response theory (IRT) scaling to derive student achievement scores for analysis and reporting. This review played a crucial role in the quality assurance of the 2011 TIMSS and PIRLS achievement data prior to scaling, making it possible to detect 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 guide 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.

Statistics for Item Review

The TIMSS & PIRLS International Study Center computed item statistics for all achievement items in the 2011 assessments, including TIMSS fourth grade (179 mathematics items and 172 science items), TIMSS eighth grade (217 mathematics items and 217 science items), PIRLS (135 items), and prePIRLS (123 items). The item statistics for each of the participating countries were then carefully reviewed. Exhibits 1 and 2 show actual samples of the statistics calculated for a multiple choice and a constructed response item, respectively.

Exhibit 1: International Item Statistics for a Multiple Choice Item

Trends in International Mathematics and Science Study - TIMSS 2011 Assessment Results
International Item Review Statistics (Unweighted) - 4th Grade

15:33 February 17, 2012 48

Confidential Review: DO NOT CITE OR CIRCULATE

Mathematics: Data Display / Applying (M04_12 - M051077) Type: MC Key: C
Label: What students did after school

Country	Cases	DIFF	DISC	Percentages P_A	P_B	P_C	P_D	P_OM	P_NR	Point Biserials PB_A	PB_B	PB_C	PB_D	PB_OM	PB_NR	RDIFF	Flags
Armenia	721	26.3	0.38	39.1	13.6	26.3	8.0	13.1	15.1	-0.13	-0.07	0.38	-0.08	-0.13	-0.04	0.80	__H_F_
*Australia	881	50.6	0.54	34.7	8.2	50.6	3.8	2.6	2.3	-0.32	-0.19	0.54	-0.17	-0.16	-0.08	0.15	_____F_
*Austria	668	51.0	0.52	32.6	8.2	51.0	4.4	3.8	1.6	-0.40	-0.07	0.52	-0.04	-0.18	-0.10	0.07	_____F_
Azerbaijan	696	43.9	0.49	30.4	8.1	43.9	7.5	10.1	10.1	-0.23	-0.15	0.49	-0.17	-0.10	-0.08	0.07	_____F_
Bahrain	574	31.8	0.51	44.2	15.0	31.8	7.1	2.0	4.5	-0.23	-0.20	0.51	-0.07	-0.12	-0.07	0.28	_____F_
Belgium (Flemish)	704	59.4	0.45	29.7	6.7	59.4	2.0	2.2	2.8	-0.34	-0.21	0.45	-0.06	-0.04	-0.01	0.35	_____F_
Chile	786	41.2	0.48	38.1	12.7	41.2	6.0	2.0	6.7	-0.24	-0.18	0.48	-0.19	-0.09	-0.01	0.06	_____F_
Chinese Taipei	613	64.1	0.52	27.5	5.9	64.1	2.5	0.2	0.2	-0.42	-0.16	0.52	-0.15	0.03	-0.06	0.60	__H_F_
Croatia	646	45.4	0.53	37.5	7.3	45.4	5.9	3.9	5.0	-0.35	-0.11	0.53	-0.11	-0.16	-0.07	0.11	_____F_
*Czech Republic	663	60.1	0.56	28.4	5.3	60.1	2.4	3.8	3.9	-0.42	-0.11	0.55	-0.13	-0.13	-0.10	-0.24	__E_F_
Denmark	547	58.8	0.45	28.2	5.4	58.8	2.5	5.0	5.9	-0.28	-0.21	0.45	-0.11	-0.17	-0.02	0.09	_____F_
*England	480	57.6	0.51	29.7	7.6	57.6	3.7	1.5	3.8	-0.31	-0.19	0.51	-0.19	-0.12	-0.10	0.14	_____F_
Finland	665	69.3	0.53	20.1	5.3	69.3	3.3	2.0	1.2	-0.42	-0.15	0.53	-0.15	-0.07	-0.09	-0.37	__E_F_
Georgia	679	35.5	0.54	42.9	10.0	35.5	7.2	4.5	8.2	-0.32	-0.11	0.54	-0.13	-0.12	-0.03	0.10	_____F_
Germany	569	55.4	0.50	29.3	8.6	55.4	3.2	3.4	2.3	-0.32	-0.21	0.49	-0.05	-0.11	-0.11	0.17	_____F_
*Hong Kong SAR	575	64.3	0.46	23.8	7.1	64.3	4.0	0.7	0.0	-0.28	-0.24	0.46	-0.14	-0.14	.	0.77	__H_F_
*Hungary	730	55.6	0.57	30.2	8.5	55.6	3.7	2.0	4.4	-0.39	-0.16	0.57	-0.07	-0.06	-0.20	0.16	_____F_
*Iran, Islamic Rep. of	823	24.5	0.34	47.5	12.7	24.5	8.3	7.1	10.7	-0.07	-0.08	0.34	-0.14	-0.13	-0.04	0.70	_CH_F_
*Ireland	647	54.7	0.57	34.6	5.6	54.7	3.3	1.7	1.4	-0.45	-0.17	0.57	-0.06	-0.09	-0.11	0.07	_____F_
*Italy	605	45.1	0.54	41.8	7.3	45.1	4.3	1.4	4.8	-0.37	-0.15	0.54	-0.13	-0.12	-0.04	0.29	_____F_
*Japan	628	71.5	0.40	19.3	5.3	71.5	2.1	1.8	1.6	-0.30	-0.10	0.40	-0.17	-0.04	-0.10	0.05	_____F_
Kazakhstan	630	44.9	0.49	37.9	8.7	44.9	6.0	2.5	5.2	-0.30	-0.11	0.49	-0.11	-0.05	-0.14	0.34	_____F_
*Korea, Rep. of	625	76.7	0.50	15.4	5.6	76.7	2.1	0.2	0.5	-0.45	-0.18	0.50	-0.05	-0.01	-0.01	0.24	_____F_
Kuwait	581	16.8	0.30	43.7	19.7	16.8	13.8	5.9	6.7	0.01	-0.11	0.29	-0.12	-0.10	-0.01	0.26	_C____
Lithuania	664	58.4	0.56	30.6	7.8	58.4	2.0	1.2	2.0	-0.43	-0.21	0.56	-0.07	-0.08	-0.05	0.11	_____F_
Malta	522	36.3	0.45	42.5	11.2	36.3	8.0	2.0	2.3	-0.24	-0.17	0.45	-0.16	-0.06	0.04	0.62	__H_F_
Morocco	1092	13.4	0.15	41.4	30.9	13.4	9.7	4.5	10.7	0.07	-0.01	0.15	-0.14	-0.10	-0.05	0.69	_CH_F_
*Netherlands	453	75.2	0.42	16.2	5.3	75.2	2.2	1.1	0.4	-0.33	-0.18	0.42	-0.09	-0.10	-0.02	-0.69	__E_F_
*New Zealand	786	39.3	0.48	45.7	8.4	39.3	4.8	1.8	2.5	-0.30	-0.17	0.48	-0.10	-0.06	-0.08	0.30	_____F_
Northern Ireland	514	65.8	0.56	26.2	6.1	65.8	1.4	0.6	0.6	-0.46	-0.15	0.56	-0.12	-0.02	-0.12	0.20	_____F_
*Norway	450	44.1	0.49	40.3	8.7	44.1	4.5	2.4	5.8	-0.32	-0.16	0.49	-0.16	-0.08	0.00	0.32	_____F_
Oman	1491	21.5	0.43	43.8	15.4	21.5	14.7	4.7	7.0	-0.04	-0.14	0.43	-0.21	-0.09	-0.06	0.37	_CH____
Poland	712	47.4	0.54	39.7	5.6	47.4	3.0	4.3	5.5	-0.37	-0.12	0.54	-0.08	-0.10	-0.13	-0.12	__E_F_
*Portugal	572	56.3	0.48	30.0	8.5	56.3	3.2	2.0	3.1	-0.36	-0.18	0.48	-0.15	0.00	-0.01	0.25	_____F_
Qatar	585	26.5	0.51	42.5	16.0	26.5	11.6	3.3	2.7	-0.13	-0.22	0.51	-0.19	-0.12	0.00	0.33	_____
Romania	673	48.3	0.61	37.1	8.5	48.3	3.7	2.3	4.3	-0.41	-0.12	0.61	-0.11	-0.15	-0.16	0.17	_____F
Russian Federation	626	66.9	0.51	20.1	6.1	66.9	4.9	2.0	3.0	-0.40	-0.13	0.51	-0.12	-0.09	-0.10	-0.21	__E_F_
Saudi Arabia	513	27.7	0.42	41.2	17.5	27.7	11.8	1.8	2.1	-0.15	-0.16	0.42	-0.19	-0.07	0.09	0.28	_____
Serbia	623	55.9	0.53	28.8	5.4	55.9	5.8	4.0	8.2	-0.36	-0.08	0.53	-0.14	-0.14	-0.10	0.01	_____F
*Singapore	906	78.1	0.54	12.2	8.0	78.1	1.6	0.2	0.3	-0.43	-0.22	0.54	-0.17	-0.01	-0.06	-0.07	__E_F_
Slovak Republic	800	50.1	0.56	35.3	7.6	50.1	4.7	2.3	3.3	-0.35	-0.18	0.56	-0.15	-0.10	-0.10	0.03	_____F_
*Slovenia	644	53.4	0.51	30.8	7.7	53.4	5.6	2.6	5.1	-0.36	-0.16	0.51	-0.13	-0.12	-0.02	-0.12	__E_F_
Spain	601	43.3	0.54	38.5	9.7	43.3	5.1	3.4	5.8	-0.36	-0.09	0.54	-0.09	-0.17	-0.07	0.22	_____F_
Sweden	681	58.6	0.52	28.4	6.4	58.6	3.9	2.7	6.5	-0.38	-0.20	0.52	-0.16	-0.12	0.02	-0.43	__E_F_
Thailand	638	37.6	0.42	38.6	15.0	37.6	7.0	1.8	2.0	-0.19	-0.11	0.42	-0.22	-0.07	-0.05	0.28	_____F_
Tunisia	698	23.0	0.22	35.8	17.9	23.0	11.2	12.1	15.9	0.03	-0.05	0.22	-0.07	-0.20	0.00	0.13	_C____
Turkey	1065	38.1	0.51	41.5	10.6	38.1	8.4	1.4	2.5	-0.24	-0.18	0.51	-0.21	-0.02	-0.10	0.38	__H_F_
United Arab Emirates	2107	30.4	0.52	44.4	13.7	30.4	8.4	3.2	4.1	-0.20	-0.20	0.52	-0.14	-0.10	-0.08	0.36	__H_F_
*United States	1798	57.0	0.50	32.2	5.1	57.0	2.9	2.8	3.6	-0.39	-0.11	0.49	-0.16	-0.04	-0.05	0.26	__H_F_
Yemen	1157	19.2	0.19	25.0	31.2	19.2	17.4	7.2	5.7	0.12	-0.11	0.19	-0.09	-0.17	0.01	-0.54	_CEA__
*Reference Avg (n=18)	12934	56.4	0.50	30.3	7.4	56.4	3.7	2.2	3.1	-0.35	-0.16	0.50	-0.13	-0.09	-0.07	0.15	_____F_
International Avg (n=50)	37107	47.5	0.48	33.5	10.1	47.5	5.7	3.2	4.5	-0.29	-0.15	0.48	-0.13	-0.10	-0.06	0.17	_____F_
Botswana	604	23.4	0.41	48.9	10.4	23.4	16.6	0.7	2.5	-0.08	-0.20	0.40	-0.19	0.05	-0.03	0.46	_CH____
Honduras	560	17.4	0.43	52.7	14.6	17.4	11.5	3.8	6.8	-0.17	-0.08	0.43	-0.03	-0.07	-0.09	0.49	_CH____
Yemen (6)	703	21.1	0.34	38.6	22.9	21.1	13.4	4.0	3.7	0.02	-0.19	0.34	-0.10	-0.11	-0.07	-0.01	_C_____
Alberta, Canada	515	55.2	0.44	31.9	8.3	55.2	2.0	2.6	3.9	-0.34	-0.15	0.44	-0.13	0.00	-0.01	-0.16	__E_F_
Ontario, Canada	672	52.6	0.47	34.1	7.5	52.6	4.5	1.2	4.5	-0.36	-0.17	0.47	-0.09	-0.02	-0.01	0.16	_____F_
Quebec, Canada	597	54.4	0.43	32.1	8.2	54.4	3.7	1.7	3.9	-0.29	-0.18	0.43	-0.10	-0.08	-0.03	0.11	_____F_
Abu Dhabi, UAE	601	24.1	0.42	44.8	16.9	24.1	10.2	4.1	3.3	-0.11	-0.16	0.42	-0.10	-0.12	-0.06	0.46	_CH____
Dubai, UAE	873	37.0	0.56	43.0	10.3	37.0	6.9	2.9	4.0	-0.28	-0.22	0.56	-0.13	-0.11	-0.08	0.28	_____F_
Florida, US	380	54.4	0.56	32.3	8.1	54.4	3.0	2.2	2.4	-0.38	-0.27	0.56	-0.06	-0.05	-0.08	0.34	_____F_
North Carolina, US	248	57.9	0.48	33.3	3.9	57.9	2.6	2.2	8.1	-0.33	-0.09	0.48	-0.13	-0.11	-0.12	0.43	_____F_
Netherlands (Natl)	136	68.7	0.55	24.6	4.5	68.7	0.7	1.5	1.5	-0.47	-0.18	0.55	-0.02	-0.09	-0.03	-0.45	__E_F_
Norway (5)	189	58.7	0.47	27.7	9.2	58.7	2.2	2.2	2.6	-0.31	-0.27	0.47	-0.06	0.03	-0.08	0.20	_____F_

Keys: DIFF= Percent correct score; DISC= Item discrimination; P_A...P_D= Percentage choosing each option; P_OM, P_NR= Percentage Omitted, Not Reached;
PB_A...PB_D= Point Biserial for each option; PB_OM, PB_NR= Point Biserial for Omitted, Not Reached; RDIFF= Rasch difficulty.
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 less than 70%; V= Difficulty greater than 95%.

Exhibit 2: International Item Statistics for a Constructed Response Item

Acquire and Use Information - Honey (R31W02C) How the Boran and honeyguide help
Interpret and Integrate Ideas and Information - Type: CR 2 Points

Country	Cases	DIFF	DISC	Percentages P_0	P_1	P_2	P_3	P_OM	P_NR	Point Biserials PB_0	PB_1	PB_2	PB_3	PB_OM	PB_NR	RDIFF	Reliability N	Agr	Avg. Score Girls	Boys	Flags
Australia	1202	50.6	0.66	33.5	22.8	39.2	.	4.5	0.0	-0.49	0.04	0.58	.	-0.32	.	0.62	226	81.0	54.2	47.3	__E___G
Austria	933	41.4	0.60	36.7	16.4	33.2	.	13.7	0.0	-0.35	0.03	0.56	.	-0.30	.	1.09	204	90.7	41.1	41.8	__H____
Azerbaijan	950	15.8	0.45	53.1	17.7	6.9	.	22.3	0.0	-0.12	0.26	0.33	.	-0.30	.	1.39	185	100.0	14.7	16.7	__H_F__
Belgium (French)	742	38.1	0.58	34.9	19.5	28.3	.	17.3	0.0	-0.30	-0.01	0.56	.	-0.29	.	0.89	205	94.6	38.7	37.5	_____
*Bulgaria	1036	42.0	0.58	38.6	24.0	30.0	.	7.3	0.0	-0.35	0.03	0.52	.	-0.33	.	1.20	214	86.4	42.9	41.2	__H____
*Canada	3671	53.0	0.61	29.7	26.4	39.6	.	4.3	0.0	-0.46	-0.01	0.53	.	-0.24	.	0.73	207	94.2	52.4	53.6	__E____
Chinese Taipei	857	65.3	0.66	18.8	23.7	53.4	.	4.1	0.0	-0.46	-0.12	0.59	.	-0.32	.	0.42	227	97.4	68.5	62.4	__E___G
*Colombia	777	23.4	0.61	56.1	19.0	13.9	.	10.9	0.0	-0.43	0.21	0.52	.	-0.16	.	0.76	185	98.9	20.9	25.7	_____
Croatia	916	58.0	0.60	23.3	25.8	45.1	.	5.9	0.0	-0.41	-0.07	0.54	.	-0.27	.	0.72	241	84.2	57.5	58.4	__E____
*Czech Republic	910	64.5	0.67	16.3	22.6	53.2	.	7.9	0.0	-0.46	-0.08	0.59	.	-0.34	.	0.38	209	93.3	63.5	65.4	__E____
Denmark	896	46.8	0.54	37.5	21.5	36.0	.	4.9	0.0	-0.36	0.00	0.50	.	-0.29	.	1.25	236	89.8	45.0	48.6	__H____
*England	757	53.5	0.67	29.1	24.3	41.3	.	5.3	0.0	-0.47	-0.04	0.60	.	-0.30	.	0.86	116	94.8	53.8	53.2	_____
Finland	931	69.5	0.60	14.6	23.8	57.6	.	4.0	0.0	-0.40	-0.21	0.57	.	-0.24	.	0.47	239	87.9	71.5	67.5	__E____
*France	880	36.9	0.60	32.5	19.9	26.9	.	20.7	0.0	-0.27	0.10	0.53	.	-0.36	.	1.08	220	86.8	38.1	35.7	__H____
Georgia	950	28.2	0.51	48.2	25.2	15.6	.	11.1	0.0	-0.27	0.14	0.44	.	-0.27	.	1.20	227	80.2	26.7	29.5	__H____
*Germany	801	44.4	0.56	31.6	17.1	35.8	.	15.5	0.0	-0.27	-0.08	0.55	.	-0.31	.	1.15	195	91.8	43.7	45.0	__H____
*Hong Kong SAR	771	64.2	0.59	16.9	27.8	50.3	.	5.1	0.0	-0.34	-0.14	0.53	.	-0.34	.	0.71	174	98.9	65.8	62.8	__E____
*Hungary	1034	48.2	0.67	35.8	19.8	38.3	.	6.1	0.0	-0.52	0.04	0.60	.	-0.23	.	1.01	224	95.5	47.1	49.2	_____
Indonesia	929	20.3	0.56	64.4	18.2	11.2	.	6.2	0.0	-0.42	0.20	0.48	.	-0.12	.	0.69	223	98.2	21.3	19.2	__E____
*Iran, Islamic Rep. of	1144	23.1	0.57	52.0	18.4	13.9	.	15.6	0.0	-0.30	0.18	0.49	.	-0.25	.	1.03	224	96.4	22.9	23.3	__H____
Ireland	890	58.3	0.66	25.7	22.1	47.2	.	4.9	0.0	-0.46	-0.07	0.60	.	-0.32	.	0.74	213	93.4	56.7	59.6	__E____
*Israel	843	51.2	0.66	31.8	20.9	40.8	.	6.5	0.0	-0.47	-0.03	0.61	.	-0.29	.	0.83	204	91.2	46.9	55.5	_____B
*Italy	836	44.2	0.59	36.0	22.6	32.9	.	8.5	0.0	-0.37	0.02	0.54	.	-0.32	.	1.09	206	92.2	42.2	46.0	__H____
*Lithuania	917	53.1	0.58	28.4	26.1	40.0	.	5.6	0.0	-0.40	-0.05	0.53	.	-0.26	.	0.64	210	97.1	52.8	53.2	__E____
Malta	717	29.5	0.62	44.1	19.9	19.5	.	16.5	0.0	-0.36	0.22	0.51	.	-0.31	.	0.96	198	88.4	27.9	31.0	_____
*Morocco	1502	4.3	0.43	75.5	6.1	1.3	.	17.1	0.0	-0.16	0.34	0.25	.	-0.11	.	1.20	200	84.5	4.5	4.1	HAF__
*Netherlands	800	42.8	0.55	42.5	24.6	30.5	.	2.4	0.0	-0.46	0.05	0.48	.	-0.11	.	1.25	194	95.4	42.1	43.6	__H____
*New Zealand	1107	52.0	0.67	31.3	24.7	39.7	.	4.3	0.0	-0.48	-0.06	0.62	.	-0.27	.	0.66	250	89.6	53.5	50.5	__E____
Northern Ireland	711	59.6	0.64	24.8	23.3	48.0	.	3.9	0.0	-0.48	-0.06	0.57	.	-0.25	.	0.81	106	93.4	62.0	57.4	_____
*Norway	633	36.7	0.54	43.6	22.0	25.8	.	8.7	0.0	-0.36	0.09	0.48	.	-0.24	.	0.95	147	88.4	39.9	33.3	_____G
Oman	2052	12.5	0.55	65.2	16.2	4.4	.	14.2	0.0	-0.31	0.31	0.42	.	-0.15	.	1.14	175	93.1	14.3	10.8	__H_F_G
Poland	984	44.9	0.69	26.0	23.5	33.1	.	17.4	0.0	-0.38	0.05	0.61	.	-0.37	.	0.98	210	99.0	44.9	44.8	_____
Portugal	800	45.4	0.59	36.0	27.0	31.9	.	5.1	0.0	-0.44	0.00	0.54	.	-0.17	.	1.09	184	99.5	45.3	45.5	__H____
Qatar	827	22.8	0.67	56.5	20.0	12.8	.	10.8	0.0	-0.48	0.21	0.59	.	-0.13	.	0.70	218	95.0	25.6	20.2	__E___G
*Romania	904	43.9	0.62	34.7	23.6	32.1	.	9.6	0.0	-0.46	0.16	0.50	.	-0.28	.	0.80	202	97.0	43.8	43.9	_____
*Russian Federation	889	54.3	0.59	33.1	20.9	43.9	.	2.1	0.0	-0.49	-0.02	0.54	.	-0.19	.	1.24	208	94.7	55.7	53.1	__H____
Saudi Arabia	887	21.1	0.57	58.5	19.5	11.4	.	10.6	0.0	-0.39	0.26	0.45	.	-0.18	.	0.81	206	73.3	22.8	19.4	_____
*Singapore	1258	60.9	0.71	26.8	22.3	49.7	.	1.2	0.0	-0.64	-0.02	0.61	.	-0.14	.	0.82	218	93.6	61.6	60.1	_____
*Slovak Republic	1121	46.1	0.62	34.3	22.3	35.0	.	8.5	0.0	-0.43	0.05	0.55	.	-0.28	.	0.94	204	93.6	44.3	47.9	_____
*Slovenia	884	41.9	0.60	32.5	29.0	27.4	.	11.2	0.0	-0.35	0.05	0.53	.	-0.30	.	1.12	184	96.2	43.4	40.4	__H____
Spain	1680	40.1	0.62	38.6	22.9	28.7	.	9.8	0.0	-0.40	0.06	0.56	.	-0.29	.	0.97	203	95.1	40.5	39.8	_____
*Sweden	932	48.6	0.56	31.8	24.5	36.4	.	7.4	0.0	-0.36	0.04	0.49	.	-0.32	.	0.94	226	90.7	51.8	45.8	_____G
Trinidad and Tobago	785	33.4	0.68	47.9	23.4	21.7	.	7.0	0.0	-0.51	0.19	0.57	.	-0.23	.	0.78	178	77.5	34.6	32.1	_____
United Arab Emirates	2893	26.5	0.67	50.2	20.3	16.3	.	13.2	0.0	-0.42	0.20	0.58	.	-0.25	.	0.69	584	83.0	26.5	26.5	__E____
*United States	2554	61.1	0.67	25.0	25.1	48.6	.	1.4	0.0	-0.57	-0.08	0.60	.	-0.16	.	0.62	630	84.9	60.8	61.4	__E____
*Reference Avg (n=24)	26961	45.6	0.60	35.2	22.2	34.5	.	8.0	0.0	-0.41	0.03	0.53	.	-0.26	.	0.92	5251	92.8	45.6	45.6	_____
International Avg (n=45)	49493	42.7	0.61	37.4	21.9	31.7	.	8.9	0.0	-0.40	0.05	0.53	.	-0.26	.	0.90	9939	91.6	43.0	42.4	_____
Botswana	825	20.4	0.64	67.4	21.2	9.8	.	1.6	0.0	-0.52	0.21	0.56	.	-0.07	.	0.64	207	86.0	22.6	18.1	__E_F_G
Honduras	762	17.1	0.52	65.9	16.4	8.9	.	8.8	0.0	-0.35	0.25	0.42	.	-0.15	.	1.21	180	98.3	16.1	18.1	__H_F__
Kuwait	638	23.4	0.62	45.0	25.1	10.8	.	19.1	0.0	-0.32	0.34	0.45	.	-0.32	.	0.89	74	93.2	27.4	17.9	_____G
Morocco (6)	1430	11.2	0.43	70.3	14.7	3.8	.	11.1	0.0	-0.22	0.29	0.29	.	-0.19	.	1.56	189	73.5	12.3	10.1	__H_F__
Alberta, Canada	748	57.8	0.59	27.3	25.0	45.3	.	2.4	0.0	-0.50	0.02	0.50	.	-0.22	.	0.66	29	93.1	53.9	61.3	__E___B
Ontario, Canada	894	51.3	0.62	31.3	26.0	38.4	.	4.4	0.0	-0.46	-0.01	0.55	.	-0.23	.	0.73	47	95.7	50.4	52.3	__E____
Quebec, Canada	842	51.1	0.59	31.1	24.1	39.1	.	5.7	0.0	-0.42	-0.04	0.54	.	-0.23	.	0.74	60	93.3	50.8	51.4	__E____
Maltese - Malta	713	22.4	0.62	50.5	18.0	13.5	.	18.1	0.0	-0.27	0.21	0.54	.	-0.34	.	0.98	190	88.9	23.2	21.7	_____
Eng/Afr (5) - RSA	699	24.0	0.56	60.2	19.9	14.0	.	5.9	0.0	-0.40	0.09	0.53	.	-0.10	.	0.64	217	73.3	24.8	23.3	__E____
Andalusia, Spain	849	38.3	0.63	42.9	20.7	27.9	.	8.5	0.0	-0.42	0.05	0.58	.	-0.25	.	1.05	40	95.0	38.6	38.0	__H____
Abu Dhabi, UAE	822	23.7	0.64	52.4	20.1	13.6	.	13.9	0.0	-0.41	0.23	0.54	.	-0.21	.	0.59	186	84.9	25.1	22.3	__E____
Dubai, UAE	1198	33.5	0.70	44.2	21.1	23.0	.	11.7	0.0	-0.46	0.16	0.61	.	-0.30	.	0.65	192	78.6	32.1	35.1	__E____
Florida, US	516	66.6	0.65	21.9	21.5	55.8	.	0.8	0.0	-0.57	-0.10	0.58	.	-0.14	.	0.68	124	85.5	68.2	64.8	__E____
Netherlands (Natl)	190	40.3	0.57	44.2	22.6	28.9	.	4.2	0.0	-0.41	0.01	0.54	.	-0.22	.	1.39	101	94.1	41.1	39.6	__H____
Norway (5)	240	50.6	0.68	36.7	16.3	42.5	.	4.6	0.0	-0.52	-0.02	0.64	.	-0.27	.	0.96	114	84.2	49.6	51.8	_____

Keys: DIFF= Percent correct score; DISC= Item discrimination; P_0...P_3= Percentage obtaining score level; P_OM, P_NR= Percentage Omitted, Not Reached;
PB_0...PB_3= Point Biserial for score level; PB_OM, PB_NR= Point Biserial for Omitted, Not Reached; RDIFF= Rasch difficulty;
Reliability: N= Responses double scored; Agr= Percentage agreement.
Flags: A= Point Biserial not ordered; B= Boys outperform girls; C= Difficulty less than chance; D= Negative/low discrimination; E= Easier than average;
F= Score obtained by less than 10%; G= Girls outperform boys; H= Harder than average; R= Scoring reliability less than 70%; V= Difficulty greater than 95%.

For all items, regardless of format (i.e., multiple choice or constructed response), 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 total score).[1] Also provided was an estimate of the difficulty of the item using a Rasch one-parameter IRT model. Statistics for each item were displayed alphabetically by country, together with an international average – i.e., based on all participating countries listed above the international average – and a reference average – based on a pool of countries that have participated regularly in the TIMSS and PIRLS assessments – for each statistic. The reference countries are shown with an asterisk next to their names. The international and reference averages of the item difficulties and item discriminations served as guides to the overall statistical properties of the items. The item review outputs also listed countries that participated at higher grades, as well as all the benchmarking participants.

Statistics displayed for multiple choice items included the percentage of students that chose each response option – as well as the percentage of students that omitted or did not reach the item – and the point-biserial correlations for each response option. Statistics displayed for constructed response items (which could have 1, 2, or 3 score points) included the difficulty and discrimination of each score level. Constructed response item displays also provided information about the reliability with which each item was scored in each country, showing the total number of double-scored responses, the percentage of score agreement between the scorers, and – in the case of TIMSS with its 2-digit scoring scheme – the percentage of code agreement between scorers.

During item review, "not reached" responses (i.e., items toward the end of the booklet that that the student did not attempt)[2] were treated as "not administered" and thus did not contribute to the calculation of the item statistics. However, the percentage of students not reaching each item was reported. Omitted responses, although treated as incorrect, were tabulated separately from incorrect responses for the sake of distinguishing students who provided no form of response from students who attempted a response.

The definitions and detailed descriptions of the statistics that were calculated are given below. The statistics were calculated separately for each assessment and, in the case of TIMSS, separately by grade and subject. The

1 For computing point-biserial correlations, the total score is the percentage of points a student has scored on the items (s) he was administered. In the context of TIMSS, a separate total score is computed for mathematics and for science. Not-reached responses are not included in the total score.

2 An item was considered "not reached" if the item itself and the item immediately preceding it were not answered and no subsequent items had been attempted. The decision as to whether an item was not reached was made separately for part 1 and part 2 of each assessment booklet.

statistics are listed in order of their appearance in the item review outputs:

Cases: This is the number of students to whom the item was administered. Not-reached responses were not included in this count.

DIFF: The item difficulty is the average percent correct on an item. For a 1-point item, including all multiple choice items, it is the percentage of students providing a fully correct response to the item. For 2-point and 3-point items, it is the average percentage of points; for example, if 25 percent of students scored 2 points and 50 percent scored 1 point on a 2-point item, then the average percent correct for such an item would be 50 percent. For this statistic, not-reached responses were not included.

DISC: The item discrimination is computed as the correlation between the response to an item and the total score on all items administered to a student. Items exhibiting good measurement properties should have a moderately positive correlation, indicating that the more able students get the item right, the less able get it wrong. For this statistic, not-reached items were not included.

PCT_A, PCT_B, PCT_C, and PCT_D: Available for multiple choice items. Each column indicates the percentage of students choosing the particular response option for the item (A, B, C, or D). Not-reached responses were excluded from the denominator.

PCT_0, PCT_1, PCT_2 and PCT_3: Available for constructed response items. Each column indicates the percentage of students responding at that particular score level, up to and including the maximum score level for the item. Not-reached items were excluded from the denominator.

PCT_OM: Percentage of students who, having reached the item, did not provide a response. Not reached responses were excluded from the denominator.

PCT_NR: Percentage of students who did not reach the item. This statistic is the number of students who did not reach an item as a percentage of all students who were administered that item, including those who omitted or did not reach that item.

PB_A, PB_B, PB_C, and PB_D: Available for multiple choice items. These columns show the point-biserial correlations between choosing each of the response options (A, B, C, or D) and the total score on all of the items administered to a student. Items with good psychometric properties have moderately positive correlations for the correct option

and negative correlations for the distracters (the incorrect options). Not-reached responses were not included in these calculations.

PB_0, PB_1, PB_2, and PB_3: Available for constructed response items. These columns present the point-biserial correlations between the score levels on the item (0, 1, 2, or 3) and the overall score on all of the items the student was administered. For items with good measurement properties, the correlation coefficients should monotonically increase from negative to positive as the score on the item increases. Not-reached responses were not included in these calculations.

PB_OM: The point-biserial correlation between a binary variable indicating an omitted response to the item, and the total score on all items administered to a student. This correlation should be negative or near zero. Not-reached responses were not included in this statistic.

PB_NR: The point-biserial correlation between a binary variable indicating a not-reached response to the item, and the total score on all items administered to a student. This correlation should be negative or near zero.

RDIFF: An estimate of the difficulty of an item based on a Rasch one-parameter IRT model applied to the achievement data for a given country. 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 across all items within each country was zero.

Reliability (N): To provide a measure of the reliability of the scoring of the constructed response items, items in approximately 25 percent of the test booklets in each country were independently scored by two scorers. This column indicates the number of responses that were double-scored for a given item in a country.

Reliability (Agr): Available for PIRLS items. This column contains the percentage of agreement on the scores assigned by the two independent PIRLS scorers.

Reliability (Score): Available for TIMSS items. This column contains the percentage of agreement on the score value of the two-digit diagnostic codes assigned by the two independent TIMSS scorers.

Reliability (Code): Available for TIMSS items. This column contains the percentage of agreement on the two-digit diagnostic codes assigned by the two independent TIMSS scorers.

Avg. Score (Girls/Boys): Available for PIRLS items because of the concern over gender differences in reading ability. These two numbers are the average percent correct, similar to the DIFF statistic, calculated separately for girls and boys.

As an aid to the reviewers, the item-review displays included a series of flags signaling the presence of one or more conditions that might indicate a problem with an item. The following conditions were flagged:

- The item discrimination (DISC) was less than 0.10 (flag D).
- The item difficulty (DIFF) was less than 25 percent for multiple choice items (flag C).
- The item difficulty (DIFF) exceeded 95 percent (flag V).
- The Rasch difficulty estimate (RDIFF) for a given country made the item either easier (flag E) or more difficult (flag H) relative to the international average for that item.
- The point-biserial correlation for at least one distracter in a multiple choice item was positive, or the point-biserial correlations across the score levels of a constructed response item were not ordered (flag A).
- The percentage of students selecting one of the response options for a multiple choice item, or one of the score values for a constructed response item, was less than 10 percent (flag F).
- Scoring reliability for agreement on the score value of a constructed response item was less than 70 percent (flag R).

Although not all of these conditions necessarily indicated a problem, the flags were a useful tool to draw attention to potential sources of concern.

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 assessment should perform relatively well on each of the items, and low-scoring countries should do less well on each of the items. When this does not occur (e.g., when a high-performing 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 that an item is flawed in some way and that steps should be taken to address the problem. To assist in detecting sizeable item-by-country interactions, the TIMSS & PIRLS

International Study Center produced a graphical display for each item showing the difference between each country's Rasch item difficulty and the international average Rasch item difficulty across all countries. Examples of the graphical displays are contained in Exhibits 3 and 4.

Exhibit 3: Example Plot of Item-by-Country Interaction for a TIMSS 2011 Item

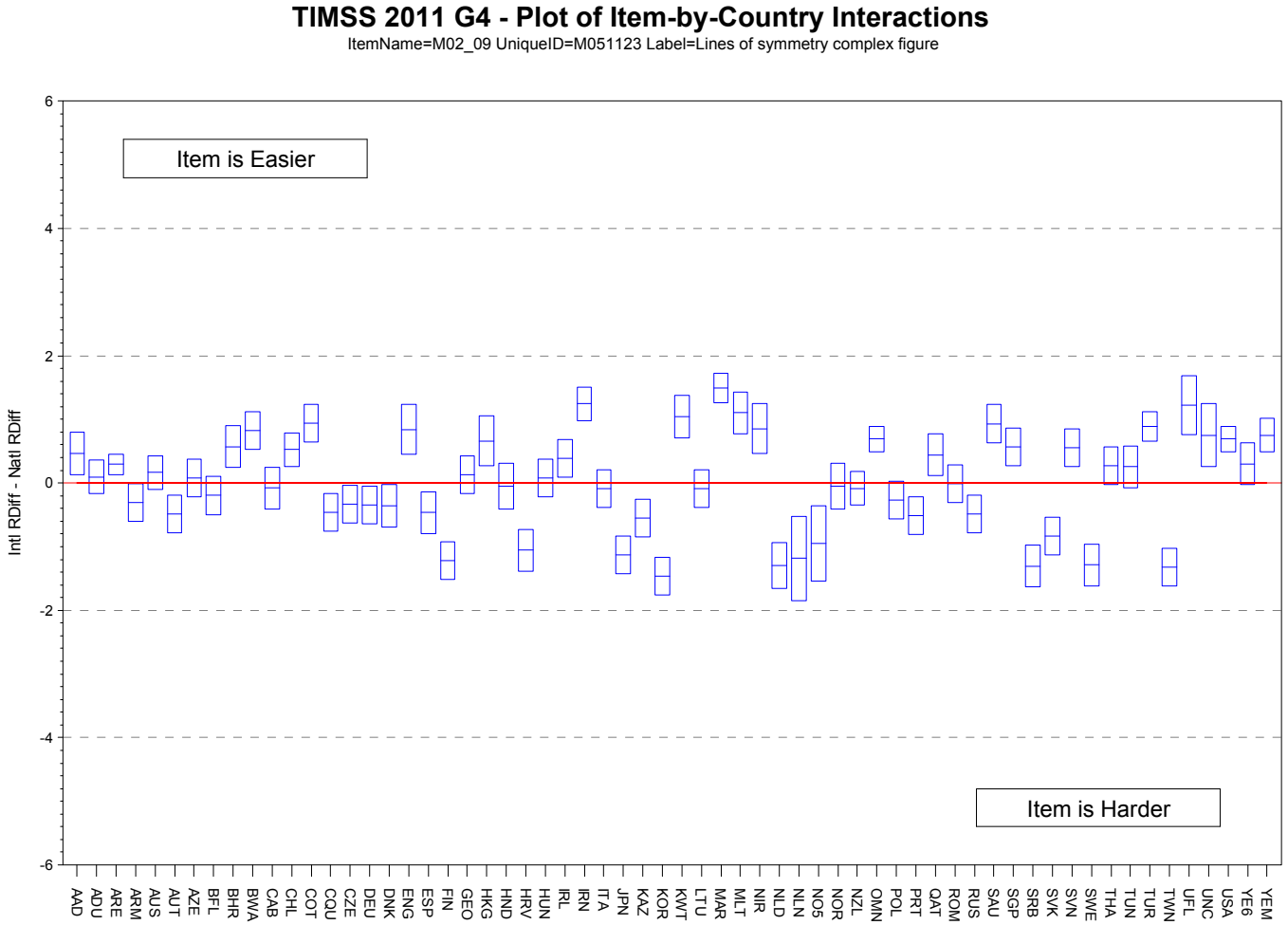
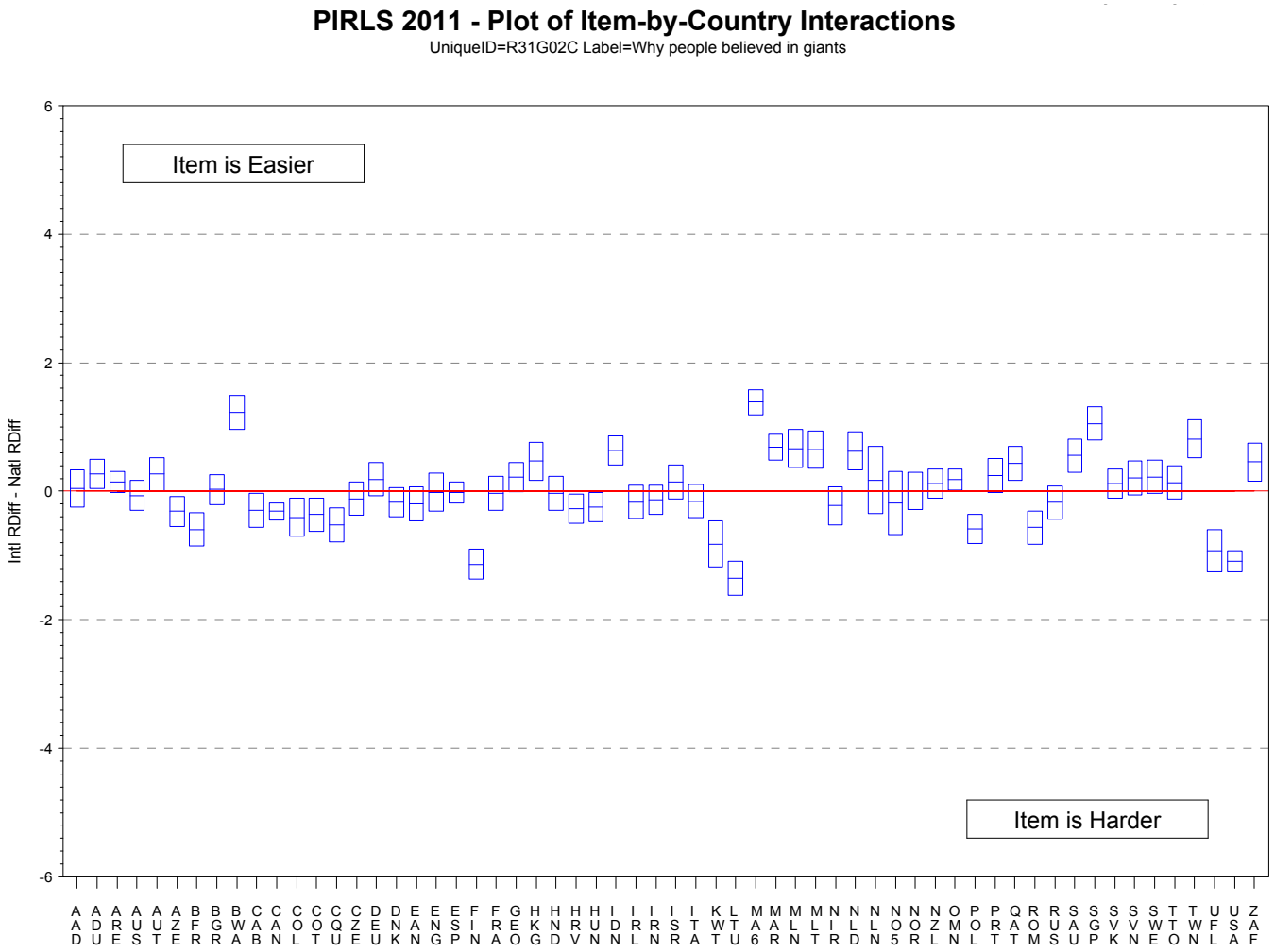


Exhibit 4: Example Plot of Item-by-Country Interaction for a PIRLS 2011 Item



In each of these item-by-country interaction displays, the difference in Rasch item difficulty for each country is presented as a 95 percent confidence interval, which includes a built-in Bonferroni correction for multiple comparisons across the participating countries. The limits for this confidence interval were computed as follows:

$$\text{Upper Limit} = RDIFF_{i.} - RDIFF_{ik} + SE(RDIFF_{ik}) \cdot Z_b$$

$$\text{Lower Limit} = RDIFF_{i.} - RDIFF_{ik} - SE(RDIFF_{ik}) \cdot Z_b$$

Where $RDIFF_{ik}$ is the Rasch difficulty of item i in country k, $RDIFF_{i.}$ is the international average Rasch difficulty of item i, $SE(RDIFF_{ik})$ is the standard error of the Rasch difficulty of item i in country k, and Z_b is the 95% critical value from the Z distribution corrected for multiple comparisons using the Bonferroni procedure.

Trend Item Review

In order to measure trends, TIMSS and PIRLS in 2011 included achievement items from previous assessments as well as items developed for use for the first time in 2011. Accordingly, the TIMSS assessments included items from 2003 and 2007 as well as 2011, and PIRLS included passages and items from 2001 and 2006 as well as from 2011. An important review step, therefore, was to check that these "trend items" had statistical properties in 2011 similar to those they had in the previous assessments (e.g., a TIMSS item that was relatively easy in 2007 should still be relatively easy in 2011).

As can be seen in the examples in Exhibit 5 and Exhibit 6, the trend item review focused on statistics for trend items from the current and previous assessments (2011 and 2006 for PIRLS, 2011 and 2007 for TIMSS) for countries that participated in both. For each country, trend item statistics included the percentage of students in each score category (or response option for multiple choice items) for each assessment, as well as the difficulty of the item and the percent correct by gender. In reviewing these item statistics, the aim was to detect any unusual changes in item difficulties between administrations, which might indicate a problem in using the item to measure change.

Exhibit 5: Example Item Statistics for a TIMSS Trend Item

Trends in International Mathematics and Science Study - TIMSS 2011 Assessment Results
Trend Achievement Data Almanac for Science Items (Weighted) - 4th Grade

16:32 Monday, October 15, 2012

S09_05 (S041183): Life Science / Applying
Label: Maintaining good physical health - Points: 2

COUNTRY	YEAR	N	20 %	29 %	10 %	19 %	79 %	OMITTED %	NOT REACHED %	V1 %	V2 %	1.GIRL % RIGHT	2.BOY % RIGHT
Australia	2007	579	17.5	1.5	51.0	10.1	11.4	8.3	0.3	80.1	19.0	18.2	19.8
	2011	877	8.8	3.1	57.3	5.5	14.5	10.1	0.7	74.7	11.9	11.3	12.4
Austria	2007	688	14.8	2.9	59.0	7.2	8.0	8.2	0.0	83.8	17.6	21.2	14.0
	2011	672	10.1	4.0	59.5	10.8	6.7	9.0	0.0	84.3	14.1	15.7	12.5
Chinese Taipei	2007	590	28.7	2.1	48.7	6.1	8.1	6.3	0.0	85.6	30.8	31.7	30.0
	2011	614	38.5	6.2	37.9	8.6	3.2	5.6	0.0	91.2	44.7	49.4	40.6
Czech Republic	2007	611	19.3	0.8	29.6	2.4	23.5	24.3	0.1	52.1	20.1	22.2	18.7
	2011	647	21.8	0.6	40.1	4.7	16.4	16.5	0.0	67.1	22.4	20.7	23.9
Denmark	2007	505	14.1	0.0	68.3	1.6	7.4	8.3	0.4	84.0	14.1	17.5	10.5
	2011	580	13.7	0.3	68.2	0.0	8.4	8.4	1.0	82.2	14.0	14.5	13.6
England	2007	623	20.5	2.4	44.0	1.3	20.8	11.0	0.0	68.2	22.9	30.2	16.1
	2011	481	15.3	7.4	41.6	5.1	17.3	12.8	0.4	69.5	22.7	24.1	21.5
Georgia	2007	571	12.7	1.3	15.6	8.5	32.4	29.2	0.2	38.1	14.0	15.2	13.1
	2011	689	17.6	2.7	35.1	10.2	15.2	18.4	0.8	65.6	20.3	20.3	20.3
Germany	2007	748	16.6	5.7	48.7	11.4	9.7	7.9	0.0	82.4	22.3	21.0	23.4
	2011	569	8.1	5.3	48.5	17.6	13.6	6.4	0.5	79.5	13.4	14.7	12.2
Hong Kong SAR	2007	543	26.5	7.9	43.6	9.4	7.6	5.0	0.0	87.4	34.3	39.5	29.9
	2011	561	24.5	4.0	43.6	9.3	12.5	6.1	0.0	81.4	28.6	29.5	27.8
Hungary	2007	574	29.8	5.7	36.9	2.8	11.8	12.6	0.3	75.3	35.5	41.6	29.1
	2011	740	32.0	4.1	39.0	2.5	12.8	9.0	0.5	77.7	36.2	37.7	34.8
Iran, Islamic Rep. of	2007	542	7.0	5.0	39.1	15.6	15.4	17.7	0.2	66.7	12.0	10.1	13.8
	2011	824	6.3	4.0	41.4	11.4	20.0	15.7	1.1	63.1	10.3	12.0	8.7
Italy	2007	635	15.5	0.4	60.3	6.9	5.7	10.7	0.5	83.1	16.0	16.9	15.1
	2011	601	11.3	0.7	52.1	4.0	21.8	10.1	0.0	68.2	12.0	12.3	11.7
Japan	2007	632	7.0	4.7	59.1	16.5	8.6	4.1	0.0	87.2	11.7	11.8	11.6
	2011	634	14.9	7.4	56.6	11.5	6.2	3.4	0.0	90.4	22.3	21.0	23.7
Kuwait	2007	534	9.5	0.3	19.9	2.5	12.7	53.6	1.4	32.3	9.8	12.4	7.2
	2011	587	5.7	0.0	19.8	1.7	31.7	39.8	1.2	27.3	5.7	7.8	3.3
Lithuania	2007	561	18.7	2.3	61.1	1.7	7.8	8.5	0.0	83.7	21.0	26.7	15.9
	2011	693	15.9	0.2	63.4	2.8	10.5	7.0	0.2	82.3	16.2	17.5	15.0
Netherlands	2007	481	28.8	2.0	60.4	2.2	3.9	2.8	0.0	93.3	30.8	35.9	25.9
	2011	467	20.8	4.4	52.9	1.7	16.0	3.9	0.2	79.8	25.2	24.1	26.3
New Zealand	2007	711	15.9	2.3	48.9	8.3	14.9	9.3	0.6	75.3	18.1	22.1	14.4
	2011	804	12.8	3.8	45.0	3.3	27.9	6.5	0.7	64.8	16.5	16.5	16.6
Norway	2007	570	7.4	1.5	60.1	7.6	13.7	8.8	0.9	76.6	8.9	10.7	6.9
	2011	445	7.3	1.5	66.3	4.3	12.3	8.0	0.3	79.4	8.8	11.3	5.5
Russian Federation	2007	638	12.0	0.4	71.4	6.8	5.8	3.5	0.0	90.6	12.4	13.0	11.8
	2011	644	14.4	1.2	70.7	4.0	3.5	6.2	0.0	90.3	15.7	18.5	12.8
Singapore	2007	723	34.5	2.0	48.5	0.8	11.8	2.5	0.0	85.7	36.4	39.1	34.0
	2011	906	32.0	2.2	53.0	2.5	7.9	2.4	0.0	89.7	34.2	36.7	31.8
Slovak Republic	2007	715	19.9	6.2	44.3	8.8	7.4	13.4	0.0	79.2	26.1	30.5	21.5
	2011	803	17.7	2.5	44.6	4.0	19.8	10.9	0.5	68.8	20.2	24.8	15.8
Slovenia	2007	619	17.4	1.6	52.6	3.5	11.2	13.6	0.0	75.2	19.0	20.2	17.9
	2011	637	16.6	1.5	41.2	5.0	13.7	21.7	0.4	64.3	18.0	18.9	17.3

V1 = Percent scoring 1 or better; V2 = Percent scoring 2 or better
Percent right for boys and girls corresponds to the percent obtaining full credit on the item.
Because of missing gender information, some totals may appear inconsistent.

Exhibit 5: Example Item Statistics for a TIMSS Trend Item (Continued)

Trends in International Mathematics and Science Study - TIMSS 2011 Assessment Results
Trend Achievement Data Almanac for Science Items (Weighted) - 4th Grade

16:32 Monday, October 15, 2012 146

S09_05 (S041183): Life Science / Applying
Label: Maintaining good physical health - Points: 2

COUNTRY	YEAR	N	20 %	29 %	10 %	19 %	79 %	OMITTED %	NOT REACHED %	V1 %	V2 %	1.GIRL % RIGHT	2.BOY % RIGHT
Sweden	2007	659	6.0	1.6	48.1	10.0	22.5	11.2	0.7	65.7	7.6	7.5	7.7
	2011	686	10.5	2.6	49.4	9.2	17.8	9.7	0.8	71.6	13.0	16.0	9.7
Tunisia	2007	591	13.0	2.5	17.7	5.9	31.2	29.0	0.8	39.1	15.5	16.8	14.3
	2011	692	5.6	0.5	26.9	0.7	40.6	20.9	4.8	33.7	6.1	7.6	5.0
United States	2007	1131	34.0	4.1	36.8	5.9	13.4	5.8	0.0	80.9	38.1	40.0	36.2
	2011	1785	24.7	4.4	45.1	4.4	17.6	3.6	0.3	78.5	29.0	28.2	29.8
International Avg. (25)	2007	15774	17.9	2.7	46.9	6.5	13.1	12.6	0.3	74.1	20.6	22.9	18.3
	2011	17638	16.3	3.0	48.0	5.8	15.5	10.9	0.6	73.0	19.3	20.4	18.1
Alberta, Canada	2007	590	18.5	3.6	58.8	3.4	9.2	6.4	0.1	84.3	22.1	23.9	20.5
	2011	518	18.6	2.3	42.1	13.3	15.1	8.1	0.4	76.4	20.9	24.7	17.4
Ontario, Canada	2007	507	28.7	2.3	48.5	3.8	8.4	8.1	0.4	83.2	30.9	34.9	26.8
	2011	639	25.6	4.1	42.4	8.6	12.4	6.1	0.8	80.7	29.7	30.9	28.5
Quebec, Canada	2007	557	20.9	1.2	65.3	5.4	3.3	3.8	0.0	92.8	22.1	19.6	24.9
	2011	615	19.6	3.5	60.9	5.6	4.3	6.1	0.0	89.6	23.1	26.0	20.5
Dubai, UAE	2007	433	17.1	1.8	38.2	4.6	16.7	21.5	0.1	61.7	18.9	24.5	13.6
	2011	870	14.2	1.5	37.3	2.1	22.2	22.5	0.2	55.0	15.7	15.8	15.6

V1 = Percent scoring 1 or better; V2 = Percent scoring 2 or better
Percent right for boys and girls corresponds to the percent obtaining full credit on the item.
Because of missing gender information, some totals may appear inconsistent.

Exhibit 6: Example Item Statistics for a PIRLS Trend Item

Progress in International Reading Literacy Study - PIRLS 2011 Assessment Results
Trend Achievement Data Almanac for Acquire and Use Information Items (Weighted)

Sharks: Acquire and Use Information / Interpret and Integrate Ideas and Information
R21K12C: Shark information table - Points: 3

COUNTRY	YEAR	N	0 %	1 %	2 %	3 %	OMITTED %	NOT REACHED %	V1 %	V2 %	V3 %	1.GIRL % RIGHT	2.BOY % RIGHT
Austria	2006	1010	8.1	23.7	23.7	32.8	7.9	3.8	80.2	56.5	32.8	33.3	32.3
	2011	921	9.9	21.3	27.3	32.5	7.9	1.1	81.0	59.7	32.5	34.6	30.4
Belgium (French)	2006	893	13.1	19.3	19.0	23.8	10.2	14.7	62.1	42.8	23.8	23.8	23.7
	2011	753	10.0	22.1	22.1	25.2	12.7	7.8	69.5	47.3	25.2	25.7	24.8
Bulgaria	2006	768	4.6	15.8	25.4	43.8	6.5	3.9	85.0	69.2	43.8	51.9	34.9
	2011	1034	10.8	19.8	26.6	30.3	9.3	3.3	76.6	56.8	30.3	35.2	25.4
Chinese Taipei	2006	927	6.1	21.5	33.4	34.4	2.3	2.3	89.2	67.8	34.4	31.9	36.6
	2011	861	3.2	14.4	30.4	47.8	4.0	0.2	92.6	78.2	47.8	49.6	46.2
Denmark	2006	802	6.6	14.7	30.9	35.5	5.7	6.7	81.0	66.4	35.5	37.8	33.0
	2011	910	6.2	16.2	31.7	34.9	6.5	4.6	82.7	66.6	34.9	38.3	31.4
England	2006	807	11.1	22.7	24.1	31.8	8.3	2.1	78.5	55.9	31.8	32.7	30.9
	2011	764	7.6	20.5	25.4	37.7	6.7	2.0	83.7	63.1	37.7	42.8	32.9
France	2006	881	9.3	23.4	22.8	31.7	8.1	4.6	77.9	54.6	31.7	33.9	29.6
	2011	877	8.4	24.7	19.3	34.1	10.4	3.1	78.1	53.4	34.1	33.5	34.9
Georgia	2006	863	28.3	18.8	11.0	10.8	22.3	8.8	40.6	21.9	10.8	12.9	8.8
	2011	949	23.0	25.0	13.8	18.1	14.4	5.7	56.9	31.9	18.1	23.2	13.8
Germany	2006	1605	9.7	25.0	24.9	28.7	6.7	5.0	78.6	53.6	28.7	26.4	31.2
	2011	781	9.6	20.7	26.9	35.8	4.8	2.1	83.5	62.7	35.8	39.9	31.6
Hong Kong SAR	2006	940	3.9	12.3	30.8	47.6	4.5	0.8	90.7	78.4	47.6	51.0	44.6
	2011	772	2.1	9.5	25.8	57.0	5.0	0.6	92.4	82.8	57.0	61.7	53.1
Hungary	2006	819	6.7	16.1	25.5	38.1	8.0	5.6	79.6	63.6	38.1	41.9	34.1
	2011	1037	9.4	18.0	25.2	41.4	4.4	1.6	84.6	66.7	41.4	46.3	36.5
Indonesia	2006	945	17.3	17.6	22.0	10.0	28.4	4.7	49.7	32.0	10.0	9.3	10.7
	2011	946	21.1	23.8	18.0	15.3	17.0	4.7	57.1	33.3	15.3	18.8	11.9
Iran, Islamic Rep. of	2006	1069	24.0	16.5	10.1	8.7	18.5	22.1	35.3	18.8	8.7	8.0	9.2
	2011	1132	25.7	27.8	19.1	7.1	12.9	7.4	54.0	26.2	7.1	6.9	7.3
Italy	2006	713	8.5	18.8	30.2	33.7	6.2	2.5	82.8	64.0	33.7	37.6	29.7
	2011	837	8.0	24.4	25.3	34.7	4.5	3.2	84.4	60.0	34.7	33.2	36.3
Lithuania	2006	935	8.5	27.1	24.7	28.8	6.9	4.0	80.6	53.5	28.8	34.3	23.7
	2011	925	13.6	24.4	26.8	28.7	4.5	2.0	79.9	55.5	28.7	36.4	21.1
Netherlands	2006	844	4.5	14.6	34.0	38.9	5.7	2.3	87.5	72.9	38.9	43.4	34.6
	2011	781	6.1	18.4	33.1	33.2	7.1	2.1	84.7	66.3	33.2	36.1	30.6
New Zealand	2006	1247	11.4	16.9	22.8	36.1	7.8	5.0	75.8	58.8	36.1	39.3	32.8
	2011	1129	10.0	17.6	22.3	42.4	4.9	2.8	82.3	64.7	42.4	48.5	35.5
Norway	2006	762	17.6	18.7	17.2	16.7	12.1	17.8	52.5	33.8	16.7	16.8	16.5
	2011	636	13.3	28.5	26.1	18.9	7.0	6.2	73.5	45.0	18.9	17.5	20.2
Poland	2006	975	14.3	31.5	20.5	20.0	7.4	6.3	72.0	40.5	20.0	21.8	17.9
	2011	994	12.4	28.1	20.2	31.6	5.8	1.9	79.8	51.8	31.6	33.9	29.5
Romania	2006	858	16.9	21.8	18.5	19.5	14.9	8.3	59.9	38.1	19.5	20.3	18.9
	2011	923	16.6	23.5	19.3	24.4	11.5	4.7	67.2	43.7	24.4	27.7	21.1
Russian Federation	2006	950	7.3	17.1	24.1	42.7	5.4	3.5	83.9	66.8	42.7	46.9	38.3
	2011	882	4.6	19.1	22.9	48.9	3.8	0.7	90.9	71.8	48.9	56.5	42.5
Singapore	2006	1267	4.7	11.2	23.3	57.9	2.7	0.2	92.4	81.3	57.9	59.5	56.5
	2011	1257	3.8	12.5	21.7	59.6	2.3	0.1	93.8	81.3	59.6	63.3	56.3

V1 = Percent scoring 1 or better; V2 = Percent scoring 2 or better; V3 = Percent scoring 3 or better
Percent right for boys and girls corresponds to percent obtaining full credit on the item.
Because of missing gender information, some totals may appear inconsistent.

Exhibit 6: Example Item Statistics for a PIRLS Trend Item (Continued)

Progress in International Reading Literacy Study - PIRLS 2011 Assessment Results 16:21 Monday, October 15, 2012 48
Trend Achievement Data Almanac for Acquire and Use Information Items (Weighted)

Sharks: Acquire and Use Information / Interpret and Integrate Ideas and Information
R21K12C: Shark information table - Points: 3

COUNTRY	YEAR	N	0 %	1 %	2 %	3 %	OMITTED %	NOT REACHED %	V1 %	V2 %	V3 %	1.GIRL % RIGHT	2.BOY % RIGHT
Slovak Republic	2006	1081	11.6	24.8	25.1	26.0	8.0	4.7	75.8	51.1	26.0	29.4	22.8
	2011	1116	11.4	22.4	25.6	34.4	4.6	1.5	82.5	60.1	34.4	37.8	31.1
Slovenia	2006	1078	9.1	20.3	23.5	34.5	7.4	5.2	78.3	58.0	34.5	38.1	31.3
	2011	903	10.0	25.1	27.1	31.5	4.6	1.6	83.8	58.6	31.5	36.2	27.3
Spain	2006	822	12.0	22.4	22.7	20.4	11.9	10.6	65.4	43.1	20.4	16.7	23.8
	2011	1701	11.4	26.6	25.6	23.5	8.7	4.2	75.7	49.1	23.5	23.7	23.3
Sweden	2006	870	9.0	14.1	21.6	39.5	7.9	7.8	75.2	61.1	39.5	42.0	37.0
	2011	912	11.0	20.8	22.9	35.5	6.4	3.4	79.2	58.4	35.5	41.4	29.7
Trinidad and Tobago	2006	786	20.9	20.1	9.9	11.8	19.3	17.9	41.8	21.7	11.8	13.6	10.2
	2011	779	19.4	26.0	19.5	15.6	13.9	5.6	61.1	35.1	15.6	18.0	13.3
United States	2006	1018	7.6	20.7	25.2	33.8	7.6	5.1	79.7	59.0	33.8	30.7	36.7
	2011	2530	5.2	17.7	24.7	44.5	4.7	3.2	86.9	69.2	44.5	46.0	43.0
International Avg. (28)	2006	26535	11.2	19.5	23.1	29.9	9.6	6.7	72.6	53.0	29.9	31.6	28.2
	2011	28042	10.9	21.4	24.1	33.0	7.5	3.1	78.5	57.1	33.0	36.2	30.0
Alberta, Canada	2006	857	9.3	21.9	29.4	29.9	6.9	2.6	81.2	59.2	29.9	28.6	31.1
	2011	755	7.1	19.6	30.1	34.8	5.4	3.0	84.5	64.9	34.8	35.7	33.9
Ontario, Canada	2006	799	8.5	17.0	28.5	34.7	6.0	5.3	80.2	63.2	34.7	36.0	33.5
	2011	896	8.2	19.0	26.5	38.1	5.4	2.8	83.7	64.7	38.1	41.9	34.6
Quebec, Canada	2006	740	10.0	26.6	27.7	20.2	6.7	8.8	74.5	47.9	20.2	21.8	18.8
	2011	842	4.8	18.0	26.7	42.7	5.8	2.0	87.5	69.4	42.7	46.7	38.2

V1 = Percent scoring 1 or better; V2 = Percent scoring 2 or better; V3 = Percent scoring 3 or better
Percent right for boys and girls corresponds to percent obtaining full credit on the item.
Because of missing gender information, some totals may appear inconsistent.

While some changes in item difficulties were anticipated as countries' overall achievement may have improved or declined, items were noted if the difference between the Rasch difficulties across the two assessments for a particular country was greater than 2 logits. The TIMSS & PIRLS International Study Center used two different graphical displays to examine the differences in item difficulties. The first of these, shown for example items in Exhibits 7 and 8, displays the difference in Rasch item difficulty of the item between 2011 and the previous assessment for each country. A positive difference for a country indicates that the item was relatively easier in 2011, and a negative difference indicates that the item was relatively more difficult.

Exhibit 7: Example Plot of Differences in Rasch Item Difficulties Between 2011 and 2007 for a TIMSS Trend Item

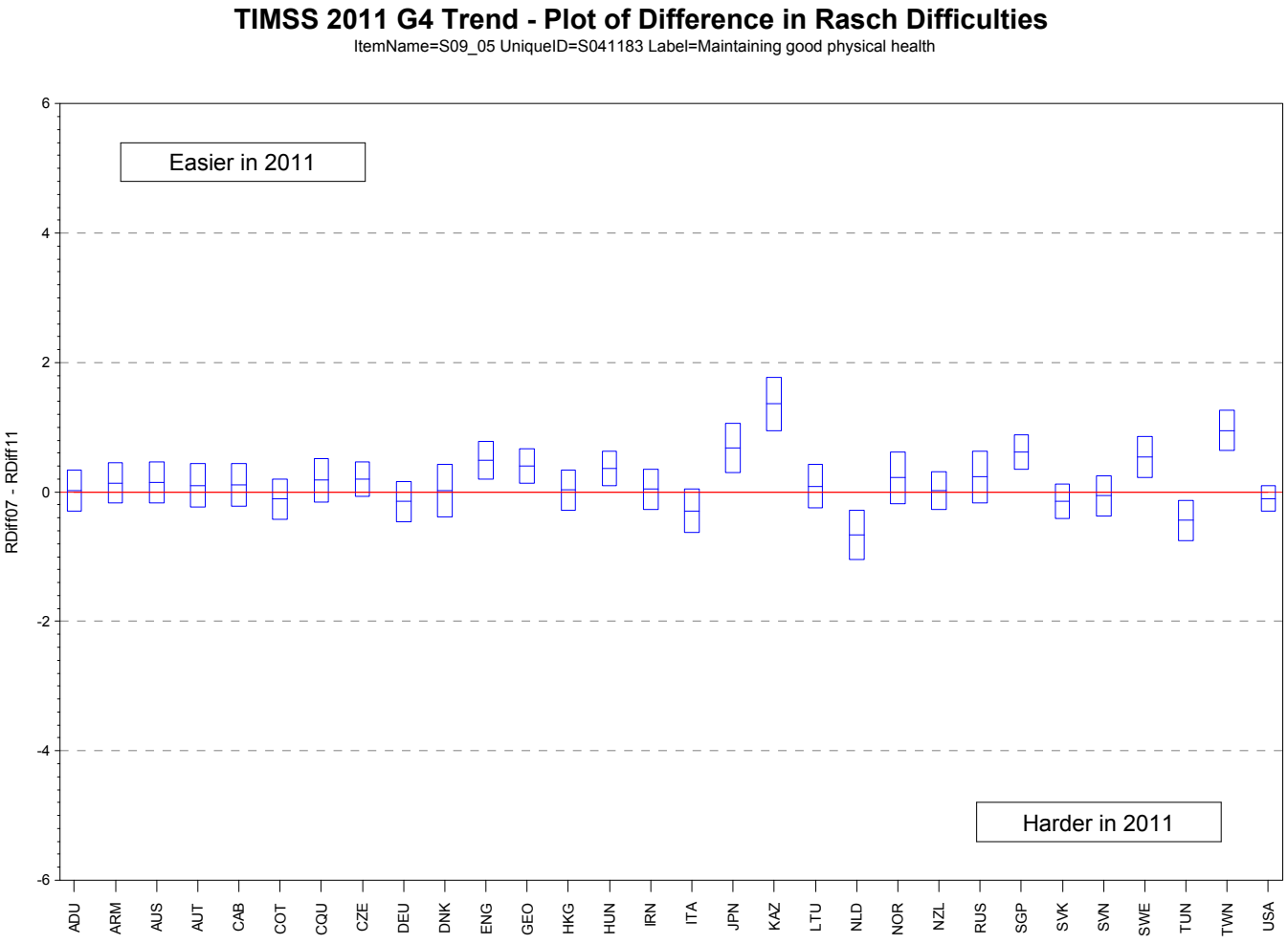
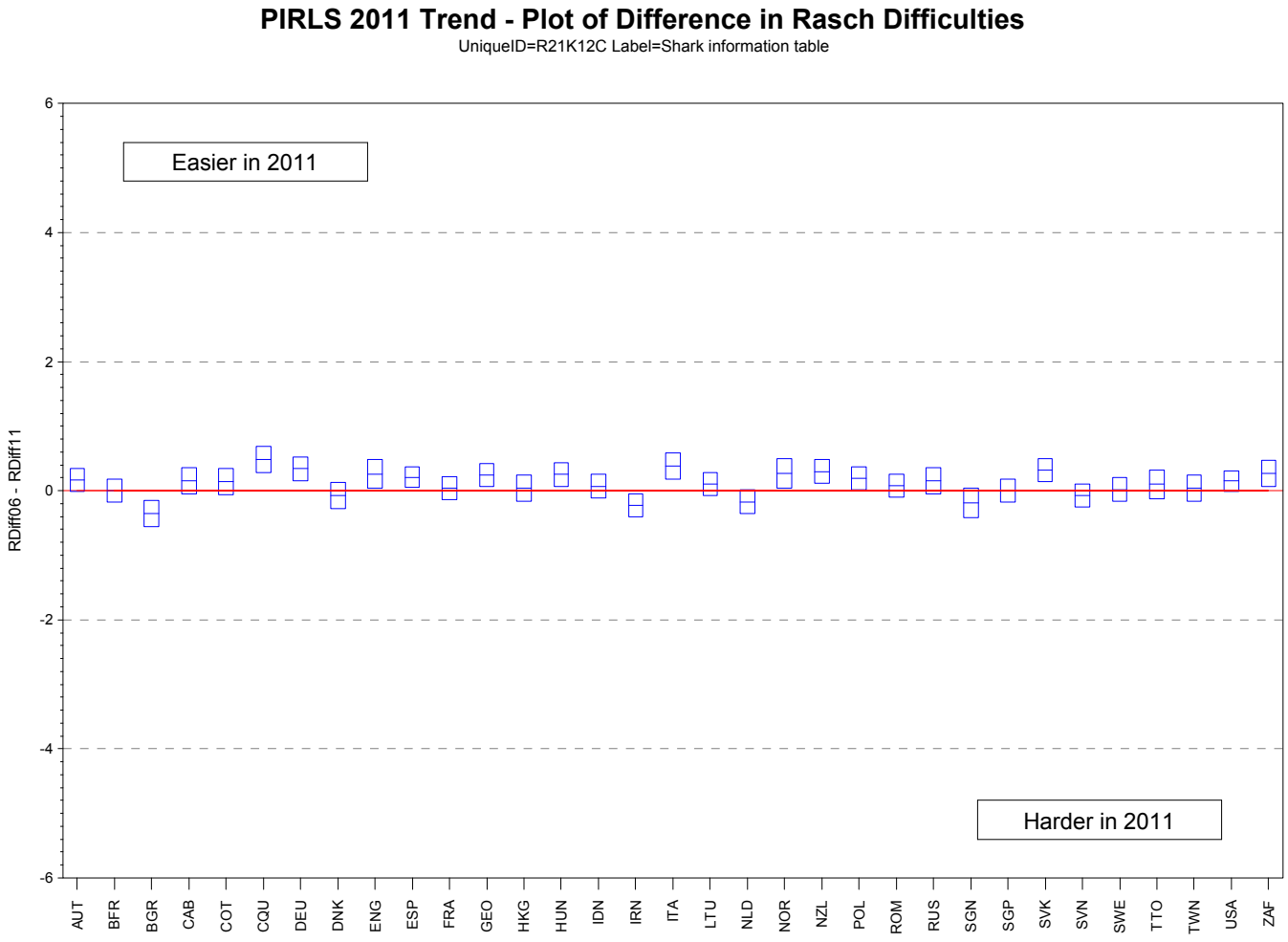


Exhibit 8: Eample Plot of Differences in Rasch Item Difficulties Between 2011 and 2006 for a PIRLS Trend Item



The second set of graphical displays, presented in Exhibits 9 and 10, showed the performance of a given country on all trend items simultaneously. For each country, the graph plotted the 2011 Rasch difficulty of every trend item against its Rasch difficulty in the previous administration (2006 for PIRLS, 2007 for TIMSS). Where there were no differences between the difficulties in the two successive administrations, the data points aligned on or near the diagonal.

Exhibit 9: Example Plot of Rasch Item Difficulties across TIMSS Trend Items by Country

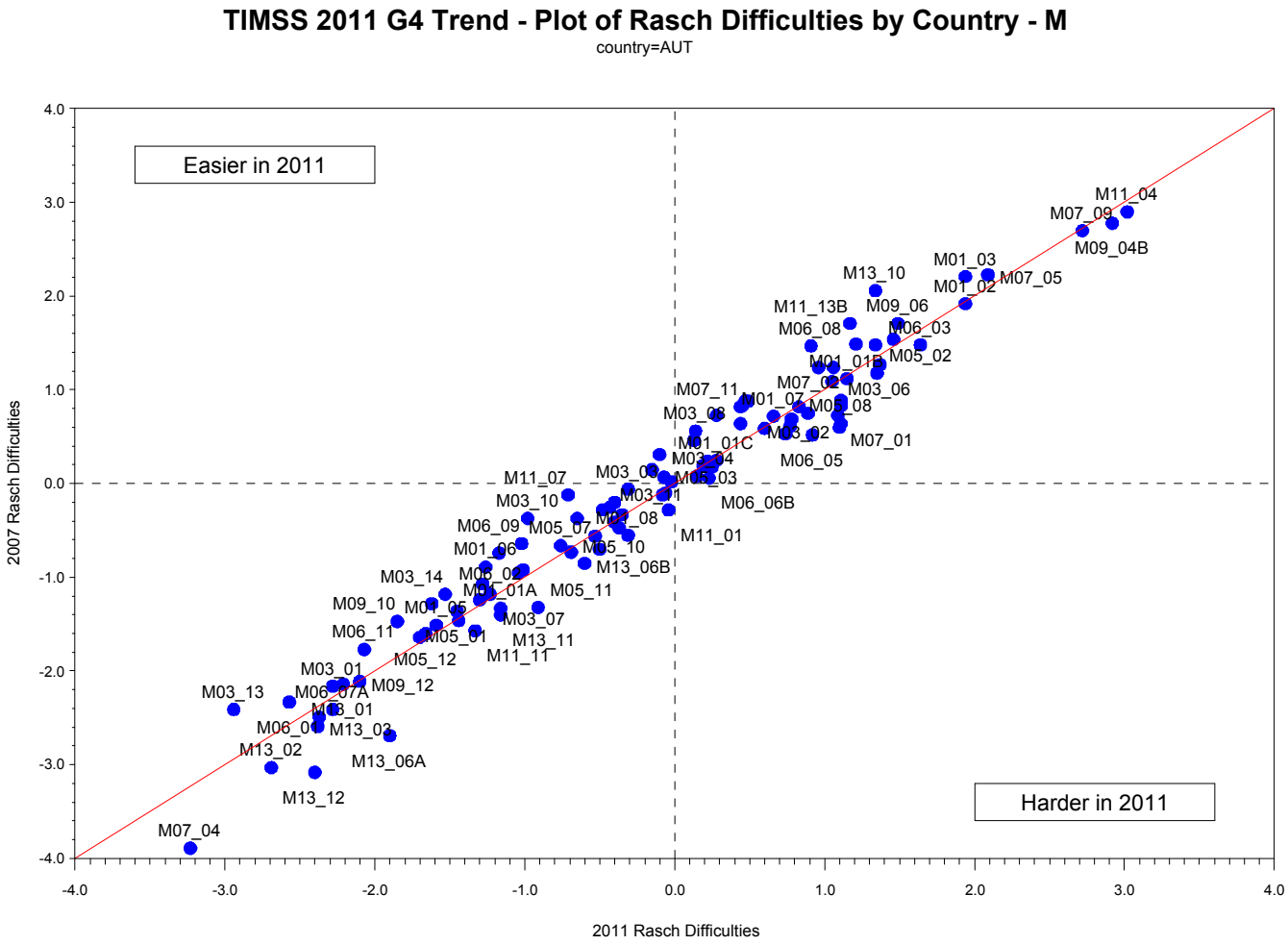
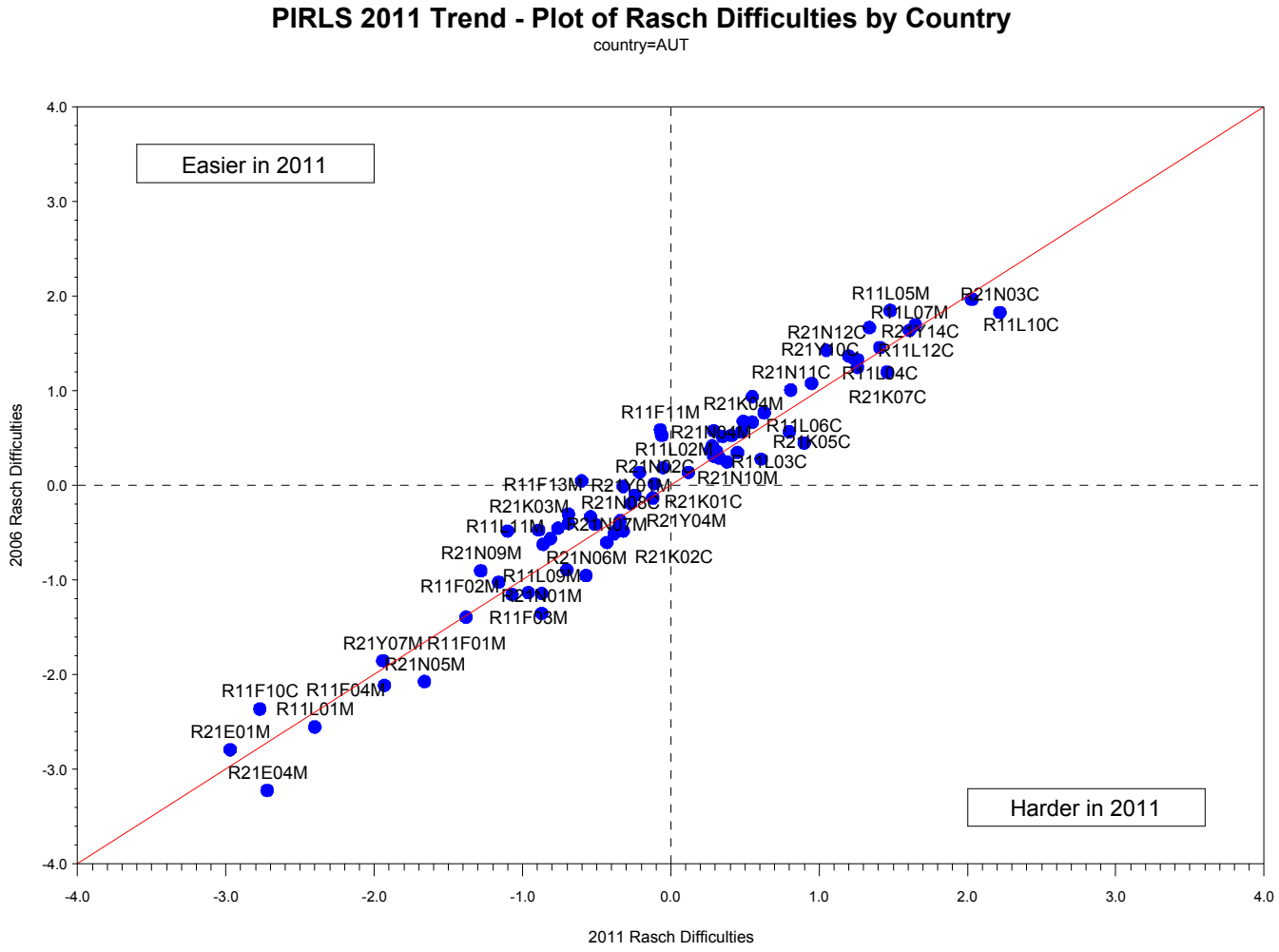


Exhibit 10: Example Plot of Rasch Item Difficulties across PIRLS Trend Items by Country



Reliability

Documenting the reliability of the TIMSS and PIRLS 2011 assessments was a critical quality control step in reviewing the items. As one indicator of reliability, the review considered Cronbach's Alpha coefficient of reliability calculated at the assessment booklet level. Secondly, the scoring of the constructed response items had to meet specific reliability criteria in terms of consistent within-country scoring, cross-country scoring, and across assessment – or trend – scoring.

Test Reliability

Exhibits 11, 12, and 13 display the TIMSS 2011 fourth and eighth grade mathematics and science test reliability coefficients for every country and PIRLS reading test reliability coefficients, respectively. These coefficients are the median Cronbach's alpha reliability across all assessment booklets in TIMSS and PIRLS. In general, reliabilities were relatively high. For TIMSS at the fourth grade, the international median reliability (the median of the reliability coefficients for all countries) was 0.82 for mathematics and 0.78 for science, and at the eighth grade, 0.87 for mathematics and 0.83 for science. The international median reliability for PIRLS reading was 0.88.

Exhibit 11: Cronbach's Alpha Reliability Coefficient - TIMSS 2011 Fourth Grade

Country	Reliability Coefficient	
	Mathematics	Science
Armenia	0.82	0.74
Australia	0.86	0.79
Austria	0.79	0.76
Azerbaijan	0.86	0.79
Bahrain	0.82	0.80
Belgium (Flemish)	0.78	0.70
Chile	0.83	0.77
Chinese Taipei	0.84	0.77
Croatia	0.80	0.71
Czech Republic	0.82	0.75
Denmark	0.82	0.76
England	0.89	0.79
Finland	0.82	0.74
Georgia	0.82	0.75
Germany	0.81	0.75
Hong Kong SAR	0.80	0.76
Hungary	0.87	0.82
Iran, Islamic Rep. of	0.84	0.80
Ireland	0.84	0.79
Italy	0.82	0.76
Japan	0.84	0.73
Kazakhstan	0.85	0.80
Korea, Rep. of	0.81	0.74
Kuwait	0.71	0.78
Lithuania	0.83	0.74
Malta	0.83	0.78
Morocco	0.76	0.67
Netherlands	0.75	0.67
New Zealand	0.84	0.80
Northern Ireland	0.87	0.76
Norway	0.80	0.70
Oman	0.80	0.80
Poland	0.82	0.78
Portugal	0.83	0.75
Qatar	0.86	0.81
Romania	0.88	0.85
Russian Federation	0.83	0.78
Saudi Arabia	0.83	0.79
Serbia	0.85	0.80
Singapore	0.86	0.83
Slovak Republic	0.84	0.79
Slovenia	0.82	0.78
Spain	0.80	0.75
Sweden	0.80	0.77
Thailand	0.81	0.80
Tunisia	0.75	0.76
Turkey	0.86	0.80
United Arab Emirates	0.85	0.81
United States	0.85	0.80
Yemen	0.57	0.62
International Median	**0.82**	**0.78**
Sixth Grade Countries		
Botswana	0.80	0.81
Honduras	0.75	0.76
Yemen	0.71	0.62
Benchmarking Participants		
Alberta, Canada	0.79	0.76
Ontario, Canada	0.84	0.78
Quebec, Canada	0.79	0.69
Abu Dhabi, UAE	0.83	0.80
Dubai, UAE	0.87	0.82
Florida, US	0.84	0.77
North Carolina, US	0.83	0.78

Exhibit 12: Cronbach's Alpha Reliability Coefficient - TIMSS 2011 Eighth Grade

Country	Reliability Coefficient	
	Mathematics	Science
Armenia	0.88	0.83
Australia	0.90	0.85
Bahrain	0.86	0.85
Chile	0.83	0.78
Chinese Taipei	0.94	0.87
England	0.91	0.86
Finland	0.86	0.81
Georgia	0.89	0.78
Ghana	0.67	0.67
Hong Kong SAR	0.91	0.83
Hungary	0.90	0.84
Indonesia	0.78	0.70
Iran, Islamic Rep. of	0.87	0.84
Israel	0.91	0.87
Italy	0.87	0.82
Japan	0.91	0.84
Jordan	0.84	0.84
Kazakhstan	0.87	0.82
Korea, Rep. of	0.91	0.85
Lebanon	0.84	0.80
Lithuania	0.88	0.83
Macedonia, Rep. of	0.89	0.84
Malaysia	0.87	0.83
Morocco	0.76	0.71
New Zealand	0.89	0.85
Norway	0.83	0.80
Oman	0.84	0.84
Palestinian Nat'l Auth.	0.85	0.83
Qatar	0.89	0.86
Romania	0.90	0.82
Russian Federation	0.90	0.83
Saudi Arabia	0.83	0.79
Singapore	0.91	0.89
Slovenia	0.87	0.83
Sweden	0.85	0.83
Syrian Arab Republic	0.81	0.77
Thailand	0.85	0.79
Tunisia	0.81	0.72
Turkey	0.92	0.87
Ukraine	0.89	0.84
United Arab Emirates	0.87	0.85
United States	0.88	0.85
International Median	**0.87**	**0.83**
Ninth Grade Countries		
Botswana	0.78	0.79
Honduras	0.66	0.69
South Africa	0.81	0.80
Benchmarking Participants		
Alberta, Canada	0.85	0.81
Ontario, Canada	0.87	0.79
Quebec, Canada	0.85	0.79
Abu Dhabi, UAE	0.87	0.84
Dubai, UAE	0.90	0.87
Alabama, US	0.86	0.84
California, US	0.88	0.84
Colorado, US	0.88	0.84
Connecticut, US	0.91	0.87
Florida, US	0.89	0.86
Indiana, US	0.88	0.84
Massachusetts, US	0.89	0.86
Minnesota, US	0.88	0.83
North Carolina, US	0.90	0.85

Exhibit 13: Cronbach's Alpha Reliability Coefficient - PIRLS 2011 Overall Reading

Country	Reliability Coefficient
Australia	0.90
Austria	0.87
Azerbaijan	0.84
Belgium (French)	0.87
Bulgaria	0.90
Canada	0.88
Chinese Taipei	0.88
Colombia	0.88
Croatia	0.85
Czech Republic	0.86
Denmark	0.88
England	0.91
Finland	0.85
France	0.88
Georgia	0.89
Germany	0.89
Hong Kong SAR	0.86
Hungary	0.90
Indonesia	0.83
Iran, Islamic Rep. of	0.89
Ireland	0.89
Israel	0.91
Italy	0.87
Lithuania	0.88
Malta	0.91
Morocco	0.79
Netherlands	0.84
New Zealand	0.91
Northern Ireland	0.90
Norway	0.86
Oman	0.88
Poland	0.89
Portugal	0.86
Qatar	0.91
Romania	0.91
Russian Federation	0.88
Saudi Arabia	0.87
Singapore	0.91
Slovak Republic	0.88
Slovenia	0.88
Spain	0.89
Sweden	0.87
Trinidad and Tobago	0.90
United Arab Emirates	0.91
United States	0.90
International Median	**0.88**
Sixth Grade Countries	
Botswana	0.89
Honduras	0.89
Kuwait	0.91
Morocco	0.88
Benchmarking Participants	
Alberta, Canada	0.89
Ontario, Canada	0.90
Quebec, Canada	0.84
Maltese - Malta	0.89
Eng/Afr (5) - RSA	0.92
Andalusia, Spain	0.87
Abu Dhabi, UAE	0.90
Dubai, UAE	0.92
Florida, US	0.89
prePIRLS Countries	
Botswana	0.90
Colombia	0.89
South Africa	0.93

Scoring Reliability for Constructed Response Items

A sizeable proportion of the items in the TIMSS and PIRLS 2011 assessments were constructed response items, comprising about half of the assessment score points. 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 and PIRLS, measures taken to ensure that the constructed response items were scored reliably in all countries included developing scoring guides for each constructed response question (that provided descriptions of acceptable responses for each score point value) and providing extensive training in the application of the scoring guides. See **Assessment Framework and Instrument Development** for more information on the scoring guides and see **Operations and Quality Assurance** for information on the scoring process.

Within-Country Scoring Reliability

To gather and document information about the within-country agreement among scorers for TIMSS and PIRLS 2011, a random sample of approximately 25 percent of the assessment booklets was selected to be scored independently by two scorers. The inter-scorer agreement for each item in each country was examined as part of the item review process. Exact percent agreement across items was high on average across countries—95 percent or above, on average internationally, in both studies. In TIMSS there also was high agreement at the diagnostic score level, where percent agreement ranged from 94 percent in science at the eighth grade to 98 percent in mathematics at both grades, on average. See **Item Review Details** for the average and range of the within-country percentage of correctness score agreement across all items. The TIMSS Within-country Scoring Reliability documents also provide the average and range of the within-country percentage of diagnostic score agreement.

Trend Item Scoring Reliability

The TIMSS & PIRLS International Study Center also took steps to show that the 2011 constructed response items used in previous administrations (2006 for PIRLS and 2007 for TIMSS) were scored in the same way in both assessments. In anticipation of this, countries that participated in PIRLS 2006 or TIMSS 2007 sent samples of scored student booklets from the 2006 or 2007 data collections to the IEA Data Processing and Research Center, where they were digitally scanned and stored in presentation software for later use. As a check on scoring consistency from one administration to the next, staff members working in

each country on scoring the 2011 data were asked also to score these 2006 or 2007 responses using the **Trend Reliability Scoring Software** developed by the IEA Data Processing and Research Center (IEA DPC). Each country scored 200 responses for each of 22 mathematics, 22 science, and 33 reading items at the fourth grade, and 25 mathematics and 25 science items at the eighth grade.

There was a very high degree of scoring consistency in both TIMSS and PIRLS. In TIMSS exact agreement between the scores awarded in 2007 and those given by the 2011 scorers ranged from 93 percent in science at the eighth grade to 98 percent in mathematics at both grades, on average internationally. Overall, the percent of agreement across the trend constructed response items in PIRLS was also high—95 percent on average across countries. There also was high agreement in TIMSS at the diagnostic score level, although somewhat less in science than in mathematics, on average. The average and range of scoring consistency over time can be found in **Item Review Details**.

Cross-country Scoring Reliability Study

It also was important to document the consistency of scoring across countries. Because of the many different languages in use in TIMSS and PIRLS 2011, establishing the reliability of constructed response scoring across all countries was not feasible. However, the TIMSS & PIRLS International Study Center did conduct a cross-country study of scoring reliability among Northern Hemisphere countries that had scorers who were proficient in English. A sample of student responses was provided by the English-speaking Southern Hemisphere countries. Cross-country scoring included 200 student responses for each of 20 mathematics, 19 science, and 26 reading items at the fourth grade, and 25 mathematics and 25 science items at the eighth grade. This set of student responses in English was then scored independently in each country that had two scorers proficient in English, using the **Cross-country Scoring Reliability Software** provided by the IEA DPC. In all, scorers from 43 countries at fourth grade and 33 countries at eighth grade in TIMSS, and 39 countries in PIRLS participated in the study. Scoring for this study took place shortly after the other scoring reliability activities were completed. Making all possible comparisons among scorers gave 7,800 comparisons at fourth grade and 6,600 comparisons at eighth grade in TIMSS and 5,200 comparisons in PIRLS for each student response to each item. This resulted in more than 100,000 total comparisons at each grade and subject when aggregated across all 200 student responses to that item. Agreement across countries was defined in terms of the percentage

of these comparisons that were in exact agreement.

On average internationally, scorer reliability across countries was high in both studies. In TIMSS exact agreement between the scores awarded across countries ranged from 89 percent in science to 97 percent in mathematics at the fourth grade and from 86 percent in science to 94 percent in mathematics at the eighth grade, on average internationally. In PIRLS, there was 85 percent agreement, on average, with variation across the individual items. There also was high agreement in TIMSS at the diagnostic score level, where percent agreement ranged from 82 percent in science at the eighth grade to 96 percent in mathematics at the fourth grade, on average. See **Item Review Details** for the results of the cross-country scoring reliability study.

Item Review Procedures

Using the information from the comprehensive collection of item analyses and reliability data that were computed and summarized for TIMSS and PIRLS 2011, the TIMSS & PIRLS International Study Center thoroughly reviewed all item statistics for every participating country and benchmarking participant to ensure that the 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 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 1 score point, point biserials that did not increase with each score level;
- 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; and
- For trend items, an item performed substantially differently in 2011 compared to the previous administration (2006 for PIRLS or 2007 for TIMSS), or an item was not included in the previous assessment for a particular country.

When the item statistics indicated a problem with an item, the

documentation from the translation verification 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 was consulted before deciding how the item should be treated.

The checking of the TIMSS and PIRLS 2011 achievement data involved review of more than 1,000 items 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. See **Country Adaptations to Items and Item Scoring** for a list of deleted items, as well as a list of recodes made to constructed response item codes. There also were a number of items in each study that were combined, or derived, for scoring purposes. See **Derived Items in TIMSS and PIRLS 2011** for details about how score points were awarded for each derived item.