

Appendix F

The Test-Curriculum Matching Analysis—Science

TIMSS went to great lengths to ensure that comparisons of student achievement across countries would be as fair and equitable as possible. The *TIMSS 2011 Assessment Frameworks* were designed to specify the important aspects of science that participating countries agreed should be the focus of an international assessment of science achievement, and the assessment items were developed through a collaborative process with national representatives to faithfully represent the specifications in the frameworks and field tested extensively in participating countries. Finalizing the TIMSS 2011 assessments involved a series of reviews by representatives of the participating countries, experts in science, and testing specialists. At the end of this process, the National Research Coordinators (NRCs) from each country formally approved the TIMSS 2011 assessments, thus accepting them as being sufficiently fair to compare their students' science achievement with that of students from other countries.

Although the assessments were developed to represent an agreed-upon framework and were intended to have as much in common across countries as possible, it was unavoidable that the match between the TIMSS 2011 assessment (or test) and the science curriculum would not be the same in all countries. To restrict test items to just those topics included in the curricula of all participating countries and covered in the same sequence would severely limit test coverage and restrict the research questions that the study is designed to address. The tests, therefore, inevitably have some items measuring topics unfamiliar to some students in some countries.

The Test-Curriculum Matching Analysis (TCMA) was conducted to investigate the extent to which the TIMSS 2011 science assessment was relevant to each country's curriculum. The TCMA also investigates the impact on a country's performance of including only achievement items that were judged to be relevant to its own curriculum.¹

To gather data about the extent to which the TIMSS 2011 tests were relevant to the curricula of the TIMSS countries and benchmarking participants, NRCs were asked to examine each achievement item and indicate whether the item was in their country's intended curriculum at the grade tested (fourth or eighth grade). The NRCs were asked to choose persons very familiar with the curriculum at these grades to make this determination. In some countries, the curriculum was prescribed for a range of grades and was not explicit about what was to be covered by the end of the fourth or eighth grades. For example, in Sweden the curriculum specifies the curricular goals to be achieved by the end of the fifth and ninth grades, but does not provide a grade-by-grade specification.

¹ Because there also may be curriculum areas covered in some countries that are not covered by the TIMSS 2011 tests, the TCMA does not provide complete information about how well the tests cover the curricula of the countries.

In such situations, coordinators were asked to make the best judgment possible.² Since an item might be in the curriculum for some but not all students in a country, coordinators were asked to consider an item included if it was in the intended curriculum for more than 50 percent of the students. All TIMSS 2011 participants took part in the TCMA analysis except Bahrain, Georgia, Saudi Arabia, Honduras (sixth grade participant), and the US benchmarking states at the fourth grade, and Bahrain, Georgia, Ghana, Indonesia, Saudi Arabia, Syrian Arab Republic, Honduras (ninth grade participant), and the US benchmarking states at the eighth grade.

Exhibits F.1 through F.4 present the TCMA results for the TIMSS 2011 science test at the fourth and eighth grades. Exhibits F.1 and F.2 show the average percent correct on the science items judged appropriate by each country at the fourth and eighth grades, respectively. Exhibits F.3 and F.4 show the standard errors corresponding to the percentages presented in Exhibits F.1 and F.2.

In Exhibit F.1, the bottom row of the exhibit shows the number of items, in terms of score points, identified as appropriate in each country. At the fourth grade, the maximum number of score points in the assessment was 181 points.³ Reading along the bottom row, it can be seen that only eight participants—Singapore, Korea, Japan, Chinese Taipei, the Russian Federation, Chile, Tunisia, and Yemen—judged less than half of the science items to be included in their curricula, although interestingly, five of the eight were among the highest performers on the TIMSS 2011 assessment. Two countries, Thailand and Armenia, judged 100 percent of the items (all 181 score points) to be included in their curricula. A further 29 countries, including one sixth grade participant, and two benchmarking participants, judged 75 percent or more (136 score points) to be appropriate.

At the eighth grade, the percentage of items judged appropriate was somewhat higher; five countries and one benchmarking participant accepted 100 percent of the items (all 233 score points), and a further 23 countries, two ninth grade participants, and two benchmarking participants judged 75 percent or more (175 score points) to be appropriate. Only Morocco, with 116 score points, judged less than half of the score points to be appropriate.

Because most countries indicated that some items were not included in their intended curriculum at the grade tested, the data were analyzed

2 Exhibit 6 of the *TIMSS 2011 Encyclopedia* provides information on the grade-to-grade structure of the science curriculum for each TIMSS 2011 participant.

3 The TIMSS 2011 fourth grade science assessment contained 172 items, yielding 184 score points. However, following item review, three items were deleted, resulting in data for reporting on 169 items and 181 score points. Similarly, following item review, the 217 items and 234 score points in the eighth grade assessment were reduced to 216 items and 233 score points.

Appendix F.2: Average Percent Correct for the Test-Curriculum Matching Analysis (Continued)

Read across the row to compare that country's performance based on the test items included by each of the countries across the top. Read down the column under a country name to compare the performance of the country down the left on the items included by the country listed on the top. Read along the diagonal to compare performance for each different country based on its own decisions about the test items to include.

										Benchmarking Participants					Average Percent Correct on All Items		Country
Malaysia	Palestinian Nat'l Auth.	Tunisia	Macedonia, Rep. of	Lebanon	Morocco	Botswana (9)	South Africa (9)	Alberta, Canada	Ontario, Canada	Quebec, Canada	Dubai, UAE	Abu Dhabi, UAE					
64	65	65	64	65	65	64	64	66	64	64	64	65	64	(0.9)	Singapore		
59	60	58	59	58	58	58	59	60	59	59	59	59	59	(0.5)	Chinese Taipei		
58	60	59	58	57	58	58	58	61	59	58	58	58	58	(0.4)	Korea, Rep. of		
58	57	57	57	56	57	57	57	58	55	57	57	57	57	(0.5)	Japan		
56	56	54	56	55	52	56	56	57	56	56	56	56	56	(0.5)	Finland		
54	55	54	54	54	54	54	54	55	54	54	54	54	54	(0.7)	Russian Federation		
54	55	55	54	55	54	54	54	55	54	54	54	54	54	(0.5)	Slovenia		
52	53	51	52	52	51	53	52	54	52	52	52	52	52	(0.7)	Hong Kong SAR		
52	52	50	52	52	50	52	52	54	52	52	52	52	52	(1.0)	England		
50	50	49	50	50	49	50	50	52	50	50	50	50	50	(0.5)	United States		
50	51	50	50	50	49	50	50	51	50	50	50	50	50	(0.6)	Hungary		
50	50	48	49	50	49	49	49	51	49	49	49	50	49	(0.8)	Israel		
49	48	46	49	48	47	49	49	51	49	49	49	49	49	(1.0)	Australia		
48	48	46	48	47	45	47	48	49	48	48	48	48	48	(0.5)	Lithuania		
48	47	46	48	48	47	47	48	49	48	48	48	48	48	(0.9)	New Zealand		
47	48	45	47	46	44	47	47	49	46	47	47	47	47	(0.5)	Sweden		
45	47	43	45	45	43	45	45	47	45	46	45	46	45	(0.4)	Italy		
46	46	45	45	45	45	45	45	46	45	46	45	46	45	(0.7)	Ukraine		
43	44	42	43	43	42	43	43	46	44	44	43	44	43	(0.5)	Norway		
43	45	44	43	43	42	43	42	44	43	42	43	44	43	(0.7)	Turkey		
43	44	44	43	43	44	42	43	43	43	43	43	43	43	(0.9)	Kazakhstan		
40	42	41	41	41	40	40	40	41	40	41	41	40	41	(0.7)	Iran, Islamic Rep. of		
39	40	39	39	40	40	39	39	40	39	40	39	39	39	(0.4)	United Arab Emirates		
39	39	38	38	39	38	38	38	39	38	39	38	39	38	(0.7)	Romania		
37	38	36	37	37	36	37	37	39	37	37	37	38	37	(0.4)	Chile		
37	38	37	37	38	37	37	37	37	37	37	37	38	37	(0.6)	Jordan		
36	37	35	36	36	34	36	35	38	36	36	36	36	36	(0.7)	Thailand		
35	37	36	35	37	34	35	34	38	35	35	35	35	35	(0.5)	Armenia		
34	35	34	34	35	33	34	33	35	32	34	34	34	34	(0.5)	Qatar		
33	35	33	33	34	33	33	33	35	33	33	33	33	33	(0.4)	Oman		
33	34	34	33	34	33	33	33	34	34	33	33	34	33	(0.9)	Malaysia		
34	35	34	33	35	33	33	33	33	32	33	33	34	33	(0.5)	Palestinian Nat'l Auth.		
33	34	33	33	33	33	33	32	34	32	33	33	33	33	(0.4)	Tunisia		
32	34	31	32	33	30	32	32	33	31	32	32	33	32	(0.8)	Macedonia, Rep. of		
30	31	31	29	31	30	29	29	30	28	29	29	30	29	(0.7)	Lebanon		
25	25	24	25	26	25	25	25	25	23	25	25	25	25	(0.2)	Morocco		
44	45	44	44	44	43	44	44	45	44	44	44	44	44	(0.1)	International Avg.		
31	32	31	30	31	29	31	30	31	31	30	30	31	30	(0.4)	Botswana (9)		
22	23	22	22	23	22	22	22	23	21	22	22	23	22	(0.3)	South Africa (9)		
Benchmarking Participants																	
54	54	51	54	53	53	54	54	59	56	54	54	54	54	(0.5)	Alberta, Canada		
49	48	46	49	48	47	49	48	52	50	49	49	49	49	(0.5)	Ontario, Canada		
48	48	47	48	48	47	48	48	51	48	48	48	48	48	(0.6)	Quebec, Canada		
44	44	43	43	44	43	44	43	45	43	43	43	44	43	(0.4)	Dubai, UAE		
39	39	38	38	39	39	38	38	40	38	39	38	39	38	(0.7)	Abu Dhabi, UAE		
220	181	124	233	162	116	219	212	154	146	205	233	200	233		Number of Items (Score Points) Identified*		

SOURCE: IEA's Trends in International Mathematics and Science Study – TIMSS 2011

to determine whether the inclusion of these items had any effect on the international performance comparisons.⁴

The first column of data in Exhibits F.1 and F.2 show the average percent correct on all test items for each participant, together with its standard error. Subsequent columns show the performance of each participant on those items judged appropriate by the participant listed at the head of the column. Participants are presented in order of their performance based on average percent correct on all items, from highest to lowest. To interpret these exhibits, choosing a country and reading across its row provides the average percent correct for the students in that country on the items selected by each of the countries listed along the top of the exhibit. For example, at the fourth grade, Singapore, where the average percent correct was 77 percent on its own set of items, had 67 percent correct on the items selected by Korea, 70 percent on the items selected by Finland, 75 percent on the items selected by Japan, and so forth. The column for a country listed at the top shows how each of the other participants performed on the set of items selected as appropriate for that country's students. Using the set of items selected by the Russian Federation as an example, 58 percent of these items, on average, were answered correctly by students in Singapore, 57 percent by students in Korea, 63 percent by students in Finland, 61 percent by students in Japan, 58 percent by those in Chinese Taipei, and so forth. The shaded diagonal element in the exhibit shows how each country performed on the set of items that it selected based on its own curriculum. Thus, Russian students averaged 68 percent correct on the set of items identified by the Russian Federation for the analysis.

For each country's selected items, the international averages across participating countries are presented in the lower part of the exhibit. These show that the selection of items by the participating countries varied somewhat in average difficulty, ranging at the fourth grade from 47 percent correct for those chosen by Singapore, Korea, Hong Kong SAR, and Croatia to 55 percent correct for those chosen by the Russian Federation. Similarly at the eighth grade, the average percent correct ranged from 42 percent for those items chosen by Singapore to 47 percent for those chosen by Jordan.

Comparing the diagonal element for a country with the overall average percent correct shows the difference between performance on the set of items chosen as appropriate for that country and performance on the test as a whole. In general, countries performed better on their own item sets than on the

4 It should be noted that the science achievement presented in Exhibits F.1 and F.2 is based on average percent correct (the percentage of students in a country, averaged across all items), which is different from the average scale scores that are presented in Chapter 1.

items overall, although usually not by much. Singapore had one of the greatest differences. The average percent correct for Singapore across all fourth grade science items was 66 percent. The diagonal element shows that Singaporean students had a greater average percent correct (77 percent) across the set of items selected as appropriate for Singapore than they did overall. However, most participants had a difference of one or two percentage points between the two performance measures. In addition to Singapore, with a difference of eleven percentage points, other exceptions included Korea (a difference of 10 points), Japan and the Russian Federation (9 points), the Slovak Republic (6 points), and Chinese Taipei and Poland (5 points). At the eighth grade, the differences were generally less; the largest being in province of Alberta (5 points) and Japan, Slovenia, and Jordan (4 percentage points).

It is clear that the selection of items does not have a major effect on the relative performance among TIMSS participants. Participants that had relatively high or low performance across all the science items also had relatively high or low performance on each of the various sets of items selected for the TCMA. For example, at the fourth grade, Singapore had the highest average percent correct not only on the test as a whole, but also on all of the different item selections, with Korea, Finland, and Japan next in order of performance on practically all selections of items. Although there are some changes in the ordering of countries based on the items selected for the TCMA, most of these differences are within the boundaries of sampling error.⁵

Even when countries performed better on the items judged by them to be included in their curriculum than they did overall, their performance relative to other participants was changed little. As an example, consider the 200 score points selected by Slovenia at the eighth grade. The students in Slovenia did better on these items (58% correct) than on the test as a whole (54% correct). However, most other countries also did better on these particular items, with an international average of 46 percent correct compared with 44 percent correct overall. In general, the TIMSS participants that performed as well or better than Slovenia on the overall test also performed as well or better on the items selected by Slovenia.

The TCMA results provide evidence that the TIMSS 2011 science assessment provides a reasonable basis for comparing achievement of the participating countries and benchmarking entities. This result is not unexpected; making the assessment as fair as possible was a major consideration in test

5 Small differences in performance between adjacent countries shown in this exhibit usually are not statistically significant. The standard errors for the average percent correct statistics based on the TIMSS 2011 sample are provided in Exhibits F.3 and F.4. For any sample average shown in Exhibits F.1 and F.2, it can be said with 95 percent confidence that the corresponding value in the population falls between the sample estimate plus or minus two standard errors.

Appendix F.4: Standard Errors for the Test-Curriculum Matching Analysis (Continued)

Read across the row to compare that country's performance based on the test items included by each of the countries across the top. Read down the column under a country name to compare the performance of the country down the left on the items included by the country listed on the top. Read along the diagonal to compare performance for each different country based on its own decisions about the test items to include.

Malaysia	Palestinian Nat'l Auth.	Tunisia	Macedonia, Rep. of	Lebanon	Morocco	Botswana (9)	South Africa (9)	Benchmarking Participants					Average Percent Correct on All Items	Country
								Alberta, Canada	Ontario, Canada	Quebec, Canada	Dubai, UAE	Abu Dhabi, UAE		
0.9	0.9	0.9	0.9	0.9	1.0	0.9	0.9	1.0	0.9	0.9	0.9	0.9	64 (0.9)	Singapore
0.5	0.5	0.6	0.5	0.5	0.6	0.5	0.5	0.5	0.5	0.5	0.5	0.5	59 (0.5)	Chinese Taipei
0.4	0.4	0.4	0.4	0.4	0.4	0.5	0.4	0.4	0.4	0.4	0.4	0.4	58 (0.4)	Korea, Rep. of
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	57 (0.5)	Japan
0.5	0.6	0.6	0.5	0.6	0.6	0.5	0.5	0.6	0.6	0.5	0.5	0.5	56 (0.5)	Finland
0.7	0.8	0.8	0.7	0.8	0.7	0.7	0.7	0.7	0.8	0.7	0.7	0.7	54 (0.7)	Russian Federation
0.5	0.6	0.6	0.5	0.6	0.6	0.5	0.5	0.6	0.6	0.5	0.5	0.5	54 (0.5)	Slovenia
0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	52 (0.7)	Hong Kong SAR
1.0	1.0	1.1	1.0	1.0	1.1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	52 (1.0)	England
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	50 (0.5)	United States
0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	50 (0.6)	Hungary
0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	49 (0.8)	Israel
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	49 (1.0)	Australia
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	48 (0.5)	Lithuania
0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	48 (0.9)	New Zealand
0.5	0.5	0.5	0.5	0.5	0.6	0.5	0.5	0.5	0.5	0.5	0.5	0.5	47 (0.5)	Sweden
0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	45 (0.4)	Italy
0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	45 (0.7)	Ukraine
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	43 (0.5)	Norway
0.6	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.6	0.7	0.7	0.7	0.6	43 (0.7)	Turkey
0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	43 (0.9)	Kazakhstan
0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	41 (0.7)	Iran, Islamic Rep. of
0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	39 (0.4)	United Arab Emirates
0.7	0.7	0.7	0.7	0.7	0.6	0.7	0.7	0.6	0.7	0.7	0.6	0.6	38 (0.7)	Romania
0.4	0.4	0.4	0.4	0.4	0.5	0.4	0.4	0.4	0.4	0.4	0.4	0.4	37 (0.4)	Chile
0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	37 (0.6)	Jordan
0.7	0.7	0.7	0.7	0.7	0.8	0.7	0.7	0.7	0.7	0.7	0.7	0.7	36 (0.7)	Thailand
0.5	0.6	0.6	0.5	0.6	0.5	0.5	0.5	0.5	0.6	0.5	0.5	0.5	35 (0.5)	Armenia
0.6	0.5	0.5	0.5	0.5	0.6	0.6	0.6	0.5	0.6	0.6	0.5	0.6	34 (0.5)	Qatar
0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	33 (0.4)	Oman
0.9	1.0	1.0	0.9	0.9	1.0	1.0	0.9	0.9	0.9	0.9	0.9	0.9	33 (0.9)	Malaysia
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	33 (0.5)	Palestinian Nat'l Auth.
0.4	0.4	0.4	0.4	0.4	0.4	0.5	0.4	0.4	0.4	0.4	0.4	0.4	33 (0.4)	Tunisia
0.8	0.8	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	32 (0.8)	Macedonia, Rep. of
0.7	0.7	0.7	0.7	0.7	0.8	0.7	0.7	0.7	0.7	0.7	0.7	0.7	29 (0.7)	Lebanon
0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	25 (0.2)	Morocco
0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	44 (0.1)	International Avg.
0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	30 (0.4)	Botswana (9)
0.3	0.3	0.4	0.3	0.4	0.4	0.3	0.3	0.4	0.4	0.3	0.3	0.3	22 (0.3)	South Africa (9)
Benchmarking Participants														
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	54 (0.5)	Alberta, Canada
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	49 (0.5)	Ontario, Canada
0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	48 (0.6)	Quebec, Canada
0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	43 (0.4)	Dubai, UAE
0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	38 (0.7)	Abu Dhabi, UAE
Benchmarking Participants														
220	181	124	233	162	116	219	212	154	146	205	233	200	233	Number of Items (Score Points) Identified*

SOURCE: IEA's Trends in International Mathematics and Science Study—TIMSS 2011

development. The fact that the majority of countries indicated that most items were appropriate for their students means that the different average percent correct estimates were based on many of the same items. Insofar as countries rejected items that would be difficult for their students, these items tended to be difficult for students in other countries as well. The analysis shows that omitting such items tends to improve the results for that country, but also tends to improve the results for all other countries, so that the overall pattern of relative performance is largely unaffected.

