Appendix B

Overview of Procedures TIMSS 2003 Developmental Project

Process for Establishing the Mathematics Cognitive Domains for Scaling and Reporting

As explained in Chapter 1, developing reliable and valid achievement scales in the cognitive domains began with conducting a meeting of mathematics experts to examine the classification of the TIMSS 2003 items. Hosted by the IEA Secretariat in Amsterdam, 10 participants (see below) met in February 2005.

Participants in Mathematics Expert Meeting

Amsterdam, February 2005

Khattab Mohammad Abu Lebdeh – *Jordan*
Yu-Hsien Chang – *Chinese Taipei*
Tandi Clausen-May – *England*
Robert Garden – *New Zealand*
Barbara Japelj – *Slovenia*
Michael Martin – *TIMSS Study Director*
Ina Mullis – *TIMSS Study Director*
Peter Nystrom – *Sweden*
David Robitaille – *Canada*
Graham Ruddock – *England*

Mathematics Participants in TIMSS 2007 Science and Mathematics Item Review Committee Meeting

London, April 2005

Khattab Mohammad Abu Lebdeh – *Jordan*

Alka Arora – *TIMSS Research Associate*

Kiril Bankov – *Bulgaria*

Robert Garden – *New Zealand*

Liv Sissel Gronmo – *Norway*

Chen-yung Lin – *Chinese Taipei*

Mary Lindquist – *United States*

Ina Mullis – *TIMSS Study Director*

Graham Ruddock – *TIMSS 2007 Mathematics Coordinator*

Hanako Senuma – *Japan*

Characteristics of Items Within Cognitive Domains

IEA's TIMSS & PIRLS International Study Center (ISC) examined the spread of the items within the three domains according to item type (constructed-response or multiple-choice), content domain (algebra, geometry, etc.), and average difficulty (mean percent correct) to ensure there was sufficient coverage within each domain. As shown in Exhibit B.1, the classification resulted in a substantial number of items in each cognitive domain at both eighth grade (first page) and fourth grade (second page). Of the 194 items at the eighth grade, 65 were classified in the knowing cognitive domain, 93 in the applying cognitive domain, and 36 in the reasoning cognitive domain. Of the 159 items at the fourth grade, 58 were classified in the knowing cognitive domain, 63 in the applying cognitive domain, and 38 in the reasoning cognitive domain.

Within each cognitive domain, there was a very good spread of items in terms of item type (constructed-response or multiple-choice) at both eighth and fourth grades. Equivalent percentages of applying items were multiple-choice and constructed-response. As would

Exhibit B.1: Characteristics of Items Within Cognitive Domains

MATHEMATICS Grade 4

Number of Items by Item Type and Cognitive Domains

Item Type	Cognitive Domains			Total
	Knowing	Applying	Reasoning	
Constructed Response	&	! '	!#	%(
Multiple Choice	#	"$	#	(
Total	**58**	**63**	**38**	**159** [1]

Percent of Score Points by Item Type and Cognitive Domains

Item Type	Cognitive Domains			Total Score Points
	Knowing	Applying	Reasoning	
Constructed Response	!#	" '	" '	&%
Multiple Choice	#%	"(	%	(
Total	**36%**	**39%**	**26%**	**166**

Number of Items by Content Domain and Cognitive Domain

Content Domain	Cognitive Domains			Total
	Knowing	Applying	Reasoning	
Number	!$	(	(	%"
Patterns and Relationships	!	"	'	!"
Measurement		'	#	"!
Geometry	&	%		!#
Data	#	&	%	&
Total	**58**	**63**	**38**	**159**

Percent of Score Points by Content Domain and Cognitive Domain

Content Domain	Cognitive Domains			Total Score Points
	Knowing	Applying	Reasoning	
Number	"&	!(	"#	%'
Patterns and Relationships	'	$#	" '	!#
Measurement	"	$%	"	"!
Geometry	&!	!#	#	!$
Data	!#	#	"$	&
Total	**36%**	**39%**	**26%**	**166**

Mean Percent Correct by Content Domain and Cognitive Domain

Item Difficulties (Mean Percent Correct)	Cognitive Domains			Total
	Knowing	Applying	Reasoning	
Number	%"	$%	"&	$"
Patterns and Relationships	%"	$&	"%	$
Measurement	%$	#&	"(	$!
Geometry	%	$"	#"	$&
Data	%(	$%	$'	%
Total	**53%**	**53%**	**40%**	**54**

be expected, however, at both grades a relatively higher percentage of items in the knowing domain were multiple-choice, and a commensurately higher percentage of items in the reasoning domain were constructed-response. Often, the multiple-choice format is a cost-effective way to assess specific knowledge, while the constructed-response format may be required in complex problem-solving situations involving multiple strategies.

Despite some unevenness, there was good spread across content domains within each of the three cognitive domains. At eighth grade, it would have been preferable to have a higher proportion of number items in the reasoning domain (an effort is being made to address this in TIMSS 2007). That the distribution for measurement is concentrated in the applying domain makes some sense, since by eighth grade students should know about basic measurement tools and units. (In the TIMSS 2007 Framework, aspects of measurement were incorporated into the number and geometry content domains because there is little emphasis on measurement in eighth-grade mathematics curricula around the world).

Because algebra is generally not taught as a formal subject in primary school, only introductory concepts about patterns and relationships are assessed at the fourth grade. As such, a higher proportion of patterns and relationship items in the knowing category would have been preferable at the fourth grade. (In the TIMSS 2007 Framework, the patterns and relationships content domain has been incorporated into the number content domain.) Also, a higher proportion of measurement items in the reasoning domain would have been better. The low coverage of geometry in the reasoning domain is understandable, since this is a subject little emphasized at the fourth grade. (In the TIMSS 2007 Framework, the geometry content domain, now called geometric shapes and measures, has been recast to better describe the fourth-grade curricula of participating countries.)

Finally, Exhibit B.1 also shows a good range in item difficulty (mean percentage correct) internationally, on average, within each of

the probability that a student will respond in a specific way to an item in terms of the respondent'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, which were scored as correct or incorrect, and a two-parameter model for constructed-response items with just two response options, which also were scored as correct or incorrect. Since each of these item types has just 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 score points.

Item Calibration

The first step in constructing the cognitive domain scales was to estimate the IRT model parameters for each item on each of the cognitive domain scales. This procedure, known as item calibration, was implemented using the PARSCALE software applied to a self-weighting random sample of 1000 students from each country's TIMSS 2003 student sample. Using student samples of equal size ensured that the data from each country contributed equally to the item calibration, while keeping the amount of data to be analyzed to a reasonable size.

At the fourth and eighth grades, separate calibrations were conducted for each of the three mathematics cognitive domains: knowing, applying, and reasoning (abbreviated labels). At the eighth grade, the calibrations were based on 46,000 student records; 1,000 from each of the 46 countries that participated in the 2003 assessment. At the fourth grade, the calibrations were based on 26,000 student records, 1,000 from each of the 26 countries that participated in the 2003 assessment at the fourth grade.

Reliability

Exhibit B.2 displays the reliability coefficient for each country for the mathematics test overall and for the knowing, applying, and reasoning cognitive domains. The first page shows the reliabilities for the eighth grade and the second page shows the reliabilities for the fourth grade. Reliability was measured as the ratio of sampling variance to sampling variance plus imputation variance. This approach is more suitable for multiple-matrix-sampling designs where students respond to relatively few items than classical reliability methods (such as the well-known Kuder-Richardson formulas) that are affected by the number of items taken by the student. Reliability coefficients greater than .80 are generally considered acceptable for such designs.

At both grade levels, despite some variation, reliabilities generally were high for most countries. The international median (the median of the reliability coefficients for all countries) was .96 at the eighth grade and .97 at the fourth grade for the overall mathematics assessment. At the eighth grade, the median reliabilities for the cognitive domains were .93 for knowing, .96 for applying, and .88 for reasoning. At the fourth grade, they were .92 for knowing, .93 for applying, and .91 for reasoning.

Exhibit B.2: Reliabilities of Overall Mathematics and Cognitive Domains

MATHEMATICS Grade 8

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

Exhibit B.2: Reliabilities of Overall Mathematics and Cognitive Domains

	Overall	Knowing	Applying	Reasoning
, N J DK H@ ÀŒ@ #-	0.99 ñ Kñ à ' 4H6ñ	qt¿ñN H2D ` P À à	0.99	#"+)
, QOPN@IH@	0.97	0.98	0.98	#"+$
- DIF HQ J 0ID J HOG	0.99	0.83	0.76	#"+$
. GHKD0D =@HMDH	0.94	0.88	0.89	#")+
4 31 43				
97 0.83 99 1 ø(HOÐ ` P À à				0.98

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

Correlations

Exhibit B.3 presents the Pearson correlation coefficient indicating the linear relationship between achievement in each cognitive domain and

Exhibit B.3: Correlations of Mathematics Cognitive Domains with Overall Mathematics

MATHEMATICS Grade 8

	Knowing	Applying	Reasoning
Armenia	0.85	0.87	0.79
Australia	0.89	0.90	0.81
Bahrain	0.81	0.81	0.69
Belgium (Flemish)	0.91	0.91	0.83
Botswana	0.75	0.76	0.64
Bulgaria	0.88	0.88	0.76
Chile	0.85	0.85	0.71
Chinese Taipei	0.92	0.93	0.86
Cyprus	0.86	0.87	0.76
Egypt	0.84	0.84	0.73
England	0.89	0.90	0.79
Estonia	0.88	0.89	0.79
Ghana	0.65	0.68	0.54
Hong Kong, SAR	A Ţ Ã ¾(F	(FleminAR	((FC

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

Exhibit B.3: Correlations of Mathematics Cognitive Domains with Overall Mathematics

MATHEMATICS Grade 4

Countries	Pearson Correlations of Mathematics Cognitive Domains with Overall Mathematics		
	Knowing	Applying	Reasoning
Armenia	0.81	0.84	0.77
Australia	0.86	0.87	0.84
Belgium (Flemish)	0.80	0.83	0.78
Chinese Taipei	0.82	0.84	0.81
Cyprus	0.85	0.88	0.84
England	0.87	0.89	0.85
Hong Kong, SAR	0.81	0.84	0.81
Hungary	0.85	0.88	0.83
Iran, Islamic Rep. of	0.78	0.80	0.71
Italy	0.86	0.88	0.83
Japan	0.83	0.86	0.82
Latvia	0.84	0.87	0.83
Lithuania	0.85	0.87	0.83
Moldova, Rep. of	0.85	0.88	0.83
Morocco	0.72	0.74	0.63
Netherlands	0.77	0.82	0.76
New Zealand	0.87	0.88	0.86
Norway	0.82	0.85	0.79
Philippines	0.82	0.83	0.77
Russian Federation	0.85	0.88	0.85
Scotland	0.84	0.86	0.81
Singapore	0.85	0.89	0.87
Slovenia	0.84	0.86	0.83
Tunisia	0.75	0.77	0.66
United States	0.85	0.88	0.85
International Median	0.84	0.86	0.83
Benchmark Participants:			
Ontario Province, Can.	0.84	0.86	0.83
Quebec Province, Can.	0.82	0.84	0.80
Indiana State, US	0.77	0.79	0.75

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

Exhibit B.4: Correlations of Mathematics Cognitive Domains

MATHEMATICS Grade 4

Countries	Pearson Correlations for Mathematics Cognitive Domains		
	Knowing Applying	Knowing Reasoning	Applying Reasoning
Armenia	0.84	0.74	0.86
Australia	0.92	0.89	0.91
Belgium (Flemish)	0.89	0.80	0.84
Chinese Taipei	0.92	0.87	0.91
Cyprus	0.92	0.87	0.91
England	0.94	0.89	0.91
Hong Kong, SAR	0.91	0.85	0.90
Hungary	0.90	0.82	0.89
Iran, Islamic Rep. of	0.86	0.73	0.81
Italy	0.92	0.84	0.88
Japan	0.91	0.84	0.89
Latvia	0.91	0.85	0.88
Lithuania	0.93	0.86	0.90
Moldova, Rep. of	0.89	0.82	0.89
Morocco	0.80	0.63	0.74
Netherlands	0.87	0.80	0.85
New Zealand	0.93	0.88	0.90
Norway	0.92	0.80	0.86
Philippines	0.90	0.83	0.86
Russian Federation	0.88	0.85	0.90
Scotland	0.91	0.85	0.87
Singapore	0.92	0.86	0.94
Slovenia	0.91	0.87	0.92
Tunisia	0.80	0.69	0.73
United States	0.93	0.88	0.92
International Median	0.91	0.85	0.89
Benchmark Participants:			
Ontario Province, Can.	0.91	0.87	0.90
Quebec Province, Can.	0.91	0.83	0.86
Indiana State, US	0.90	0.83	0.88

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003