CHAPTER 11

Reviewing the TIMSS 2015 Achievement Item Statistics

Pierre Foy
Michael O. Martin
Ina V.S. Mullis
Liqun Yin
Victoria A.S. Centurino
Katherine A. Reynolds

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 TIMSS 2015 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 TIMSS 2015 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 2015 assessments, including TIMSS fourth grade (169 mathematics items and 176 science items), TIMSS eighth grade (212 mathematics items and 220 science items), and TIMSS Numeracy (124 items). The item statistics for each of the participating countries were then carefully reviewed. Exhibits 11.1 and 11.2 show actual samples of the statistics calculated for a multiple-choice and a constructed response item, respectively.

Exhibit 11.1: Example International Item Statistics for a TIMSS 2015 Multiple-Choice Item

Trends in International Mathematics and Science Study - TIMSS 2015 Assessment Results - 4th Grade
International Item Review Statistics (Unweighted)

Mathematics: Number / Applying (M08_04 - M061040) Type: MC Key: A
Label: Shaded fraction of a square

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
*Australia	856	47.8	0.52	47.8	24.0	10.4	15.6	2.2	0.7	0.52	-0.24	-0.16	-0.27	-0.06	-0.09	0.04	
Bahrain	600	21.0	0.34	21.0	26.4	11.5	38.6	2.4	1.7	0.34	-0.03	-0.00	-0.25	-0.04	-0.01	0.81	C
Belgium (Flemish)	768	50.3	0.53	50.3	25.5	11.0	12.0	1.3	0.3	0.53	-0.30	-0.18	-0.23	-0.03	-0.03	0.41	
Bulgaria	616	31.4	0.47	31.4	27.7	13.2	13.7	14.0	0.3	0.47	-0.20	-0.12	-0.16	-0.09	-0.04	1.06	
*Canada	1777	39.1	0.50	39.1	23.6	10.8	23.8	2.7	1.0	0.50	-0.17	-0.11	-0.30	-0.06	-0.04	0.48	
Chile	676	31.9	0.43	31.9	26.0	8.4	30.2	3.5	1.6	0.43	-0.10	-0.10	-0.26	-0.07	-0.02	0.55	F
Chinese Taipei	611	66.3	0.56	66.3	11.0	8.5	13.9	0.3	0.0	0.56	-0.32	-0.24	-0.28	-0.01	.	0.44	F
Croatia	570	25.2	0.28	25.2	25.9	18.8	17.1	13.0	1.2	0.28	-0.16	0.00	-0.13	-0.02	-0.10	1.13	
*Cyprus	594	54.1	0.56	54.1	15.3	11.9	16.1	2.7	0.7	0.56	-0.24	-0.24	-0.30	-0.03	-0.11	-0.12	
*Czech Republic	741	37.7	0.47	37.7	26.7	10.6	19.1	6.0	0.4	0.47	-0.20	-0.08	-0.22	-0.11	-0.03	0.93	
Denmark	534	52.6	0.55	52.6	15.3	12.0	15.3	4.8	2.1	0.55	-0.27	-0.13	-0.29	-0.14	0.03	0.36	
*England	569	44.0	0.56	44.0	22.2	17.4	14.1	2.3	0.2	0.56	-0.26	-0.12	-0.31	-0.11	0.00	0.57	
Finland	708	53.5	0.45	53.5	17.2	12.5	13.0	3.8	0.4	0.45	-0.18	-0.19	-0.21	-0.12	0.02	0.54	
France	688	30.4	0.49	30.4	23.8	14.8	24.9	6.1	2.5	0.49	-0.19	-0.07	-0.25	-0.05	-0.07	0.05	
Georgia	564	23.9	0.35	23.9	22.8	14.5	26.7	12.1	3.5	0.35	-0.16	0.05	-0.16	-0.09	-0.08	0.72	C
Germany	561	32.6	0.31	32.6	23.4	20.8	14.5	8.8	0.4	0.31	-0.19	-0.04	-0.10	-0.05	-0.04	1.23	
*Hong Kong SAR	516	79.5	0.48	79.5	9.1	6.6	4.5	0.4	0.0	0.48	-0.30	-0.27	-0.21	-0.02	.	-0.07	F
*Hungary	719	51.8	0.56	51.8	18.0	11.2	16.8	2.2	0.4	0.56	-0.31	-0.15	-0.27	-0.06	-0.04	0.29	
Indonesia	577	27.8	0.25	27.8	28.9	17.4	24.5	1.4	0.3	0.25	-0.10	0.00	-0.16	0.00	-0.05	-0.21	
*Iran, Islamic Rep. of	546	40.7	0.48	40.7	25.9	10.6	18.7	4.1	1.1	0.48	-0.22	-0.12	-0.24	-0.06	-0.01	-0.44	
*Ireland	625	52.7	0.49	52.7	26.4	10.7	9.0	1.1	0.2	0.49	-0.27	-0.20	-0.19	-0.08	-0.05	0.26	F
*Italy	627	33.1	0.42	33.1	28.1	8.9	25.0	5.0	1.1	0.42	-0.16	-0.07	-0.22	-0.05	-0.02	0.75	F
*Japan	631	75.1	0.51	75.1	7.9	9.4	7.0	0.6	0.0	0.51	-0.19	-0.29	-0.31	-0.05	.	0.10	F
Kazakhstan	673	57.0	0.51	57.0	17.0	7.7	17.3	1.0	0.1	0.51	-0.28	-0.13	-0.28	-0.09	-0.10	0.01	F
*Korea, Rep. of	662	76.1	0.47	76.1	6.9	7.1	9.7	0.2	0.0	0.47	-0.19	-0.21	-0.34	0.02	.	0.14	F
Kuwait	503	17.1	0.27	17.1	32.3	13.4	29.2	8.0	3.4	0.27	-0.20	-0.04	0.03	-0.03	0.13	0.12	C A
Lithuania	662	34.5	0.53	34.5	21.3	18.7	22.3	3.2	0.6	0.53	-0.13	-0.16	-0.31	-0.06	-0.09	1.17	
Morocco	719	22.8	0.30	22.8	31.8	18.5	22.9	4.0	2.4	0.30	-0.12	-0.13	-0.03	-0.06	-0.04	-0.17	C
*Netherlands	641	51.9	0.43	51.9	17.4	13.5	15.0	2.2	0.5	0.43	-0.17	-0.15	-0.24	-0.07	-0.03	0.24	
*New Zealand	915	44.8	0.55	44.8	25.4	13.9	14.4	1.5	0.4	0.55	-0.27	-0.16	-0.29	-0.03	-0.06	-0.15	
Northern Ireland	433	56.8	0.51	56.8	17.6	14.5	10.9	0.2	0.0	0.51	-0.26	-0.20	-0.26	-0.05	.	0.35	
*Norway (5)	612	51.4	0.56	51.4	19.3	9.7	18.0	1.6	0.8	0.56	-0.25	-0.13	-0.33	-0.11	-0.02	0.58	F
Oman	1300	30.7	0.43	30.7	24.0	16.4	26.8	2.1	0.5	0.43	-0.18	-0.12	-0.16	-0.08	0.02	-0.09	
Poland	686	54.5	0.57	54.5	20.1	10.1	13.6	1.6	0.6	0.57	-0.28	-0.25	-0.26	-0.06	0.04	-0.01	
*Portugal	668	40.3	0.58	40.3	31.4	10.3	16.2	1.8	1.2	0.58	-0.22	-0.10	-0.39	-0.05	-0.02	0.64	C
Qatar	750	24.8	0.49	24.8	31.4	15.0	32.6	3.0	1.6	0.49	-0.13	-0.06	-0.27	-0.04	-0.07	0.50	C
Russian Federation	698	53.5	0.52	53.5	18.6	12.1	14.0	1.9	0.4	0.52	-0.27	-0.16	-0.25	-0.13	-0.10	0.78	
Saudi Arabia	618	24.5	0.31	24.5	26.2	19.3	25.9	4.1	1.1	0.31	-0.18	-0.03	-0.06	-0.08	-0.03	-0.10	C
Serbia	573	40.9	0.56	40.9	25.7	10.4	16.8	6.2	1.4	0.56	-0.27	-0.16	-0.23	-0.09	-0.08	0.73	
*Singapore	936	70.5	0.46	70.5	10.3	9.8	9.1	0.3	0.0	0.46	-0.15	-0.26	-0.30	-0.01	.	0.25	F
Slovak Republic	815	27.8	0.50	27.8	30.1	13.4	22.4	6.3	0.9	0.50	-0.25	-0.05	-0.18	-0.06	-0.08	1.02	
*Slovenia	650	46.7	0.59	46.7	21.8	11.0	15.6	4.9	0.5	0.59	-0.27	-0.15	-0.30	-0.12	-0.01	0.48	
Spain	1104	33.2	0.43	33.2	29.8	9.4	23.6	4.0	0.8	0.43	-0.14	-0.08	-0.24	-0.03	-0.03	0.75	F
Sweden	501	44.5	0.51	44.5	21.1	13.9	14.4	6.1	0.7	0.51	-0.24	-0.10	-0.22	-0.19	-0.01	0.37	
Turkey	931	53.3	0.49	53.3	15.2	12.5	17.8	1.3	0.2	0.49	-0.25	-0.10	-0.32	-0.01	-0.05	-0.42	
United Arab Emirates	3025	29.7	0.49	29.7	23.0	12.2	32.4	2.7	0.4	0.49	-0.18	-0.08	-0.25	-0.04	-0.01	0.46	
*United States	1439	48.8	0.56	48.8	21.4	9.7	16.4	3.7	1.0	0.56	-0.19	-0.19	-0.35	-0.08	-0.06	0.38	F
*Reference Avg (n=19)	14724	51.9	0.51	51.9	20.1	10.7	14.9	2.4	0.5	0.51	-0.23	-0.17	-0.28	-0.06	-0.04	0.28	
International Avg (n=47)	36268	43.4	0.47	43.4	22.0	12.4	18.5	3.7	0.8	0.47	-0.21	-0.13	-0.24	-0.06	-0.04	0.38	
Buenos Aires, Argentina	443	20.9	0.41	20.9	28.2	8.0	20.7	22.3	3.8	0.41	-0.10	-0.01	-0.12	-0.17	-0.13	0.45	C F
Ontario, Canada	659	39.0	0.47	39.0	23.9	12.3	21.9	2.9	1.5	0.47	-0.19	-0.08	-0.29	-0.03	-0.06	0.56	F
Quebec, Canada	401	46.0	0.50	46.0	20.9	9.3	22.1	1.8	0.7	0.50	-0.10	-0.23	-0.32	-0.06	-0.05	0.65	F
Norway (4)	586	36.4	0.43	36.4	25.9	12.6	21.2	4.0	1.0	0.43	-0.21	-0.11	-0.15	-0.08	0.05	0.48	
Abu Dhabi, UAE	713	24.7	0.43	24.7	25.8	15.3	30.9	3.2	0.1	0.43	-0.19	-0.10	-0.15	0.01	-0.02	0.34	C
Dubai, UAE	1074	40.5	0.56	40.5	18.4	11.7	27.0	2.3	0.5	0.56	-0.17	-0.13	-0.36	-0.05	0.01	0.44	
Florida, US	282	54.8	0.45	54.8	19.4	13.3	9.3	3.2	1.1	0.45	-0.26	-0.17	-0.19	-0.06	0.02	0.15	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= 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; V= Difficulty greater than 95%.

Exhibit 11.2: Example International Item Statistics for a TIMSS 2015 Constructed Response Item

Trends in International Mathematics and Science Study - TIMSS 2015 Assessment Results - 8th Grade
International Item Review Statistics (Unweighted)

Science: Biology / Knowing (S14_04 - S062101) Type: CR 2 Points
Label: Stomach tissue and function

Country	Cases	DIFF	DISC	Percentages					Point Biserials					RDIFF	Reliability			Flags
				P_0	P_1	P_2	P_OM	P_NR	PB_0	PB_1	PB_2	PB_OM	PB_NR		N	Score	Code	
*Australia	1455	59.0	0.59	16.5	41.6	38.2	3.7	0.2	-0.40	-0.08	0.49	-0.26	-0.05	-0.57	317	99.1	98.7	
Bahrain	698	35.8	0.53	42.5	34.0	18.8	4.7	0.1	-0.41	0.22	0.38	-0.24	-0.07	0.13	196	88.8	88.3	
Botswana (9)	850	29.2	0.55	48.7	35.5	11.4	4.4	0.0	-0.42	0.22	0.43	-0.16	.	0.17	210	95.2	95.2	
*Canada	1245	52.5	0.48	22.1	43.6	30.7	3.6	0.2	-0.34	0.01	0.39	-0.24	-0.06	-0.10	220	95.9	95.5	
*Chile	690	45.3	0.53	27.8	32.7	28.9	10.6	0.3	-0.31	0.03	0.46	-0.27	0.02	-0.35	181	96.1	96.1	
*Chinese Taipei	821	73.3	0.47	9.9	32.2	57.2	0.7	0.0	-0.46	-0.04	0.34	-0.17	.	-0.59	220	98.2	98.2	F
Egypt	1114	17.0	0.44	60.6	21.1	6.5	11.8	0.4	-0.26	0.33	0.26	-0.23	-0.02	0.35	209	99.0	98.6	AF
*England	675	58.0	0.52	19.9	41.1	37.5	1.5	0.4	-0.41	-0.05	0.43	-0.16	-0.08	-0.26	237	97.0	97.0	
Georgia	576	52.1	0.60	23.9	27.4	38.4	10.3	0.5	-0.39	-0.01	0.53	-0.28	-0.07	-0.92	206	97.6	97.6	
*Hong Kong SAR	592	68.6	0.48	13.7	34.0	51.6	0.7	0.2	-0.41	-0.06	0.38	-0.20	-0.09	-0.70	194	100.0	100.0	
*Hungary	682	64.7	0.59	16.6	34.3	47.5	1.6	0.0	-0.47	-0.11	0.50	-0.15	.	-0.61	190	98.4	98.4	
*Iran, Islamic Rep. of	879	38.7	0.54	35.7	37.1	20.2	7.1	0.2	-0.38	0.15	0.42	-0.23	-0.08	-0.10	214	99.5	99.5	
*Ireland	680	59.7	0.51	16.4	40.7	39.4	3.5	0.3	-0.31	-0.11	0.44	-0.24	-0.04	-0.47	247	98.4	98.4	
*Israel	790	47.2	0.58	30.8	32.5	31.0	5.7	0.6	-0.43	0.09	0.47	-0.26	-0.11	0.11	195	98.5	98.5	
*Italy	648	50.8	0.53	23.5	41.8	29.9	4.8	0.0	-0.32	-0.05	0.47	-0.25	.	-0.31	199	96.0	96.0	
*Japan	671	73.4	0.53	11.0	29.0	59.0	1.0	0.1	-0.40	-0.18	0.46	-0.20	-0.12	-0.63	213	100.0	100.0	
*Jordan	1133	23.0	0.46	56.7	28.9	8.6	5.9	0.0	-0.33	0.30	0.29	-0.23	.	0.43	209	99.5	97.1	AF
Kazakhstan	704	48.9	0.55	30.2	25.8	36.0	8.0	0.3	-0.34	0.05	0.48	-0.35	-0.05	0.02	221	86.0	86.0	
*Korea, Rep. of	762	67.7	0.55	18.4	26.0	54.7	0.9	0.0	-0.45	-0.11	0.48	-0.20	.	-0.51	267	96.3	95.9	
Kuwait	661	21.1	0.56	50.3	22.2	10.0	17.5	1.4	-0.23	0.26	0.44	-0.32	-0.02	0.30	188	97.9	97.3	F
Lebanon	542	25.4	0.47	44.4	28.2	11.3	16.2	1.8	-0.28	0.34	0.27	-0.27	-0.11	0.15	192	93.8	92.7	A
*Lithuania	621	58.5	0.59	18.2	34.1	41.4	6.3	0.0	-0.36	-0.11	0.53	-0.29	.	-0.44	197	100.0	100.0	
*Malaysia	1383	39.8	0.58	37.5	38.4	20.6	3.5	0.0	-0.47	0.19	0.43	-0.22	.	0.47	231	99.1	98.7	
Malta	540	39.8	0.60	37.8	36.3	21.7	4.3	0.0	-0.50	0.23	0.43	-0.23	.	0.07	210	91.0	91.0	
*Morocco	1899	19.5	0.40	57.2	24.9	7.1	10.8	0.3	-0.25	0.23	0.29	-0.16	0.01	0.12	206	95.1	91.7	F
*New Zealand	1149	54.3	0.56	22.4	39.0	34.8	3.8	0.2	-0.41	-0.01	0.46	-0.23	-0.06	-0.27	204	99.0	99.0	
Norway (9)	666	57.0	0.58	25.5	24.8	44.6	5.1	0.3	-0.40	-0.06	0.52	-0.28	-0.02	-0.49	229	97.4	97.4	
Oman	1282	30.6	0.54	46.6	29.7	15.7	8.0	0.3	-0.33	0.19	0.43	-0.29	-0.06	0.22	229	97.4	96.5	
Qatar	765	36.3	0.60	38.8	36.8	17.9	6.4	0.7	-0.45	0.22	0.45	-0.24	-0.07	0.24	237	98.7	97.9	
*Russian Federation	685	75.0	0.49	9.2	25.4	62.3	3.1	0.0	-0.33	-0.16	0.42	-0.24	.	-0.86	249	98.8	98.4	F
Saudi Arabia	544	26.1	0.58	55.0	28.3	11.9	4.8	0.0	-0.44	0.25	0.45	-0.18	.	-0.14	190	100.0	99.5	
*Singapore	875	68.5	0.57	13.9	33.4	51.8	0.9	0.0	-0.48	-0.09	0.46	-0.17	.	-0.13	218	96.8	96.8	
*Slovenia	605	67.2	0.59	17.2	26.3	54.0	2.5	0.0	-0.38	-0.23	0.57	-0.24	.	-0.41	211	100.0	100.0	
South Africa (9)	1787	18.7	0.55	57.2	24.8	6.3	11.7	0.2	-0.34	0.27	0.44	-0.16	0.01	0.12	194	95.9	92.8	F
*Sweden	582	47.4	0.51	31.0	35.6	29.6	3.8	0.2	-0.35	-0.05	0.47	-0.16	-0.10	0.09	199	95.5	95.5	
*Thailand	929	47.1	0.53	21.4	42.6	25.9	10.1	0.1	-0.30	0.08	0.41	-0.31	0.02	-0.58	207	100.0	100.0	
*Turkey	865	29.8	0.48	46.1	39.8	9.9	4.2	0.0	-0.38	0.26	0.32	-0.18	.	0.81	173	98.3	98.3	F
United Arab Emirates	2574	34.1	0.59	43.1	29.4	19.4	8.0	0.2	-0.42	0.23	0.45	-0.27	0.01	0.28	722	96.4	96.4	
*United States	1457	59.8	0.53	18.0	40.0	39.8	2.2	0.2	-0.38	-0.11	0.46	-0.16	-0.04	-0.35	215	99.1	99.1	
*Reference Avg (n=25)	22773	54.0	0.53	24.4	35.0	36.5	4.1	0.1	-0.38	-0.01	0.43	-0.22	-0.06	-0.25	5413	98.2	97.9	
International Avg (n=39)	36076	46.7	0.54	31.2	32.8	30.3	5.7	0.3	-0.38	0.06	0.43	-0.23	-0.05	-0.15	8846	97.2	96.8	
Buenos Aires, Argentina	440	25.1	0.52	32.1	25.1	12.6	30.2	2.3	-0.09	0.24	0.40	-0.43	-0.08	-0.12	151	98.0	97.4	
Ontario, Canada	640	55.7	0.50	20.2	41.0	35.2	3.6	0.2	-0.32	-0.03	0.42	-0.28	-0.06	-0.30	126	94.4	94.4	
Quebec, Canada	565	49.0	0.47	24.2	46.9	25.6	3.4	0.4	-0.35	0.06	0.36	-0.19	-0.07	0.22	90	97.8	96.7	
Norway (8)	689	55.1	0.61	24.7	27.3	41.5	6.6	0.6	-0.43	-0.06	0.55	-0.22	-0.11	-0.61	231	95.2	94.8	
Abu Dhabi, UAE	690	24.6	0.60	47.5	25.7	11.8	15.1	0.1	-0.29	0.30	0.44	-0.36	0.04	0.43	227	96.0	96.0	
Dubai, UAE	871	47.9	0.57	31.5	34.3	30.8	3.4	0.1	-0.48	0.13	0.43	-0.19	-0.00	0.13	279	97.8	97.8	
Florida, US	291	56.7	0.61	22.1	35.3	39.1	3.5	0.7	-0.44	-0.13	0.56	-0.14	-0.02	-0.35	52	96.2	96.2	

Keys: DIFF= Percent correct score; DISC= Item discrimination; P_0...P_2= Percentage obtaining score level; P_OM, P_NR= Percentage Omitted, Not Reached;
PB_0...PB_2= Point Biserials for score level; PB_OM, PB_NR= Point Biserials for Omitted, Not Reached; RDIFF= Rasch difficulty;
Reliability: N= Responses double scored; Score= Percentage agreement on score; Code= Percentage agreement on code.
Flags: A= Point Biserials not ordered; 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 less than 85%; 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 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 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 or 2 score points) included the percent correct and point-biserial of each score level. Constructed response item tables 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—because TIMSS has a 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 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 by grade and subject, and within each table 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 items, it is the average percentage of points. For example, if 25 percent of students scored 2 points, 50 percent scored 1 point on a 2-point item, and 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.

other 25 percent score 0 points, 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, and PCT_2: 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, and PB_2: Available for constructed response items. These columns present the point-biserial correlations between the score levels on the item (0, 1, or 2) 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 of 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 (Score): 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): This column contains the percentage of agreement on the two-digit diagnostic codes assigned by the two independent TIMSS scorers.

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:

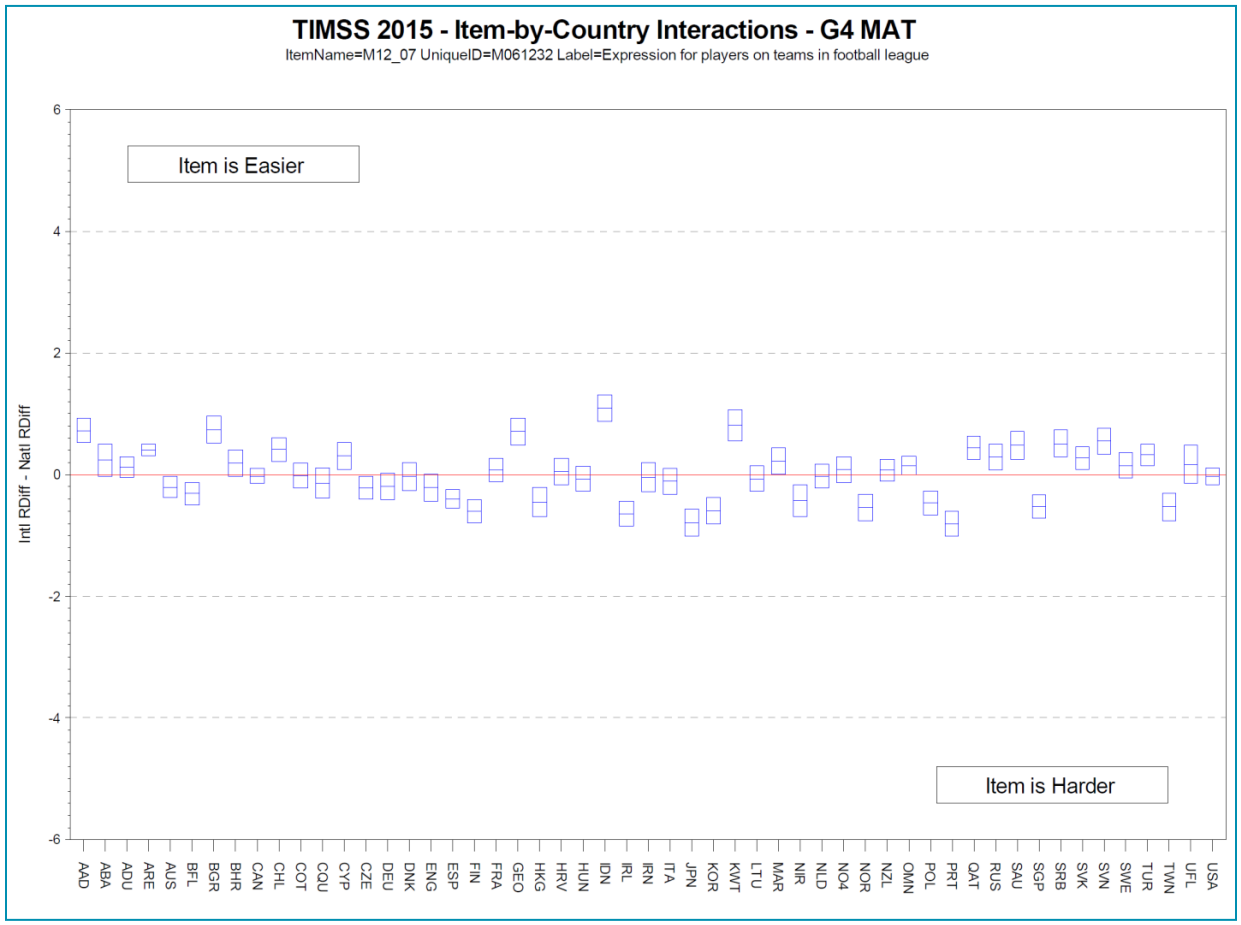
- The item discrimination (DISC) was less than 0.10 (flag D)
- The item difficulty (DIFF) was less than .25 for multiple-choice items (flag C)
- The item difficulty (DIFF) exceeded .95 (flag V)
- The Rasch difficulty estimate (RDIFF) for a given country made the item either easier (flag E) or more difficult (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 85 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. An example of the graphical displays is provided in Exhibit 11.3.

Exhibit 11.3: Example Plot of Item-by-Country Interaction for a TIMSS 2015 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 2015 included achievement items from previous assessments as well as items developed for use for the first time in 2015. Accordingly, the TIMSS 2015 assessments included items from 2007, 2011, and 2015. An important review step, therefore, was to check that these "trend items" had statistical properties in 2015 similar to those they had in the previous assessments (e.g., a TIMSS item that was relatively easy in 2011 should still be relatively easy in 2015).

As can be seen in the example in Exhibit 11.4, the trend item review focused on statistics for trend items from the current and previous assessments (2015 and 2011) 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 trends.

Exhibit 11.4: Example Item Statistics for a TIMSS 2015 Trend Item

Trends in International Mathematics and Science Study - TIMSS 2015 Assessment Results - 8th Grade
Trend Achievement Data Almanac for Science Items (Weighted)

S06_08 (S052049): Chemistry / Applying Type: CR 2 Points
Label: Separating iron and copper

COUNTRY	Year	N	20 %	10 %	79 %	OMITTED %	NOT REACHED %	V1 %	V2 %	GIRL PCT RIGHT	BOY PCT RIGHT
Australia	2011	1078	13.8	34.7	48.8	2.6	0.0	48.5	13.8	12.1	15.5
	2015	1468	12.1	32.2	52.1	3.3	0.4	44.3	12.1	13.5	10.5
Bahrain	2011	654	21.7	27.8	46.3	3.2	1.0	49.5	21.7	27.8	15.3
	2015	702	11.4	34.2	50.5	3.3	0.7	45.5	11.4	15.4	7.5
Botswana	2011	764	7.3	18.3	70.0	3.7	0.7	25.6	7.3	7.8	6.8
	2015	847	6.6	20.7	68.4	3.5	0.8	27.3	6.6	5.9	7.3
Chile	2011	828	4.2	26.7	57.1	10.8	1.3	30.9	4.2	4.3	4.2
	2015	676	11.1	18.5	53.2	16.7	0.5	29.6	11.1	12.1	10.1
Chinese Taipei	2011	718	28.3	42.3	27.1	2.2	0.1	70.6	28.3	30.9	25.9
	2015	821	24.7	47.2	25.3	2.7	0.2	71.9	24.7	27.1	22.1
England	2011	542	24.3	34.2	36.3	4.3	0.8	58.5	24.3	24.6	24.1
	2015	692	22.1	32.4	41.1	4.0	0.4	54.5	22.1	24.4	19.2
Georgia	2011	653	14.8	28.5	38.3	17.4	0.9	43.4	14.8	11.7	17.7
	2015	572	19.7	35.4	36.5	8.1	0.4	55.0	19.7	19.1	20.2
Hong Kong SAR	2011	576	12.5	34.6	49.8	3.0	0.2	47.0	12.5	15.5	9.3
	2015	597	11.6	35.7	50.9	1.8	0.0	47.3	11.6	11.8	11.3
Hungary	2011	730	17.1	46.0	33.8	3.0	0.0	63.2	17.1	15.1	18.9
	2015	704	17.7	48.6	30.8	2.9	0.0	66.3	17.7	21.4	13.8
Iran, Islamic Rep. of	2011	874	5.7	41.9	47.9	4.5	0.1	47.5	5.7	6.4	5.0
	2015	886	3.9	40.8	50.2	4.7	0.3	44.8	3.9	5.2	2.7
Israel	2011	664	18.1	35.0	43.5	2.6	0.8	53.1	18.1	21.9	14.6
	2015	787	18.5	37.9	38.6	4.5	0.5	56.3	18.5	22.6	14.0
Italy	2011	563	15.5	40.0	33.2	11.1	0.2	55.6	15.5	18.7	12.5
	2015	627	9.4	46.4	36.3	7.9	0.0	55.8	9.4	8.9	9.9
Japan	2011	628	32.4	25.6	38.2	3.8	0.0	57.9	32.4	29.6	35.3
	2015	681	19.7	43.5	33.2	3.1	0.5	63.2	19.7	19.1	20.3
Jordan	2011	1092	12.7	36.1	45.2	5.3	0.8	48.8	12.7	17.3	8.4
	2015	1134	9.0	31.3	52.2	7.4	0.2	40.3	9.0	9.4	8.6
Kazakhstan	2011	617	30.6	32.8	24.5	11.7	0.4	63.4	30.6	31.1	30.1
	2015	697	32.6	38.7	24.8	3.3	0.6	71.3	32.6	31.0	34.4
Korea, Rep. of	2011	741	21.5	47.2	29.7	1.5	0.1	68.7	21.5	24.6	18.2
	2015	757	13.9	43.0	41.5	1.6	0.0	56.9	13.9	16.4	12.0
Lebanon	2011	557	11.7	19.1	56.5	11.5	1.2	30.8	11.7	10.9	12.6
	2015	548	10.3	25.5	45.8	13.9	4.3	35.9	10.3	11.5	9.1
Lithuania	2011	673	5.9	45.7	44.2	4.0	0.3	51.5	5.9	7.1	4.8
	2015	621	6.6	41.7	48.6	3.1	0.1	48.3	6.6	9.2	3.7

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

Exhibit 11.4: Example Item Statistics for a TIMSS 2015 Trend Item (Continued)

Trends in International Mathematics and Science Study - TIMSS 2015 Assessment Results - 8th Grade
Trend Achievement Data Almanac for Science Items (Weighted)

S06_08 (S052049): Chemistry / Applying Type: CR 2 Points
Label: Separating iron and copper

COUNTRY	Year	N	20 %	10 %	79 %	OMITTED %	NOT REACHED %	V1 %	V2 %	GIRL PCT RIGHT	BOY PCT RIGHT
Malaysia	2011	833	12.0	26.4	53.4	6.7	1.6	38.4	12.0	12.6	11.4
	2015	1377	14.8	34.3	43.9	6.6	0.5	49.0	14.8	15.3	14.3
Morocco	2011	1335	3.0	22.8	66.3	6.8	1.1	25.7	3.0	3.4	2.6
	2015	1879	0.3	17.4	74.8	7.0	0.6	17.6	0.3	0.5	0.1
New Zealand	2011	755	21.1	29.9	43.6	5.1	0.4	51.0	21.1	20.9	21.2
	2015	1122	23.1	26.5	46.8	3.2	0.4	49.5	23.1	25.4	20.6
Oman	2011	1363	14.9	23.4	55.0	6.1	0.6	38.3	14.9	21.7	7.7
	2015	1255	12.2	25.2	55.6	6.6	0.4	37.4	12.2	14.6	10.0
Qatar	2011	624	15.0	25.4	55.1	4.3	0.2	40.3	15.0	19.0	11.3
	2015	768	11.7	31.1	50.9	5.6	0.7	42.8	11.7	12.3	11.2
Russian Federation	2011	709	15.8	49.3	28.4	5.2	1.3	65.0	15.8	13.3	18.2
	2015	677	35.8	32.5	28.0	3.6	0.1	68.3	35.8	33.6	37.8
Saudi Arabia	2011	628	4.1	19.8	72.4	3.4	0.3	23.9	4.1	3.3	4.9
	2015	531	3.7	24.0	69.4	2.5	0.5	27.6	3.7	4.0	3.3
Singapore	2011	850	33.1	31.8	33.6	1.3	0.1	65.0	33.1	36.7	29.7
	2015	872	43.4	30.6	25.4	0.6	0.0	74.0	43.4	43.5	43.3
Slovenia	2011	633	20.8	47.3	29.8	2.1	0.0	68.0	20.8	21.3	20.4
	2015	604	33.9	35.5	28.5	2.1	0.0	69.4	33.9	35.7	32.1
South Africa	2011	1704	6.0	7.0	77.8	7.7	1.5	13.0	6.0	6.0	5.9
	2015	1791	1.3	14.4	78.3	4.4	1.6	15.7	1.3	1.6	0.9
Sweden	2011	797	11.2	38.0	46.2	4.3	0.2	49.3	11.2	11.2	11.3
	2015	585	15.9	35.0	42.4	6.0	0.8	50.9	15.9	18.6	13.9
Thailand	2011	859	9.6	33.8	39.6	16.4	0.6	43.4	9.6	10.9	8.0
	2015	918	6.9	32.1	48.9	11.8	0.2	39.0	6.9	10.7	2.3
Turkey	2011	990	10.8	41.0	37.8	9.7	0.7	51.9	10.8	12.7	8.8
	2015	869	22.0	26.8	45.1	6.1	0.0	48.8	22.0	25.9	18.4
United Arab Emirates	2011	1999	12.0	33.8	50.0	3.3	0.8	45.8	12.0	12.6	11.5
	2015	2581	10.2	33.9	50.7	4.9	0.3	44.1	10.2	11.3	9.2
United States	2011	1490	13.9	37.0	45.7	3.2	0.3	50.8	13.9	13.7	14.0
	2015	1475	10.3	37.5	48.5	2.7	0.9	47.8	10.3	10.8	9.8
International Avg (n=33)	2011	28521	15.2	32.8	45.6	5.8	0.6	48.0	15.2	16.3	14.1
	2015	31121	15.3	33.0	46.0	5.1	0.5	48.4	15.3	16.6	14.1
Ontario, Canada	2011	682	19.8	28.5	44.9	6.2	0.7	48.3	19.8	18.9	20.4
	2015	644	10.0	26.9	56.3	6.1	0.8	36.8	10.0	9.7	10.2

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

Exhibit 11.4: Example Item Statistics for a TIMSS 2015 Trend Item (Continued)

Trends in International Mathematics and Science Study - TIMSS 2015 Assessment Results - 8th Grade
Trend Achievement Data Almanac for Science Items (Weighted)

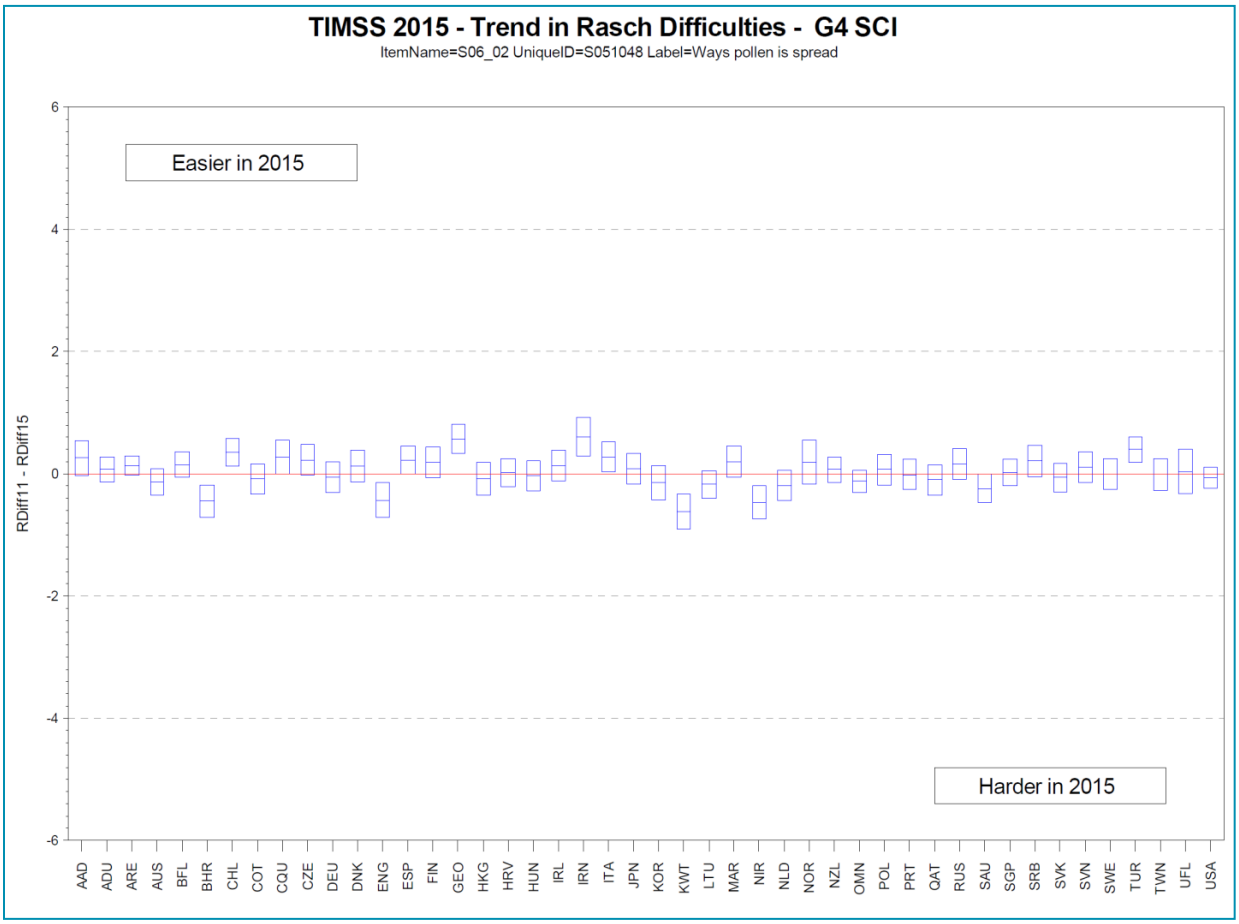
S06_08 (S052049): Chemistry / Applying Type: CR 2 Points
Label: Separating iron and copper

COUNTRY	Year	N	20 %	10 %	79 %	OMITTED %	NOT REACHED %	V1 %	V2 %	GIRL PCT RIGHT	BOY PCT RIGHT
Quebec, Canada	2011	880	20.2	33.2	41.5	4.5	0.6	53.4	20.2	20.9	19.5
	2015	555	13.9	33.1	46.8	5.0	1.2	47.0	13.9	15.0	12.7
Norway (8)	2011	540	8.6	37.2	48.9	5.0	0.3	45.8	8.6	11.1	5.9
	2015	685	10.3	35.4	46.7	7.2	0.4	45.8	10.3	10.1	10.5
Abu Dhabi, UAE	2011	624	12.2	31.7	52.5	2.7	0.9	43.9	12.2	9.8	14.3
	2015	692	7.1	30.0	56.4	6.0	0.5	37.1	7.1	7.8	6.5
Dubai, UAE	2011	789	14.0	32.8	48.1	4.0	1.2	46.8	14.0	15.3	12.7
	2015	879	16.8	37.4	42.1	3.3	0.4	54.3	16.8	17.7	15.8
Florida, US	2011	246	10.9	34.9	52.1	2.2	0.0	45.8	10.9	11.4	10.4
	2015	298	8.4	33.1	57.0	1.5	0.0	41.5	8.4	9.3	7.6

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

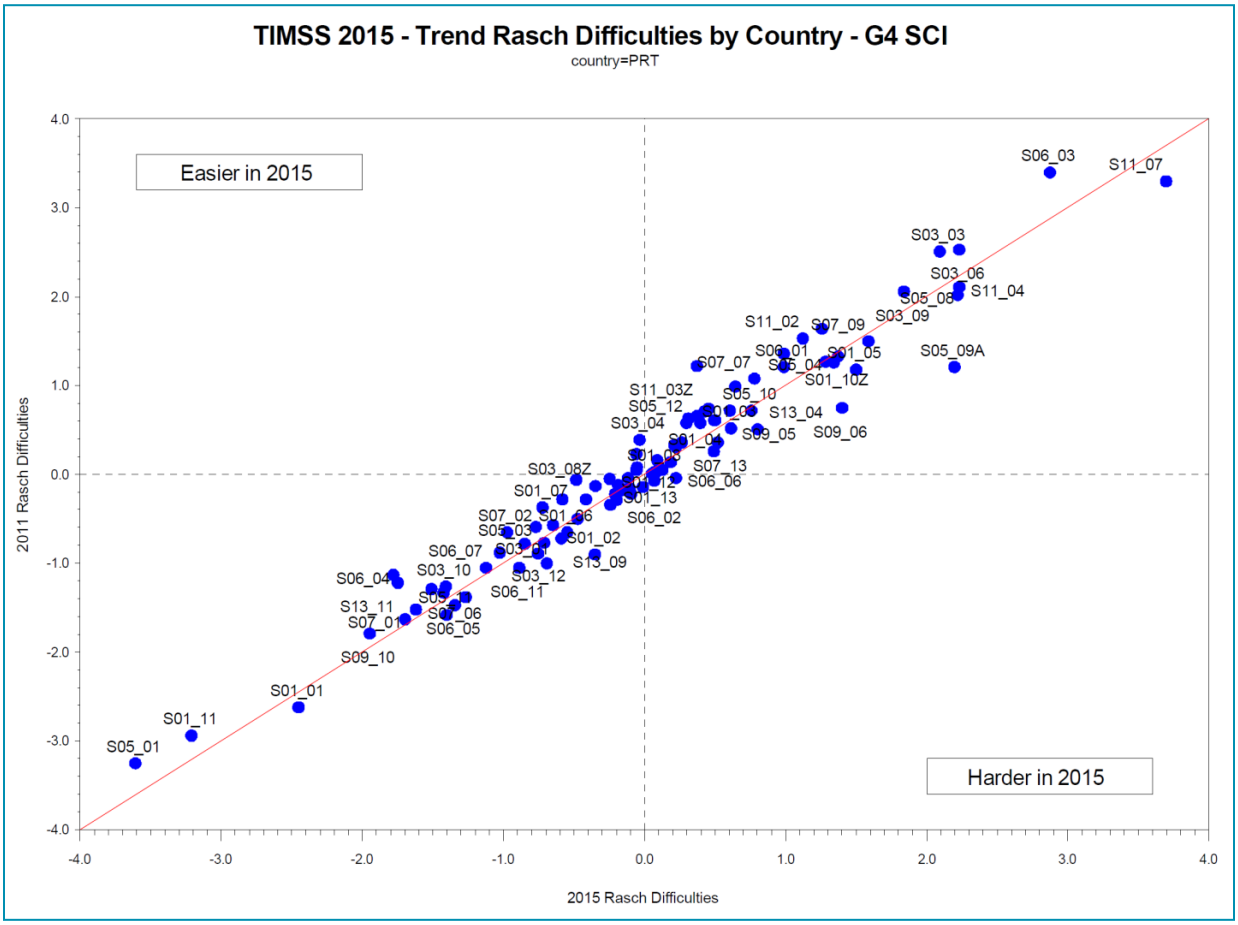
While some changes in item difficulties were anticipated as countries' overall achievement may have improved or declined, items were noted if the difference between the Rasch difficulties across the two assessments for a particular country was greater than 2 logits. The TIMSS & PIRLS International Study Center used two different graphical displays to examine the differences in item difficulties. The first of these, shown for an example item in Exhibit 11.5, displays the difference in Rasch item difficulty of the item between 2015 and 2011 for each country. A positive difference for a country indicates that the item was relatively easier in 2015, and a negative difference indicates that the item was relatively more difficult.

Exhibit 11.5: Example Plot of Differences in Rasch Item Difficulties Between 2015 and 2011 for a TIMSS 2015 Trend Item



The second graphical display, presented in Exhibit 11.6, shows the performance of a given country on all trend items simultaneously. For each country, the graph plots the 2015 Rasch difficulty of every trend item against its Rasch difficulty in 2011. Where there were no differences between the difficulties in the two successive administrations, the data points aligned on or near the diagonal.

Exhibit 11.6: Example Plot of Rasch Item Difficulties Across TIMSS Trend Items by Country



Reliability

Documenting the reliability of the TIMSS 2015 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 11.7 and 11.8 display the TIMSS 2015 fourth and eighth grade mathematics and science test reliability coefficients for every country, respectively. Exhibit 11.7 also displays the test reliability coefficients for TIMSS Numeracy. These coefficients are the median Cronbach's alpha reliability across all TIMSS 2015 assessment booklets. 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 0.83 for mathematics and 0.78 for science, and at the eighth

grade, 0.88 for mathematics and 0.83 for science. The international median reliability for TIMSS Numeracy was 0.92.

Exhibit 11.7: Cronbach's Alpha Reliability Coefficient – TIMSS 2015 Fourth Grade

Country	Reliability Coefficient		
	Mathematics	Numeracy	Science
Australia	0.86	—	0.79
Bahrain	0.81	0.93	0.82
Belgium (Flemish)	0.80	—	0.73
Bulgaria	0.86	—	0.85
Canada	0.82	—	0.79
Chile	0.80	—	0.76
Chinese Taipei	0.83	—	0.77
Croatia	0.81	—	0.73
Cyprus	0.85	—	0.77
Czech Republic	0.83	—	0.78
Denmark	0.84	—	0.76
England	0.86	—	0.77
Finland	0.81	—	0.74
France	0.82	—	0.78
Georgia	0.82	—	0.76
Germany	0.82	—	0.77
Hong Kong SAR	0.81	—	0.77
Hungary	0.88	—	0.82
Indonesia	0.76	0.91	0.76
Iran, Islamic Rep. of	0.83	0.94	0.80
Ireland	0.84	—	0.77
Italy	0.82	—	0.75
Japan	0.83	—	0.77
Jordan	—	0.92	—
Kazakhstan	0.86	—	0.81
Korea, Rep. of	0.82	—	0.75
Kuwait	0.76	0.92	0.78
Lithuania	0.83	—	0.77
Morocco	0.76	0.92	0.78
Netherlands	0.77	—	0.71
New Zealand	0.85	—	0.82

Exhibit 11.7: Cronbach's Alpha Reliability Coefficient – TIMSS 2015 Fourth Grade (Continued)

Country	Reliability Coefficient		
	Mathematics	Numeracy	Science
Northern Ireland	0.87	—	0.77
Norway (5)	0.83	—	0.72
Oman	0.83	—	0.84
Poland	0.83	—	0.78
Portugal	0.84	—	0.72
Qatar	0.84	—	0.82
Russian Federation	0.84	—	0.77
Saudi Arabia	0.76	—	0.80
Serbia	0.87	—	0.80
Singapore	0.88	—	0.83
Slovak Republic	0.84	—	0.82
Slovenia	0.82	—	0.78
South Africa (5)	—	0.93	—
Spain	0.80	—	0.77
Sweden	0.81	—	0.79
Turkey	0.87	—	0.81
United Arab Emirates	0.87	—	0.85
United States	0.87	—	0.82
International Median	**0.83**	**0.92**	**0.78**
Benchmarking Participants			
Buenos Aires, Argentina	0.78	0.91	0.78
Ontario, Canada	0.83	—	0.79
Quebec, Canada	0.80	—	0.73
Norway (4)	0.81	—	0.74
Abu Dhabi, UAE	0.86	—	0.85
Dubai, UAE	0.87	—	0.85
Florida, US	0.85	—	0.81

Exhibit 11.8: Cronbach's Alpha Reliability Coefficient – TIMSS 2015 Eighth Grade

Country	Reliability Coefficient	
	Mathematics	Science
Australia	0.89	0.84
Bahrain	0.83	0.86
Botswana (9)	0.75	0.79
Canada	0.87	0.80
Chile	0.82	0.79
Chinese Taipei	0.92	0.87
Egypt	0.81	0.79
England	0.90	0.85
Georgia	0.87	0.78
Hong Kong SAR	0.89	0.81
Hungary	0.91	0.86
Iran, Islamic Rep. of	0.87	0.82
Ireland	0.88	0.83
Israel	0.92	0.88
Italy	0.86	0.81
Japan	0.91	0.83
Jordan	0.77	0.80
Kazakhstan	0.91	0.85
Korea, Rep. of	0.91	0.84
Kuwait	0.82	0.83
Lebanon	0.80	0.80
Lithuania	0.88	0.83
Malaysia	0.88	0.85
Malta	0.88	0.87
Morocco	0.72	0.74
New Zealand	0.90	0.85
Norway (9)	0.87	0.83
Oman	0.82	0.84
Qatar	0.88	0.87
Russian Federation	0.89	0.83
Saudi Arabia	0.76	0.79
Singapore	0.91	0.87
Slovenia	0.87	0.84
South Africa (9)	0.80	0.82

Exhibit 11.8: Cronbach's Alpha Reliability Coefficient – TIMSS 2015 Eighth Grade (Continued)

Country	Reliability Coefficient	
	Mathematics	Science
Sweden	0.86	0.84
Thailand	0.86	0.80
Turkey	0.91	0.87
United Arab Emirates	0.89	0.87
United States	0.89	0.85
International Median	**0.88**	**0.83**
Benchmarking Participants		
Buenos Aires, Argentina	0.82	0.79
Ontario, Canada	0.87	0.81
Quebec, Canada	0.84	0.78
Norway (8)	0.83	0.80
Abu Dhabi, UAE	0.88	0.86
Dubai, UAE	0.90	0.86
Florida, US	0.89	0.86

Scoring Reliability for Constructed Response Items

A sizeable proportion of the items in the TIMSS 2015 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 2015, 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 Chapter 1: Developing the TIMSS 2015 Achievement Items for more information on the scoring guides and see Chapter 6: Survey Operations Procedures for information on the scoring process.

Within-Country Scoring Reliability

To gather and document information about the within-country agreement among scorers for TIMSS 2015, 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—96 percent or above, on average internationally. In TIMSS 2015 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 the fourth grade, on average. See Appendix 11A 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 2015 constructed response items used in TIMSS 2011 were scored in the same way in both assessments. In anticipation of this, countries that participated in TIMSS 2011 sent samples of scored student booklets from the 2011 data collections to the IEA Data Processing and Research Center (IEA DPC), where they were digitally scanned and stored for later use. As a check on scoring consistency from one administration to the next, staff members working in each country on scoring the 2015 data were asked also to score these 2011 responses using the Trend Reliability Scoring Software developed by the IEA DPC. Each country scored 200 responses for each of 21 mathematics and 23 science items at the fourth grade, and 27 mathematics and 33 science items at the eighth grade.

There was a very high degree of scoring consistency in TIMSS 2015. The exact agreement between the scores awarded in 2011 and those given by the 2015 scorers ranged from 92 percent in science to 98 percent in mathematics at the fourth grade, on average internationally. 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 Appendix 11B.

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 2015, 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 11 mathematics and 10 science items at the fourth grade, and 13 mathematics and 13 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 46 countries at fourth grade and 37 countries at eighth grade 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 1,035 comparisons at fourth grade and 666 comparisons at eighth grade for each student response to each item. This resulted in more than 130,000 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 in TIMSS 2015 was high. The exact agreement between the scores awarded across countries ranged from 86 percent in science to 97 percent in mathematics at the fourth grade and from 83 percent in science to 93 percent in mathematics at the eighth grade, on average internationally. There also was high agreement at the diagnostic score level, where percent agreement ranged from 79 percent in science at the eighth grade to 97 percent in mathematics at the fourth grade, on average. See Appendix 11C 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 2015, 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
- For trend items, an item performed substantially differently in 2015 compared to the TIMSS 2011 administration, 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 2015 achievement data involved review of more than 750 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 Appendix 11D: 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 Appendix 11E for details about how score points were awarded for each derived item.

Appendix 11A: TIMSS 2015 Within–Country Scoring Reliability for the Constructed Response Items

TIMSS 2015 Within-Country Scoring Reliability for the Fourth Grade Constructed Response Mathematics Items

Country	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	
		Minimum	Maximum		Minimum	Maximum
Australia	98	87	100	97	86	100
Bahrain	99	90	100	99	89	100
Belgium (Flemish)	98	90	100	97	80	100
Bulgaria	99	96	100	98	94	100
Canada	97	80	100	95	77	100
Chile	99	91	100	98	87	100
Chinese Taipei	99	89	100	99	88	100
Croatia	99	89	100	98	81	100
Cyprus	100	99	100	100	98	100
Czech Republic	98	89	100	97	86	100
Denmark	97	89	100	95	84	100
England	99	95	100	99	92	100
Finland	99	90	100	99	89	100
France	98	84	100	97	68	100
Georgia	99	96	100	98	85	100
Germany	98	78	100	98	78	100
Hong Kong SAR	100	98	100	100	98	100
Hungary	99	96	100	99	95	100
Indonesia	99	92	100	96	68	100
Iran, Islamic Rep. of	99	89	100	97	85	100
Ireland	99	94	100	99	94	100
Italy	98	92	100	97	86	100
Japan	99	96	100	99	96	100
Kazakhstan	93	84	99	93	82	98
Korea, Rep. of	100	95	100	99	95	100
Kuwait	99	96	100	98	93	100
Lithuania	100	98	100	100	97	100
Morocco	95	43	100	91	42	99

TIMSS 2015 Within-Country Scoring Reliability for the Fourth Grade Constructed Response Mathematics Items (Continued)

Country	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	
		Minimum	Maximum		Minimum	Maximum
Netherlands	98	83	100	96	76	100
New Zealand	99	88	100	98	81	100
Northern Ireland	99	90	100	98	88	100
Norway (5)	98	86	100	97	77	100
Oman	98	90	100	96	72	100
Poland	99	90	100	98	82	100
Portugal	100	99	100	100	97	100
Qatar	99	97	100	98	93	100
Russian Federation	99	97	100	99	97	100
Saudi Arabia	98	81	100	96	77	100
Serbia	97	79	100	94	66	100
Singapore	99	94	100	99	93	100
Slovak Republic	100	100	100	100	99	100
Slovenia	99	97	100	99	96	100
Spain	99	95	100	98	92	100
Sweden	98	86	100	97	81	100
Turkey	100	98	100	100	98	100
United Arab Emirates	98	85	100	96	80	100
United States	98	81	100	97	78	100
International Avg.	**99**	**90**	**100**	**98**	**85**	**100**
Benchmarking Participants						
Buenos Aires, Argentina	97	87	100	94	80	100
Ontario, Canada	96	69	100	94	69	100
Quebec, Canada	98	84	100	97	83	100
Norway (4)	98	86	100	97	74	100
Abu Dhabi, UAE	98	86	100	96	78	100
Dubai, UAE	97	85	100	96	79	100
Florida, US	98	83	100	97	80	100

TIMSS Numeracy 2015 Within-Country Scoring Reliability for the Constructed Response Items

Country	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	
		Minimum	Maximum		Minimum	Maximum
Bahrain	100	97	100	99	97	100
Indonesia	98	87	100	96	78	100
Iran, Islamic Rep. of	99	94	100	98	94	100
Jordan	99	98	100	98	93	100
Kuwait	99	95	100	98	95	100
Morocco	94	53	100	92	53	100
South Africa (5)	100	97	100	99	97	100
Benchmarking Participants						
Buenos Aires, Argentina	96	83	100	94	83	100

TIMSS 2015 Within-Country Scoring Reliability for the Fourth Grade Constructed Response Science Items

Country	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	
		Minimum	Maximum		Minimum	Maximum
Australia	95	85	100	94	85	100
Bahrain	92	80	99	90	71	98
Belgium (Flemish)	95	79	100	94	79	99
Bulgaria	97	80	100	97	78	100
Canada	95	82	100	94	82	99
Chile	95	86	100	94	79	100
Chinese Taipei	94	85	100	94	76	100
Croatia	97	91	100	96	82	100
Cyprus	99	97	100	99	97	100
Czech Republic	94	83	100	93	70	100
Denmark	93	81	99	91	73	98
England	97	78	100	97	78	100
Finland	96	87	100	95	87	100
France	94	59	100	93	56	100
Georgia	97	90	100	96	81	100
Germany	95	83	100	95	81	100
Hong Kong SAR	100	98	100	99	97	100
Hungary	98	91	100	97	88	100
Indonesia	95	66	100	93	64	100
Iran, Islamic Rep. of	98	92	100	96	86	100
Ireland	98	89	100	98	88	100
Italy	95	86	100	95	86	100
Japan	99	93	100	99	93	100
Kazakhstan	94	89	98	94	89	98
Korea, Rep. of	97	93	100	97	93	100
Kuwait	99	96	100	98	92	100
Lithuania	100	98	100	99	98	100
Morocco	91	65	100	88	61	99
Netherlands	92	78	99	91	69	99
New Zealand	96	82	100	95	81	100
Northern Ireland	95	86	100	94	86	99
Norway (5)	89	64	100	88	64	100

TIMSS 2015 Within-Country Scoring Reliability for the Fourth Grade Constructed Response Science Items (Continued)

Country	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	
		Minimum	Maximum		Minimum	Maximum
Oman	93	76	100	91	76	99
Poland	93	75	100	93	75	99
Portugal	99	97	100	99	96	100
Qatar	99	96	100	97	93	100
Russian Federation	98	94	100	98	94	100
Saudi Arabia	97	88	100	95	86	100
Serbia	90	72	99	88	70	98
Singapore	97	90	100	96	88	100
Slovak Republic	100	98	100	100	98	100
Slovenia	98	91	100	97	90	100
Spain	98	90	100	98	86	100
Sweden	93	81	100	93	81	100
Turkey	99	92	100	99	91	100
United Arab Emirates	92	80	99	90	78	98
United States	95	85	100	95	80	100
International Avg.	**96**	**85**	**100**	**95**	**82**	**100**
Benchmarking Participants						
Buenos Aires, Argentina	93	76	100	90	75	98
Ontario, Canada	94	82	100	93	79	100
Quebec, Canada	96	81	100	95	81	100
Norway (4)	91	71	100	90	71	100
Abu Dhabi, UAE	93	80	99	91	77	98
Dubai, UAE	90	75	100	89	73	98
Florida, US	95	85	100	95	77	100

TIMSS 2015 Within-Country Scoring Reliability for the Eighth Grade Constructed Response Mathematics Items

Country	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	
		Minimum	Maximum		Minimum	Maximum
Australia	98	87	100	97	86	100
Bahrain	99	97	100	99	96	100
Botswana (9)	98	74	100	96	60	100
Canada	97	87	100	95	81	100
Chile	98	85	100	96	77	100
Chinese Taipei	98	87	100	97	63	100
Egypt	99	95	100	97	88	100
England	99	95	100	99	95	100
Georgia	99	93	100	98	88	100
Hong Kong SAR	100	98	100	100	98	100
Hungary	99	93	100	98	89	100
Iran, Islamic Rep. of	99	93	100	97	89	100
Ireland	98	87	100	98	84	100
Israel	98	92	100	96	87	100
Italy	98	86	100	97	85	100
Japan	100	93	100	100	93	100
Jordan	99	97	100	98	90	100
Kazakhstan	89	71	98	88	70	96
Korea, Rep. of	99	89	100	98	88	100
Kuwait	99	95	100	98	93	100
Lebanon	96	75	100	93	74	99
Lithuania	100	99	100	100	98	100
Malaysia	99	95	100	98	93	100
Malta	98	90	100	97	79	100
Morocco	97	45	100	93	44	100
New Zealand	98	91	100	97	86	100
Norway (9)	97	79	100	95	70	100
Oman	98	85	100	96	77	100
Qatar	99	96	100	98	92	100
Russian Federation	99	95	100	99	91	100
Saudi Arabia	100	97	100	99	90	100
Singapore	98	86	100	97	84	100

TIMSS 2015 Within-Country Scoring Reliability for the Eighth Grade Constructed Response Mathematics Items (Continued)

Country	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	
		Minimum	Maximum		Minimum	Maximum
Slovenia	99	97	100	99	96	100
South Africa (9)	100	94	100	99	89	100
Sweden	98	81	100	96	79	100
Thailand	100	99	100	99	85	100
Turkey	99	98	100	99	96	100
United Arab Emirates	98	87	100	96	75	100
United States	98	81	100	97	75	100
International Avg.	**98**	**89**	**100**	**97**	**84**	**100**
Benchmarking Participants						
Buenos Aires, Argentina	99	96	100	98	93	100
Ontario, Canada	97	85	100	95	77	100
Quebec, Canada	97	79	100	96	76	100
Norway (8)	97	83	100	96	77	100
Abu Dhabi, UAE	98	86	100	96	81	100
Dubai, UAE	97	81	100	95	65	100
Florida, US	99	87	100	97	83	100

TIMSS 2015 Within-Country Scoring Reliability for the Eighth Grade Constructed Response Science Items

Country	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	
		Minimum	Maximum		Minimum	Maximum
Australia	95	82	100	94	80	100
Bahrain	92	72	100	88	60	100
Botswana (9)	91	72	100	88	62	100
Canada	94	73	100	92	69	99
Chile	94	85	100	92	78	99
Chinese Taipei	95	85	100	94	77	100
Egypt	99	93	100	97	87	100
England	98	93	100	98	93	100
Georgia	98	86	100	97	86	100
Hong Kong SAR	100	98	100	100	98	100
Hungary	97	93	100	97	91	100
Iran, Islamic Rep. of	98	91	100	97	86	100
Ireland	96	77	100	95	77	100
Israel	98	92	100	96	85	100
Italy	95	87	100	94	86	100
Japan	99	84	100	99	84	100
Jordan	98	94	100	96	81	100
Kazakhstan	89	73	97	89	73	97
Korea, Rep. of	96	87	100	95	87	100
Kuwait	99	94	100	97	91	100
Lebanon	93	78	100	88	57	99
Lithuania	100	98	100	99	96	100
Malaysia	98	94	100	97	91	100
Malta	92	71	100	89	71	100
Morocco	91	73	100	84	52	100
New Zealand	96	84	100	95	84	100
Norway (9)	92	63	100	91	63	100
Oman	93	78	100	91	72	100
Qatar	99	97	100	98	89	100
Russian Federation	98	92	100	97	79	100
Saudi Arabia	98	84	100	96	81	100
Singapore	96	82	100	95	81	100

TIMSS 2015 Within-Country Scoring Reliability for the Eighth Grade Constructed Response Science Items (Continued)

Country	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	
		Minimum	Maximum		Minimum	Maximum
Slovenia	99	97	100	99	93	100
South Africa (9)	98	87	100	97	79	100
Sweden	94	74	100	92	74	100
Thailand	100	99	100	99	92	100
Turkey	97	89	100	96	80	100
United Arab Emirates	92	74	99	90	71	99
United States	94	75	100	93	75	100
International Avg.	**96**	**85**	**100**	**94**	**80**	**100**
Benchmarking Participants						
Buenos Aires, Argentina	99	96	100	98	92	100
Ontario, Canada	94	82	100	93	71	100
Quebec, Canada	94	71	100	92	69	100
Norway (8)	93	73	100	91	72	100
Abu Dhabi, UAE	93	77	100	91	70	99
Dubai, UAE	91	70	100	88	66	100
Florida, US	95	79	100	94	71	100

Appendix 11B: Trend Scoring Reliability for the Constructed Response Items

TIMSS 2015 Trend Scoring Reliability for the Fourth Grade Constructed Response Mathematics Items

Country	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	
		Minimum	Maximum		Minimum	Maximum
Australia	99	95	100	97	88	100
Bahrain	98	82	100	94	68	100
Belgium (Flemish)	97	89	100	95	62	100
Canada	98	85	100	95	78	100
Chile	97	71	100	94	69	99
Chinese Taipei	97	77	100	96	69	100
Croatia	98	89	100	97	85	100
Czech Republic	98	82	100	95	78	99
Denmark	97	80	100	95	77	100
England	98	85	100	96	52	100
Finland	98	84	100	97	83	100
Georgia	98	87	100	96	82	100
Germany	99	93	100	98	92	100
Hungary	98	76	100	97	76	100
Iran, Islamic Rep. of	98	90	100	96	85	99
Ireland	98	80	100	96	76	100
Italy	97	82	100	95	82	100
Japan	98	87	100	97	75	100
Kazakhstan	94	72	99	91	66	99
Korea, Rep. of	99	89	100	99	85	100
Kuwait	95	72	100	89	60	99
Lithuania	98	88	100	97	81	100
Netherlands	97	79	99	95	79	99
New Zealand	97	78	100	95	77	100
Northern Ireland	98	80	100	97	79	100
Norway	97	70	100	95	69	100
Oman	97	82	100	93	74	99
Poland	98	89	100	96	86	99

TIMSS 2015 Trend Scoring Reliability for the Fourth Grade Constructed Response Mathematics Items (Continued)

Country	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	
		Minimum	Maximum		Minimum	Maximum
Portugal	98	91	100	98	90	100
Qatar	98	86	100	95	74	100
Russian Federation	97	81	100	95	48	100
Serbia	98	87	100	96	70	100
Singapore	98	87	100	98	82	100
Slovak Republic	98	96	100	97	91	100
Slovenia	96	84	99	93	73	99
Spain	96	72	100	93	67	100
Sweden	98	80	100	96	79	100
Turkey	97	78	100	95	75	100
United Arab Emirates	97	86	100	94	59	99
United States	97	84	100	96	83	100
International Avg.	**98**	**83**	**100**	**95**	**76**	**100**
Benchmarking Participant						
Dubai, UAE	98	83	100	94	63	100

TIMSS 2015 Trend Scoring Reliability for the Fourth Grade Constructed Response Science Items

Country	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	
		Minimum	Maximum		Minimum	Maximum
Australia	94	88	100	93	85	99
Bahrain	88	73	96	85	70	96
Belgium (Flemish)	89	61	98	87	61	98
Canada	91	81	99	89	77	99
Chile	91	75	99	89	72	99
Chinese Taipei	87	62	99	84	55	99
Croatia	92	74	100	91	74	99
Czech Republic	92	71	100	90	69	100
Denmark	87	69	98	85	67	97
England	92	74	100	91	74	100
Finland	94	84	100	93	81	100
Georgia	92	78	99	89	71	99
Germany	95	89	99	94	88	99
Hungary	94	84	100	93	83	99
Iran, Islamic Rep. of	92	75	99	90	72	99
Ireland	91	67	99	89	66	99
Italy	95	85	100	93	85	99
Japan	89	55	100	88	53	100
Kazakhstan	83	60	95	76	50	95
Korea, Rep. of	94	80	100	94	78	100
Kuwait	93	85	99	88	76	96
Lithuania	93	56	100	91	56	99
Netherlands	89	65	99	88	65	99
New Zealand	94	82	100	93	80	99
Northern Ireland	94	78	100	93	78	100
Norway	91	69	99	90	69	99
Oman	94	84	99	89	77	99
Poland	90	65	99	87	65	98
Portugal	95	83	99	93	83	99
Qatar	92	83	99	90	80	98

TIMSS 2015 Trend Scoring Reliability for the Fourth Grade Constructed Response Science Items (Continued)

Country	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	
		Minimum	Maximum		Minimum	Maximum
Russian Federation	95	82	100	94	81	100
Serbia	89	68	98	86	67	98
Singapore	93	85	100	93	83	100
Slovak Republic	97	90	99	96	89	99
Slovenia	89	62	99	86	62	97
Spain	88	70	99	85	68	99
Sweden	92	75	99	91	75	99
Turkey	92	68	98	90	68	98
United Arab Emirates	92	76	98	88	72	97
United States	92	77	100	92	76	100
International Avg.	**92**	**75**	**99**	**90**	**73**	**99**
Benchmarking Participant						
Dubai, UAE	91	78	100	89	77	100

TIMSS 2015 Trend Scoring Reliability for the Eighth Grade Constructed Response Mathematics Items

Country	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	
		Minimum	Maximum		Minimum	Maximum
Australia	98	91	100	96	86	100
Bahrain	97	75	100	93	60	100
Botswana	96	77	100	92	63	100
Canada	95	83	100	92	77	98
Chinese Taipei	95	67	100	94	67	100
England	97	80	100	94	76	100
Georgia	97	86	100	93	66	100
Hong Kong SAR	97	82	100	95	76	100
Hungary	97	79	100	96	78	100
Iran, Islamic Rep. of	97	76	100	94	67	100
Israel	97	83	100	95	81	100
Italy	98	86	100	96	83	100
Japan	97	84	100	95	77	100
Jordan	97	84	100	93	61	100
Kazakhstan	91	74	100	87	60	100
Korea, Rep. of	98	91	100	97	88	100
Lithuania	98	85	100	97	81	100
Malaysia	97	86	100	92	70	99
New Zealand	96	75	100	94	68	100
Norway	97	76	100	94	68	100
Oman	97	84	100	91	60	98
Qatar	97	88	100	95	83	99
Russian Federation	97	81	100	94	76	100
Singapore	97	79	100	96	73	100
Slovenia	96	71	100	93	71	100
South Africa	96	87	99	91	71	99
Sweden	97	79	100	95	74	100
Thailand	98	91	100	96	82	100
Turkey	96	83	100	92	72	100
United Arab Emirates	97	85	100	94	81	100
United States	96	69	100	94	62	100
International Avg.	**97**	**81**	**100**	**94**	**73**	**100**
Benchmarking Participant						
Dubai, UAE	97	85	100	95	78	100

TIMSS 2015 Trend Scoring Reliability for the Eighth Grade Constructed Response Science Items

Country	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	
		Minimum	Maximum		Minimum	Maximum
Australia	94	85	100	92	85	100
Bahrain	90	71	99	86	60	98
Botswana	90	67	98	84	61	96
Canada	90	66	100	85	64	100
Chinese Taipei	93	75	100	90	75	100
England	92	57	100	89	57	100
Georgia	94	76	99	89	58	99
Hong Kong SAR	93	67	100	91	66	100
Hungary	95	70	100	93	70	100
Iran, Islamic Rep. of	92	68	99	88	67	98
Israel	93	73	100	90	72	100
Italy	94	86	99	91	77	99
Japan	93	73	100	89	54	99
Jordan	94	76	99	88	70	99
Kazakhstan	85	52	100	77	52	99
Korea, Rep. of	95	79	100	93	76	100
Lithuania	96	81	100	95	77	100
Malaysia	92	66	99	82	59	97
New Zealand	94	73	100	91	73	100
Norway	93	64	100	91	64	100
Oman	93	65	99	87	61	98
Qatar	92	79	99	87	76	98
Russian Federation	94	68	100	92	68	100
Singapore	93	66	100	91	66	100
Slovenia	90	57	99	88	56	98
South Africa	95	82	100	88	47	100
Sweden	94	72	100	91	69	100
Thailand	96	81	100	93	77	100
Turkey	94	77	100	91	77	99
United Arab Emirates	93	66	100	90	64	99
United States	94	61	100	90	60	99
International Avg.	**93**	**71**	**100**	**89**	**66**	**99**
Benchmarking Participant						
Dubai, UAE	93	72	100	90	72	100

Appendix 11C: TIMSS 2015 Cross-Country Scoring Reliability for the Constructed Response Items

TIMSS 2015 Cross-Country Scoring Reliability for the Fourth Grade Constructed Response Mathematics Items

Item Label	Total Valid Comparisons	Exact Percent Agreement	
		Correctness Score Agreement	Diagnostic Score Agreement
M09_01 - M051206	207000	97	97
M09_04 - M051045	207000	99	98
M09_06 - M051030	206820	98	98
M09_11 - M051533	206910	99	99
M09_12 - M051080	206865	91	88
M11_01 - M051401	206865	99	99
M11_03 - M051402	207000	99	99
M11_05 - M051131	206955	98	98
M11_07 - M051217	206955	97	96
M11_08 - M051079	207000	97	97
M11_11 - M051009	207000	98	98
Average Percent Agreement		**97**	**97**

TIMSS 2015 Cross-Country Scoring Reliability for the Fourth Grade Constructed Response Science Items

Item Label	Total Valid Comparisons	Exact Percent Agreement	
		Correctness Score Agreement	Diagnostic Score Agreement
S09_01 - S051044	198000	88	88
S09_04 - S051168	198000	81	77
S09_05 - S051010	198000	86	83
S09_07 - S051059	198000	71	71
S09_10 - S051151	198000	98	98
S11_04 - S051194	198000	89	89
S11_06 - S051077	198000	95	95
S11_07 - S051200	198000	86	86
S11_08 - S051075	198000	84	84
S11_12 - S051175	198000	77	77
Average Percent Agreement		**86**	**85**

TIMSS 2015 Cross-Country Scoring Reliability for the Eighth Grade Constructed Response Mathematics Items

Item Label	Total Valid Comparisons	Exact Percent Agreement	
		Correctness Score Agreement	Diagnostic Score Agreement
M09_05A - M052174A	133128	97	97
M09_05B - M052174B	132675	96	94
M09_08 - M052110	132948	98	98
M09_09 - M052105	133164	88	88
M09_11 - M052036	133200	87	87
M09_12 - M052502	133056	95	95
M09_13 - M052117	133092	86	71
M11_03 - M052364	133200	98	98
M11_04 - M052215	133020	98	98
M11_08 - M052087	131767	94	94
M11_09 - M052048	133056	95	83
M11_10 - M052039	133164	98	98
M11_14 - M052421	132984	80	80
Average Percent Agreement		**93**	**91**

TIMSS 2015 Cross-Country Scoring Reliability for the Eighth Grade Constructed Response Science Items

Item Label	Total Valid Comparisons	Exact Percent Agreement	
		Correctness Score Agreement	Diagnostic Score Agreement
S09_02 - S052272	133200	90	82
S09_03A - S052085A	133128	78	68
S09_03B - S052085B	133200	78	78
S09_04 - S052094	133200	95	95
S09_06 - S052146	133200	91	88
S09_10 - S052214	133200	98	98
S09_12 - S052101	132948	82	82
S11_01B - S052090B	133200	80	66
S11_04 - S052273	133056	52	52
S11_06 - S052051	133092	83	83
S11_10 - S052189	133128	79	74
S11_13 - S052099	133164	80	80
S11_14 - S052118	133164	90	84
Average Percent Agreement		**83**	**79**

Appendix 11D: Country Adaptations to Items and Item Scoring

TIMSS Fourth Grade Mathematics

Deleted Items

BELGIUM (FLEMISH)

M041200, M05_13 (printing error)

BULGARIA

M051125B, M06_11B (translation error)

FRANCE

M061239, M02_10 (printing error)

IRAN, ISLAMIC REP. OF

M061041, M14_04 (transcription error)

LITHUANIA

M041034, M01_03 (Russian only; translation error)

M051236, M06_10 (Polish only; translation error)

TURKEY

M051502, M09_07 (printing error)

Constructed Response Items with Category Recodes

ALL COUNTRIES

M061239, M02_10 (recode 20 to 10, 10 to 70, 11 to 71)

M061084, M08_11 (recode 20 to 10, 10 to 70)

M051080, M09_12 (recode 20 to 10, 10 to 71, 11 to 72)

M061254, M14_02 (recode 20 to 10, 10 to 70)

M061224, M14_08 (recode 70 to 12)

TIMSS Fourth Grade Mathematics – Numeracy

Constructed Response Items with Category Recodes

ALL COUNTRIES

M061239, N04_10 (recode 20 to 10, 10 to 70, 11 to 71)

M061084, N08_11 (recode 20 to 10, 10 to 70)

TIMSS Fourth Grade Science

Deleted Items

ALL COUNTRIES

S041193, S01_09 (poor discrimination)

S041002, S05_07 (faulty distracters)

S051079, S06_09 (attractive distracter)

TIMSS Fourth Grade Science

Deleted Items (Continued)

S041080, S07_08 (attractive distracter)

S041171, S07_10 (faulty distracters)

S051020, S09_02 (poor discrimination)

S061166, S10_05 (poor discrimination)

S051138C, S11_03C (poor discrimination)

S061125, S14_01 (poor discrimination)

FRANCE

S051106, S13_10 (printing error)

INDONESIA

S051191, S11_10 (negative discrimination)

LITHUANIA

S041052, S07_06 (Polish only; translation error)

NORWAY

S061081, S02_06 (translation error)

TIMSS Eighth Grade Mathematics

Deleted Items

ALL COUNTRIES

M062345B, M04_12B (poor discrimination)

M062345BA, M04_12BA (poor discrimination)

M062345BB, M04_12BB (poor discrimination)

M062345BC, M04_12BC (poor discrimination)

M062345BD, M04_12BD (poor discrimination)

M062342, M10_07 (poor discrimination)

M062048, M14_12 (poor discrimination)

M062048A, M14_12A (poor discrimination)

M062048B, M14_12B (poor discrimination)

M062048C, M14_12C (poor discrimination)

KAZAKHSTAN

M062106, M02_12 (negative discrimination)

KUWAIT

M062271, M12_01 (translation error)

LITHUANIA

M052125, M13_03 (Russian only; translation error)

MOROCCO

M052090, M06_07 (negative discrimination)

TIMSS Eighth Grade Mathematics

Deleted Items (Continued)

SWEDEN

M062237, M02_04 (transcription error)

M052090, M06_07 (transcription error)

Constructed Response Items with Category Recodes

ALL COUNTRIES

M042302C, M01_06C (recode 11 to 71)

M042229B, M05_10B (recode 11 to 71)

M052095, M06_04 (recode 20 to 10 and 10 to 70)

M062254, M08_13 (recode 20 to 10)

M052087, M11_08 (recode 20 to 10 and 10 to 70)

TIMSS Eighth Grade Science

Deleted Items

ALL COUNTRIES

S042401, S01_13 (faulty distracters)

S062189C, S02_01C (poor discrimination)

S062272, S08_12 (poor discrimination)

S052221, S09_11 (poor discrimination)

S062036, S12_12 (attractive distracter)

S062242C, S12_15C (poor discrimination)

S062266, S14_05 (attractive distracter)

BOTSWANA

S062032, S10_05 (negative discrimination)

S052134, S13_06 (negative discrimination)

JORDAN

S052134, S13_06 (negative discrimination)

KAZAKHSTAN

S062090, S10_01 (negative discrimination)

KUWAIT

S052134, S13_06 (negative discrimination)

LITHUANIA

S062190, S04_13 (Polish only; translation error)

MOROCCO

S052134, S13_06 (negative discrimination)

SAUDI ARABIA

S062225, S08_04 (not administered)

TIMSS Eighth Grade Science
Deleted Items (Continued)
SOUTH AFRICA
S062032, S10_05 (negative discrimination)
THAILAND
S052141, S06_12 (translation error)

Appendix 11E: Derived Items in TIMSS 2015

TIMSS Fourth Grade Mathematics

M051061Z, M06_08 – Item parts A, B, and C are combined to create a 1-point item, where 1 score point is awarded if all parts are correct

M061018, M10_01 – Item parts A, B, C, and D are combined to create a 1-point item, where 1 score point is awarded if all parts are correct

M061240, M14_01 – Item parts A, B, and C are combined to create a 1-point item, where 1 score point is awarded if all parts are correct

TIMSS Numeracy

MN11042, N01_10 – Item parts B, C, D, and E are combined to create a 1-point item, where 1 score point is awarded if all parts are correct (part A is an example)

TIMSS Fourth Grade Science

S041149Z, S01_10 – Item parts A and B are combined to create a 2-point item, where 2 score points are awarded if both parts A and B are correct, 1 score point is awarded if either part A or part B is correct, and 0 score points are awarded if both parts A and B are incorrect

S051026Z, S03_05 – Item parts A, B, C, and D are combined to create a 1-point item, where 1 score point is awarded if all parts are correct

S051121Z, S03_08 – Item parts A, B, C, D, and E are combined to create a 1-point item, where 1 score point is awarded if all parts are correct

S051188Z, S06_08 – Item parts A, B, C, D, and E are combined to create a 1-point item, where 1 score point is awarded if all parts are correct

S061083, S10_06 – Item parts B, C, and D are combined to create a 1-point item, where 1 score point is awarded if all parts are correct (part A is an example)

S061142A, S10_09 – Item parts A, B, and C are combined to create a 1-point item, where 1 score point is awarded if all parts are correct

S051138Z, S11_03 – Item parts A and B are combined to create a 1-point item, where 1 score point is awarded if both parts are correct (part C was deleted)

S061124, S14_11 – Item parts B, C, D, and E are combined to create a 1-point item, where 1 score point is awarded if all parts are correct (part A is an example)

S061116, S14_12 – Item parts B, C, D, and E are combined to create a 1-point item, where 1 score point is awarded if all parts are correct (part A is an example)

TIMSS Eighth Grade Mathematics

M062208, M02_01 – Item parts A, B, C, and D are combined to create a 1-point item, where 1 score point is awarded if all parts are correct
M042229Z, M05_10 – Item parts A and B are combined to create a 2-point item, where 2 score points are awarded if both parts A and B are correct, 1 score point is awarded if only part A or only part B is correct, and 0 score points are awarded if both parts A and B are incorrect

TIMSS Eighth Grade Science

S062189, S02_01 – Item parts A, B, D, and E are combined to create a 2-point item, where 2 score points are awarded if all parts are correct, 1 score point is awarded if 3 parts are correct, and 0 score points are awarded if 2 or fewer parts are correct (part C was deleted)
S062010, S02_05 – Item parts A, B, C, and D are combined to create a 1-point item, where 1 score point is awarded if all parts are correct
S052092Z, S03_02 – Item parts A, B, C, and D were combined to create a 2-point item, where 2 score points are awarded if all parts are correct, 1 score point is awarded if 2 or 3 parts are correct, and 0 score points are awarded if 1 or 0 parts are correct
S052043Z, S03_07 – Item parts A, B, C, and D are combined to create a 1-point item, where 1 score point is awarded if all parts are correct
S062018, S04_08 – Item parts A, B, C, D, and E are combined to create a 2-point item, where 2 points are awarded if all parts are correct, 1 point is awarded if 4 parts are correct, and 0 score points are awarded if 3 or fewer parts are correct
S062173A, S10_13 – Item parts A, B, C, and D are combined to create a 1-point item, where 1 score point is awarded if all parts are correct
S052015Z, S11_05 – Item parts A, B, C, D, E, and F are combined to create a 1-point item, where 1 score point is awarded if all parts are correct
S062242, S12_15 – Item parts A, B, D, and E are combined to create a 1-point item, where 1 score point is awarded if all parts are correct (part C was deleted)
S052095Z, S13_05 – Item parts B, C, and D are combined to create a 1-point item, where 1 score point is awarded if all parts are correct (part A is an example)
S062047, S14_07 – Item parts A, B, and C are combined to create a 1-point item, where 1 score point is awarded if all parts are correct
S062022, S14_14 – Item parts A, B, C, and D are combined to create a 1-point item, where 1 score point is awarded if all parts are correct