Chapter 8

Scaling the Data from the TIMSS Advanced 2008 Assessments

Pierre Foy, Joseph Galia, and Isaac Li

8.1 Overview

The TIMSS Advanced 2008 goals of broad coverage of the advanced mathematics and physics curricula and of measuring trends across assessments necessitated the adoption of a complex matrix-sampling booklet design,[1] with individual students responding to a subset of the advanced mathematics and physics items in the assessment, and not the entire assessment item pool. Given the complexities of the data collection and the need to have student scores on the entire assessment for analysis and reporting purposes, TIMSS Advanced relied on Item Response Theory (IRT) scaling to describe student achievement on the assessment and to provide accurate measures of trends from the previous assessments. The TIMSS IRT scaling approach used multiple imputation—or "plausible values"—methodology to obtain proficiency scores in advanced mathematics and physics for all students, even though each student responded to only a part of the assessment item pool. To enhance the reliability of the student scores, the TIMSS

1 The TIMSS Advanced 2008 assessment design is described in the *TIMSS Advanced 2008 Assessment Frameworks* (Garden, et al., 2006).

advanced scaling combined student responses to the items they were administered with information about students' backgrounds, a process known as "conditioning."

This chapter describes the steps that produced scaled scores of student achievement in advanced mathematics and physics. First, it explains the process of reviewing item statistics to validate the statistical properties of the achievement items used in the TIMSS Advanced 2008 assessments. It then provides a general explanation of the methodology used to scale the TIMSS Advanced 2008 data, and describes how this approach was applied to the 2008 assessment data and to the data from the previous TIMSS Advanced 1995 study in order to measure trends in achievement. The TIMSS Advanced scaling was conducted by the TIMSS & PIRLS International Study Center at Boston College, using software from Educational Testing Service.[2] This chapter also provides a description of the scale anchoring methodology used to describe student performance at various points on the TIMSS Advanced mathematics and physics achievement scales, and the methodology used to estimate standard errors of the estimates published for TIMSS Advanced 2008.

8.2 Item Review

For TIMSS Advanced 2008, as in the TIMSS assessments at the fourth and eighth grades, 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 in the ten countries that participated in TIMSS Advanced 2008. This review of item statistics was conducted before applying item response theory (IRT) scaling to the TIMSS Advanced 2008 achievement data to derive student achievement scores in advanced mathematics and physics for analysis and reporting. The review of item statistics played

2 TIMSS is indebted to Matthias Von Davier, Ed Kulick, and John Barone of Educational Testing Service for their advice and support.

a crucial role in the quality assurance of the TIMSS Advanced 2008 data, 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.

8.2.1 Statistics for Item Analysis

To begin the review process, the TIMSS & PIRLS International Study Center computed item statistics for all 143 advanced mathematics and physics achievement items that were administered in the TIMSS Advanced 2008 assessments. The properties of the items in each of the ten countries that participated were then carefully reviewed. Exhibits 8.1 and 8.2 show actual samples of the statistics calculated for a multiple-choice and a constructed-response item, respectively.

For all items, regardless of format, 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 a total score).[3] Also provided was an estimate of the difficulty of the item using a Rasch one-parameter IRT model. The international means of the item difficulties and item discriminations served as guides to the overall statistical properties of the items. Statistics for each item were displayed alphabetically by country,

3 For computing point-biserial correlations, the total score was the percentage of advanced mathematics or physics items a student answered correctly.

Exhibit 8.1 International Item Statistics for a Multiple-choice Item

Trends in International Mathematics and Science Study - TIMSS 2008 Assessment Results
International Item Statistics (Unweighted) - Review Version - 12th Grade
For Internal Review Only: DO NOT CITE OR CIRCULATE

Mathematics: Calculus / Reasoning (MA23206 - M4_06)
Label: Sign of derivative function
Type: MC Key: D

				Percentages							Point Biserials							
Country	N	Diff	Disc	Pct_A	Pct_B	Pct_C	Pct_D	Pct_E	Pct_OM	Pct_NR	PB_A	PB_B	PB_C	PB_D	PB_E	PB_OM	RDIFF	Flags
Armenia	439	41.5	0.45	32.8	9.1	8.4	41.5	.	8.2	0.2	-0.14	-0.22	-0.20	0.45	.	-0.13	-0.77	____F_
Iran, Islamic Rep. of	1195	59.7	0.54	13.6	5.4	9.5	59.7	.	11.7	0.1	-0.28	-0.10	-0.22	0.54	.	-0.26	-0.78	____F_
Italy	1080	44.3	0.47	44.4	2.2	4.1	44.3	.	5.1	0.1	-0.39	-0.09	-0.06	0.47	.	-0.06	-0.70	____F_
Lebanon	801	80.9	0.37	15.7	0.6	2.1	80.9	.	0.6	0.0	-0.32	-0.08	-0.10	0.36	.	-0.07	-1.36	__E_F_
Netherlands	763	87.5	0.25	5.4	1.3	5.6	87.5	.	0.1	0.3	-0.17	-0.05	-0.16	0.25	.	-0.03	-1.94	__E_F_
Norway	966	52.3	0.38	28.8	8.6	8.1	52.3	.	2.3	0.0	-0.33	-0.03	-0.09	0.38	.	-0.08	-1.15	____F_
Philippines	2057	17.6	0.23	61.1	9.7	10.3	17.6	.	1.3	0.0	-0.16	-0.07	0.05	0.23	.	-0.02	0.15	_CH_F_
Russian Federation	1588	68.1	0.54	24.5	0.9	5.5	68.1	.	1.0	0.0	-0.42	-0.08	-0.22	0.54	.	-0.11	-0.64	__H_F_
Slovenia	1090	38.3	0.46	41.2	4.4	12.3	38.3	.	3.9	0.0	-0.36	-0.04	-0.10	0.46	.	-0.04	-0.26	__H_F_
Sweden	1141	67.3	0.48	17.0	4.8	9.6	67.3	.	1.3	0.3	-0.35	-0.11	-0.19	0.48	.	-0.11	-2.01	__E_F_
International Avg.	1112	55.8	0.42	28.4	4.7	7.5	55.8	.	3.6	0.1	-0.29	-0.09	-0.13	0.42	.	-0.09	-0.95	____F_

Keys: Diff: Percent correct score; Disc: Item discrimination; Pct_A...E: Percent choosing option; Pct_OM, NR: Percent Omitted and Not Reached;
PB_A...E: Point Biserial for option; PB_OM: Point Biserial for Omittted. RDIFF= Rasch difficulty.

Flags: A= Ability not ordered/Attractive distractor; C= Difficulty less than chance; D= Negative/low discrimination; E= Easier than average;
F= Distractor chosen by less than 10%; H= Harder than average; R= Scoring reliability < 80%; V= Difficulty greater than 95.

Exhibit 8.2 International Item Statistics for a Constructed-response Item

Trends in International Mathematics and Science Study - TIMSS 2008 Assessment Results
International Item Statistics (Unweighted) - Review Version - 12th Grade
For Internal Review Only: DO NOT CITE OR CIRCULATE

Mathematics: Calculus / Reasoning (MA23157 - M5_06)
Label: Find point of maxima and inflection
Type: CR Key: X

				Percentages					Point Biserials					Reliability			
Country	N	Diff	Disc	Pct_0	Pct_1	Pct_2	Pct_OM	Pct_NR	PB_0	PB_1	PB_2	PB_OM	RDIFF	Cases	Score	Code	Flags
Armenia	407	11.5	0.57	24.1	13.3	4.9	57.7	4.9	-0.14	0.43	0.36	-0.32	1.02	91	100.0	100.0	____F_
Iran, Islamic Rep. of	1132	33.4	0.68	27.0	23.8	21.6	27.7	6.8	-0.25	0.13	0.61	-0.41	0.52	294	100.0	98.3	______
Italy	1032	16.2	0.55	27.8	17.6	7.4	47.2	3.4	-0.12	0.20	0.49	-0.29	0.83	139	100.0	100.0	____F_
Lebanon	784	39.5	0.47	26.4	50.5	14.3	8.8	2.1	-0.32	0.12	0.38	-0.18	0.82	188	99.5	99.5	______
Netherlands	757	43.8	0.34	22.3	58.0	14.8	4.9	1.2	-0.19	0.04	0.28	-0.18	0.60	198	94.9	93.9	______
Norway	936	28.6	0.39	28.5	45.0	6.1	20.4	2.6	-0.21	0.10	0.39	-0.17	0.41	235	100.0	99.1	__E_F_
Philippines	2033	16.7	0.38	50.2	31.0	1.2	17.6	0.2	-0.21	0.24	0.36	-0.12	0.89	507	95.1	93.9	__H_F_
Russian Federation	1582	57.2	0.47	12.2	41.2	36.6	10.1	0.9	-0.19	-0.10	0.42	-0.30	0.02	402	98.8	98.8	__E___
Slovenia	1058	20.1	0.43	44.6	31.8	4.3	19.4	1.9	-0.19	0.26	0.32	-0.21	1.04	283	100.0	99.3	__H_F_
Sweden	1112	20.7	0.54	46.8	21.8	9.8	21.7	4.1	-0.31	0.21	0.46	-0.24	0.36	289	96.9	94.5	__E_F_
International Avg.	1083	28.8	0.48	31.0	33.4	12.1	23.5	2.8	-0.21	0.16	0.41	-0.24	0.65	.	98.5	97.7	______

Keys: Diff: Percent correct score; Disc: Item discrimination; Pct_0...3: Percent obtaining score level; Pct_OM, NR: Percent Omitted and Not Reached;
PB_0...3: Point Biserial for score level; PB_OM: Point Biserial for Omittted; RDIFF= Rasch difficulty;
Reliability: (Cases) Responses double scored; (Score) Percent agreement on score; (Code) Percent agreement on code.
Flags: A= Ability not ordered/Attractive distractor; 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 < 80%; V= Difficulty greater than 95.

together with the international average for each statistic in the last row of the exhibits.

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 difficulty and discrimination of each score level. Constructed-response item displays also provided information about the reliability with which each item was scored in each country, showing the total number of double-scored cases, the percentage of code agreement between the scorers, and the percentage of score agreement between scorers.

The definitions and detailed descriptions of the statistics that were calculated are given below for the examples shown in Exhibits 8.1 and 8.2. The statistics were calculated separately for advanced mathematics and physics. The statistics are listed in order of their appearance in the item analysis outputs:

N: 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 item analysis.[4]

Diff: The item difficulty is the average percent correct. For a 1-point item, including 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; i.e., if all students scored 1 point on a 2-point item, then the average percent correct for such an item would be 50 percent. For this statistic, not-reached items were treated as not administered.

4 For item review and scaling, items not reached by a student were treated as if they had not been administered. For estimating student proficiency, however, not reached items were treated as incorrect.

Disc: The item discrimination was computed as the correlation between a correct response to the item and the overall score on all of the advanced mathematics or physics items administered to a student.[5] 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.

PCT_A, PCT_B, PCT_C, PCT_D, and PCT_E: Used for multiple-choice items only (see Exhibit 8.1). Each column indicates the percentage of students choosing the particular response option for the item (A, B, C, D, or E). Not-reached items were excluded from the denominator.

PCT_0, PCT_1, and PCT_2: Used for constructed-response items only (see Exhibit 8.2). 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.

PCT_OM: Percentage of students who, having reached the item, did not provide a response. No reached items were excluded from the denominator.

PCT_NR: Percentage of students who did not reach the item. 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 also omitted.

PB_A, PB_B, PB_C, PB_D, and PB_E: Used for multiple-choice items only. These columns show the point-biserial correlations between choosing each of the response options (A, B, C, D, or E) and the overall score on all of the advanced mathematics or physics items administered to a student. Items with good

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

psychometric properties have moderately positive correlations for the correct option and negative correlations for the distractors (the incorrect options).

PB_0, PB_1, and PB_2: Used for constructed-response items only. 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 mathematics or science items the student was administered. For items with good measurement properties, the correlation coefficients should increase from negative to positive as the score on the item increases.

PB_OM: The point-biserial correlation between a binary variable, indicating an omitted response to the item, and the overall score on all of the mathematics or physics 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 the achievement data for a given country. The difficulty estimate is expressed in the logit metric (with a positive logit indicating a difficult item) and was scaled so that the average Rasch item difficulty across all items within each country was zero.

Reliability (Cases): To provide a measure of the reliability of the scoring of 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 codes assigned by the two independent scorers.

Reliability (Code): This column contains the percentage of exact agreement on the two-digit codes assigned by the two independent scorers.

As an aid to the reviewers, the item-analysis 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:

- Item discrimination was less than 0.15 (flag D).
- Item difficulty was less than 25 percent for four-option multiple-choice items, or less than 20 percent for five-option multiple choice items (flag C).
- Item difficulty exceeded 95 percent (flag V).
- The Rasch difficulty estimate for a given country made the item either easier (flag E) or harder (flag H) relative to the international average Rasch difficulty of that item.
- The point-biserial correlation for at least one distractor in a multiple choice item was positive, or the estimated mean abilities 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 of 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 80 percent (flag R).

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

In order to measure trends, TIMSS Advanced 2008 included items from the 1995 assessments.[6] For these trend items, the review included an examination of changes in item statistics between the 1995 and 2008 administrations. An example is shown in Exhibit 8.3. The information in this exhibit is different from that presented in Exhibits 8.1 and 8.2, and includes countries' statistics from both the 1995 and 2008 assessments. In reviewing these item statistics, the aim was to detect any unusual changes in item properties between assessments that might indicate a problem in using the item to measure trends.

Exhibit 8.3 International Item Statistics for a Trend Item

Trends in International Mathematics and Science Study - TIMSS Advanced 2008 Assessment Results
Percent of Responses by Item Category (Mathematics) - Trend Items - Final Year of Secondary School

Mathematics: Geometry / Applying (MA13026A - M3_06A)
Label: Triangle abc/reflection
Type: CR Key: X

COUNTRY	Year	N	10	70	71	79	99	INVALID	OMIT	NOT REACHED	V1	1.GIRL % Right	2.BOY % Right
Italy	1995	126	65.8	4.3	5.1	11.8	13.1	1.5	5.0	6.6	65.8	65.8	65.7
	2008	1070	54.9	2.5	6.7	21.2	14.8	0.0	8.8	6.0	54.9	56.5	54.0
Russian Federation	1995	468	75.4	3.8	5.3	9.2	6.3	0.0	1.2	5.1	75.4	74.4	76.3
	2008	1588	75.0	2.1	6.2	10.7	5.9	0.0	4.4	1.6	75.0	74.5	75.5
Slovenia	1995	452	69.2	1.4	0.3	21.9	7.3	2.0	2.2	3.1	69.2	66.7	71.6
	2008	1083	69.7	7.6	2.7	16.9	3.1	0.0	0.9	2.2	69.7	68.7	71.2
Sweden	1995	244	39.5	10.0	1.6	25.4	23.5	0.4	13.6	9.5	39.5	25.4	47.1
	2008	1148	18.8	8.3	5.2	34.5	33.2	0.0	26.0	7.2	18.8	17.9	19.4
International Avg.	1995	323	62.4	4.8	3.1	17.1	12.6	1.0	5.5	6.1	62.4	58.1	65.2
	2008	1222	54.6	5.1	5.2	20.8	14.3	0.0	10.0	4.2	54.6	54.4	55.0

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

6 Information on trend items is available in Chapter 2.

8.2.2 Item-by-Country Interaction

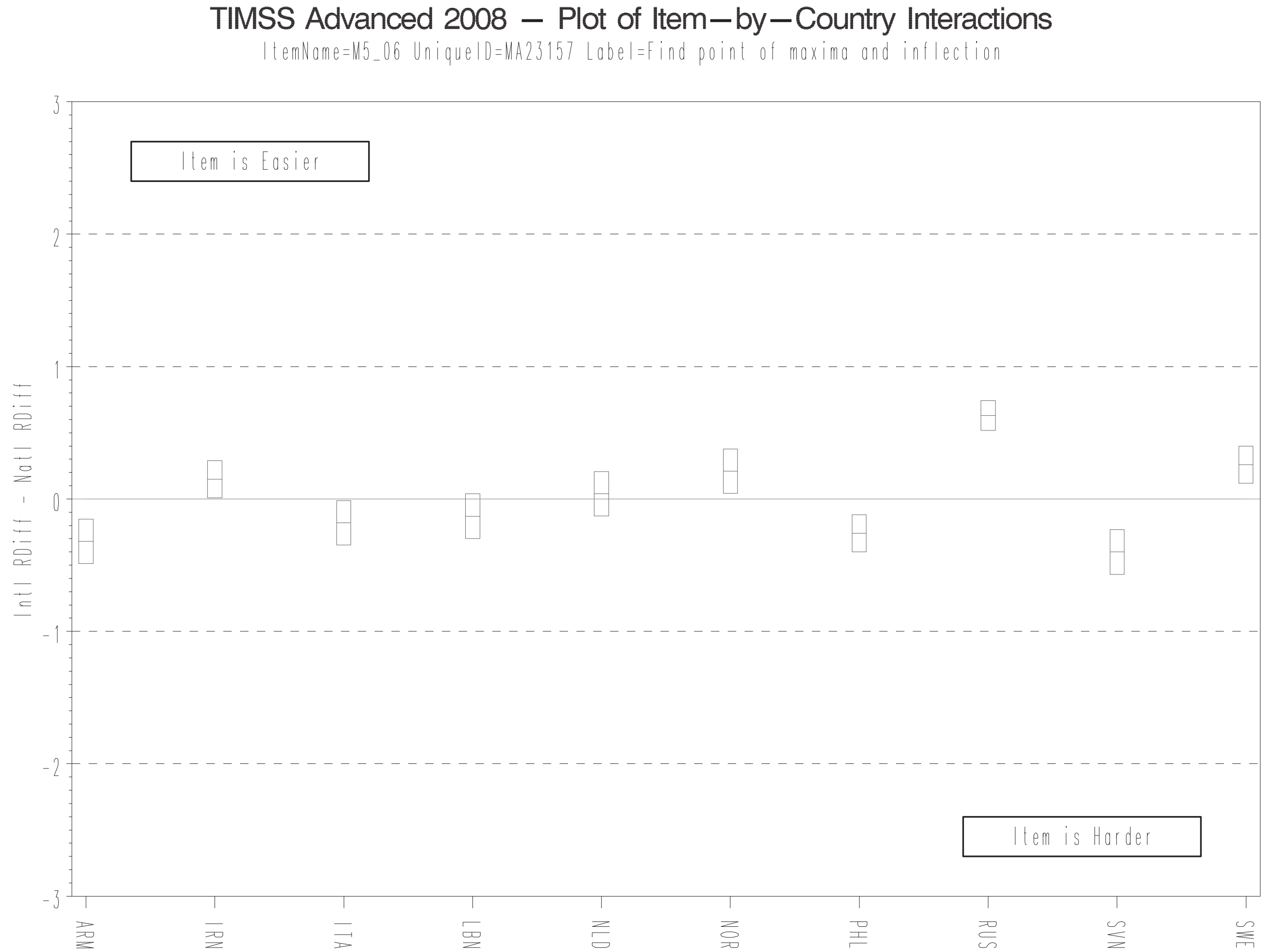
Although countries are expected to exhibit some variation in performance across items, in general countries with high average performance on the assessment should perform relatively well on all of the items, and low-scoring countries should do less well on all of items. When this does not occur (i.e., when a high-performing country has a 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 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 average Rasch item difficulty across all countries. Exhibit 8.4 provides an example of a TIMSS Advanced 2008 item-by-country interaction display. 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. 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 $RDIFF_{ik}$ was the Rasch difficulty of item i in country k, $RDIFF_{i}.$ was the average difficulty of item i across all countries, SE($RDIFF_{ik}$) was the standard error of the Rasch difficulty of item i in country k , and Z_b was the critical value from the Z distribution corrected for multiple comparisons using the Bonferroni procedure.

Exhibit 8.4 Sample Plot of Item-by-Country Interaction for a TIMSS Advanced 2008 Item



8.2.3 Trend Item Analysis

Because an important part of the TIMSS Advanced 2008 assessment was measuring trends across the 1995 and 2008 assessment cycles, an additional review step ensured that the trend items had similar characteristics in both cycles (i.e., an item that was relatively easy in 1995 should have been relatively easy in 2008). The comparison between cycles was made in a number of ways. For each trend country, almanacs of item statistics displayed the percentage of students within each score category (or response option for multiple-choice items) for each cycle, as well as the difficulty of the item and the percent correct by gender. While some changes were anticipated as countries' overall achievement may have improved or declined, items were noted if the difference between the Rasch difficulties across the two cycles for a particular country was greater than 2 logits.

The TIMSS & PIRLS International Study Center used two different graphical displays to examine the differences item difficulties. The first of these, shown in Exhibit 8.5, displays the difference in Rasch item difficulties between 1995 and 2008. A positive difference indicated that an item was relatively easier in a country in 2008, and a negative difference indicated that an item was relatively more difficult. The second, Exhibit 8.6, shows the performance of a given country on all trend items simultaneously. For each country, the graph plotted the 1995 Rasch difficulty of every trend item in 1995 against its 2008 Rasch difficulty. Where there were no differences between the difficulties in 1995 and 2008, the data points aligned on or near the diagonal.

Exhibit 8.5 Sample Plot of Difference in Rasch Item Difficulties for a Trend Item

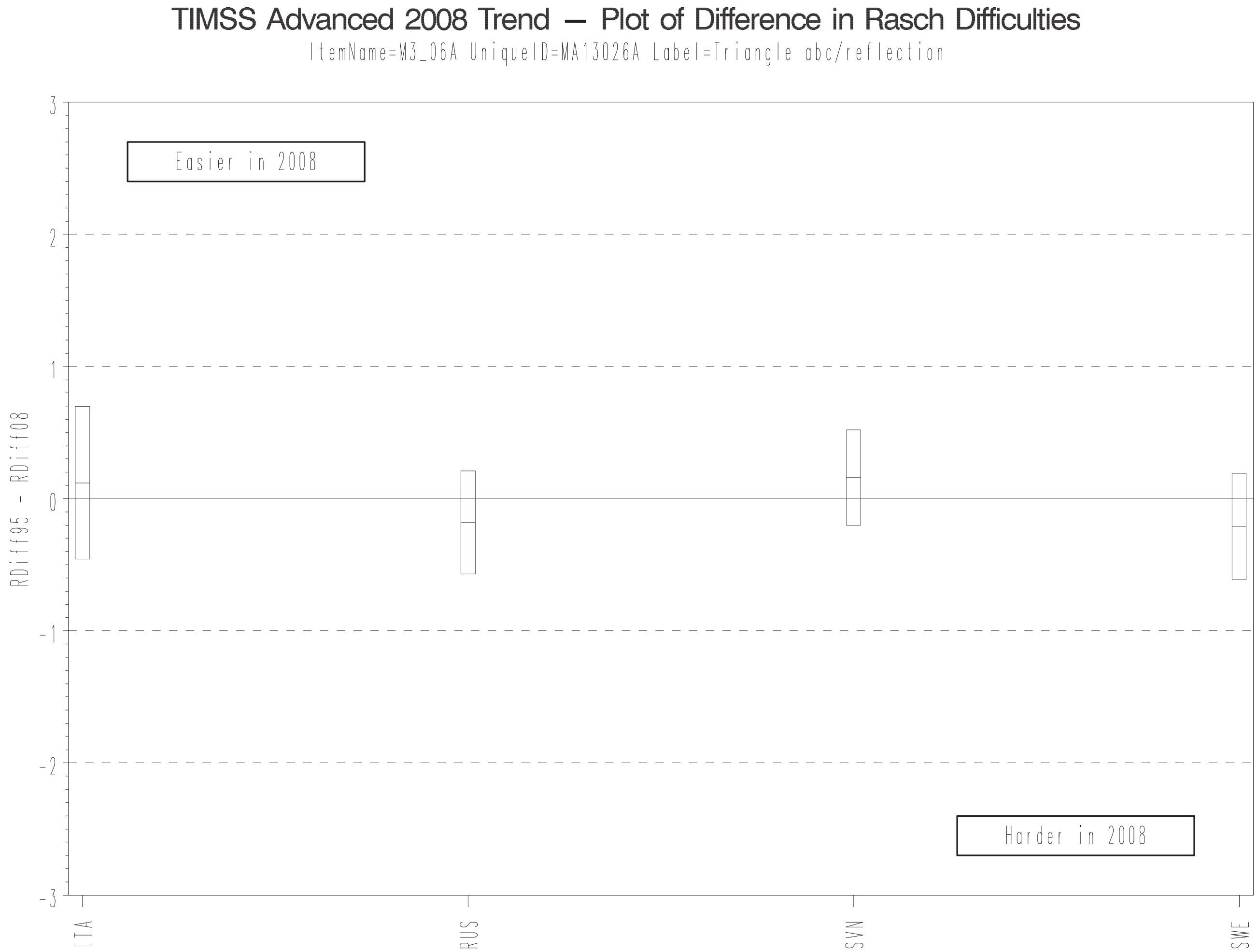
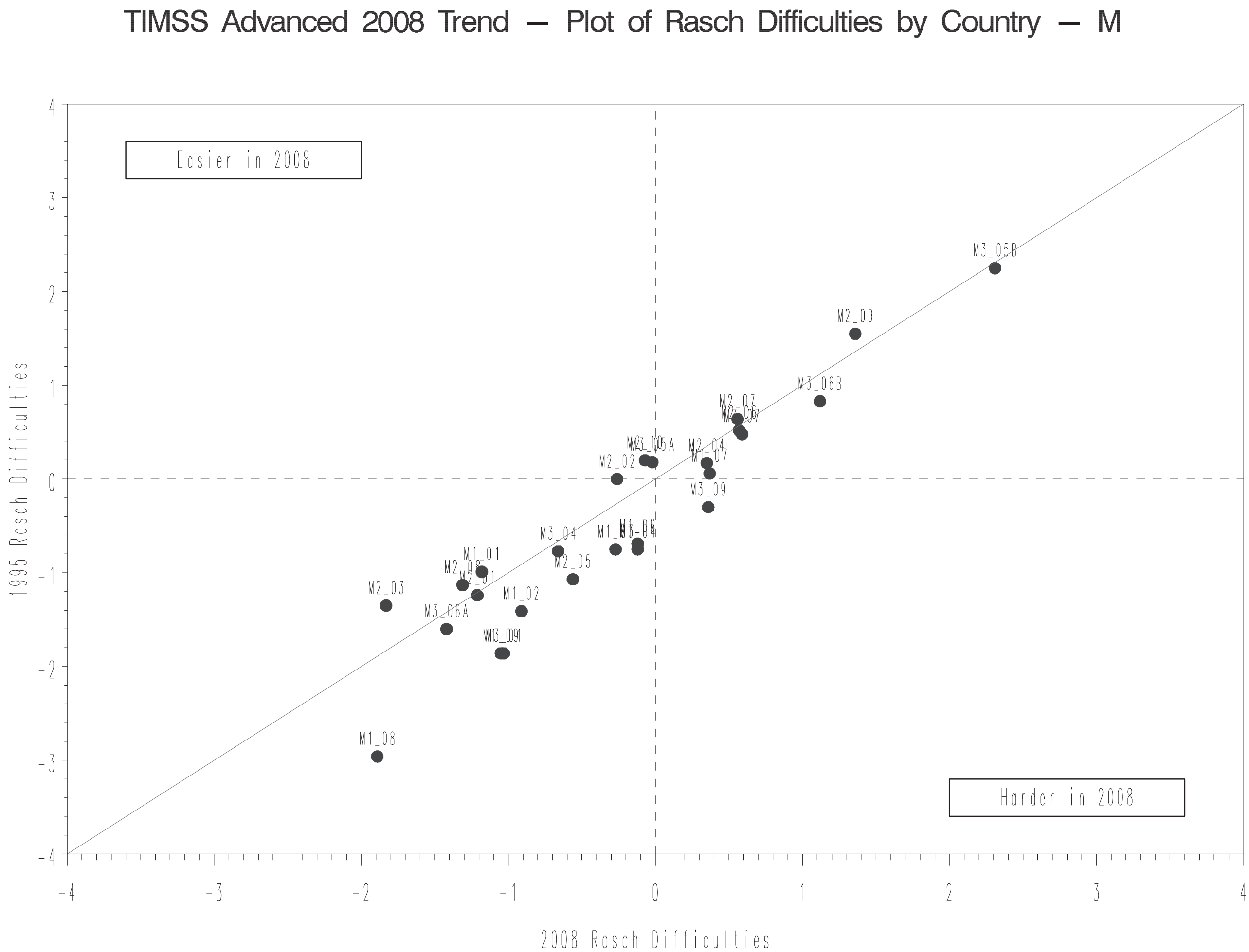


Exhibit 8.6 Sample Plot of Rasch Item Difficulties across Trend Items by Country



8.2.4 Reliability

Gauging the reliability of the TIMSS Advanced 2008 assessments was a critical quality control step in reviewing the items. There were two aspects of reliability under review. The set of items selected as part of the advanced mathematics and physics assessments needed to constitute a cohesive whole measuring their respective domains, a quality known as test reliability. Also, the scoring of the constructed-response items had to meet specific reliability criteria in terms of consistent within-country and cross-country scoring.

8.2.4.1 Test Reliability

Exhibit 8.7 displays the advanced mathematics and physics test reliability coefficients for every country. These coefficients are the median Cronbach's alpha reliability across the four test booklets of advanced mathematics and physics. In general, median reliabilities were relatively high in both subjects with the international median at 0.82 for advanced mathematics and 0.80 for physics. All median reliabilities were at least 0.70, except for physics in Lebanon, where the median reliability was 0.68.

Exhibit 8.7 Cronbach's Alpha Reliability Coefficients for TIMSS Advanced 2008

Country	Reliability Coefficient	
	Advanced Mathematics	Physics
Armenia	0.87	0.82
Iran, Islamic Rep. of	0.90	0.85
Italy	0.84	0.75
Lebanon	0.80	0.68
Netherlands	0.70	0.74
Norway	0.78	0.79
Philippines	0.79	—
Russian Federation	0.88	0.88
Slovenia	0.83	0.81
Sweden	0.80	0.80
International Median	0.82	0.80

8.2.4.2 Scoring Reliability for Constructed-response Items

About one third of the items in the TIMSS Advanced 2008 assessment were constructed-response items, comprising nearly half of the score points for the assessment.[7] 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 Advanced 2008, measures taken to ensure that the constructed-response items were scored reliably in all countries, and these measures included developing scoring guides for each constructed-response question (that provided descriptions of acceptable responses for each score point value)[8] as well as providing extensive training in the application of the scoring guides.

Within-Country Scoring Reliability

To gather and document information about the within-country agreement among scorers, 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. The average and range of the within-country percentage agreement across all items for both grades are presented in Exhibit 8.8 for both advanced mathematics and physics.

Scoring reliability was high on average across countries. The percent agreement on the correctness score across all countries was 98 percent in advanced mathematics and 97 percent in physics. All countries had an average percent agreement on the correctness score above 94 percent in advanced mathematics and above 91 percent in physics.

7 The development of the TIMSS Advanced 2008 assessment items is described in Chapter 2.

8 A discussion of the development of the scoring guides for constructed-response items is provided in Chapter 2.

Exhibit 8.8 Within-country Scoring Reliability for TIMSS Advanced 2008 Constructed-Response Items

Advanced Mathematics

Countries	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Percent Agreement Across Items	Range of Percent Agreement		Average of Percent Agreement Across Items	Range of Percent Agreement	
		Min	Max		Min	Max
Armenia	100	98	100	97	86	100
Iran, Islamic Rep. of	98	90	100	95	89	100
Italy	100	97	100	98	94	100
Lebanon	100	99	100	99	97	100
Netherlands	94	72	100	91	65	99
Norway	99	98	100	98	95	100
Philippines	98	93	100	95	85	100
Russian Federation	97	86	100	95	86	100
Slovenia	100	99	100	99	97	100
Sweden	97	88	100	93	83	99
International Average	98	92	100	96	88	100

Physics

Countries	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Percent Agreement Across Items	Range of Percent Agreement		Average of Percent Agreement Across Items	Range of Percent Agreement	
		Min	Max		Min	Max
Armenia	99	93	100	97	93	100
Iran, Islamic Rep. of	96	91	100	90	71	99
Italy	99	94	100	97	86	100
Lebanon	99	93	100	98	92	100
Netherlands	91	80	99	85	71	97
Norway	97	90	100	94	87	100
Russian Federation	96	89	100	93	83	99
Slovenia	100	98	100	99	95	100
Sweden	97	89	99	93	83	99
International Average	97	91	100	94	85	99

Cross-Country Scoring Reliability

Because of the different languages used by the countries participating in TIMSS Advanced 2008, establishing the reliability of constructed-response scoring across all countries was not feasible. However, TIMSS Advanced 2008 did conduct a cross-country study of scoring using English as a common language. A sample of student responses from a pilot study carried out in English was provided to countries. It included 100 student responses to each of nine advanced mathematics items and nine physics items. This set of 1,800 student responses in English was then scored independently in each country that had two scorers proficient in English. In all, 14 scorers from 7 countries participated in the study. Scoring for this study took place shortly after the other scoring reliability activities were completed. Making all possible pair-wise comparisons among scorers gave 91 comparisons for each student response to each item. This resulted in 9,100 total comparisons when aggregated across all 100 student responses to an item. Agreement across countries was defined in terms of the percentage of these comparisons that were in agreement.

Exhibit 8.9 shows that scorer reliability across countries was high for advanced mathematics, with the percent agreement averaging 94 percent across the nine items for the correctness score and 90 percent for the diagnostic score. For physics, the percent agreement averaged 88 percent across the nine items for the correctness score and 80 percent for the diagnostic score.

Exhibit 8.9 Cross-Country Scoring Reliability for TIMSS Advanced 2008

Advanced Mathematics

Item Label	Total Valid Comparisons	Percent Agreement	
		Correctness Score Agreement	Diagnostic Score Agreement
M4_04 - MA23201	9100	88	88
M4_08 - MA23043	9100	95	90
M5_03 - MA23054	9100	94	92
M5_05 - MA23131A	9100	99	99
M5_05 - MA23131B	9100	98	97
M5_10 - MA23094	9100	89	81
M6_07 - MA23198	9100	94	92
M7_03 - MA23141	9100	98	83
M7_11 - MA23170	9100	90	84
Average Percent Agreement		94	90

Physics

Item Label	Total Valid Comparisons	Percent Agreement	
		Correctness Score Agreement	Diagnostic Score Agreement
P4_07 - PA23053	9100	86	82
P4_09 - PA23119	9100	80	67
P4_11 - PA23066	9100	83	81
P5_03 - PA23035	9100	95	85
P5_05 - PA23012	9100	88	74
P5_07 - PA23051	9100	94	87
P6_05 - PA23022	9100	83	76
P7_05 - PA23034	9100	97	95
P7_06 - PA23044	9100	90	70
Average Percent Agreement		88	80

8.2.5 Summary of TIMSS Advanced 2008 Item Statistics Review

Based on the information from the comprehensive collection of item analyses and reliability data that were computed and summarized for TIMSS Advanced 2008, the TIMSS & PIRLS International Study Center thoroughly reviewed all item statistics for every participating country to ensure that the items were performing comparably across countries. Specifically, items with the following problems were considered for possible deletion from the international database:

- An error was detected during the TIMSS Advanced 2008 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 an item to have a negative point-biserial, or, for an item with more than 1 score point, a non-monotonic relationship between score level and total score.
- The item-by-country interaction results showed a large negative interaction for a particular country.
- For constructed-response items, the within-country scoring reliability data showed a score agreement of less than 70 percent.
- For trend items, an item performed substantially differently in 2008 compared to 1995, or an item was not included in the 1995 assessment for a particular country.

When the item statistics indicated a problem with an item, the documentation from the translation verification[9] 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

9 Chapter 3 describes the process of translation verification applied to the TIMSS Advanced 2008 instruments.

treated. If a problem was detected by the TIMSS & PIRLS 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.

The checking of the TIMSS Advanced 2008 achievement data involved a review of 143 items for 10 countries and resulted in the detection of very few items that were inappropriate for international comparisons. The few items singled out in the review process were mostly items with differences attributable to either translation or printing problems. The following is a list of deleted items as well as a list of recodings made to constructed-response items.

Advanced Mathematics

Items deleted

ALL COUNTRIES

M2_04 (MA13014) – attractive distractor

LEBANON

M5_08 (MA23082) – printing error

Constructed-response items needing category recoding

ALL COUNTRIES

M3_06A (MA13026A) – recode 11 to 71
M3_06B (MA13026B) – recode 11 to 72
M3_08 (MA13028) – recode 20 to 10, 10 to 70, 11 to 71, 12 to 72
M4_07 (MA23166) – recode 20 to 10, 21 to 11, 10 to 70, 11 to 71, 70 to 72

Physics

Items deleted

ALL COUNTRIES

P1_07 (PA13007) – attractive distractor
P2_10 (PA13020) – low discrimination
P3_07B (PA13027B) – percent omitted too high

SWEDEN

P4_04 (PA23104) – negative discrimination

Constructed-response items needing category recoding

ALL COUNTRIES

P3_03 (PA13023) – recode 11 to 72
P3_05 (PA13025) – recode 29 to 19
P3_06 (PA13026) – recode 20 to 10, 29 to 19, 10 to 72, 19 to 79
P3_07A (PA13027A) – recode 20 to 10, 21 to 11, 22 to 19, 29 to 19, 10 to 70, 11 to 71
P4_10 (PA23088) – recode 11 to 71
P6_04 (PA23072) – recode 20 to 10, 10 to 11
P6_07 (PA23078) – recode 20 to 10, 10 to 11, 11 to 12

8.3 The TIMSS Advanced 2008 Scaling Methodology[10]

The IRT scaling approach used by TIMSS was developed originally by Educational Testing Service for use in the U.S. National Assessment of Educational Progress. It is based on psychometric models that were first used in the field of educational measurement in the 1950s and have become popular since the 1970s for use in large-scale surveys, test construction, and computer adaptive testing.

10 A more detailed description of the TIMSS scaling methodology is given in Chapter 11 of the *TIMSS 2007 Technical Report* (Foy, Galia, & Li, 2008).

Three distinct IRT models, depending on item type and scoring procedure, were used in the analysis of the TIMSS Advanced assessment data. Each is a “latent variable” model that describes the probability that a student will respond in a specific way to an item in terms of the student’s proficiency, which is an unobserved, or “latent,” trait, and various characteristics (or “parameters”) of the item. A three-parameter model was used with multiple-choice items, that were scored as correct or incorrect; and a two-parameter model, for constructed-response items with two response options, that also were scored as correct or incorrect. Since each of these item types has two response categories, they are known as dichotomous items. A partial credit model was used with polytomous constructed-response items, i.e., those with more than two response options.

8.3.1 Proficiency Estimation Using Plausible Values

Most cognitive testing endeavors to assess the performance of individual students for the purposes of diagnosis, selection, or placement. Regardless of the measurement model used, whether classical test theory or item response theory, the accuracy of these measurements can be improved—that is, the amount of measurement error can be reduced—by increasing the number of items given to the individual. Thus, it is common to see achievement tests designed to provide information on individual students that contain more than 70 items. Since the uncertainty associated with estimates of individual student ability is negligible under these conditions, the distribution of student ability, or its joint distribution with other variables, can be approximated using each individual student’s estimated ability.

For the distribution of proficiencies in large populations, more efficient estimates can be obtained from a matrix-sampling design such as that used in TIMSS Advanced. This design solicits relatively few responses from each sampled student while maintaining a wide range of content representation when responses are aggregated across all students. With this approach, the advantage of estimating population characteristics more efficiently is offset to some degree by the inability to make precise statements about individuals. The uncertainty associated with individual student ability estimates becomes too large to be ignored.

Plausible values methodology was developed as a way to address this issue. Instead of first computing estimates of individual student abilities and then aggregating these to estimate population parameters, the plausible values approach uses all available data—students' responses to the items they were administered together with all background data—to estimate directly the characteristics of student populations and subpopulations. Although these directly estimated population characteristics could be used for reporting purposes, the usual plausible values approach generates multiple imputed scores, called plausible values, from the estimated ability distributions and uses these in analyses and reporting, making use of standard statistical software. By including all the available background data in the model, a process known as "conditioning," relationships between these background variables and the estimated proficiencies are appropriately accounted for in the plausible values. Because of this, analyses conducted using plausible values provide an accurate representation of these underlying relationships.

Plausible values are not intended to be estimates of individual student scores, but rather are imputed scores for similar students—students with similar response patterns and background characteristics in the sampled population—that may be used to estimate population characteristics correctly. When the underlying model is correctly specified, plausible values provide consistent estimates of population characteristics, even though they are not generally unbiased estimates of the proficiencies of the individuals with whom they are associated. Taking the average of the plausible values does not yield suitable estimates of individual student scores.[11]

8.4 Implementing the Scaling Procedures for the TIMSS Advanced Assessment Data

The application of IRT scaling and plausible values methodology to the data from the TIMSS Advanced 2008 assessments involved four major tasks: calibrating the achievement test items (estimating model parameters for each item), creating principal components from the student questionnaire data for use in conditioning, generating proficiency scores for advanced mathematics and for physics, and placing these proficiency scores on the scales—one for advanced mathematics and one for physics—used to report the results from the TIMSS Advanced assessments in 1995.

Before scaling the 2008 assessment data, however, the data from the 1995 assessments had to be rescaled from the one-parameter Rasch model used in 1995 to the multi-parameter models that have been in use in TIMSS since 1999.

8.4.1 Rescaling the Data from the TIMSS Advanced 1995 Assessments

The students' responses to the achievement items and to the questions in the student background questionnaire from the TIMSS Advanced 1995 international database provided the data for rescaling the TIMSS Advanced 1995 data. The TIMSS Advanced 1995 assessments included 68 items for advanced mathematics and 66 items for physics. These items were classified into the content and cognitive domains defined in the *TIMSS Advanced 2008 Assessment Frameworks* (Garden, et al., 2006) in preparation for trend scaling. Of the 134 items, 10 advanced

11 For further discussion, see Mislevy, Beaton, Kaplan, and Sheehan (1992) and von Davier, Gonzalez, and Mislevy (2009).

mathematics items and 4 physics items did not fit any framework classification and thus were omitted from the TIMSS Advanced 1995 rescaling since they were no longer appropriate for the domains specified in the 2008 frameworks. Also, one advanced mathematics item and three physics items were omitted for the reasons given in Section 8.2.5. Finally, one physics item released after the 1995 assessment (PA13052), was omitted because of poor discrimination. Some trend items that required recoding in the 2008 assessments also were recoded in the 1995 database.

All countries that participated in TIMSS Advanced 1995 were included in the item calibrations. Exhibit 8.10 presents the sample sizes for the countries included in the TIMSS Advanced 1995 item calibrations.[12]

Exhibit 8.10 Sample Sizes for Item Calibrations of the TIMSS Advanced 1995 Assessments

Country	Advanced Mathematics	Physics
Australia	645	661
Austria	782	777
Canada	2,781	2,367
Cyprus	391	368
Czech Republic	1,101	1,087
France	1,071	1,110
Germany	2,296	723
Greece	456	459
Italy	398	—
Latvia	—	708
Lithuania	734	—
Norway	—	1,048
Russian Federation	1,638	1,233
Slovenia	1,536	747
Sweden	1,001	1,012
Switzerland	1,404	1,371
United States	2,785	3,114
Total	**19,019**	**16,785**

12 Because Denmark and Israel failed to satisfy the 1995 sampling guidelines, they were not included in the item calibrations for the rescaling, as was also the case for the original scaling.

The item calibrations were conducted separately for each subject by the TIMSS & PIRLS International Study Center using the commercially-available Parscale software (Muraki & Bock, 1991; version 4.1). The two- and three-parameter and polytomous IRT models were fitted to the data to produce item parameter estimates. These new estimated values can be found in Exhibits D.1 and D.2 of Appendix D. These parameter estimates then became part of the input for producing proficiency scores.

A principal components analysis was run separately within each country to generate input for the conditioning step. The estimated proficiency scores are conditioned on the student background variables to improve the reliability of sub-population reporting. Principal components analysis is used to reduce the number of conditioning variables to a manageable size. The usual TIMSS approach retains the number of principal components that account for at least 90 percent of the variability in the student background data. Since most countries in 1995 had small sample sizes, the 90 percent criterion was reduced to 70 percent to minimize over-specification in the conditioning model, provided the number of components retained did not exceed 10 percent of the sample size—in which case the number of components was limited to 10 percent of the sample size. Exhibit 8.11 displays the total number of variables considered for conditioning and the number of principal components selected for each country.

The generation of IRT proficiency scores was conducted separately for each country and for each subject using Educational Testing Service's MGROUP program (Sheehan, 1985; version 3.2).[13] MGROUP takes as input the students' responses to the items they were given, the item parameters estimated at the calibration stage, and the conditioning variables, and generates as output the plausible values

13 The MGROUP program was provided by ETS under contract to the TIMSS & PIRLS International Study Center at Boston College. It is now commercially available as DESI.

Exhibit 8.11 Number of Variables and Principal Components for Conditioning in Rescaling the TIMSS Advanced 1995 Assessments

Countries	Advanced Mathematics			Physics		
	Number of Primary Conditioning Variables	Total Number of Principal Components	Number of Principal Components Retained	Number of Primary Conditioning Variables	Total Number of Principal Components	Number of Principal Components Retained
Australia	2	553	64	2	560	66
Austria	2	542	78	2	545	77
Canada	3	573	148	3	570	134
Cyprus	2	562	39	2	565	36
Czech Republic	2	565	110	2	549	108
Denmark	2	545	123	2	533	65
France	2	434	107	2	435	111
Germany	2	542	147	2	526	72
Greece	2	551	45	2	550	45
Israel	2	589	95	2	583	85
Italy	2	531	39	—	—	—
Latvia	—	—	—	2	579	70
Lithuania	2	542	73	—	—	—
Norway	—	—	—	3	579	104
Russian Federation	2	602	152	2	598	123
Slovenia	2	582	141	2	573	74
Sweden	2	576	100	2	571	101
Switzerland	4	544	127	4	546	125
United States	2	612	154	2	618	166

that represent student proficiency. Exhibit 8.12 shows the sample sizes of the countries for which proficiency scores were generated.[14]

The reporting metrics for the rescaled 1995 data were established to give the distribution of TIMSS Advanced 1995 proficiency scores in advanced mathematics and in physics a mean of 500 and a standard deviation of 100, with all 1995 countries included in the item calibrations contributing equally. Extreme scale values were truncated, giving plausible values a minimum of 5 and a maximum of 995.

14 Denmark and Israel, which had been excluded from the item calibrations, were included among the countries for which proficiency scores were produced.

Exhibit 8.12 Sample Sizes for Proficiency Estimation of the TIMSS Advanced 1995 Assessments

Country	Advanced Mathematics	Physics
Australia	645	661
Austria	782	777
Canada	2,781	2,367
Cyprus	391	368
Czech Republic	1,101	1,087
Denmark	1,388	654
France	1,071	1,110
Germany	2,296	723
Greece	456	459
Israel	953	853
Italy	398	—
Latvia	—	708
Lithuania	734	—
Norway	—	1,048
Russian Federation	1,638	1,233
Slovenia	1,536	747
Sweden	1,001	1,012
Switzerland	1,404	1,371
United States	2,785	3,114
Total	**21,360**	**18,292**

8.4.2 Calibrating the TIMSS Advanced 2008 Assessment Data

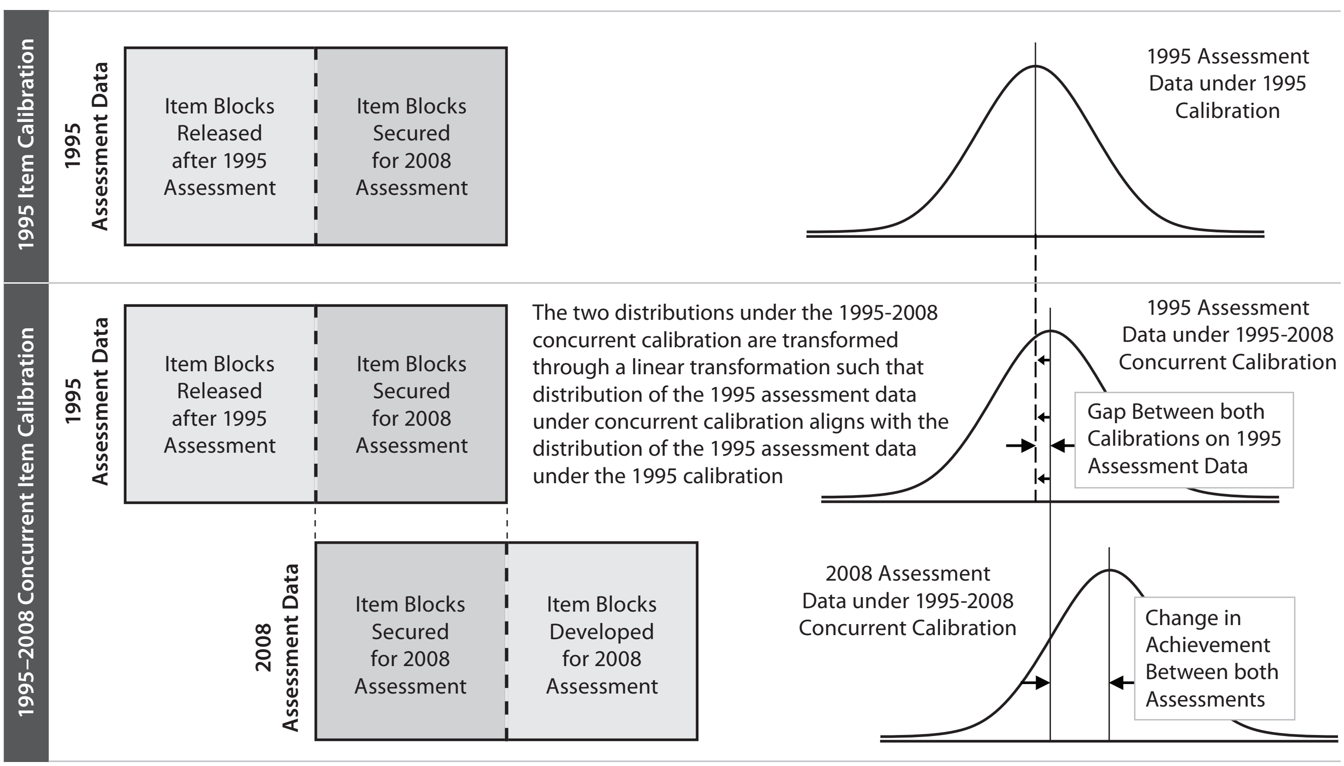
As described in the *TIMSS Advanced 2008 Assessment Frameworks* (Garden, et al., 2006), the TIMSS Advanced 2008 assessments consisted of a total of seven advanced mathematics blocks and seven physics blocks, distributed across eight assessment booklets. Each block contained either advanced mathematics or physics items, drawn from a range of content and cognitive domains. The seven mathematics blocks were designated M1 through M7, and the seven physics blocks P1 through P7. Blocks M1 through M3 and P1 through P3 contained items that were previously used in the 1995 assessments, whereas blocks M4 through M7 and P4 through P7 consisted of newly-developed items for the 2008 assessments. Each assessment booklet contained three blocks of either all advanced mathematics items or all physics items. The booklets were distributed among the students in each sampled class according to a scheme that ensured equivalent random samples of students responding to each booklet.

Separate IRT scales were constructed for reporting overall student achievement in advanced mathematics and in physics. Concurrent item calibrations were conducted by the TIMSS & PIRLS International Study Center using Parscale, and included data from the TIMSS Advanced 2008 assessments and the TIMSS Advanced 1995 assessments to measure trends from 1995. The calibrations used all available data from each country's TIMSS Advanced student samples, which were weighted such that each country contributed equally.

The first step in constructing the scales for TIMSS Advanced 2008 was to estimate the IRT model item parameters for each item on each of the scales through a concurrent calibration of both sets of assessment data—1995 and 2008. It was then possible to obtain the mean and standard deviation of the latent ability distributions of students in both assessments using item paramters from the concurrent calibration. The difference between these two distributions was the change in achievement from 1995 to 2008. The second step was to find the linear transformation that transformed the distribution of the 1995 assessment data under the 1995-2008 concurrent calibration to match the distribution of these data under the 1995 calibration. The third step was to apply this same transformation to the 2008 assessment data scaled using the concurrent calibration. This placed the 2008 assessment data on the metric of the 1995 assessment—i.e., a scale with a mean of 500 and a standard deviation of 100.

Exhibit 8.13 illustrates how the concurrent calibration approach was applied in the context of TIMSS Advanced 2008 trend scaling. The observed gap between the distribution of the 1995 data under the 1995 item calibration and the 1995 data under the 1995-2008 concurrent calibration was small and arose from slight differences in the item parameter estimations, which in turn were due mostly to the 1995 assessment data being calibrated with the 2008 assessment data. The linear transformation removed this gap by shifting the two distributions from the concurrent calibration such that the distribution of the 1995 assessment data from the concurrent calibration aligned with the distribution of the 1995 assessment data from the 1995 calibration, while preserving the gap between the 1995 and 2008 assessment data under the concurrent calibration. This latter gap was the change in achievement between the previous and current assessments that TIMSS Advanced set out to measure as its trend.

Exhibit 8.13 Concurrent Calibration Model Used for TIMSS Advanced 2008



Having estimated the item parameters from the 1995-2008 concurrent calibration, new achievement distributions were generated by applying these item parameters to the 1995 assessment and to the 2008 assessment data. Following the procedure outlined above, the next step was to identify the linear transformation that transformed the 1995 assessment distribution generated by the 1995-2008 concurrent calibration item parameters to match the 1995 assessment distribution generated by the item parameters from the 1995 rescaling, and to apply this same transformation to the 2008 assessment data distribution (also generated by the concurrent calibration item parameters).

Exhibit 8.14 shows the distribution of items included in the TIMSS Advanced 2008 concurrent calibrations for reporting trends in overall advanced mathematics and physics. All the data from both the 1995 and 2008 assessments were included. Items were categorized as unique to the 1995 assessment, common to both assessments, or unique to the 2008 assessment. For advanced mathematics, the 2008 assessment contributed 45 items worth 51 score points that were unique to 2008 and 26 items worth 28 score points that also were included in the 1995 assessment. The 1995 assessment also contributed 31 items worth 40 score points that were released in 1995. For physics, the 2008 assessment contributed 45 items worth 51 score points that were unique to 2008 and 23 items worth 26 score points that also were included in the 1995 assessment. The 1995 assessment also contributed 35 items worth 42 score points that were released in 1995.

Exhibit 8.14 Items Included in the TIMSS Advanced 2008 Concurrent Item Calibrations

TIMSS 2008 Trend Scales	Items Unique to the TIMSS Advanced 2008 Assessments		Items Common to the TIMSS Advanced 2008 and 1995 Assessments		Items Unique to the TIMSS Advanced 1995 Assessments		TOTAL	
	Number	Points	Number	Points	Number	Points	Number	Points
Advanced Mathematics	45	51	26	28	31	40	102	119
Physics	45	51	23	26	35	42	103	119

Because of the small number of countries that participated in both TIMSS Advanced assessments, concurrent item calibrations were conducted using data from all the countries that participated in either the 1995 assessments or the 2008 assessments. To construct the advanced mathematics scale, the calibration included 19,019 students from 15 countries in the 1995 assessment and 22,242 students from 10 countries in the 2008 assessment. The item parameters established in this calibration were used subsequently for estimating student proficiency scores in advanced mathematics for the 10 countries that participated in TIMSS Advanced 2008. The national samples included in the calibration for reporting trends in advanced mathematics are presented in Exhibit 8.15.

Similarly, to construct the physics scale, the item calibration was conducted using data from countries that participated in either the 1995 assessment or the 2008 assessment. The physics concurrent calibration included 16,785 students from 15 countries in the 1995 assessment and 16,489 students from 9 countries in the 2008 assessment. The item parameters obtained in this calibration were used subsequently for estimating student proficiency scores in physics for the nine countries that participated in TIMSS Advanced 2008. Exhibit 8.16 presents the national samples included in the calibration for reporting trends in Physics.

Exhibits D.3 and D.4 of Appendix D display the item parameters for advanced mathematics and physics, respectively, generated from the concurrent calibration of the 1995 and the 2008 data. As a by-product of the calibrations, interim scores in advanced mathematics and physics were produced for use in constructing conditioning variables.

Exhibit 8.15 Sample Sizes for Concurrent Item Calibration of Advanced Mathematics for the TIMSS Advanced 1995 and 2008 Assessments

Country	1995 Assessment	2008 Assessment
Countries in Both Cycles		
Italy	398	2,143
Russian Federation	1,638	3,185
Slovenia	1,536	2,156
Sweden	1,001	2,303
Countries in 1995		
Australia	645	—
Austria	782	—
Canada	2,781	—
Cyprus	391	—
Czech Republic	1,101	—
France	1,071	—
Germany	2,296	—
Greece	456	—
Lithuania	734	—
Switzerland	1,404	—
United States	2,785	—
Countries in 2008		
Armenia	—	858
Iran, Islamic Rep. of	—	2,425
Lebanon	—	1,612
Netherlands	—	1,537
Norway	—	1,932
Philippines	—	4,091
Total	**19,019**	**22,242**

Exhibit 8.16 Sample Sizes for Concurrent Item Calibration of Physics for the TIMSS Advanced 1995 and 2008 Assessments

Country	1995 Assessment	2008 Assessment
Countries in Both Cycles		
Norway	1,048	1,640
Russian Federation	1,233	3,166
Slovenia	747	1,097
Sweden	1,012	2,291
Countries in 1995		
Australia	661	—
Austria	777	—
Canada	2,367	—
Cyprus	368	—
Czech Republic	1,087	—
France	1,110	—
Germany	723	—
Greece	459	—
Latvia	708	—
Switzerland	1,371	—
United States	3,114	—
Countries in 2008		
Armenia	—	894
Iran, Islamic Rep. of	—	2,434
Italy	—	1,861
Lebanon	—	1,595
Netherlands	—	1,511
Total	**16,785**	**16,489**

8.4.3 Omitted and Not-Reached Responses

Apart from missing data on items that by design were not administered to a student, missing data could also occur because a student did not answer an item—whether because the student did not know the answer, omitted it by mistake, or did not have time to attempt the item. An item was considered not reached when the item itself and the item immediately preceding it were not answered, and there were no other items completed in the remainder of the booklet.

In TIMSS Advanced 2008, as in TIMSS Advanced 1995 and previous TIMSS assessments, not-reached items were treated differently in estimating item parameters than they were in generating student proficiency scores. In estimating the values of the item parameters, items in the TIMSS Advanced assessment booklets that were considered not to have been reached by students were treated as if they had not been administered. This approach was considered optimal for parameter estimation. However, not-reached items were always considered as incorrect responses when student proficiency scores were generated.

8.4.4 Evaluating the Fit of IRT Models to the TIMSS Advanced 2008 Data

After the concurrent item calibrations were completed, checks were performed to verify that the item parameters obtained from Parscale adequately reproduced the observed distribution of student responses across the proficiency continuum. The fit of the IRT models to the TIMSS Advanced data was examined by comparing the item response function curves generated using the item parameters estimated from the data with the empirical item response functions calculated from the posterior distributions of the proficiencies for each student that responded to an item. When the empirical results fall near the fitted curves for any given item, the IRT model fits the data well and leads to more accurate and reliable measurement of the underlying proficiency scale. Graphical plots of these response function curves are called item characteristic curves (ICC).

Exhibit 8.17 shows an ICC of the empirical and fitted item response functions for a dichotomous multiple-choice item. In the graph, the horizontal axis represents the proficiency scale; and the vertical axis, the probability of a correct response. The fitted curve based on the estimated item parameters is shown as a solid line. Empirical results are represented by circles. The empirical results were obtained by dividing the proficiency scale into intervals of equal size and then counting the number of students responding to the item whose estimated a-priori (EAP) scores from Parscale fell in each interval. Then the proportion of students in each interval that responded correctly to the item was calculated. In the exhibit, the center of each circle represents this empirical proportion of correct responses. The size of each circle is proportional to the number of students contributing to the estimation of its empirical proportion correct.

Exhibit 8.17 Example Item Response Function for a TIMSS Advanced 2008 Dichotomous Item

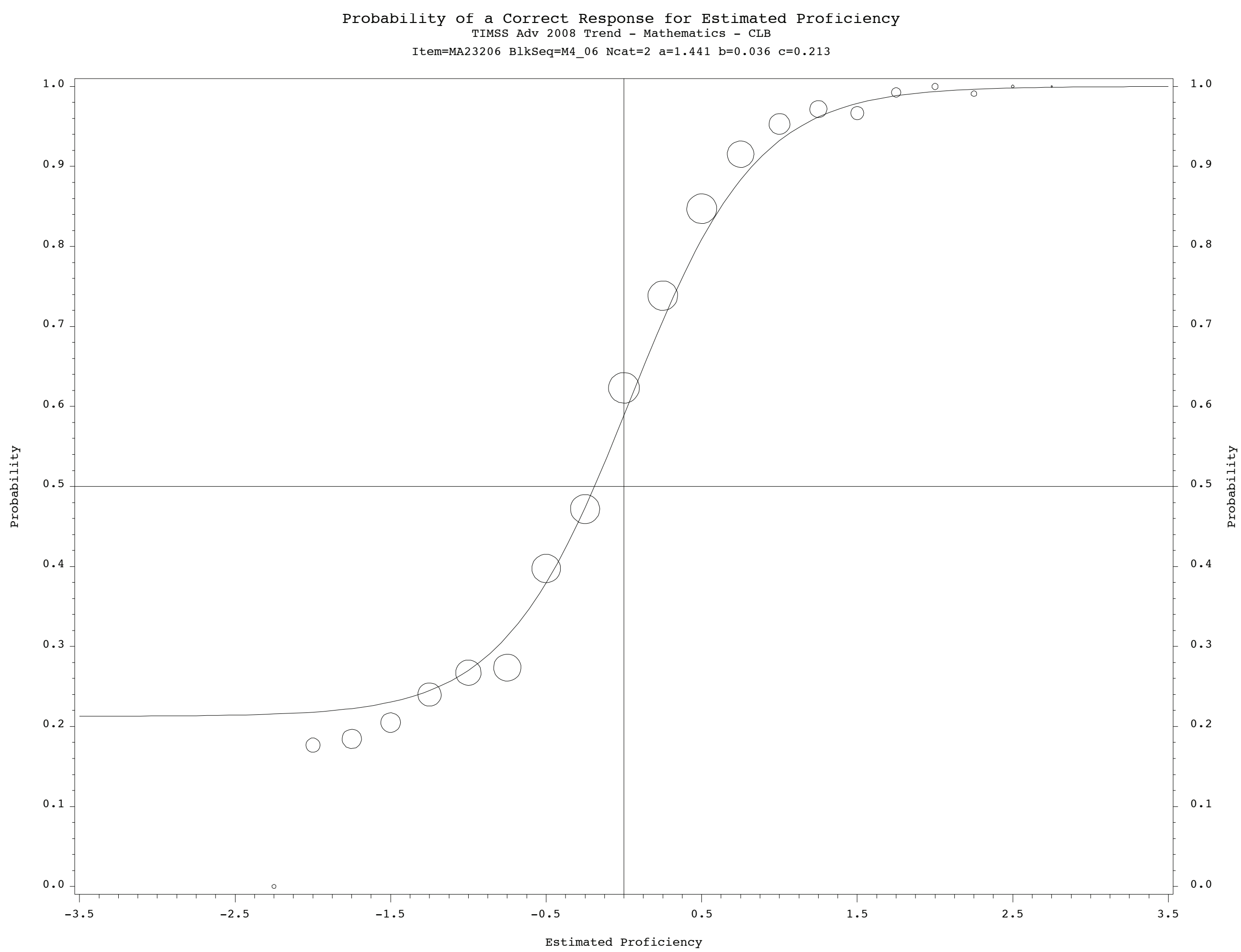
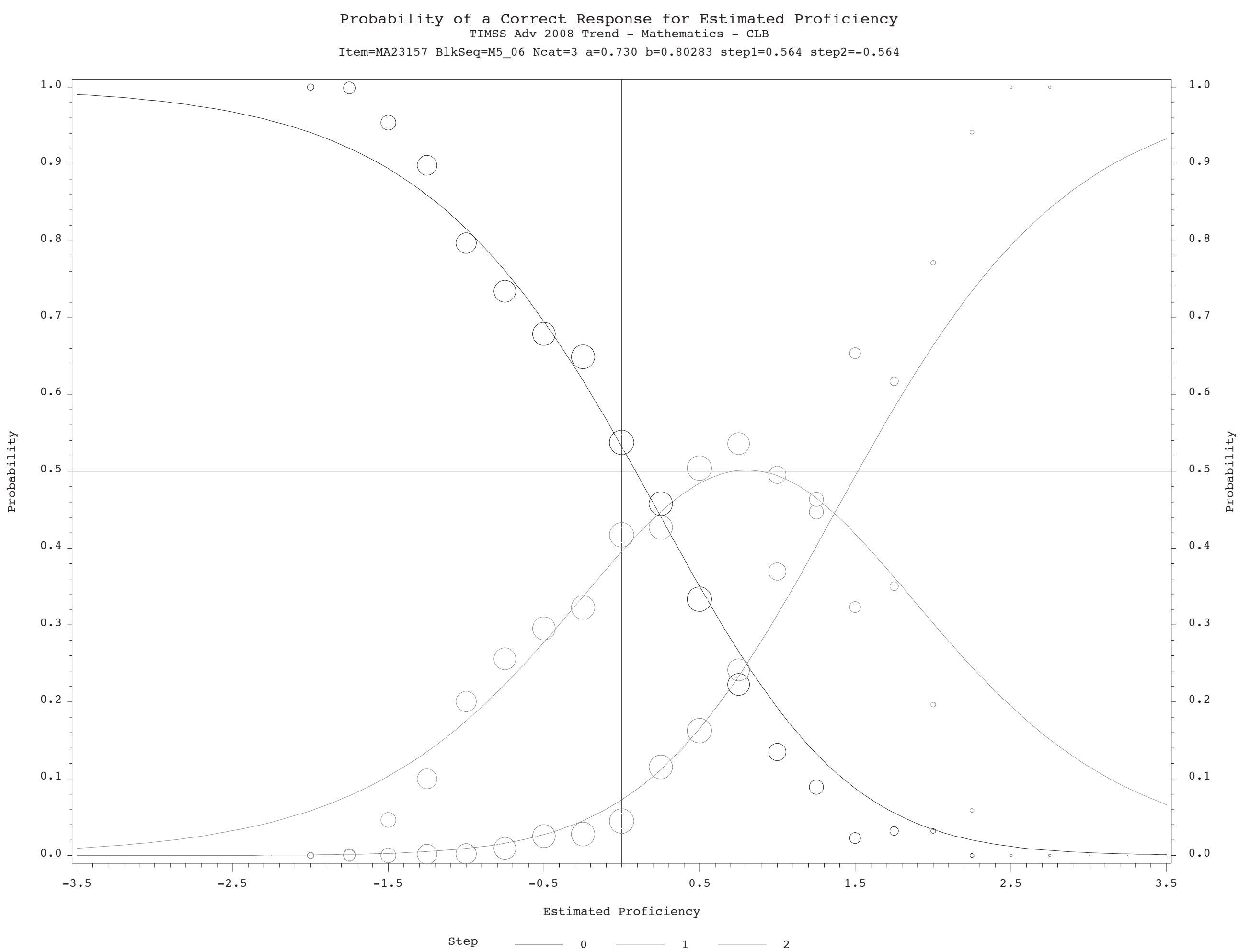


Exhibit 8.18 contains an ICC of the empirical and fitted item response functions for a polytomous constructed-response item with three response categories—0, 1, and 2 points. As for the dichotomous item plot, the horizontal axis represents the proficiency scale, but the vertical axis represents the probability of having a response in a given response category. The fitted curves based on the estimated item parameters are shown as solid lines. Empirical results are represented by circles. The interpretation of the circles is the same as in Exhibit 8.17. The curve starting at the top left of the chart plots the probability of a score of zero on the item, which decreases as proficiency increases.

Exhibit 8.18 Example Item Response Function for a TIMSS Advanced 2008 Polytomous Item



The bell-shaped curve shows the probability of a score of 1 point—starting low for low-ability students, reaching a maximum for medium-ability students, and decreasing for high-ability students. The curve ending at the top right corner of the chart shows the probability of a score of 2 points—full credit, starting low for low-ability students and increasing as proficiency increases.

8.4.5 Variables for Conditioning the TIMSS Advanced 2008 Data

Because there were so many background variables that could be used in conditioning, TIMSS Advanced followed the practice established by NAEP and followed by other large-scale studies of using principal components analysis to reduce the number of variables while explaining most of their common variance. Principal components for the TIMSS Advanced background data were constructed as follows:

- For categorical variables (questions with a small number of fixed response options), a "dummy coded" variable was created for each response option, with a value of 1 if the option was chosen and zero otherwise. If a student omitted or was not administered a particular question, all dummy coded variables associated with that question were assigned the value zero.
- Background variables with numerous categories (such as year of birth or time spent doing homework) were recoded using criterion scaling.[15] This was done by replacing each response option with an interim achievement score. For the overall advanced mathematics scale, the interim achievement score was the advanced mathematics score produced from the item calibration. For the overall physics scale, the interim achievement score was the physics score produced from the item calibration.
- Separately for each TIMSS country, all the dummy-coded and criterion-scaled variables were included in a principal components

15 The process of generating criterion-scaled variables is described in Beaton (1969).

analysis. Those principal components accounting for 90 percent of the variance of the background variables were retained for use as conditioning variables. However, if the selected number of principal components exceeded 5 percent of the student sample size, the number of selected principal components was reduced to 5 percent of the student sample size. Because the principal components analysis was performed separately for each country, different numbers of principal components were required to account for 90 percent of the common variance in each country's background variables.

In addition to the principal components, student gender (dummy coded), the language of the test (dummy coded), and an indicator of the class in the school to which the student belonged (criterion scaled) were included as primary conditioning variables, thereby accounting for most of the variance among students and preserving the between- and within-class variance structure in the scaling model. Conditioning variables were needed for all the TIMSS Advanced 2008 participants, as well as for all the TIMSS Advanced 1995 countries. Exhibits 8.19 and 8.20 show the total number of variables that were considered for conditioning and the number of principal components selected for each country for advanced mathematics and physics, respectively.

Exhibit 8.19 Number of Variables and Principal Components for Conditioning Advanced Mathematics in TIMSS Advanced 2008

Countries	1995 Assessment			2008 Assessment		
	Number of Primary Conditioning Variables	Total Number of Principal Components	Number of Principal Components Retained	Number of Primary Conditioning Variables	Total Number of Principal Components	Number of Principal Components Retained
Armenia	—	—	—	2	271	42
Australia	2	553	64	—	—	—
Austria	2	542	78	—	—	—
Canada	3	573	148	—	—	—
Cyprus	2	562	39	—	—	—
Czech Republic	2	565	110	—	—	—
Denmark	2	545	123	—	—	—
France	2	434	107	—	—	—
Germany	2	542	147	—	—	—
Greece	2	551	45	—	—	—
Iran, Islamic Rep. of	—	—	—	2	279	121
Israel	2	589	95	—	—	—
Italy	2	531	39	2	270	107
Lebanon	—	—	—	3	277	80
Lithuania	2	542	73	—	—	—
Netherlands	—	—	—	2	267	76
Norway	—	—	—	2	270	96
Philippines	—	—	—	2	276	156
Russian Federation	2	602	152	2	277	157
Slovenia	2	582	141	2	270	107
Sweden	2	576	100	2	268	115
Switzerland	4	544	127	—	—	—
United States	2	612	154	—	—	—

Exhibit 8.20 Number of Variables and Principal Components for Conditioning Physics in TIMSS Advanced 2008

Countries	1995 Assessment			2008 Assessment		
	Number of Primary Conditioning Variables	Total Number of Principal Components	Number of Principal Components Retained	Number of Primary Conditioning Variables	Total Number of Principal Components	Number of Principal Components Retained
Armenia	—	—	—	2	275	44
Australia	2	560	66	—	—	—
Austria	2	545	77	—	—	—
Canada	3	570	134	—	—	—
Cyprus	2	565	36	—	—	—
Czech Republic	2	549	108	—	—	—
Denmark	2	533	65	—	—	—
France	2	435	111	—	—	—
Germany	2	526	72	—	—	—
Greece	2	550	45	—	—	—
Iran, Islamic Rep. of	—	—	—	2	282	121
Israel	2	583	85	—	—	—
Italy	—	—	—	2	205	93
Lebanon	—	—	—	3	281	79
Latvia	2	579	70	—	—	—
Netherlands	—	—	—	2	272	75
Norway	3	579	104	2	270	82
Russian Federation	2	598	123	2	283	158
Slovenia	2	573	74	2	272	54
Sweden	2	571	101	2	268	114
Switzerland	4	546	125	—	—	—
United States	2	618	166	—	—	—

8.4.6 Generating IRT Proficiency Scores for the TIMSS Advanced 2008 Data

MGROUP was used to generate the IRT proficiency scores. Exhibit 8.21 shows the student sample sizes—from the 1995 assessments and the 2008 assessments—for which proficiency scores, using the item parameters obtained from the concurrent calibration, were generated on the overall advanced mathematics and physics scales.

Exhibit 8.21 Sample Sizes for TIMSS Advanced Proficiency Estimation

Country	Advanced Mathematics		Physics	
	1995 Assessment	2008 Assessment	1995 Assessment	2008 Assessment
Armenia	—	858	—	894
Australia	645	—	661	—
Austria	782	—	777	—
Canada	2,781	—	2,367	—
Cyprus	391	—	368	—
Czech Republic	1,101	—	1,087	—
France	1,071	—	1,110	—
Germany	2,296	—	723	—
Greece	456	—	459	—
Iran, Islamic Rep. of	—	2,425	—	2,434
Italy	398	2,143	—	1,861
Latvia	—	—	708	—
Lebanon	—	1,612	—	1,595
Lithuania	734	—	—	—
Netherlands	—	1,537	—	1,511
Norway	—	1,932	1,048	1,640
Philippines	—	4,091	—	—
Russian Federation	1,638	3,185	1,233	3,166
Slovenia	1,536	2,156	747	1,097
Sweden	1,001	2,303	1,012	2,291
Switzerland	1,404	—	1,371	—
United States	2,785	—	3,114	—
Total	**19,019**	**22,242**	**16,785**	**16,489**

8.4.7 Transforming the Advanced Mathematics and Physics Scores to Measure Trends

As part of rescaling the data from the TIMSS Advanced 1995 assessments using 2- and 3-parameter models as described in Section 8.4.1, the TIMSS Advanced reporting scales were established by setting the average of the mean scores of the countries included in the rescaling item calibrations to 500 and the standard deviation to 100. To provide results for the 2008 assessments that would be directly comparable to the results from the 1995 assessments, the 2008 proficiency scores (plausible values) for advanced mathematics and physics had to be transformed to the TIMSS Advanced scales established with the 1995 data. This was accomplished through a linear transformation of the proficiency scores from the 1995-2008 concurrent calibration such that the 1995 proficiency distribution from the concurrent calibration aligned itself with the 1995 proficiency distribution from the 1995 rescaling calibration.

The means and standard deviations of the 1995 advanced mathematics and physics scores produced in 2008—the plausible values from the TIMSS Advanced 1995 assessment data based on the 1995-2008 concurrent item calibrations—were made to match the means and standard deviations of the scores calculated for the TIMSS Advanced 1995 assessment—the plausible values produced using the item calibration from scaling the 1995 assessment data—by applying the appropriate linear transformations. These linear transformations were given by:

$$PV_{k,i}^{*} = A_{k,i} + B_{k,i} \cdot PV_{k,i}$$

where

$PV_{k,i}$ was plausible value i for scale k prior to transformation,

$PV_{k,i}^{*}$ was plausible value i for scale k after transformation, and $A_{k,i}$ and $B_{k,i}$ were the linear transformation constants for plausible value i of scale k.

The linear transformation constants were obtained by first computing the international means and standard deviations of the proficiency scores for the overall advanced mathematics and physics scales using the plausible values from the 1995 scaling for the 1995 countries included in the concurrent calibration. Next, the same calculation was done using the plausible values from the 1995 assessment data based on the 1995-2008 concurrent calibration for the same set of countries. The linear transformation constants were thus defined as:

$$B_{k,i} = \sigma^{*}_{k,i} / \sigma_{k,i}$$
$$A_{k,i} = \mu^{*}_{k,i} - B_{k,i}\,\mu_{k,i}$$

where

$\mu_{k,i}^{*}$ was the international mean of scale k based on plausible value i obtained from scaling the 1995 assessment data;

$\mu_{k,i}$ was the international mean of scale k based on plausible value i from the TIMSS Advanced 1995 assessment data based on the 1995-2008 concurrent item calibrations;

$\sigma_{k,i}^{*}$ was the international standard deviation of scale k based on plausible value i obtained from scaling the 1995 assessment data;

$\sigma_{k,i}$ was the international standard deviation of scale k based on plausible value i from the TIMSS Advanced 1995 assessment data based on the 1995-2008 concurrent item calibrations.

Exhibit 8.22 shows the linear transformation constants that were computed for TIMSS Advanced 2008 for the advanced mathematics and physics scales. Once these linear transformation constants had been established, all of the 2008 advanced mathematics and physics proficiency scores—the plausible values generated from the 2008 assessment data—for all the participating countries were transformed by applying the linear transformations. This provided advanced mathematics and physics student achievement scores for the 2008 assessments that were directly comparable to the rescaled scores from the 1995 assessment data.

Exhibit 8.22 Linear Transformation Constants Applied to the TIMSS Advanced 2008 Proficiency Scores

Scale	Plausible Value	TIMSS Advanced 1995 Scores Using 1995 Item Calibrations		TIMSS Advanced 1995 Scores Using 1995–2008 Concurrent Calibrations		$A_{k,i}$	$B_{k,i}$
		Mean	Standard Deviation	Mean	Standard Deviation		
Advanced Mathematics	PV1	500.00109	99.99451	0.11465	0.88983	487.11733	112.37442
	PV2	500.00010	99.99951	0.10189	0.89226	488.58124	112.07392
	PV3	500.00047	99.99754	0.11694	0.89312	486.90788	111.96396
	PV4	500.00004	99.99981	0.10578	0.88989	488.11346	112.37280
	PV5	500.00107	99.99447	0.10317	0.89075	488.41976	112.25904
Physics	PV1	500.00000	100.00000	-0.02718	0.97905	502.77604	102.14010
	PV2	500.00000	100.00000	-0.02970	0.98882	503.00383	101.13043
	PV3	500.00000	100.00000	-0.02392	0.97793	502.44621	102.25631
	PV4	500.00000	100.00000	-0.02718	0.98387	502.76209	101.63923
	PV5	500.00000	100.00000	-0.03582	0.97809	503.66228	102.23961

Note: The means and standard deviations for advanced mathematics based on the 1995 item calibrations are affected by rare cases of very low scores that were truncated.

8.5 The TIMSS Advanced 2008 International Benchmarks of Student Achievement in Advanced Mathematics and Physics

To describe student performance at various points along the TIMSS 2008 advanced mathematics and physics achievement scales, TIMSS Advanced 2008 used scale anchoring to summarize and describe student achievement at three points on the advanced mathematics and physics scales—Advanced International Benchmark (625), High

International Benchmark (550), and Intermediate International Benchmark (475). For a description of performance at the international benchmarks, please see the *TIMSS Advanced 2008 International Report* (Mullis, Martin, Robitaille, & Foy, 2009).

In brief, scale anchoring involves selecting benchmarks (scale points) on the TIMSS achievement scales to be described in terms of student performance and then identifying items that students scoring at those anchor points (the international benchmarks) can answer correctly. The items so identified were grouped by content domain within benchmarks for review by mathematics and physics experts. The committee members[16] examined the content of each item and determined the kind of mathematics or physics knowledge or skill demonstrated by students who responded correctly to the item. They then summarized the detailed list of item competencies in a brief description of achievement at each international benchmark. This procedure resulted in a content-referenced interpretation of the achievement results that can be considered in light of the TIMSS Advanced 2008 advanced mathematics and physics frameworks.

As the first step, students scoring within 20 scale score points of each benchmark were identified for the benchmark analysis. The score ranges around each international benchmark and the number of students scoring in each range for advanced mathematics and physics are shown in Exhibit 8.23. The range of 20 points above and below a benchmark provided an adequate sample in each group, yet was small enough so that performance at each benchmark anchor point was still distinguishable from the next.

Exhibit 8.23 Range Around Each International Benchmark and Number of Students Within Each Range

	Intermediate (475)	High (550)	Advanced (625)
Range of Scale Scores	455–495	530–570	605–645
Advanced Mathematics	2,826	2,752	1,138
Physics	2,201	2,369	1,327

16 In addition to Robert A. Garden, the TIMSS Advanced Mathematics Coordinator, and Svein Lie, the TIMSS Physics Coordinator, committee members included Carl Angell, Wolfgang Dietrich, Liv Sissel Gronmo, Torgeir Onstad, and David F. Robitaille.

Having identified the number of students scoring within each benchmark range, the next step was conducting the data analysis to determine which items anchored at each of the international benchmarks. An important feature of the scale anchoring method is that it yields descriptions of the performance demonstrated by students reaching each of the international benchmarks on the scales, and that the descriptions reflect demonstrably different accomplishments by students reaching each successively higher benchmark. Because the process entails the delineation of sets of items that students at each international benchmark are likely to answer correctly and that discriminate between one benchmark and the next, the criteria for identifying the items that anchor considers performance at more than one benchmark.

For *multiple-choice* items, a criterion of 65 percent was used for each benchmark being analyzed, since students would be likely (about two thirds of the time) to answer the item correctly. A criterion of less than 50 percent was used for the next lower benchmark, because with this response probability, students were more likely to have answered the item incorrectly than correctly. The criteria for each benchmark are outlined below.

- A multiple-choice item anchored at the Intermediate International Benchmark (475) if at least 65 percent of students scoring in the range answered the item correctly. Because this was the lowest benchmark described, there were no further criteria.
- A multiple-choice item anchored at the High International Benchmark (550) if at least 65 percent of students scoring in the range answered the item correctly, and less than 50 percent of students at the Intermediate International Benchmark answered the item correctly.

- A multiple-choice item anchored at the Advanced International Benchmark (625) if at least 65 percent of students scoring in the range answered the item correctly, and less than 50 percent of students at the High International Benchmark answered the item correctly.

To include all of the multiple-choice items in the anchoring process and provide information about content domains and cognitive processes that might not otherwise have had many anchor items, the concept of items that "almost anchored" was introduced. These were items that met slightly less stringent criteria for being answered correctly. The criteria to identify multiple-choice items that "almost anchored" were that at least 55 percent of students scoring in the range answered the item correctly and less than 50 percent of students at the next lowest benchmark answered correctly. To be completely inclusive for all items, items that met only the criterion that at least 55 percent of the students answered correctly (regardless of the performance of students at the next lower point) were also identified. The categories of items were mutually exclusive, and ensured that all of the items were available to inform the descriptions of student achievement at the anchor levels. A multiple-choice item was considered to be "too difficult" to anchor if less than 55 percent of students at the advanced benchmark answered the item correctly.

A somewhat less strict criterion was used for all the *constructed-response* items, because students had much less scope for guessing. For constructed-response items, the criterion of 50 percent was used for the benchmark without any discrimination criterion for the next lower benchmark. A constructed-response item anchored at one of the international benchmarks if at least 50 percent of students at that benchmark answered the item correctly. A constructed-response item was considered to be "too difficult" to anchor if less than 50 percent of students at the advanced benchmark answered the item correctly.

For students scoring in the range around each international benchmark, the percentage of those students that answered each item correctly was computed. To compute these percentages, students in each country were weighted to contribute proportional to the size of the student population in a country. Most of the TIMSS Advanced 2008 items were scored 1 point for a correct answer and 0 points for other answers. For these items, the percentage of students at each benchmark who answered each item correctly was computed. For the relatively few constructed-response items scored for partial or full credit, percentages were computed for the students receiving full credit. Then the criteria described above were applied to identify the items that anchored, almost anchored, and met only the 55 to 65 percent criteria. Exhibit 8.24 presents the number of advanced mathematics and physics items that anchored at each international benchmark.

Exhibit 8.24 Number of Items Anchoring at Each International Benchmark

	Intermediate (475)	High (550)	Advanced (625)	Too Difficult to Anchor	Total
Advanced Mathematics	16	23	21	11	71
Physics	17	14	22	15	68

In preparation for the committee review, the advanced mathematics and physics items were organized into separate binders. The items were grouped by international benchmark and, within benchmark, the items were sorted by content area and then by the anchoring criteria they met: items that anchored, followed by items that almost anchored, followed by items that met only the 55 to 65 percent criteria. The following information was included for each item: content area, cognitive domain, maximum points, answer key, release status, percent correct at each benchmark, and overall international percent correct. For constructed-response items, the scoring guides were included.

The TIMSS & PIRLS International Study Center staff convened the committee for a three-day meeting in Boston to complete three tasks: The committee (1) worked through each item and arrived at a short description of the knowledge, understanding, or skills demonstrated by students who answered the item correctly; (2) developed a description (in detailed and summary form) of the level of advanced mathematics or physics proficiency demonstrated by students at each of the three international benchmarks to publish in the TIMSS Advanced 2008 International Report; and (3) selected example items that supported and illustrated the benchmark descriptions to publish together with the descriptions.

8.6 Capturing the Uncertainty in the TIMSS Advanced 2008 Student Achievement Scores

To obtain estimates of students' proficiency in advanced mathematics and physics that were both accurate and cost-effective, TIMSS Advanced 2008 made extensive use of probability sampling techniques to sample students from national student populations, and applied matrix sampling methods to target individual students with a subset of the entire set of assessment materials. Statistics computed from these student samples were used to estimate population parameters. This approach made efficient use of resources, in particular keeping student response burden to a minimum, but at a cost of some variance, or uncertainty, in the statistics. To quantify this uncertainty, each statistic in the *TIMSS Advanced 2008 International Report* (Mullis, Martin, Robitaille, & Foy, 2009) is accompanied by an estimate of its standard error. These standard errors incorporate components reflecting the uncertainty due to generalizing from student samples to the entire student populations (sampling variance), and to inferring students' performance on the entire assessment from their performance on the subset of items that they took (imputation variance).

8.6.1 Estimating the Sampling Variance

The TIMSS Advanced 2008 sample design applied a stratified multi-stage cluster-sampling technique to the problem of selecting efficient and accurate samples of students while working with schools and classes. This design capitalized on the structure of the student population (i.e., students grouped in classes within schools) to derive student samples that permitted efficient and economical data collection. Unfortunately, such a complex sample design complicates the task of computing standard errors to quantify sampling variability.

When, as in TIMSS Advanced, the sample design involves multi-stage cluster sampling, there are several options for estimating sampling errors that avoid the assumption of simple random sampling (Wolter, 1985). The jackknife repeated replication technique (JRR) was chosen by TIMSS because it is computationally straightforward and provides approximately unbiased estimates of the sampling errors of means, totals, and percentages.

The variation on the JRR technique used in TIMSS Advanced 2008 is described in Johnson and Rust (1992). It assumes that the primary sampling units (PSUs) can be paired in a manner consistent with the sample design, with each pair regarded as members of a pseudo-stratum for variance estimation purposes. When used in this way, the JRR technique appropriately accounts for the combined effect of the between- and within-PSU contributions to the sampling variance. The general use of JRR entails systematically assigning pairs of schools to sampling zones, and randomly selecting one of these schools to have its contribution doubled and the other to have its contribution zeroed, so as to construct a number of "pseudo-replicates" of the original sample. The statistic of interest is computed once for the entire original sample, and once again for each jackknife pseudo-replicate sample. The variation between the estimates for each of the jackknife replicate samples and the original sample estimate is the jackknife estimate of the sampling error of the statistic.

8.6.2 Constructing Sampling Zones

To apply the JRR technique used in TIMSS Advanced 2008, successive sampled schools were paired and assigned to a series of groups known as sampling zones. This was done at Statistics Canada by working through the list of sampled schools in the order in which they were selected and assigning the first and second participating schools to the first sampling zone, the third and fourth participating schools to the second zone, and so on. A maximum of 75 zones were used, although most countries had fewer because they generally sampled less than 150 schools. When more than 75 zones were constructed, as was the case in Lebanon, they were collapsed to keep the total number to 75.

Sampling zones were constructed within explicit strata. When there was an odd number of schools in an explicit stratum, either by design or because of school non-response, the students in the not-paired school were randomly divided to make up two "quasi" schools for the purposes of calculating the jackknife standard error.[17] Each sampling zone then consisted of a pair of schools or "quasi" schools. Exhibit 8.25 shows the number of sampling zones in each country.

Within each sampling zone, each school was assigned an indicator (u_j), coded randomly to 0 or 1, such that one school had a value of zero, and the other a value of 1. This indicator determined whether the weights for the sampled students in the school in this zone were to be doubled ($u_j = 1$) or zeroed ($u_j = 0$) for the purposes of creating the pseudo-replicate samples.

17 If the not-paired school consisted of 2 sampled classrooms, each classroom became a "quasi" school.

Exhibit 8.25 Number of Sampling Zones Used in the TIMSS Advanced 2008 Countries

Country	Advanced Mathematics	Physics
Armenia	55	55
Italy	46	46
Iran, Islamic Rep. of	60	60
Lebanon	75	75
Netherlands	56	58
Norway	55	52
Philippines	61	—
Russian Federation	45	47
Slovenia	53	66
Sweden	59	61

8.6.3 Computing the Sampling Variance Using the JRR Method

The formula for the sampling variance of a statistic t, based on the JRR algorithm used in TIMSS Advanced 2008, is given by the following equation:

$$Var_{jrr}(t) = \sum_{h=1}^{H} [t(J_h) - t(S)]^2$$

where H is the total number of sampling zones in the sample of the country under consideration. The term $t(S)$ corresponds to the statistic of interest for the whole sample computed with the overall sampling weights.[18] The term $t(J_h)$ denotes the same statistic using the h^{th} jackknife replicate sample J_h and its set of replicate sampling weights, which are identical to the overall sampling weights, except for the students in the h^{th} sampling zone. In the h^{th} zone, all students belonging to one of the randomly selected schools of the pair were removed, and the students belonging to the other school in the zone were included twice. In practice, this was accomplished by recoding to zero the sampling weights for the students in the school to be excluded

18 The sampling weights are described in Chapter 4.

from the replication, and multiplying by 2 the sampling weights of the remaining students within the h^{th} pair. Each sampled student was assigned a vector of 75 replicate sampling weights W_{hi}, where h took values from 1 to 75. If W_{oi} was the overall sampling weight of student i, the h replicate weights for that student were computed as:

$$W_{hi} = W_{0i} \cdot k_{hi}$$

where

$$k_{hi} = \left\{ \begin{array}{ll} 2 \cdot u_j & \text{if student } i \text{ is in school } j \text{ of sampling zone } h \\ 1 & \text{otherwise} \end{array} \right\}$$

The school-level indicators u_j determined which students in a sampling zone would get zero weights and which ones would get double weights, on the basis of the school within the pair from which the students were sampled. The process of setting the k_{hi} values for all sampled students and across all sampling zones is illustrated in Exhibit 8.26. Thus, the computation of the JRR variance estimate for any statistic in TIMSS Advanced 2008 required the computation of the statistic up to 76 times for any given country: once to obtain the statistic for the full sample based on the overall weights W_{oi}, and up to 75 times to obtain the statistics for each of the jackknife replicate samples J_h using a set of replicate weights W_{hi}.

Exhibit 8.26 Construction of Replicate Weights Across Sampling Zones in TIMSS Advanced 2008

Sampling Zone	u_j	k_{hi} for Computing JRR Replicate Sampling Weights						
		1	2	3	…	h	…	75
1	0	0	1	1	…	1	…	1
	1	2						
2	0	1	0	1	…	1	…	1
	1		2					
3	0	1	1	0	…	1	…	1
	1			2				
⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮
h	0	1	1	1	…	0	…	1
	1					2		
⋮	⋮	⋮	⋮	⋮	…	⋮	⋱	⋮
75	0	1	1	1	…	1	…	0
	1							2

In the TIMSS Advanced 2008 analyses, 75 replicate weights were computed for each country regardless of the number of actual zones within the country. If a country had fewer than 75 zones, then the additional replicate weights where h was greater than the number of zones within the country were all set equal to the overall sampling weight. Although this involved some redundant computations, having 75 replicate weights for each country had no effect on the magnitude of the error variance computed using the jackknife formula, and it simplified the computation of standard errors for numerous countries at the same time. All standard errors presented in the TIMSS Advanced 2008 international report were computed using SAS programs developed at the TIMSS & PIRLS International Study Center.

8.6.4 Estimating the Imputation Variance

The TIMSS Advanced 2008 item pool was far too extensive to be administered in its entirety to any one student, and so a matrix-sampling test design was developed whereby each student was given a single test booklet containing only a part of the entire assessment.[19] The results for all of the booklets were then aggregated using item response theory to provide results for the entire assessment. Because each student responded to just a subset of the assessment items, it was necessary to use multiple imputation (the generation of plausible values) to derive reliable estimates of student performance on the assessment as a whole. Since every student's proficiency estimate incorporates some uncertainty arising from this imputation, TIMSS Advanced followed the customary procedure of generating five estimates for each student and using the variability among them as a measure of the imputation uncertainty, or error. In the TIMSS Advanced 2008 international report, the imputation error for each variable has been combined with the sampling error for that variable to provide a standard error that incorporates both.

The general procedure for estimating the imputation variance using plausible values is described in Mislevy, Beaton, Kaplan, and Sheenan (1992). First, compute the statistic t for each set of M plausible values. The statistics t_m, where $m = 1, 2, \ldots, M$, can be anything estimable from the data, such as a mean, the difference between means, percentiles, and so forth.

Once the statistics t_m are computed, the imputation variance of the statistic t is then calculated as:

$$Var_{imp}(t) = \left(1 + \frac{1}{M}\right) Var(t_1, \ldots t_M)$$

19 The TIMSS Advanced 2008 assessment design is described in the *TIMSS Advanced 2008 Assessment Frameworks* (Garden, et al., 2006).

where M is the number of plausible values used in the calculation, and $Var(t_1,\ldots,t_M)$ is the usual variance of the M estimates computed using each plausible value.

8.6.5 Combining the Sampling and Imputation Variance

The standard errors of all the proficiency statistics reported by TIMSS Advanced include both sampling and imputation variance components. These standard errors were computed using the following formula:

$$Var(t_{pv}) = Var_{jrr}(t_1) + Var_{imp}(t)$$

where $Var_{jrr}(t_1)$ is the sampling variance computed for the first plausible value[20] and $Var_{imp}(t)$ is the imputation variance. The *TIMSS Advanced 2008 User Guide for the International Database* (Foy & Arora, 2009) contains programs in SAS and SPSS that compute each of these variance components for the TIMSS Advanced 2008 data. Furthermore, the IEA IDB Analyzer—software provided with the international database—automatically computes standard errors as described in this section.

Exhibit 8.27 shows basic summary statistics for overall advanced mathematics and physics achievement in the TIMSS Advanced 2008 assessments. It presents the student sample size, the mean and standard deviation averaged across the five plausible values, the jackknife sampling error for the mean, and the overall standard error for the mean, which includes the imputation error.

20 Under ideal circumstances and with unlimited computing resources, the JRR sampling variance would be computed for each of the plausible values and the imputation variance as described here. This would require computing the same statistic up to 380 times (once overall for each of the five plausible values using the overall sampling weights, and then 75 times more for each plausible value using the complete set of replicate weights). An acceptable shortcut, however, is to compute the JRR sampling variance component using only one plausible value (the first one), and then the imputation variance using the five plausible values. Using this approach, a statistic needs to be computed only 80 times.

Exhibit 8.27 Summary Statistics and Standard Errors for Proficiency in TIMSS Advanced 2008

Advanced Mathematics

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Armenia	858	432.760	95.466	3.090	3.675
Iran, Islamic Rep. of	2,425	496.750	98.767	6.306	6.369
Italy	2,143	448.779	95.468	6.870	7.142
Lebanon	1,612	544.726	60.472	2.258	2.318
Netherlands	1,537	552.470	45.797	1.917	2.647
Norway	1,932	439.224	86.910	4.679	4.990
Philippines	4,091	355.189	105.545	5.374	5.522
Russian Federation	3,185	560.984	90.972	7.138	7.213
Slovenia	2,156	457.316	84.850	3.890	4.151
Sweden	2,303	412.806	103.265	5.370	5.571

Physics

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Armenia	894	495.067	100.284	5.125	5.363
Iran, Islamic Rep. of	2,434	459.856	115.728	7.145	7.204
Italy	1,861	422.238	102.558	7.510	7.624
Lebanon	1,595	443.542	78.324	2.615	2.990
Netherlands	1,511	582.474	53.723	3.547	3.703
Norway	1,640	534.142	78.147	4.182	4.212
Russian Federation	3,166	521.220	120.490	10.140	10.172
Slovenia	1,097	534.941	80.247	1.507	1.941
Sweden	2,291	496.950	91.865	5.509	5.651

References

Beaton, A.E. (1969). Criterion scaling of questionnaire items. *Socio-Economic Planning Sciences*, 2, 355–362.

Foy, P., & Arora, A. (2009). *TIMSS advanced 2008 user guide for the international database*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Foy, P., Galia, J., & Li, I. (2008). Scaling the data from the TIMSS 2007 mathematics and science assessments. Chapter 11 in J.F. Olson, M.O. Martin, and I.V.S. Mullis (Eds.), *TIMSS 2007 technical report* (pp. 225-280). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Garden, R.A., Lie, S., Robitaille, D.F., Angell, C., Martin, M.O., Mullis, I.V.S., Foy, P., & Arora, A. (2006), *TIMSS advanced 2008 assessment frameworks.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Johnson, E.G., & Rust, K.F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics*, 17(2), 175–190.

Mislevy, R.J., Beaton, A., Kaplan, B.A., & Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29(2), 133–161.

Mullis, I.V.S., Martin, M.O., Robitaille, D.F., & Foy, P. (2009), *TIMSS advanced 2008 international report: Findings from IEA's of achievement in advanced mathematics and physics in the final year of secondary school*, Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Muraki, E., & Bock, R.D. (1991). *PARSCALE: Parameter scaling of rating data* [Software Version 4.1]. Chicago, IL: Scientific Software, Inc.

Sheehan, K. M. (1985). *M-GROUP*: *Estimation of group effects in multivariate models* [Software Version 3.2]. Princeton, NJ: Educational Testing Service.

von Davier, M., Gonzalez, E., & Mislevy, R.J. (2009). What are plausible values and why are they useful? *IERI monograph series: Issues and methodologies in large-scale assessments*, (Vol. 2, pp. 9-36). Hamburg/Princeton NJ: IEA-ETS Research Institute.

Wolter, K. M. (1985). *Introduction to variance estimation*. New York: Springer-Verlag.