6 Item Analysis and Review

Ina V.S. Mullis
Michael O. Martin
Boston College

6.1 CROSS-COUNTRY ITEM STATISTICS

In order to assess the statistical properties of the Population 3 (final year of secondary school) items before proceeding with item response theory (IRT) scaling (see Chapter 7), TIMSS computed a series of statistics for every item in every country. These basic item statistics (see Figure 6.1 for an example item) were produced by the IEA Data Processing Center. For each item, the 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).[1] For multiple-choice items 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), the display presents the difficulty and discrimination of each score level. As a prelude to the main IRT scaling, the display presents some statistics from a preliminary Rasch analysis, the Rasch item difficulty for each item, the standard error of this difficulty estimate, and an index of the goodness-of-fit of the item to the Rasch model (Wu, 1997).

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 presents the international item difficulty (the mean of the item difficulties across countries) and the international item discrimination (the mean of the item discriminations).

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

- Item difficulty exceeds 95 percent in the sample as a whole
- Item difficulty is less than 25 percent for 4-option multiple-choice items in the sample as a whole (20 percent for 5-option items)

[1] For the purpose of computing the discrimination index, the total score was the percentage of items a student answered correctly in mathematics or science.

Figure 6.1 Example of Cross-Country Item Analysis

Population: '3' (C) Subject: 'General' Cluster:A Item: CSMGA01 Page 1

Country	Correct answer			Flags	Percentages for each alternative								Point biserials for each alternative								Rasch			Group difficulties				Internatinal Mean	
	N	DIFF	DISCR		A	B	C	D	E	W	OMIT	NR	A	B	C	D	E	W	OMIT	NR	RDIFF	SE	FIT	MAL	FEM	LOW	UPP	IDIFF	IDISCR
AUS	1958	42.2	0.30	.....sS.	5.6	42.2*	10.2	39.7			2.1	0.1	-0.12	.30*	-0.12	-0.14			-0.11	-0.02	1.11	0.05	1.09	49.7	37.1	.	.	39.5	0.26
AUT	2040	49.1	0.29	...q.....sS.	3.8	49.1*	5.8	38.2			2.1	0.1	-0.01	.29*	-0.10	-0.20			-0.12	0.00	0.53	0.05	1.07	60.0	39.6	.	.	39.5	0.26
CAN	5361	40.8	0.26	....Q....sS.	6.6	40.8*	10.2	40.9			1.5	0.1	-0.01	.26*	-0.11	-0.17			-0.05	-0.04	1.22	0.03	1.10	46.2	35.0	.	.	39.5	0.26
CHE	3458	46.9	0.35	.....sS.	5.1	46.9*	7.8	34.2			5.4	0.0	-0.01	.35*	-0.06	-0.26			-0.16	-0.05	1.16	0.04	1.01	56.7	36.7	.	.	39.5	0.26
CSK	2196	42.4	0.34	.....B..sS.	8.5	42.4*	5.2	39.3			4.3	0.0	0.07	.34*	0.00	-0.36			-0.07	-0.06	0.90	0.05	1.10	50.5	33.9	.	.	39.5	0.26
CYP	538	31.6	0.25	...qQ.....	4.8	31.6*	9.9	49.6			3.9	0.0	-0.04	.25*	-0.12	-0.13			-0.05	0.00	0.75	0.10	1.08	33.6	29.5	.	.	39.5	0.26
DEU	2439	43.2	0.33	...q.....sS.	3.2	43.2*	4.3	37.7			8.2	0.4	-0.04	.33*	-0.11	-0.18			-0.18	-0.04	0.80	0.05	1.05	53.9	34.4	.	.	39.5	0.26
FRA	1865	30.4	0.26	....Q....s..	5.1	30.4*	14.5	44.6			4.9	0.2	-0.03	.26*	0.00	-0.19			-0.08	-0.05	1.38	0.05	1.06	32.9	27.6	.	.	39.5	0.26
GRC	353	54.4	0.20	.....B.Fs..	7.1	54.4*	6.8	25.5			5.9	0.0	0.09	.20*	-0.06	-0.21			-0.05	0.00	0.27	0.12	1.14	57.1	48.1	.	.	39.5	0.26
HUN	5356	43.4	0.30	...q.....sS.	3.6	43.4*	4.0	42.7			6.4	0.0	-0.03	.30*	-0.06	-0.23			-0.09	0.00	0.58	0.03	1.04	46.9	39.8	.	.	39.5	0.26
ISL	1832	27.3	0.31	....Q.B..sS.	5.0	27.3*	6.4	59.1			1.9	0.1	0.00	.31*	0.04	-0.28			-0.07	-0.01	2.01	0.06	0.99	36.7	18.7	.	.	39.5	0.26
ISR	1357	20.9	0.20	p..qQ.B.FsS.	3.1	20.9*	8.5	58.1			8.0	1.4	-0.07	.20*	-0.09	0.07			-0.26	-0.06	1.91	0.07	1.16	26.6	14.9	.	.	39.5	0.26
LTU	2886	44.6	0.27	...q..B..s..	4.7	44.6*	4.4	38.6			7.2	0.4	0.04	.27*	-0.03	-0.19			-0.16	-0.04	0.38	0.04	1.10	47.1	43.4	.	.	39.5	0.26
MEX	3535	28.9	0.07	...qQbB.Fs..	4.0	28.9*	7.7	54.1			4.9	0.3	-0.12	.07*	-0.14	0.07			-0.03	0.00	0.37	0.04	1.14	31.0	26.4	.	.	39.5	0.26
NLD	1628	39.2	0.23	....Q.B.FsS.	8.8	39.2*	7.4	43.7			0.9	0.1	0.03	.23*	-0.14	-0.16			-0.06	0.03	1.47	0.05	1.14	46.2	31.6	.	.	39.5	0.26
NOR	2518	37.1	0.29	....Q....sS.	6.2	37.1*	8.4	43.3			5.0	0.0	-0.01	.29*	-0.07	-0.18			-0.14	-0.02	1.33	0.05	1.06	46.3	28.7	.	.	39.5	0.26
NZL	2308	37.5	0.27	....Q....sS.	5.5	37.5*	11.8	44.2			1.0	0.0	-0.10	.27*	-0.10	-0.14			-0.07	0.00	1.42	0.05	1.09	44.4	31.2	.	.	39.5	0.26
RUS	2599	49.6	0.36	...q.....sS.	3.2	49.6*	4.8	37.6			4.5	0.1	-0.01	.36*	-0.06	-0.32			-0.06	-0.03	0.35	0.04	1.01	60.2	42.8	.	.	39.5	0.26
SVN	1650	26.5	0.31	....Q....sS.	5.5	26.5*	6.9	50.3			10.2	0.0	-0.06	.31*	-0.03	-0.21			-0.04	0.00	1.52	0.06	1.06	32.9	19.4	.	.	38.9	0.26
SWE	3106	55.8	0.26	.....B.FsS.	9.4	55.8*	7.5	25.7			1.3	0.0	0.05	.26*	-0.11	-0.24			-0.06	0.00	0.72	0.04	1.13	62.8	49.2	.	.	39.5	0.26
USA	5997	38.5	0.19	.....b..FsS.	8.0	38.5*	13.7	38.5			1.2	0.0	-0.10	.19*	-0.16	-0.01			-0.02	0.00	0.46	0.03	1.22	44.1	33.1	.	.	39.5	0.26
ZAF	3662	24.9	0.12	p...QbB.F.S.	12.8	24.9*	11.7	45.7			4.8	0.0	-0.10	.12*	-0.06	0.03			-0.06	0.00	-0.04	0.04	1.22	24.8	25.2	.	.	39.5	0.26

- Item difficulty exceeds 95 percent or is less than 25 percent (20 percent for 5-option items)
- Item difficulty exceeds 95 percent or is less than 25 percent (20 percent for 5-option items)
- One or more of the distracter percentages is less than 5 percent
- 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 a reviewer's attention to potential sources of concern. The IEA Data Processing Center also produced information about the inter-rater agreement for the free-response items.

6.2 GRAPHICAL DISPLAYS

As a further aid to reviewing the psychometric characteristics of the items, the Australian Council for Educational Research (ACER) produced graphical representations of selected item statistics for each participating country (see Figure 6.2). This display presents, for each item, the difficulty level and discrimination for every country, together with the Rasch goodness-of-fit statistic and an indication of the item-by-country interaction. The item-by-country interaction chart plots a confidence interval for the probability of success on the item in each country against the average probability of success across all countries. The graphical representations allow comparisons of these statistics across countries at a glance.

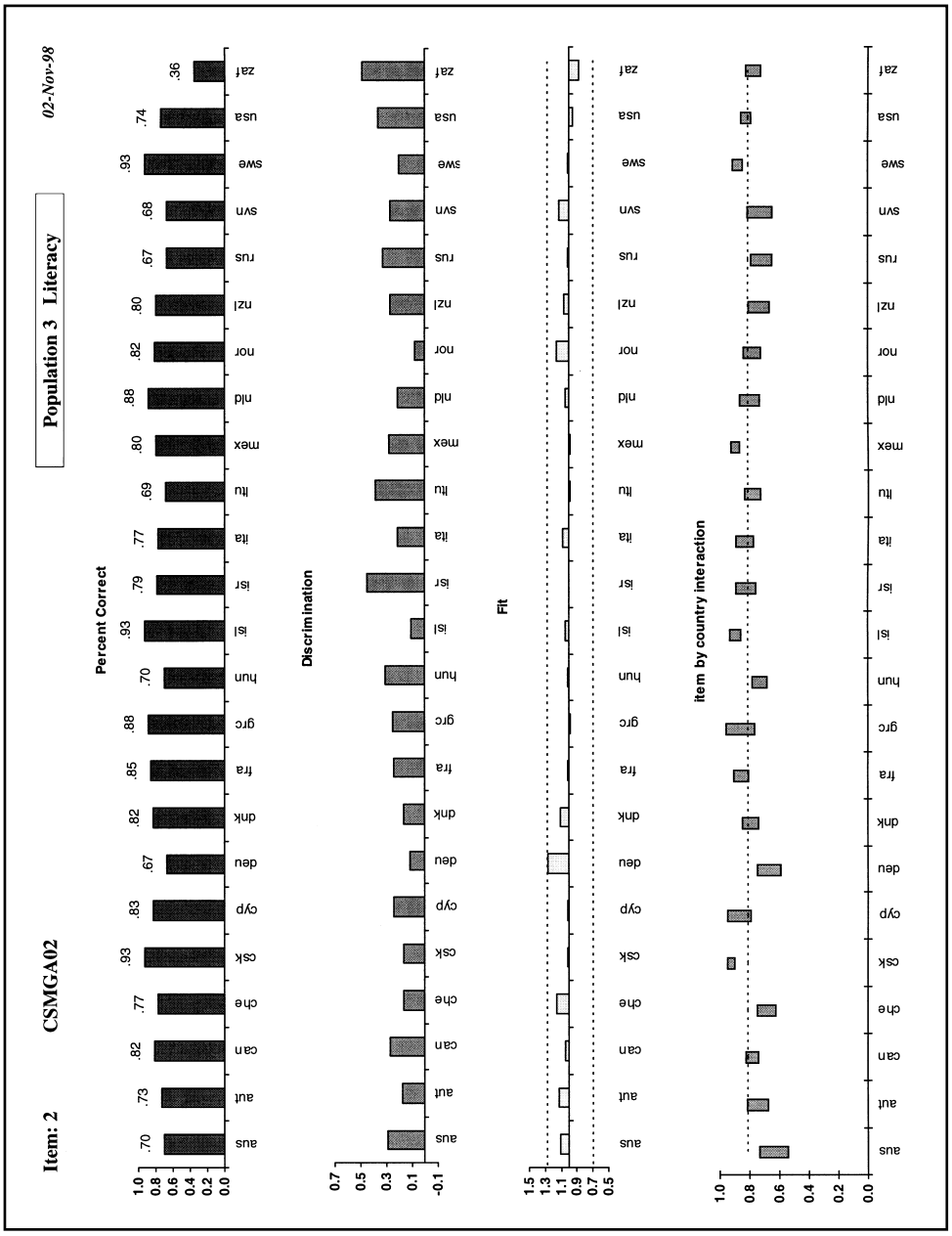

Figure 6.2 Example of Graphical Displays of Cross-Country Item Statistics

6.3 SUMMARY INFORMATION FOR POTENTIALLY PROBLEMATIC ITEMS

Although the system of flagging potentially problematic conditions and the graphical summaries were both very helpful in identifying items with possible problems, the task of reviewing the characteristics of each item in each country was still considerable. To ensure that no serious item problem would go unnoticed, ACER also provided, for each item, a list of countries that exhibited one or more potentially serious characteristics (see Figure 6.3). Countries were listed in this display if the item had a significant item-by-country interaction (i.e., students in the country found the item easier or more difficult than items in general), or if they exhibited problematic discrimination (i.e., the point-biserial for a distracter was greater than .05, the point-biserial for the correct answer was negative, or, for items with more than one score point, the point-biserial did not increase with each score level). Countries were also listed if their data showed poor fit to the Rasch model for that item.

6.4 ITEM CHECKING PROCEDURES

Prior to the international scaling of the Population 3 achievement data by ACER, 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, including omitting an item option or misprinting the graphics associated with an item. However, differences attributable to translation problems 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 the item did not fit the model
- For free-response items, the within-country scoring reliability data showed an agreement of less than 70 percent 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.

Figure 6.3 Example Summary Information for Items With Poor Statistics for Some Countries

Country	Item by Country Interactions: Easier than Expected	Item by Country Interactions: Harder than Expected	Discrimination: Non-key PB is Positive	Discrimination: Key PB is Negative	Discrimination: Ability not Ordered	Fit: Fit Large
	Tolerance = #Name					
Item: 16	***CSEGA12***					
HUN	☐	☐	☐	☐	☐	☑
RUS	☐	☐	☐	☐	☐	☑
Item: 18	***CSMGB02***					
GRC	☐	☑	☑	☐	☐	☐
Item: 19	***CSMGB03***					
USA	☑	☐	☐	☐	☐	☐
Item: 20	***CSMGB04***					
FRA	☐	☑	☐	☐	☐	☐
ISR	☐	☐	☑	☐	☐	☐
SVN	☑	☐	☐	☐	☐	☐
Item: 21	***CSMGB05***					
CYP	☐	☐	☑	☐	☐	☐

The statistics and translation verification documentation were used as pointers towards checking actual 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. However, if there was a question about potential translation or cultural issues, then the NRC was queried, and the International Study Center abided by the decision made by the NRC. In several cases, NRCs consulted mathematics or science experts before making a decision.

Considering that the checking involved approximately 200 items for more than 20 countries, very few deviations from the international format were found. Tables 6.1 and 6.2 contain a list of the changes made in the international database for Population 3.

Table 6.1 Deleted Cognitive Items - Population 3

	Country	Item	Variable Name
Mathematics and Science Literacy	All	A09, Part A	CSEGA09A
		C10	CSMGC10
	Cyprus	C05	CSMGC05
		D12	CSMGD12
	Greece	C05	CSMGC05
		D12	CSMGD12
		A11, Part C	CSEGA11C
	France	B04	CSMGB04
		B06	CSMGB06
	Hungary	B08	CSMGB08
		B21	CSMGB21
		B26	CSSGB26
		C20	CSSGC20
		D15, Part B	CSSGD15B
		D16, Part A	CSSGD16A
		D16, Part B	CSSGD16B
	Switzerland	B06	CSMGB06
	Slovenia	A11, Part C	CSEGA11C
Advanced Mathematics	Cyprus	J02	CSMMJ02
	France	J18	CSEMJ18
	Greece	J02	CSMMJ02
	Israel	J14	CSMMJ14
		J16, Part B	CSSMJ16B
		L08	CSMML08
	Lithuania	K09	CSMMK09
	Switzerland	J02	CSMMJ02
		J17	CSSMJ17
	United States	J08	CSMMJ08
Physics	All	H11	CSMPH11
	Australia	H19, Part A	CSEPH19A
	Czech Republic	F06	CSMPF06
	Denmark	F07	CSMPF07
		H14	CSSPH14
	France	F15	CSEPF15
	Germany	G16	CSEPG16
		H14	CSSPH14

Table 6.2 Recodes Made to Population 3 Free-Response Item Codes

	Item	Variable	Recodes			Comment
Mathematics and Science Literacy	B25	CSSGB25	20	→	10	Category 10 was only 1 point category and generally had less than 1 percent of the students, which made distinction between 1 and 2 points unclear.
			21	→	11	
			22	→	12	
			10	→	13	
			29	→	19	
	B26	CSSGB26	10	→	23	Categories 10 and 19 contain correct answer.
			19	→	29	
	D02	CSSGD02	20	→	12	Discrimination between 20s and 10s not clear.
			21	→	13	
	D04	CSEGD04	20	→	10	Is a link item with Y01 at Population 2 and as with Population 2 only 20s had positive point-biserials in many countries.
			21	→	11	
			22	→	12	
			29	→	19	
			10	→	73	
			11	→	74	
			19	→	75	
	D17	CSSGD17	13	→	22	In some countries 10s had almost the same or even higher point-biserials than 20s.

REFERENCES

Wu, M.L. (1997). *The development and application of a fit test for use with marginal maximum likelihood estimation and generalised item response models*. Unpublished master's dissertation, University of Melbourne.