10

Item Analysis and Review

Ina V.S. Mullis

Michael O. Martin

Ann M. Kennedy

10.1 Overview

Prior to the item response theory (IRT) scaling of the PIRLS 2001 achievement scores, the International Study Center (ISC) reviewed a range of diagnostic statistics to examine and evaluate the psychometric characteristics of each achievement item within and across the 35 countries participating in PIRLS. For constructed-response items, the review included indicators of the reliability of the scoring procedure. The review process was an important step in the quality assurance of the PIRLS 2001 data, screening items for unusual psychometric properties that could signal a problem or error for an item in a particular country. For example, an item uncharacteristically easy or difficult in a country, or with an unusually low discriminating power, could indicate a potential problem with translation or printing. In the rare instances where such items were detected, the country's translation verification documents and printed booklets were examined for flaws or inaccuracies and the items removed from the database for that country. This chapter describes the basic item statistics that were consulted, and provides examples from the assessment to illustrate the review process.

10.2 Statistics for Item Analysis

As the first stage in the item review process, the PIRLS ISC computed a set of item statistics for each achievement item, showing the properties of the item in each of the 35 countries participating in PIRLS 2001. Exhibits 10.1 and 10.2 show the statistics calculated for a multiple-choice and a constructed-response item, respectively. Statistics for each item are displayed alphabetically by country, with

Exhibit 10.1: International Item Statistics for Item R011H05M

Progress in International Reading Literacy Study - 2001 Assessment Main Survey Results — September 11, 2002 5
International Item Statistics (Unweighted) - Review Version Only - DO NOT CITE OR CIRCULATE — 10:07

Block: Hare (Literary Experience) — Item Number: H05 -R011H05M
Item Label: Lion dropped the fruit (Make Straightforward Inferences) — Type: MC Key: C
Released = Yes

Country	N	Diff	Disc	Pct_A	Pct_B	Pct_C	Pct_D	Pct_In	Pct_OM	Pct_NR	PB_A	PB_B	PB_C	PB_D	PB_In	PB_OM	RDIFF	Avg. Score Girls	Avg. Score Boys	Flags
Argentina	768	59.0	0.52	17.3	9.8	59.0	10.4	0.9	2.6	3.2	-0.23	-0.28	0.52	-0.11	-0.12	-0.13	-0.36	0.60	0.57	__H_F__
Belize	276	41.3	0.43	29.7	13.8	41.3	9.4	0.0	5.8	4.2	-0.11	-0.21	0.43	-0.21	0.00	-0.11	-0.91	0.43	0.39	____F__
Bulgaria	879	92.6	0.47	3.2	1.7	92.6	1.8	0.3	0.3	0.1	-0.30	-0.21	0.47	-0.18	-0.20	-0.12	-1.33	0.93	0.92	____F__
Canada (O,Q)	2099	92.5	0.44	3.4	1.6	92.5	2.0	0.0	0.4	0.2	-0.25	-0.24	0.44	-0.23	0.00	-0.09	-1.45	0.94	0.91	__E_F_G
Colombia	1265	68.2	0.59	13.4	9.3	68.2	6.4	0.0	2.6	1.4	-0.29	-0.33	0.59	-0.20	0.00	-0.20	-0.85	0.71	0.66	____F_G
Cyprus	746	76.8	0.53	15.1	2.9	76.8	4.4	0.0	0.7	0.1	-0.37	-0.24	0.53	-0.18	0.00	-0.15	-0.42	0.78	0.76	__H_F__
Czech Republic	748	87.0	0.44	4.9	0.9	87.0	6.1	0.0	0.9	0.0	-0.26	-0.19	0.44	-0.25	0.00	-0.15	-0.92	0.86	0.88	____F__
England	773	87.2	0.53	9.2	1.0	87.2	2.5	0.0	0.1	0.3	-0.43	-0.18	0.53	-0.17	0.00	-0.09	-0.72	0.89	0.85	____F__
France	888	89.3	0.45	5.9	0.7	89.3	2.8	0.8	0.6	0.0	-0.26	-0.16	0.45	-0.21	-0.27	-0.09	-1.12	0.90	0.89	____F__
Germany	1890	91.0	0.49	1.9	3.6	91.0	2.9	0.2	0.5	0.6	-0.20	-0.31	0.49	-0.26	-0.04	-0.09	-1.41	0.93	0.89	__E_F_G
Greece	633	85.5	0.45	10.7	0.9	85.5	2.5	0.0	0.3	0.0	-0.38	-0.11	0.45	-0.17	0.00	-0.11	-0.71	0.82	0.88	____F_B
Hong Kong	1253	93.0	0.43	2.7	2.2	93.0	1.9	0.0	0.2	0.5	-0.24	-0.26	0.43	-0.19	0.00	-0.10	-1.52	0.94	0.92	__E_F__
Hungary	1164	93.3	0.43	3.4	1.1	93.3	2.0	0.0	0.3	0.0	-0.27	-0.23	0.43	-0.24	0.00	-0.01	-1.42	0.95	0.91	__E_F_G
Iceland	349	88.0	0.44	6.0	2.3	88.0	1.7	0.0	2.0	0.9	-0.26	-0.24	0.44	-0.13	0.00	-0.15	-1.41	0.90	0.86	____F__
Iran, Islamic Rep.	1891	75.5	0.54	8.7	7.0	75.5	6.9	0.3	1.6	0.2	-0.32	-0.26	0.54	-0.19	-0.10	-0.17	-1.08	0.76	0.75	____F__
Israel	983	81.5	0.54	9.2	4.0	81.5	4.6	0.0	0.8	0.1	-0.36	-0.30	0.53	-0.18	0.00	-0.07	-0.65	0.84	0.79	__H_F_G
Italy	874	84.2	0.56	8.1	2.7	84.2	3.9	0.0	1.0	0.0	-0.42	-0.27	0.56	-0.17	0.00	-0.11	-0.34	0.86	0.82	__H_F__
Kuwait	638	61.1	0.62	15.5	14.6	61.1	5.3	0.0	3.4	1.8	-0.28	-0.35	0.62	-0.14	0.00	-0.17	-0.55	0.62	0.60	__H_F__
Latvia	747	94.6	0.35	2.8	0.7	94.6	1.3	0.4	0.1	0.0	-0.25	-0.17	0.35	-0.09	-0.18	-0.04	-1.65	0.97	0.92	__E_F_G
Lithuania	702	91.0	0.40	4.7	0.6	91.0	3.4	0.1	0.1	0.0	-0.31	-0.20	0.40	-0.16	-0.01	-0.12	-0.94	0.94	0.88	____F_G
Macedonia, Rep. of	924	65.3	0.65	18.6	5.2	65.3	7.7	0.9	2.4	1.1	-0.40	-0.29	0.65	-0.16	-0.12	-0.17	-0.56	0.67	0.64	__H_F__
Moldova, Rep. of	858	82.3	0.46	9.3	2.9	82.3	5.4	0.0	0.1	0.0	-0.31	-0.17	0.46	-0.24	0.00	-0.03	-1.19	0.84	0.80	____F__
Morocco	760	51.7	0.51	13.2	16.2	51.7	11.7	0.9	6.3	2.8	-0.11	-0.28	0.51	-0.14	-0.11	-0.21	-0.91	0.55	0.49	_______
Netherlands	1016	90.7	0.41	7.1	0.6	90.7	1.3	0.0	0.3	0.0	-0.36	-0.11	0.41	-0.12	0.00	-0.07	-0.67	0.91	0.91	____F__
New Zealand	606	86.5	0.62	8.7	2.5	86.5	1.5	0.0	0.8	0.5	-0.47	-0.20	0.62	-0.18	0.00	-0.24	-1.15	0.89	0.84	____F__
Norway	849	83.9	0.55	7.4	4.1	83.9	3.2	0.0	1.4	1.3	-0.34	-0.31	0.55	-0.17	0.00	-0.15	-1.15	0.87	0.81	____F_G
Romania	915	76.8	0.58	15.2	3.0	76.8	3.5	0.8	0.8	0.3	-0.36	-0.25	0.57	-0.16	-0.19	-0.20	-0.23	0.77	0.76	__H_F__
Russian Federation	1019	89.0	0.38	4.2	1.7	89.0	4.8	0.0	0.3	0.3	-0.25	-0.10	0.38	-0.21	0.00	-0.15	-1.23	0.88	0.90	____F__
Scotland	658	86.8	0.60	7.0	3.2	86.8	2.3	0.2	0.6	0.2	-0.42	-0.26	0.60	-0.23	-0.01	-0.13	-1.01	0.87	0.86	____F__
Singapore	1746	88.7	0.57	6.7	2.5	88.7	2.0	0.0	0.1	0.0	-0.38	-0.30	0.57	-0.28	0.00	-0.04	-1.19	0.91	0.87	____F_G
Slovak Republic	943	88.8	0.47	4.9	1.9	88.8	3.9	0.0	0.5	0.1	-0.29	-0.20	0.47	-0.24	0.00	-0.14	-1.33	0.89	0.88	____F__
Slovenia	734	83.5	0.46	10.9	2.3	83.5	2.7	0.0	0.5	0.3	-0.33	-0.18	0.46	-0.19	0.00	-0.06	-1.25	0.84	0.83	____F__
Sweden	1488	94.7	0.40	2.2	1.2	94.7	1.5	0.0	0.4	0.1	-0.24	-0.24	0.40	-0.15	0.00	-0.13	-1.52	0.96	0.94	__E_F_G
Turkey	1277	72.8	0.55	12.4	6.7	72.8	7.7	0.0	0.4	0.3	-0.28	-0.27	0.55	-0.30	0.00	-0.09	-0.66	0.77	0.69	__H_F_G
United States	924	90.4	0.54	5.3	2.4	90.4	1.2	0.0	0.8	0.2	-0.36	-0.32	0.54	-0.18	0.00	-0.09	-1.19	0.91	0.90	____F__
International Avg.	.	81.8	0.50	8.8	3.9	81.8	4.1	0.2	1.1	0.6	-0.31	-0.23	0.50	-0.19	-0.04	-0.12	-1.00	0.83	0.80	.....
Ontario (Canada)	1082	91.5	0.45	4.3	1.7	91.5	2.2	0.0	0.4	0.3	-0.25	-0.24	0.45	-0.25	0.00	-0.10	-1.44	0.94	0.89	__E_F_G
Quebec (Canada)	1017	93.6	0.41	2.6	1.5	93.6	1.9	0.0	0.5	0.1	-0.25	-0.23	0.41	-0.19	0.00	-0.10	-1.45	0.94	0.93	__E_F__
Sweden (3rd Grade)	1294	90.0	0.51	4.8	2.6	90.0	1.9	0.0	0.8	0.5	-0.36	-0.26	0.51	-0.16	0.00	-0.14	-1.47	0.91	0.89	__E_F__

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

Exhibit 10.2: International Item Statistics for Item R011R06C

Progress in International Reading Literacy Study - 2001 Assessment Main Survey Results — September 11, 2002 6
International Item Statistics (Unweighted) - Review Version Only - DO NOT CITE OR CIRCULATE — 10:07

Block: River (Acquire and Use Information) — Item Number: R06 -R011R06C
Item Label: Equipment for children (Focus on and Retrieve Explicitly Stated Information and Ideas) — Type: CR Key: X
Released = Yes

Country	N	Diff	Disc	Pct_0	Pct_1	Pct_2	Pct_3	Pct_OM	Pct_NR	PB_0	PB_1	PB_2	PB_3	PB_OM	RDIFF	Reliability Cases	Reliability Score	Avg. Score Girls	Avg. Score Boys	Flags
Argentina	770	32.9	0.64	36.1	19.5	32.9	.	11.6	5.4	-0.40	0.09	0.57	.	-0.20	-0.13	192	86.5	0.85	0.85	__E____
Belize	768	14.2	0.58	62.9	12.8	14.2	.	10.2	8.1	-0.40	0.18	0.52	.	-0.10	-0.06	149	94.6	0.45	0.39	__E____
Bulgaria	858	64.6	0.65	17.8	14.0	64.6	.	3.6	0.7	-0.47	-0.12	0.61	.	-0.29	0.22	224	92.4	1.47	1.39	_______
Canada (O,Q)	2042	51.6	0.64	30.4	14.6	51.6	.	3.4	0.0	-0.51	-0.07	0.61	.	-0.23	0.41	439	91.6	1.21	1.14	__H____
Colombia	1216	32.4	0.66	38.8	24.6	32.4	.	4.2	4.0	-0.52	0.06	0.59	.	-0.14	-0.25	299	92.6	0.92	0.87	__E____
Cyprus	753	46.1	0.67	23.1	21.6	46.1	.	9.2	1.3	-0.50	-0.02	0.60	.	-0.23	0.03	147	96.6	1.28	1.00	_____G
Czech Republic	760	56.2	0.59	28.2	12.6	56.2	.	3.0	0.0	-0.44	-0.11	0.58	.	-0.29	0.33	182	100.0	1.27	1.22	__H____
England	806	58.9	0.65	24.1	14.0	58.9	.	3.0	0.1	-0.53	-0.11	0.61	.	-0.24	0.42	196	98.5	1.37	1.27	__H____
France	880	69.4	0.59	16.5	10.2	69.4	.	3.9	0.2	-0.47	-0.09	0.55	.	-0.25	0.16	210	97.6	1.51	1.48	_______
Germany	1931	57.3	0.64	23.8	14.8	57.3	.	4.1	0.3	-0.46	-0.13	0.61	.	-0.28	0.21	226	98.2	1.34	1.25	_____G
Greece	602	37.2	0.59	37.2	15.3	37.2	.	10.3	1.1	-0.35	0.07	0.53	.	-0.34	0.72	47	100.0	1.00	0.80	__H___G
Hong Kong	1261	52.7	0.57	16.1	28.7	52.7	.	2.5	0.8	-0.37	-0.17	0.52	.	-0.27	0.41	274	91.6	1.46	1.24	__H___G
Hungary	1156	55.1	0.68	22.7	18.0	55.1	.	4.2	0.7	-0.52	-0.08	0.62	.	-0.26	0.19	232	92.7	1.32	1.24	_______
Iceland	1123	57.5	0.66	20.6	17.5	57.5	.	4.4	0.8	-0.47	-0.13	0.62	.	-0.29	-0.15	223	91.5	1.35	1.30	__E____
Iran, Islamic Rep.	1753	20.8	0.52	48.1	17.7	20.8	.	13.4	6.4	-0.25	0.05	0.50	.	-0.22	0.27	492	96.5	0.58	0.60	_______
Israel	981	46.8	0.68	21.8	24.1	46.8	.	7.3	1.2	-0.52	0.02	0.58	.	-0.29	-0.06	225	88.0	1.22	1.14	__E____
Italy	860	40.7	0.60	34.4	18.4	40.7	.	6.5	0.2	-0.41	0.04	0.53	.	-0.33	0.75	215	95.8	0.98	1.02	__H____
Kuwait	1886	11.9	0.55	35.7	31.8	11.9	.	20.6	8.5	-0.27	0.30	0.38	.	-0.23	0.39	.	.	0.63	0.47	__H___G
Latvia	748	72.3	0.66	14.4	9.9	72.3	.	3.3	0.0	-0.54	-0.16	0.63	.	-0.24	-0.13	176	94.9	1.65	1.44	__E_F_G
Lithuania	702	59.0	0.66	18.9	19.2	59.0	.	2.8	0.0	-0.50	-0.13	0.60	.	-0.28	0.08	60	96.7	1.41	1.34	_______
Macedonia, Rep. of	867	23.8	0.65	44.4	24.3	23.8	.	7.5	4.4	-0.44	0.18	0.54	.	-0.20	0.15	238	91.2	0.74	0.69	_______
Moldova, Rep. of	894	42.4	0.65	38.6	15.5	42.4	.	3.5	0.6	-0.59	0.10	0.57	.	-0.12	0.19	226	83.2	1.06	0.94	_____G
Morocco	636	9.9	0.56	44.7	17.8	9.9	.	27.2	11.0	-0.29	0.32	0.41	.	-0.18	0.47	.	.	0.43	0.33	__H_F_G
Netherlands	1044	81.4	0.58	7.5	10.3	81.4	.	0.8	0.1	-0.48	-0.21	0.53	.	-0.13	-0.42	256	97.3	1.75	1.71	__E_F__
New Zealand	622	50.6	0.71	26.5	20.1	50.6	.	2.7	0.5	-0.58	-0.03	0.63	.	-0.24	0.26	177	96.0	1.34	1.10	_____G
Norway	847	51.4	0.65	22.9	20.1	51.4	.	5.7	1.5	-0.46	-0.10	0.61	.	-0.24	-0.02	219	94.5	1.26	1.20	__E____
Romania	873	40.8	0.60	36.9	16.7	40.8	.	5.6	0.6	-0.44	0.02	0.55	.	-0.25	0.37	211	97.6	1.00	0.96	__H____
Russian Federation	1033	69.1	0.62	13.5	14.4	69.1	.	3.0	0.9	-0.43	-0.20	0.59	.	-0.26	-0.32	253	96.0	1.58	1.47	__E___G
Scotland	676	55.0	0.65	25.4	17.3	55.0	.	2.2	0.6	-0.52	-0.11	0.61	.	-0.20	0.34	204	95.1	1.30	1.25	_____G
Singapore	1741	54.2	0.61	30.0	14.5	54.2	.	1.3	0.3	-0.51	-0.09	0.58	.	-0.17	0.36	284	99.6	1.32	1.14	__H___G
Slovak Republic	945	70.4	0.60	12.2	14.0	70.4	.	3.5	0.5	-0.41	-0.18	0.56	.	-0.27	-0.67	951	99.1	1.61	1.49	__E___G
Slovenia	730	47.3	0.65	33.0	16.6	47.3	.	3.2	0.7	-0.54	-0.03	0.60	.	-0.18	0.21	198	91.4	1.15	1.07	_____G
Sweden	1488	74.7	0.64	10.4	12.2	74.7	.	2.7	0.1	-0.46	-0.19	0.59	.	-0.33	0.00	220	97.3	1.67	1.57	__E___G
Turkey	1265	18.9	0.53	59.6	12.2	18.9	.	9.3	1.4	-0.40	0.10	0.50	.	-0.09	0.46	352	99.7	0.57	0.44	__H___G
United States	928	51.1	0.65	29.4	18.6	51.1	.	0.9	0.3	-0.55	-0.12	0.62	.	-0.13	0.40	232	98.3	1.31	1.10	__H___G
International Avg.	.	48.0	0.62	28.8	17.3	48.0	.	6.0	1.8	-0.46	-0.03	0.57	.	-0.23	0.16	.	94.9	1.18	1.08	....
Ontario (Canada)	1071	46.3	0.68	32.8	16.2	46.3	.	4.8	0.1	-0.53	-0.02	0.63	.	-0.26	0.44	.	.	1.16	1.02	__H___G
Quebec (Canada)	971	57.5	0.59	27.7	13.0	57.5	.	1.9	0.0	-0.48	-0.12	0.57	.	-0.17	0.38	.	.	1.27	1.29	__H____
Sweden (3rd Grade)	1318	66.3	0.63	14.2	14.1	66.3	.	5.4	1.2	-0.42	-0.15	0.59	.	-0.28	-0.19	191	97.9	1.52	1.41	__E___G

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; B= Boys outperforming girls; C= Difficulty less than chance; D= Negative/low discrimination; E= Easier than average;
F= Distractor chosen by less than 10%; G= Girls outperforming boys; H= Harder than average; R= Scoring reliability < 80%; V= Difficulty greater than 95.

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

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

For all items, the item-analysis includes the average score for male and female students. This is the average score received by girls and boys on a scale ranging from zero to the maximum possible score point for the item. For multiple-choice items or 1-point constructed-response items, this statistic also represents the average difficulty of the item for girls and boys.

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

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

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

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

2 In calculating item statistics and in item parameter estimation for scaling, items not reached by a student were treated as if they had not been administered. In estimating student proficiency, however, not reached items were treated as answered incorrectly.

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

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

Pct_0, Pct_1, Pct_2, and Pct_3: Used for constructed-response items only (see Exhibit 10.2), each column indicates the percentage of students scoring at the particular score level, up to and including the maximum score level for the item. Not-reached items were excluded from the denominator for these calculations.

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

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

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

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

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

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

PB_OM: This is the correlation between a binary variable – indicating an omitted response to the item – and the total score. 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 total score. This correlation also should be negative or near zero.

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

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

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

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

- Item difficulty exceeds 95 percent
- Item difficulty is less than 25 percent for four-option multiple-choice items
- One or more of the distracter percentages is less than 10 percent
- One or more of the distracter percentages is greater than the percentage for the correct answer, or the point-biserial correlation for one or more of the distracters exceeds zero
- Item discrimination (i.e., the point-biserial for the correct answer) is less than 0.2
- Item discrimination does not increase with each score level (for constructed-response items with more than one score level)
- The Rasch difficulty estimate is above the average across all items

- The Rasch difficulty estimate is below the average across all items
- Difficulty levels on the item differ significantly for males and females
- Scoring reliability is less than 80 percent (for constructed-response items only).

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

10.2.1 Item-by-Country Interaction

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

To assist in detecting sizeable item-by-country interactions, the International Study Center produced a graphical display for each item showing the average probability across all countries of a correct response for a student of average proficiency internationally, compared with the probability of a correct response by a student of average proficiency in each country. Exhibit 10.3 provides an example of a PIRLS item-by-country interaction display.

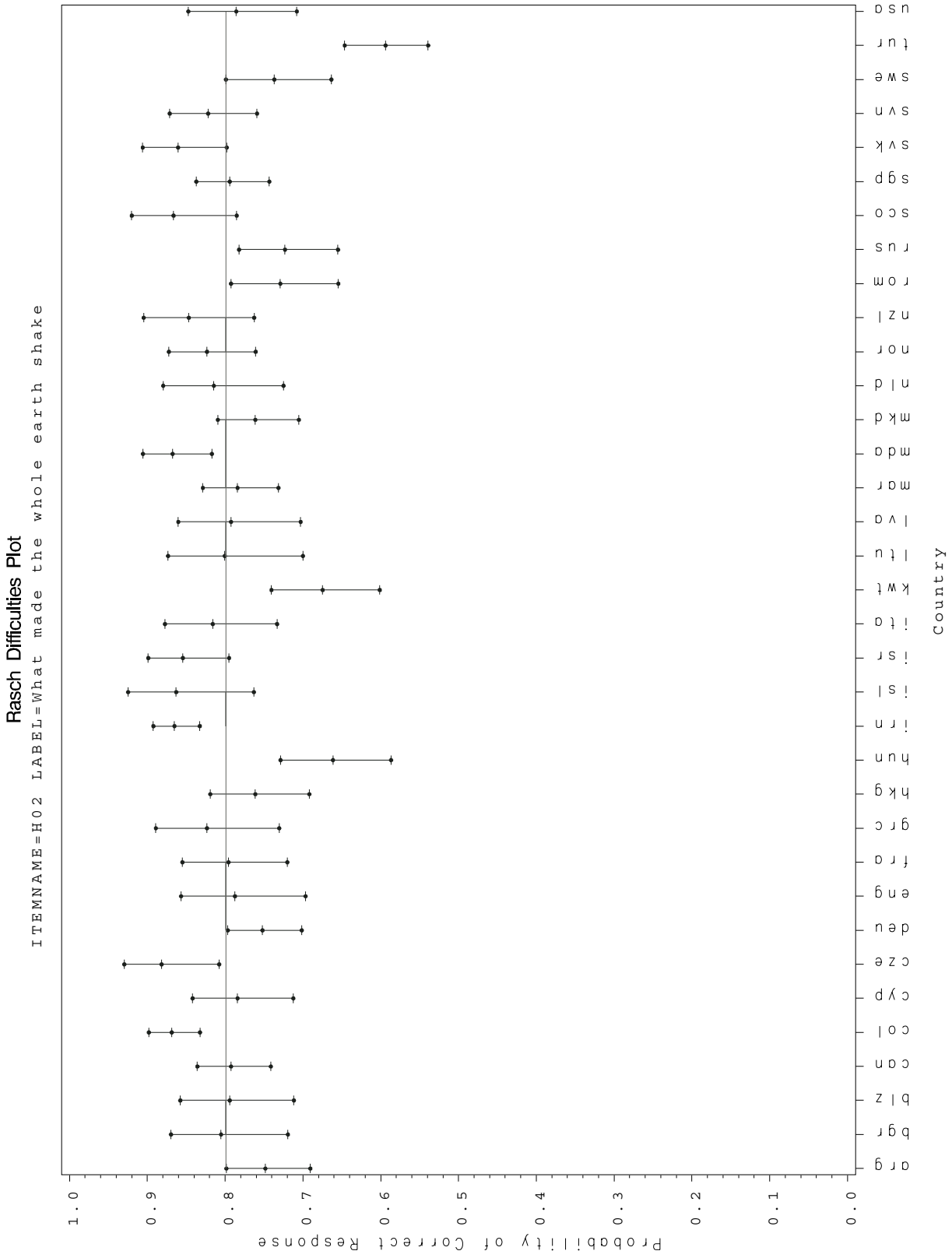
The probability for each country is presented as a 95 percent confidence interval, which includes a built-in Bonferroni correction for multiple comparisons. The limits for the confidence interval are computed as follows:

$$\text{UpperLimit} = \bar{1} - \frac{e^{RDIFF_{ik} - SE_{RDIFF_{ik}} * Z_b}}{1 + e^{RDIFF_{ik} - SE_{RDIFF_{ik}} * Z_b}}$$

$$\text{LowerLimit} = \bar{1} - \frac{e^{\text{RDIFF}_{ik} + SE_{RDIFF_{ik}} * Z_b}}{1 + e^{RDIFF_{ik} + SE_{RDIFF_{ik}} * Z_b}}$$

where $RDIFF_{ik}$ is the Rasch difficulty of item k within country i; $SE_{RDIFF_{ik}}$ 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.

Exhibit 10.3: Example Item-by-Country Interaction Display for Item R011H02M



10.3 Scoring Reliability for Constructed-Response Items

Half of the items in the PIRLS assessment were constructed-response items, comprising nearly two-thirds of the score points for the assessment.[4] 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 PIRLS, measures taken to ensure that the constructed-response items were scored reliably in all countries included developing scoring guides for each constructed-response question (which provided descriptions of acceptable responses for each score point value),[5] and providing extensive training in the application of the scoring guides. Scoring procedures for organizing and monitoring the scoring sessions were outlined in the PIRLS *Survey Operations Manual*.

10.3.1 Within-Country Scoring Reliability

To gather and document information about the agreement among scorers, a random sample of at least 200 students' responses to each item (approximately 25% of the total responses) was selected by the National Research Coordinators to be scored independently by two scorers. A measure of agreement between scorers (the percentage of times the scores of the two scorers agreed exactly) was calculated for each item in each country, and was examined as part of the item review process. Items with percentage agreement less than 70 percent were flagged for further examination. The average and range of the exact percent of agreement across all items is presented (Exhibit 10.4) for each country. The average of exact percent agreement across items was high – on average, across countries, exact percent agreement was 93 percent. All countries had an average exact percent agreement above 83 percent.

4 For details on the development of the PIRLS assessment items, see Chapter 2.

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

Exhibit 10.4: Within-Country Constructed-Response Scoring Reliability

Countries	Correctness Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent of Agreement	
		Minimum	Maximum
Argentina	86	71	95
Belize	92	86	97
Bulgaria	83	60	99
Canada (O,Q)	87	66	99
Colombia	83	65	100
Cyprus	96	86	100
Czech Republic	97	82	100
England	96	81	100
France	96	87	100
Germany	89	71	100
Greece	98	92	100
Hong Kong, SAR	88	61	97
Hungary	94	80	100
Iceland	86	70	99
Iran, Islamic Rep. of	95	90	99
Israel	91	83	97
Italy	94	68	100
Kuwait	–	–	–
Latvia	92	64	99
Lithuania	88	68	100
Macedonia, Rep. of	94	85	98
Moldova, Rep. of	94	83	99
Morocco	–	–	–
Netherlands	90	67	100
New Zealand	97	89	100
Norway	92	81	99
Romania	94	76	100
Russian Federation	98	91	100
Scotland	93	76	100
Singapore	99	98	100
Slovak Republic	99	99	100
Slovenia	92	67	100
Sweden	94	86	100
Turkey	99	98	100
United States	97	89	100
International Avg.	93	79	99

* A dash (–) indicates data not available

10.3.2 Cross-Country Scoring Reliability Study

To gather information about how consistently the scoring guides were applied across countries, the International Study Center conducted a cross-country reliability study in which a sample of student responses was scored independently by two English-proficient scorers from each participating country. Taking into consideration available resources and other feasibility issues, the cross-country scoring reliability study was conducted in English, using a core set of 200 student responses to each of 25 constructed-response questions from half of the assessment blocks – two literary and two informational.

The core set of 5,000 responses comprised student responses from Canada, England, Scotland, and the United States. A total of 55 scorers from 28 PIRLS countries participated in the study.[6] Scoring for this study took place shortly after the within-country scoring reliability activities were completed. Using the same scoring guides from the national within-country scoring activities, each scorer was asked to assign a score to each student response in the set. Each student response to an individual question resulted in 1,485 possible comparisons among scorers. When aggregated across all 200 student responses to the item, there were a total of 297,000 comparisons, provided a score was assigned by all 55 scorers.

Exhibit 10.5 shows the percentage of paired scorers that were in exact agreement across all responses to each of the items used in the reliability study. The extent of agreement varied across items. On average, across all items, 85 percent of all possible paired-scorer combinations were in exact agreement on the assigned score.

6 Only one scorer proficient in English was available in Macedonia. In the Russian Federation, resources permitted only a portion of the English-language responses to be scored.

Exhibit 10.5: Cross-Country Constructed-Response Scoring Reliability

Purpose	Item Label	Total Valid Comparisons*	Exact Percent Agreement
Literary Experience	Unreleased C01	275496	99%
	Unreleased C02	275444	89%
	Unreleased C03	275548	93%
	Unreleased C06	275341	98%
	Unreleased C08	275496	92%
	Unreleased C10	275548	66%
	Unreleased C11	275444	72%
	Hare H03	275600	90%
	Hare H04	275393	93%
	Hare H07	275444	79%
	Hare H08	275086	84%
	Hare H09	275236	84%
	Hare H10	273661	73%
Acquire and Use Information	Unreleased A01	296892	96%
	Unreleased A03	296676	98%
	Unreleased A04	296676	90%
	Unreleased A07	296892	87%
	Unreleased A08	296623	80%
	Unreleased A09	296784	81%
	Unreleased A11	296191	80%
	Pufflings N07	274724	78%
	Pufflings N08	274724	83%
	Pufflings N10	273947	84%
	Pufflings N12	274673	76%
	Pufflings N13	274621	73%
		Average Percent Agreement	85%

* Values for items differ slightly due to a small number of missing responses.

10.4 Item Analysis for the Trends in IEA's Reading Literacy Study

The review of the item statistics for each of the nine countries participating in the Trends in IEA's Reading Literacy Study followed the PIRLS approach. Statistics calculated for the trend study items were the same as those used in PIRLS (as described in Section 10.2). An example item statistics display for a trend study item is presented in Exhibit 10.6. Different from the PIRLS item statistics, the trend item statistics include countries' statistics for both 1991 and 2001. In reviewing the item statistics, comparisons in performance were made across countries within a year, as well as within countries across years.

Exhibit 10.6: International Item Statistics for Trend Item E25

Progress in International Reading Literacy Study - 2001 (10YTS) Results
International Item Statistics (Unweighted) - Review Version
For Internal Review Only: DO NOT CITE OR CIRCULATE

September 24, 2002 7
17:49

Reporting Category: Expository Prose (WALRUS)
Item Label: HOW WALRUS GETS UP ON ICE

Item Number: E25 -AEWALRU5
Type: MC Key: C

* * * * * 1991 * * * * *

Country	N	Diff	Disc	Pct_A	Pct_B	Pct_C	Pct_D	Pct_In	Pct_OM	Pct_NR	PB_A	PB_B	PB_C	PB_D	PB_In	PB_OM	RDIFF	Avg. Score Girls	Avg. Score Boys	Flags
Greece	3373	75.0	0.34	5.3	2.8	75.0	15.6	0.0	1.2	3.8	-0.25	-0.16	0.34	-0.12	0.00	-0.08	-0.39	0.74	0.76	__E_F._
Hungary	2307	75.6	0.35	6.4	1.4	75.6	15.9	0.0	0.7	22.4	-0.25	-0.09	0.35	-0.13	0.00	-0.04	-0.19	0.75	0.76	____F._
Iceland	3471	82.7	0.35	4.2	0.7	82.7	11.7	0.0	0.6	12.1	-0.24	-0.13	0.35	-0.11	0.00	-0.02	-0.58	0.86	0.79	__E_F.G
Italy	2097	73.3	0.23	1.9	1.1	73.3	23.1	0.0	0.5	5.6	-0.19	-0.17	0.23	-0.02	0.00	-0.11	0.12	0.77	0.70	__H_F.G
New Zealand	2859	74.3	0.44	6.5	1.9	74.3	16.9	0.0	0.4	4.3	-0.29	-0.22	0.44	-0.21	0.00	0.00	-0.05	0.77	0.72	__H_F.G
Singapore	7277	69.2	0.38	7.7	1.2	69.2	21.8	0.0	0.1	0.7	-0.27	-0.18	0.38	-0.20	0.00	0.00	-0.06	0.71	0.67	__H_F.G
Slovenia	2932	68.1	0.42	6.3	1.9	68.1	23.3	0.0	0.3	10.6	-0.24	-0.15	0.42	-0.21	0.00	-0.05	-0.05	0.71	0.65	__H_F.G
Sweden	3688	88.8	0.34	3.1	0.4	88.8	7.3	0.0	0.4	13.8	-0.21	-0.14	0.34	-0.15	0.00	-0.04	-0.64	0.91	0.87	__E_F.G
United States	6278	80.6	0.41	5.9	1.2	80.6	11.8	0.0	0.5	1.9	-0.29	-0.17	0.41	-0.22	0.00	-0.07	-0.33	0.81	0.80	____F._
International Avg.	.	76.4	0.36	5.3	1.4	76.4	16.4	0.0	0.5	8.4	-0.25	-0.16	0.36	-0.15	0.00	-0.05	-0.24	0.78	0.75	....

* * * * * 2001 * * * * *

Country	N	Diff	Disc	Pct_A	Pct_B	Pct_C	Pct_D	Pct_In	Pct_OM	Pct_NR	PB_A	PB_B	PB_C	PB_D	PB_In	PB_OM	RDIFF	Avg. Score Girls	Avg. Score Boys	Flags
Greece	1061	77.7	0.30	9.0	1.0	77.7	11.5	0.0	0.8	4.3	-0.30	-0.08	0.30	-0.07	0.00	-0.04	-0.21	0.78	0.77	____F._
Hungary	3750	82.0	0.29	4.8	1.7	82.0	10.7	0.0	0.7	20.3	-0.23	-0.19	0.29	-0.11	0.00	-0.02	-0.59	0.81	0.83	__E_F.B
Iceland	1663	84.9	0.29	3.1	0.9	84.9	10.6	0.0	0.4	7.4	-0.24	-0.16	0.29	-0.11	0.00	-0.11	-0.62	0.85	0.85	__E_F._
Italy	1505	67.8	0.21	2.1	0.5	67.8	28.8	0.0	0.7	5.3	-0.19	-0.11	0.21	-0.07	0.00	-0.06	0.46	0.71	0.65	__H_F.G
New Zealand	1157	73.9	0.46	7.1	1.7	73.9	16.4	0.0	0.9	2.5	-0.33	-0.27	0.46	-0.20	0.00	-0.07	0.04	0.76	0.71	__H_F.G
Singapore	3564	68.4	0.42	9.3	1.7	68.4	20.5	0.0	0.1	0.9	-0.30	-0.18	0.42	-0.21	0.00	-0.02	-0.05	0.72	0.65	__H_F.G
Slovenia	1415	74.0	0.36	5.0	1.3	74.0	19.4	0.0	0.3	5.8	-0.26	-0.11	0.36	-0.22	0.00	-0.06	-0.12	0.76	0.72	____F._
Sweden	4488	86.1	0.31	4.5	0.9	86.1	7.8	0.0	0.8	14.5	-0.29	-0.13	0.31	-0.15	0.00	-0.05	-0.52	0.86	0.86	__E_F._
United States	1817	80.5	0.40	5.2	1.0	80.5	12.9	0.0	0.4	0.4	-0.27	-0.18	0.40	-0.23	0.00	-0.03	-0.28	0.81	0.80	____F._
International Avg.	.	77.3	0.34	5.6	1.2	77.3	15.4	0.0	0.6	6.8	-0.27	-0.16	0.34	-0.15	0.00	-0.05	-0.21	0.79	0.76	....

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

10.4.1 Item-by-Country Interactions for the Trends in IEA's Reading Literacy Study

The international Study Center also produced item-by-country interaction displays for each item in the trend study, showing the results from 1991 and 2001 separately in each display. An example of an item-by-country interaction display for a trend item is presented in Exhibit 10.7. Confidence intervals for 1991 and 2001 within a country appear side-by-side in the display to compare performance from one administration to the next. At the same time, the display can be used to detect item-by-country interactions across all countries. The procedure for computing the 95 percent confidence interval limits for the probability for each country is presented in Section 10.2.1.

10.5 Item Review Procedures

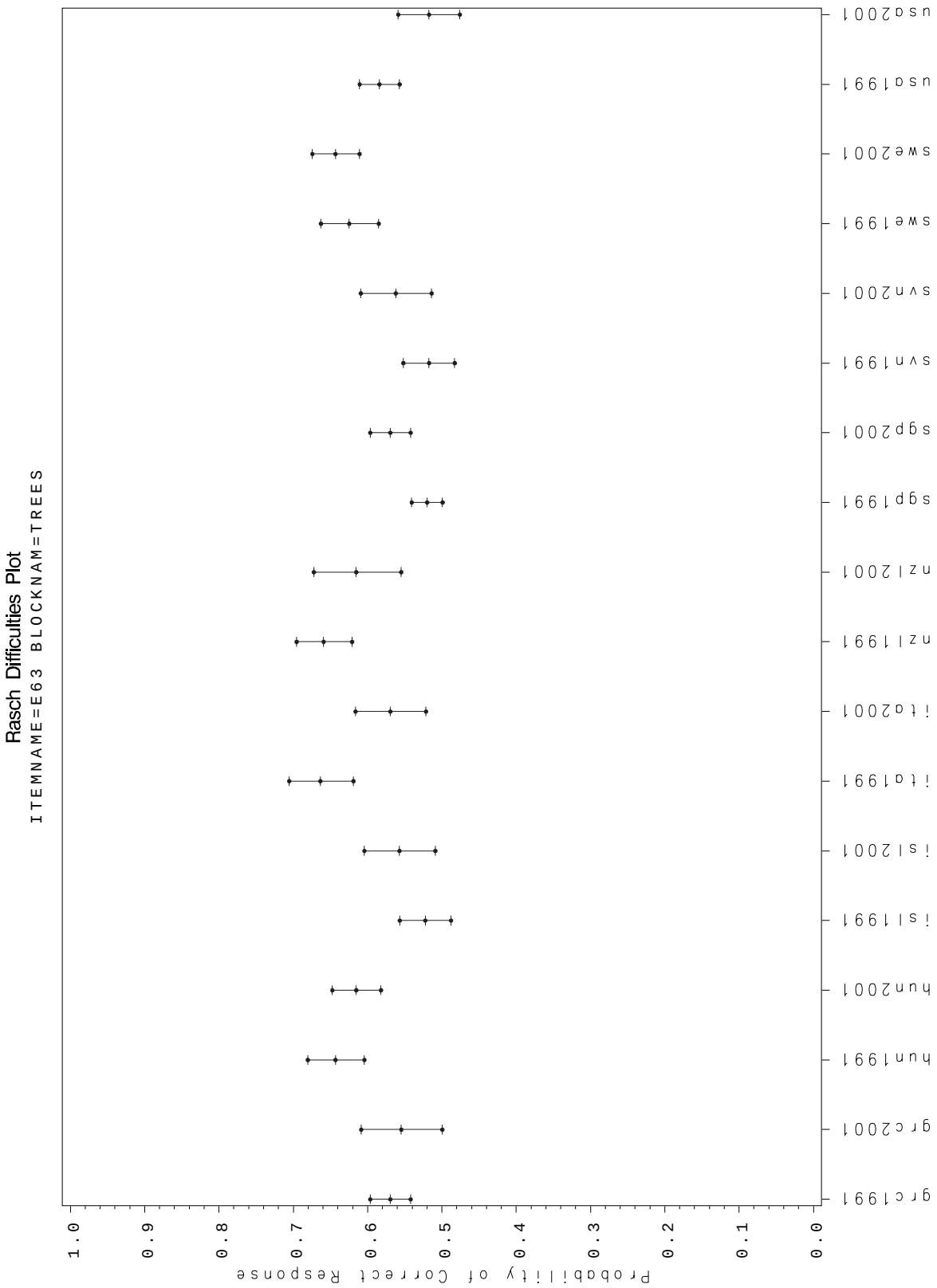
The International Study Center thoroughly reviewed the item statistics for all participating countries to ensure that items were performing comparably across countries. In particular, items with the following problems were considered for possible deletion from the international database:

- An error was detected during PIRLS 2001 translation verification but was not corrected before test administration.
- Data checking revealed a multiple-choice item with more or fewer options than in the international version.
- The item analysis showed the item to have a negative biserial, or, for an item with more than one score point, a non-monotonic relationship between score level and total score.
- The item-by-country interaction results showed a very large negative interaction for a particular country.
- For constructed-response items, the within-country scoring reliability data showed an agreement of less than 70 percent.
- For Trends in IEA's Reading Literacy Study items, an item performed substantially differently in 1991 compared to 2001, or an item was not included in the 1991 assessment for a particular country.

When the item statistics indicated a problem with an item, the documentation from the translation verification[7] was used as an aid in checking the test booklets. If a question remained about potential translation or cultural issues, however, then the National Research Coordinator (NRC) was consulted before deciding how the item should be treated. If a problem could be detected by the International Study Center (such as a negative point-biserial for a correct answer or too few options for a multiple-choice item), the item was deleted from the international scaling.

7 See chapter 5 for a description of the process for translating and verifying the PIRLS data-collection instruments.

Exhibit 10.7: Example Item-by-Country Interaction Display for Trend Item E63



The checking of the PIRLS 2001 achievement data involved 98 items for 35 countries (approximately 3,500 item-country combinations), and resulted in the detection of very few items that were inappropriate for international comparisons. Just two items had to be deleted from the international database, one for Cyprus and one for the Russian-speaking part of Moldova (see Appendix C). The checking of the Trends in IEA's Reading Literacy Study data involved 66 items for 9 countries. The items were deleted for all countries, and several items were identified in individual countries as inappropriate for international comparisons. Appendix C provides a list of deleted items as well as a list of recodes made to constructed-response item codes.