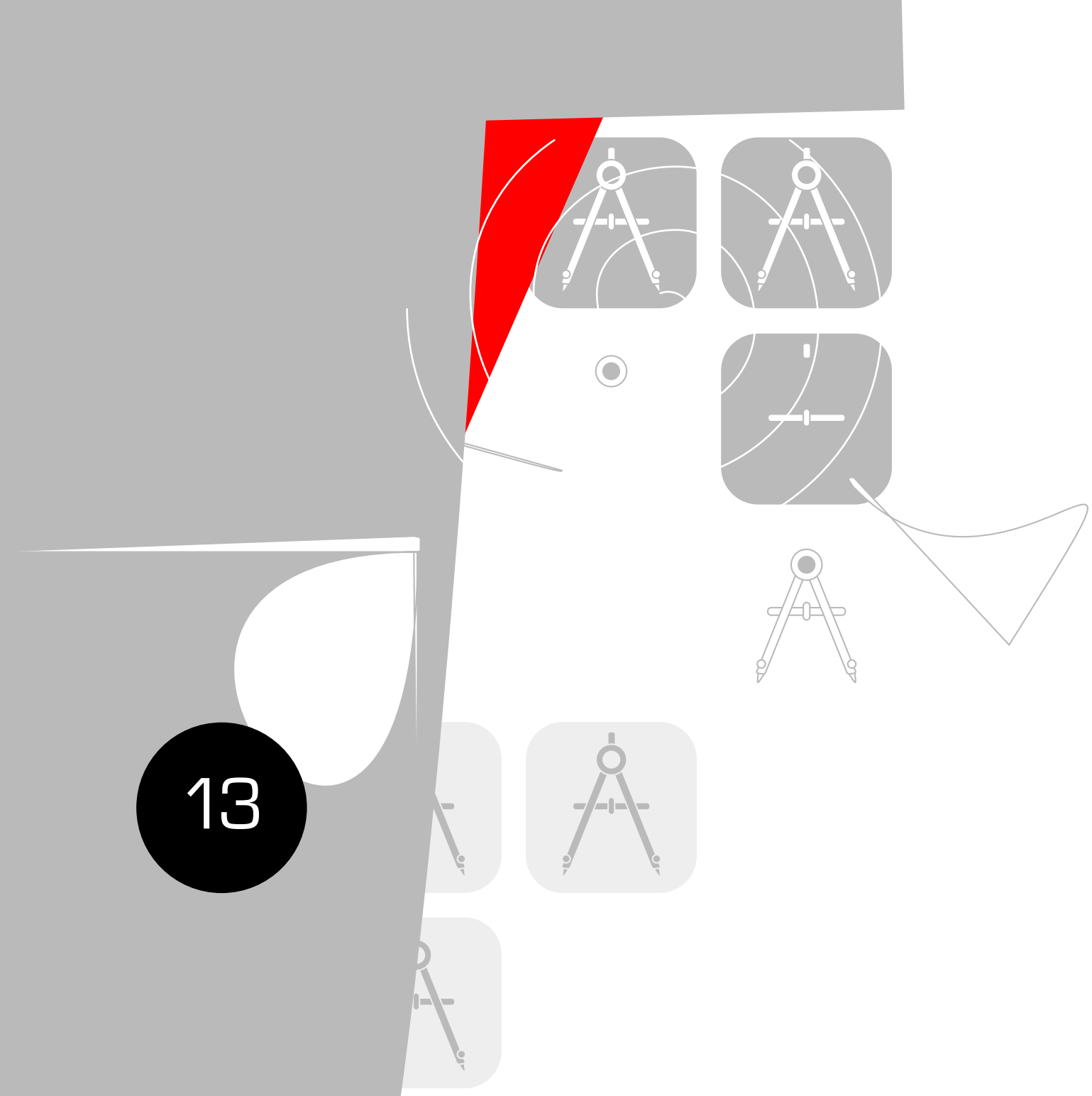
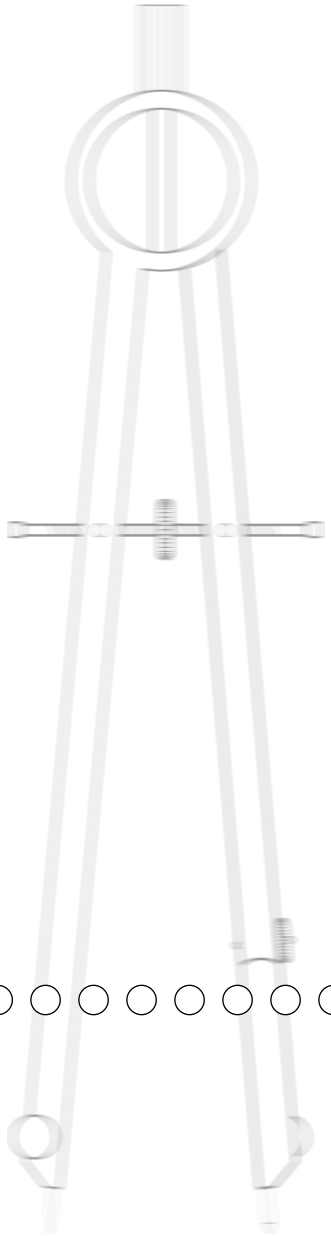


13

Item Analysis and Review

Ina V.S. Mullis
Michael O. Martin



13 Item Analysis and Review

Ina V.S. Mullis
Michael O. Martin

13.1 Overview

In order to assess the psychometric properties of the TIMSS 1999 achievement items before proceeding with item response theory (IRT) scaling,[1] TIMSS computed a series of diagnostic statistics for each item in each country. As part of the TIMSS quality assurance process, these statistics were carefully checked for any evidence of unusual item behavior. If an item was uncharacteristically easy or difficult for a particular country, or had unusually low discriminating power, this sometimes suggested a translation or printing problem. On the relatively few occasions that such items were found, the test booklets were examined for flaws, and where necessary the national research coordinator was consulted. Any item that was discovered to have a flaw in a particular country was removed from the database for that country.

13.2 Statistics for Item Analysis

The basic statistics for the item review were calculated at the IEA Data Processing Center and summarized in graphical form for review at the International Study Center. Item statistics were computed for each of the 38 TIMSS 1999 countries, and where countries tested in more than one languages, for each of the languages tested. For each item, the basic item-analysis display presents the number of students that responded in each country, the difficulty level (the percentage of students that answered the item correctly), and the discrimination index (the point-biserial correlation between success on the item and a total score).[2] For multiple-choice items (see Exhibit 13.1 for an example), the display presents the percentage of students that chose each option, including the percentage that omitted or did not reach the item, and the point-biserial correlation between each option and the total score. For free-response items (which could have more than one score level – see Exhibit 13.2 for an example), the display

○○○

1. The TIMSS 1999 IRT scaling is described in Chapter 14.
2. For the purpose of computing the discrimination index, the total score was the percentage of items a student answered correctly.

presents the difficulty and discrimination of each score level. As a prelude to the main IRT scaling, it shows some statistics from a preliminary Rasch analysis, including the Rasch item difficulty for each item and the standard error of this difficulty estimate.

The item-analysis display presents the difficulty level of each item separately for male and female students. As a guide to the overall statistical properties of the item, it also shows the international item difficulty (the mean of the item difficulties across countries) and the international item discrimination (the mean of the item discriminations).

Exhibit 13.1 International Item Statistics for a Multiple-Choice Item

Third International Mathematics and Science Study - 1999 Main Survey — May 23, 2000 47
International Item Statistics (Unweighted) - Review Version
For Internal Review Only DO NOT CITE OR CIRCULATE

Mathematics : Data Representation, Analysis & Probability (H11) Type : M Key: C Label: Defective bulbs from random sample (M012047 - BSMMH11)

Country	N	Diff	Disc	Pct_A	Pct_B	Pct_C	Pct_D	Pct_E	Pct_In	Pct_OM	Pct_NR	PB_A	PB_B	PB_C	PB_D	PB_E	PB_In	PB_OM	RDFF	Flags
Australia	1506	65.90	0.47	4.80	9.20	65.90	9.80	9.50	0.10	0.60	0.20	-0.21	-0.20	0.47	-0.24	-0.15	-0.02	-0.05	-0.22	__H_F_
Belgium (Flemish)	1985	84.50	0.42	2.60	4.00	84.50	4.00	4.70	0.00	0.20	0.10	-0.17	-0.16	0.42	-0.26	-0.18	0	-0.06	-0.68	____F_
Bulgaria	1214	64.90	0.48	8.40	6.50	64.90	7.30	10.60	0.00	2.20	0.70	-0.22	-0.17	0.48	-0.22	-0.15	0	-0.13	-0.45	____F_
Canada (English)	2352	61.50	0.45	4.60	9.00	61.50	12.70	11.60	0.00	0.60	0.10	-0.18	-0.23	0.45	-0.22	-0.11	0	-0.07	-0.15	__H_F_
Canada (French)	917	69.80	0.50	5.10	7.30	69.80	9.50	7.60	0.00	0.70	0.10	-0.16	-0.22	0.50	-0.28	-0.18	0	-0.06	-0.34	____F_
Chile	2154	46.90	0.41	11.70	10.50	46.90	10.50	15.60	0.00	4.80	2.70	-0.17	-0.14	0.41	-0.13	-0.06	0	-0.15	-0.8	__E___
Chile (7th Grade)	2187	36.70	0.35	11.50	14.00	36.70	14.40	16.60	0.00	6.80	3.70	-0.10	-0.12	0.35	-0.06	-0.07	0	-0.12	-0.66	_________
Chinese Taipei	2167	83.60	0.55	2.30	3.50	83.60	4.50	6.10	0.00	0.10	0.00	-0.23	-0.23	0.55	-0.29	-0.29	0	0.01	-0.71	____F_
Cyprus	1157	64.20	0.46	8.10	7.90	64.20	8.40	9.90	0.00	1.50	0.10	-0.19	-0.16	0.46	-0.21	-0.16	0	-0.12	-0.79	____F_
Czech Republic	1284	82.70	0.52	2.30	4.70	82.70	4.40	5.20	0.00	0.70	0.20	-0.19	-0.29	0.52	-0.26	-0.19	0	-0.12	-0.93	__E___
England	1090	59.40	0.49	6.40	8.50	59.40	15.00	10.40	0.00	0.30	0.00	-0.19	-0.20	0.49	-0.24	-0.17	0	-0.06	-0.31	____F_
Finland (Fin.)	1032	67.90	0.46	4.70	7.80	67.90	10.70	8.20	0.00	0.80	0.00	-0.14	-0.19	0.46	-0.26	-0.16	0	-0.1	-0.44	____F_
Finland (Swe.)	136	72.10	0.38	2.90	5.90	72.10	7.40	11.00	0.00	0.70	0.00	-0.11	-0.08	0.38	-0.18	-0.22	0	-0.17	-0.65	____F_
Hong Kong, SAR	1932	85.20	0.41	1.40	3.70	85.20	3.80	5.60	0.00	0.30	0.10	-0.12	-0.19	0.41	-0.23	-0.21	0	-0.08	-0.66	____F_
Hungary	1184	76.80	0.50	3.90	5.30	76.80	4.10	8.40	0.00	1.40	0.40	-0.19	-0.18	0.50	-0.22	-0.22	0	-0.19	-0.64	____F_
Indonesia	2173	49.50	0.38	13.30	10.80	49.50	10.50	14.00	0.00	1.80	0.50	-0.19	-0.11	0.38	-0.18	-0.08	0	-0.03	-0.68	_________
Iran, Islamic Rep.	1984	46.30	0.44	19.80	10.00	46.30	10.80	10.70	0.00	2.30	0.20	-0.18	-0.14	0.44	-0.17	-0.11	0	-0.13	-0.49	_________
Israel (Arabic)	290	43.80	0.36	14.50	12.80	43.80	10.30	14.10	0.00	4.50	0.70	-0.13	-0.11	0.36	-0.08	-0.11	0	-0.13	-0.67	_________
Israel (Hebrew)	1270	55.90	0.46	9.90	10.00	55.90	9.60	10.90	0.10	3.60	1.10	-0.21	-0.16	0.46	-0.20	-0.07	-0.03	-0.18	-0.23	__H_F_
Italy	1931	85.20	0.41	1.40	3.70	85.20	3.80	5.60	0.00	0.30	0.10	-0.12	-0.19	0.41	-0.23	-0.21	0	-0.08	-0.66	____F_
Japan	1184	76.80	0.50	3.90	5.30	76.80	4.10	8.40	0.00	1.40	0.40	-0.19	-0.18	0.50	-0.22	-0.22	0	-0.19	-0.64	____F_
Jordan	2173	49.50	0.38	13.30	10.80	49.50	10.50	14.00	0.00	1.80	0.50	-0.19	-0.11	0.38	-0.18	-0.08	0	-0.03	-0.68	_________
Korea, Rep. of	2292	83.00	0.43	4.10	4.50	83.00	4.00	4.40	0.00	0.00	0.00	-0.15	-0.18	0.43	-0.24	-0.23	0	-0.02	-0.48	____F_
Latvia (LSS)	1079	70.60	0.44	5.20	7.60	70.60	6.70	9.00	0.10	0.80	0.20	-0.19	-0.16	0.44	-0.22	-0.17	-0.02	-0.1	-0.72	____F_
Lithuania	894	57.60	0.52	8.60	8.60	57.60	6.30	15.20	0.00	3.70	0.20	-0.25	-0.17	0.52	-0.21	-0.16	0	-0.15	-0.23	____F_
Macedonia (Alb.)	377	37.70	0.26	18.80	12.50	37.70	9.50	11.70	0.00	9.80	1.80	-0.05	-0.09	0.26	-0.02	-0.02	0	-0.16	-0.49	____F_
Macedonia (Mac.)	1119	59.10	0.43	8.80	8.60	59.10	6.10	11.00	0.00	6.40	0.80	-0.17	-0.19	0.43	-0.13	-0.11	0	-0.17	-0.66	____F_
Malaysia	2081	70.50	0.44	4.60	5.90	70.50	6.40	11.60	0.00	0.90	0.10	-0.14	-0.17	0.44	-0.23	-0.21	0	-0.09	-0.48	____F_
Moldova (Rom.)	1147	52.50	0.38	8.40	10.70	52.50	9.20	16.50	0.20	2.50	1.10	-0.13	-0.09	0.38	-0.18	-0.13	-0.06	-0.09	-0.31	____F_
Moldova (Rus.)	232	63.80	0.38	8.20	10.30	63.80	6.00	10.30	0.00	1.30	1.70	-0.14	-0.18	0.38	-0.15	-0.16	0	-0.07	-0.8	____F_
Morocco	1981	32.00	0.26	17.10	20.00	32.00	11.70	12.40	1.80	5.00	3.00	-0.10	-0.05	0.26	-0.10	-0.03	-0.04	-0.05	-0.57	_________
Netherlands	1107	86.70	0.39	1.40	3.30	86.70	3.80	4.60	0.00	0.20	0.00	-0.14	-0.18	0.39	-0.24	-0.17	0	-0.04	-1.44	__E_F_
New Zealand	1355	57.10	0.47	5.60	11.60	57.10	15.20	10.00	0.10	0.40	0.10	-0.17	-0.19	0.47	-0.24	-0.14	-0.01	-0.02	-0.27	____F_
Philippines (Eng.)	2216	37.30	0.25	8.80	10.20	37.30	22.70	19.90	0.10	1.00	0.50	0.01	-0.06	0.25	-0.18	-0.06	-0.01	-0.03	-0.64	____F_
Philippines (Tag.)	223	27.80	0.14	20.20	11.20	27.80	20.20	17.90	0.00	2.70	0.90	-0.03	-0.11	0.14	-0.08	0.06	0	-0.07	-0.44	D_____
Romania (Hun.)	47	51.10	0.56	6.40	8.50	51.10	14.90	12.80	0.00	6.40	2.10	-0.09	-0.26	0.55	-0.39	-0.18	0	0.18	-0.27	__AF__
Romania (Rom.)	1224	59.20	0.49	7.30	11.30	59.20	7.40	12.70	0.30	1.70	0.20	-0.22	-0.15	0.49	-0.19	-0.19	-0.06	-0.13	-0.35	____F_
Russian Federation	1619	67.10	0.50	5.00	7.10	67.10	6.20	12.40	0.20	2.00	1.00	-0.16	-0.16	0.50	-0.22	-0.26	0	-0.09	-0.26	__H_F_
Singapore	1849	82.90	0.46	2.40	4.10	82.90	3.80	6.30	0.00	0.50	0.00	-0.15	-0.18	0.46	-0.28	-0.23	0	-0.1	-0.32	____F_
Slovak Republic	1315	75.70	0.51	4.70	6.10	75.70	4.90	7.80	0.00	0.90	0.20	-0.29	-0.24	0.51	-0.19	-0.19	0	-0.08	-0.39	____F_
Slovenia	1193	71.40	0.46	4.80	7.10	71.40	6.50	9.10	0.10	1.00	0.10	-0.17	-0.18	0.46	-0.20	-0.21	-0.03	-0.12	-0.47	____F_
South Africa (Afr.)	533	37.10	0.33	12.80	15.00	37.10	16.10	17.60	0.00	1.30	0.40	-0.10	-0.10	0.33	-0.06	-0.17	0	-0.02	-0.4	_________
South Africa (Eng.)	2525	29.10	0.23	11.40	9.70	29.10	20.80	26.50	0.30	2.10	2.10	-0.04	-0.17	0.23	-0.07	-0.11	-0.1	-0.05	-0.86	__E_F_
Thailand	2162	63.60	0.43	4.40	5.60	63.60	7.60	18.40	0.10	0.30	0.00	-0.17	-0.15	0.43	-0.24	-0.18	-0.03	-0.03	-0.78	__E_F_
Tunisia	1874	59.70	0.42	10.00	9.10	59.70	9.10	9.40	0.30	2.30	1.40	-0.13	-0.16	0.42	-0.19	-0.19	-0.05	-0.07	-0.88	__E_F_
Turkey	2933	54.30	0.43	9.60	11.10	54.30	10.10	13.30	0.00	1.50	0.10	-0.14	-0.19	0.43	-0.20	-0.13	0	-0.06	-0.83	__E_F_
United States	3353	56.40	0.45	4.90	9.90	56.40	13.50	14.90	0.00	0.40	0.10	-0.13	-0.18	0.45	-0.24	-0.17	-0.01	-0.03	-0.14	__H_F_
International Avg. :		65.70	0.45	7.20	7.80	65.70	7.90	9.90	0.10	1.40	0.40	-0.17	-0.17	0.45	-0.22	-0.17	-0.01	-0.08	-0.53	

Keys: Diff= Percent obtaining maximum score; Disc= Item Discrimination; RDIFF= Difficulty (1-PLi); Pct_In= Invalid Responses; Pct_NR= Not Reached; Pct_OM=Omitted
Flags: A= Ability not ordered/ Attractive distractor; C= Difficulty less than chance; D= Negative/low discrimination; E= Easier than average;
F= Distractor chosen by less than 10%; H= Harder than average; R= Scoring reliability < 80%; V= Difficulty greater than 95.

Third International Mathematics and Science Study - 1999 Main Survey May 23, 2000 60
International Item Statistics (Unweighted) - Review Version
For Internal Review Only: DO NOT CITE OR CIRCULATE

Mathematics : Fractions and Number Sense (J12) Type : S Key: X Label: Division of fractions (M022026 - BSSMJ12)

Country	N	Diff	Disc	Pct_0	Pct_1	Pct_2	Pct_3	Pct_OM	Pct_NR	PB_0	PB_1	PB_2	PB_3	PB_OM	RDIFF	Cases	Reliability Score	Code	Flags
Australia	494	23.30	0.39	62.30	23.30	.	.	14.40	0.00	-0.22	0.39	.	.	-0.16	2.10	123.00	96.70	92.70	__H___
Belgium (Flem	656	67.20	0.40	27.70	67.20	.	.	5.00	0.00	-0.28	0.40	.	.	-0.29	0.35	90.00	100.00	98.90	_____
Bulgaria	411	68.60	0.52	20.00	68.60	.	.	11.40	0.20	-0.29	0.52	.	.	-0.39	-0.60	88.00	100.00	97.70	__E___
Canada (Englis	783	32.30	0.42	57.00	32.30	.	.	10.70	0.10	-0.33	0.42	.	.	-0.10	1.36	.	.	.	__H___
Canada (Frenc	302	33.80	0.32	56.00	33.80	.	.	10.30	0.00	-0.26	0.32	.	.	-0.07	1.44	.	.	.	__H___
Chile	723	26.40	0.48	51.50	26.40	.	.	22.10	0.40	-0.20	0.48	.	.	-0.27	0.26	.	.	.	_____
Chile (7th Gra	754	12.60	0.35	56.60	12.60	.	.	30.80	0.70	-0.03	0.35	.	.	-0.20	0.89	.	.	.	_____
Chinese Taipe	725	83.70	0.64	13.40	83.70	.	.	2.90	0.00	-0.53	0.64	.	.	-0.33	-0.83	180.00	100.00	100.00	__E___
Cyprus	388	21.40	0.49	54.10	21.40	.	.	24.50	0.00	-0.14	0.49	.	.	-0.31	1.51	.	.	.	__H___
Czech Republi	406	67.70	0.52	27.30	67.70	.	.	4.90	0.00	-0.48	0.52	.	.	-0.14	0.06	121.00	99.20	94.20	_____
England	373	4.60	0.26	82.30	4.60	.	.	13.10	0.00	-0.03	0.26	.	.	-0.13	4.12	89.00	100.00	93.30	__H_F_
Finland (Fin.)	344	16.00	0.31	70.90	16.00	.	.	13.10	0.30	-0.16	0.31	.	.	-0.09	2.42	.	.	.	__H___
Finland (Swe.)	44	4.50	0.12	61.40	4.50	.	.	34.10	0.00	0.05	0.12	.	.	-0.10	3.77	.	.	.	D_H_F_
Hong Kong, SA	648	81.60	0.39	14.50	81.60	.	.	3.90	0.20	-0.23	0.39	.	.	-0.34	-0.44	171.00	98.80	97.70	__E___
Hungary	400	68.00	0.46	25.00	68.00	.	.	7.00	0.00	-0.30	0.46	.	.	-0.33	-0.13	98.00	98.00	96.90	__E___
Indonesia	730	24.10	0.59	62.20	24.10	.	.	13.70	0.00	-0.43	0.59	.	.	-0.13	0.61	173.00	98.30	94.80	_____
Iran, Islamic R	668	35.30	0.38	56.70	35.30	.	.	7.90	0.00	-0.30	0.38	.	.	-0.13	0.03	170.00	92.90	73.50	_____
Israel (Arabic)	99	16.20	0.50	64.60	16.20	.	.	19.20	0.00	-0.19	0.50	.	.	-0.24	1.09	.	.	.	_____
Israel (Hebrev	409	24.00	0.39	46.50	24.00	.	.	29.60	0.70	-0.10	0.39	.	.	-0.25	1.36	.	.	.	__H___
Italy	419	62.10	0.52	24.60	62.10	.	.	13.40	0.20	-0.35	0.52	.	.	-0.29	-0.61	115.00	99.10	96.50	__E___
Japan	589	79.30	0.47	15.60	79.30	.	.	5.10	0.00	-0.34	0.47	.	.	-0.29	-0.32	140.00	100.00	99.30	__E___
Jordan	635	32.00	0.59	57.00	32.00	.	.	11.00	0.00	-0.41	0.59	.	.	-0.23	0.31	153.00	100.00	96.10	_____
Korea, Rep. o	765	80.70	0.56	14.80	80.70	.	.	4.60	0.00	-0.40	0.56	.	.	-0.37	-0.28	123.00	97.60	95.10	__E___
Latvia (LSS)	355	67.60	0.44	22.50	67.60	.	.	9.90	0.60	-0.28	0.44	.	.	-0.29	-0.56	97.00	97.90	95.90	__E___
Lithuania	287	65.50	0.43	26.80	65.50	.	.	7.70	0.30	-0.28	0.43	.	.	-0.31	-0.57	68.00	98.50	98.50	__E___
Macedonia (A	119	8.40	0.10	43.70	8.40	.	.	47.90	0.80	0.22	0.10	.	.	-0.27	1.56	.	.	.	D___F_
Macedonia (M	375	28.00	0.45	44.50	28.00	.	.	27.50	0.50	-0.08	0.45	.	.	-0.36	0.93	.	.	.	_____
Malaysia	699	54.80	0.61	36.80	54.80	.	.	8.40	0.10	-0.45	0.61	.	.	-0.31	0.37	191.00	100.00	99.00	_____
Moldova (Rom	378	55.60	0.53	33.30	55.60	.	.	11.10	0.30	-0.38	0.53	.	.	-0.26	-0.34	.	.	.	__E___
Moldova (Rus	70	64.30	0.51	28.60	64.30	.	.	7.10	1.40	-0.36	0.51	.	.	-0.34	-0.29	.	.	.	_____
Morocco	659	4.60	0.09	61.90	4.60	.	.	33.50	0.90	0.06	0.09	.	.	-0.08	1.46	108.00	95.40	88.90	D_H_F_
Netherlands	370	13.20	0.35	78.40	13.20	.	.	8.40	0.30	-0.16	0.35	.	.	-0.16	3.21	38.00	100.00	92.10	__H___
New Zealand	456	9.90	0.31	73.50	9.90	.	.	16.70	0.00	-0.04	0.31	.	.	-0.20	2.90	114.00	99.10	95.60	__H_F_
Philippines (En	754	12.20	0.41	77.70	12.20	.	.	10.10	0.40	-0.24	0.41	.	.	-0.11	0.94	.	.	.	_____
Philippines (Ta	79	2.50	0.26	86.10	2.50	.	.	11.40	0.00	0.01	0.26	.	.	-0.14	2.30	.	.	.	____F_
Romania (Hun	15	53.30	0.72	40.00	53.30	.	.	6.70	0.00	-0.63	0.69	.	.	-0.15	-0.78	.	.	.	_____
Romania (Rom	418	69.40	0.58	21.80	69.40	.	.	8.90	0.50	-0.35	0.58	.	.	-0.41	-0.93	.	.	.	__E___
Russian Feder	527	73.60	0.50	19.70	73.60	.	.	6.60	0.00	-0.37	0.50	.	.	-0.29	-0.61	120.00	98.30	96.70	__E___
Singapore	625	83.70	0.48	13.90	83.70	.	.	2.40	0.00	-0.41	0.48	.	.	-0.23	-0.42	100.00	98.00	98.00	__E___
Slovak Republ	430	73.30	0.42	20.90	73.30	.	.	5.80	0.00	-0.32	0.42	.	.	-0.23	-0.15	101.00	99.00	99.00	__E___
Slovenia	386	74.10	0.43	22.30	74.10	.	.	3.60	0.00	-0.33	0.43	.	.	-0.26	-0.67	104.00	100.00	95.20	__E___
South Africa (	170	5.90	0.42	85.30	5.90	.	.	8.80	0.00	-0.22	0.42	.	.	-0.07	2.24	.	.	.	__H_F_
South Africa (	851	4.20	0.32	88.50	4.20	.	.	7.30	0.20	-0.19	0.32	.	.	0.00	1.45	.	.	.	__H_F_
Thailand	711	41.90	0.59	51.10	41.90	.	.	7.00	0.10	-0.52	0.59	.	.	-0.12	0.28	207.00	100.00	100.00	_____
Tunisia	625	28.60	0.36	48.30	28.60	.	.	23.00	0.20	-0.21	0.36	.	.	-0.14	0.65	157.00	99.40	96.80	_____
Turkey	984	38.60	0.56	48.20	38.60	.	.	13.20	0.20	-0.38	0.56	.	.	-0.26	-0.17	247.00	100.00	98.80	__E___
United States	1128	37.40	0.52	55.30	37.40	.	.	7.30	0.10	-0.41	0.52	.	.	-0.18	0.77	118.00	99.20	97.50	__H___
International Avg. :		49.80	0.46	39.60	49.80	.	.	10.60	0.10	-0.30	0.46	.	.	-0.24	0.42	.	98.80	95.70	

Keys: Diff = Percent obtaining maximum score; RDIFF= Difficulty (1-PL); Pct_In= Invalid Responses; Pct_NR= Not Reached; Pct_OM= Omitted
Flags: A= Ability not ordered/ Attractive distractor; C= Difficulty less than chance; D= Negative/low discrimination; E= Easier than average;
F= Distractor chosen by less than 10%; H= Harder than average; R= Scoring reliability < 80%; V= Difficulty greater than 95.

Exhibits 13.1 and 13.2 contained the statistics described below.

N: This is the number of students to whom the item was administered. If an item was not reached by a student it was considered to be not administered for the purpose of the item analysis.[3]

Diff: The item difficulty was the percentage of students that provided a fully correct response to the item. In the case of free-response items worth more than one point this was the percentage of students achieving the maximum score on the item. When computing this statistic, not reached items were treated as not administered.

Disc: The item discrimination was computed as the correlation between correctly answering the item and total score on all of the items in the subject area in the test booklet.[4] This correlation should be moderately positive for items with good measurement properties.

PCT_A, PCT_B, PCT_C, PCT_D and PCT_E: Used for multiple-choice items only (Exhibit 13.1), these represent the percentage of students choosing each response option for the item. Not reached items were excluded from the denominator for these calculations.

PCT_0, PCT_1, PCT_2 and PCT_3: Used for open-ended items only (Exhibit 13.2), these are the percentage of students scoring at each score level for the item. Not reached items were excluded from the denominator for these calculations.

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

PCT_OM: This is the percentage of students that did not provide a response to the item even though the item was administered and they had reached it. Not reached items were excluded from the denominator when calculating this statistic.

○○○

3. In TIMSS, for the purposes of item analysis and item parameter estimation in scaling, items not reached by a student were treated as if they had not been administered. For purposes of estimating student proficiency, however, not reached items were treated as incorrectly answered.
4. For free-response items, the discrimination is the correlation between the number of score points and total score.

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

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

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

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

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

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

Reliability - Cases: It was expected that the free-response items in approximately one-quarter of the test booklets would be scored by two independent scorers. This column indicates the number of times each item was double scored in a country.

Reliability - Score: This column contains the percentage of times the two independent scorers agreed on the score level for the item.

Reliability - Code: This column contains the percentage of times the two scorers agreed on the two-digit code (score and diagnostic code) for the item.

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

- Item difficulty exceeds 95% in the sample as a whole
- Item difficulty is less than 25% for 4-option multiple-choice items in the sample as a whole (20% for 5-option items)
- Item difficulty exceeds 95% or is less than 25% (20% for 5-option items)
- One or more of the distracter percentages is less than 5%
- One or more of the distracter percentages is greater than the percentage for the correct answer
- Point-biserial correlation for one or more of the distracters exceeds zero
- Item discrimination (i.e., the point-biserial for the correct answer) is less than 0.2
- Item discrimination does not increase with each score level (for an item with more than one score level)
- Rasch goodness-of-fit index is less than 0.88 or greater than 1.12
- Difficulty levels on the item differ significantly for males and females
- Difference in item difficulty levels between males and females diverge significantly from the average difference between males and females across all the items making up the total score

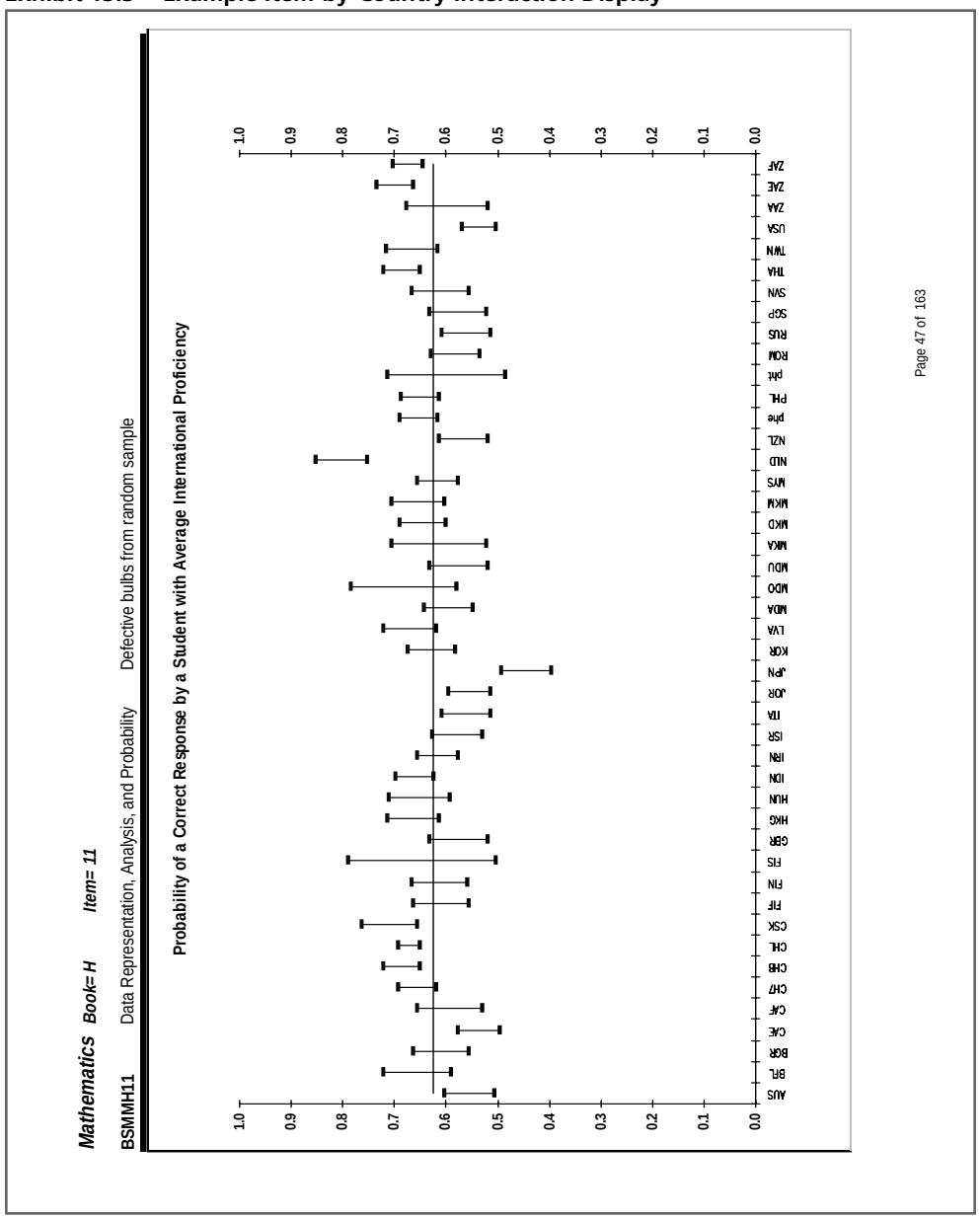
Although not all of these conditions necessarily indicate a problem, the flags are a useful way to draw attention to potential sources of concern. The IEA Data Processing Center also produced information about the inter-rater agreement for the free-response items.

13.2.1 Item-by-Country Interaction

Although there is room for variation across items, in general countries with high average performance on the achievement tests as a whole should perform relatively well on each of the items, and low-scoring countries should do less well on each of items. When this does not occur, i.e., when a high-scoring country has low performance on an item on which other countries are

doing well, there is said to be an item-by-country interaction. Since large item-by-country interactions can indicate an item that is flawed in some way, the item review also included this aspect of item performance.

Exhibit 13.3 Example Item-by-Country Interaction Display



To help examine item-by-country interactions, the International Study Center produced a graphical display for each item showing the average probability across all countries of a correct response for a student of average international proficiency, compared with the probability of a correct response by a student of average proficiency in each country (see Exhibit 13.3 for an example). The probability for each country is presented as a 95% confidence interval, which includes a built-in Bonferroni correction for multiple comparisons.

The limits for the confidence interval are computed as follows:

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

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

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

13.3 Item Checking Procedures

Prior to the IRT scaling of the TIMSS 1999 achievement data by Educational Testing Service, the International Study Center thoroughly reviewed the item statistics for all participating countries to ensure that items were performing comparably across countries. Although only a small number of items were found to be inappropriate for international comparisons, throughout the series of item-checking steps a number of reasons were discovered for differences in items across countries. Most of these were inadvertent changes in the items during printing, such as omitting an item option or misprinting the graphics associated with an item. Differences attributable to translation problems, however, were found for an item or two in several countries.

In particular, items with the following problems were considered for possible deletion from the international database:

- Errors were detected during translation verification but were not corrected before test administration

- Data cleaning revealed more or fewer options than in the original version of the item
- The item-analysis information showed the item to have a negative biserial
- The item-by-country interaction results showed a very large negative interaction for a given country
- The item-fit statistic indicated that the item did not fit the model
- For free-response items, the within-country scoring reliability data showed an agreement of less than 70% for the score level. Also, performance in items with more than one score level was not ordered by score, or correct levels were associated with negative point-biserials.

When the item statistics indicated a problem with an item, the documentation from the translation verification[5] was used as an aid in checking the test booklets and contacting National Research Coordinators (NRCs). If a problem could be detected by the International Study Center (such as a negative point-biserial for a correct answer or too few options for the multiple-choice questions), the item was deleted from the international scaling. If there was a question about potential translation or cultural issues, however, then the NRC was consulted before deciding how the item should be treated. Appendix D provides a list of deleted items as well as a list of recodes made to free-response item codes.

13.4 Summary

Considering that the checking involved more than 300 items for 38 countries (almost 12,000 item-country combinations), very few deviations from the international format were found. Appendix D summarizes the changes that were made to items in the international database before beginning the 1999 IRT scaling.

○○○

5. See Chapter 5 for a description of the translation and verification of the TIMSS data-collection instruments.