

Appendix B

THE TEST-CURRICULUM MATCHING ANALYSIS

When comparing student achievement across countries, it is important that the comparisons be as “fair” as possible. TIMSS has worked towards this goal in a number of ways, including providing detailed procedures for standardizing the population definitions, sampling, test translations, test administration, scoring, and database formation. Developing the TIMSS tests involved the interaction of experts in the sciences with representatives of the participating countries and testing specialists.¹ The National Research Coordinators (NRCs) from each country formally approved the TIMSS test, thus accepting it as being sufficiently fair to compare their students’ science achievement with that of students from other countries.

Although the TIMSS test was developed to represent a set of agreed-upon science content areas, there are differences among the curricula of participating countries that result in various science topics being taught at different grades. To restrict test items not only to those topics in the curricula of all countries but also to those covered in the same sequence in all participating countries would severely limit test coverage and restrict the research questions about international differences that TIMSS is designed to address. The TIMSS tests, therefore, inevitably contain some items measuring topics unfamiliar to some students in some countries.

The Test-Curriculum Matching Analysis (TCMA) was developed and conducted to investigate the appropriateness of the TIMSS science test for seventh- and eighth-grade students in the participating countries, and to show how student performance for individual countries varied when based only on the test questions that were judged to be relevant to their own curriculum.²

To gather data about the extent to which the TIMSS tests were relevant to the curriculum of the participating countries, TIMSS asked the NRC of each country to report whether or not each item was in their country’s intended curriculum at each of the two grades being tested. The NRC was asked to choose a person or persons who were very familiar with the curricula at the grades being tested to make the determination. Since an item might be in the curriculum for some but not all students in a country, an item was determined appropriate if it was in the intended curriculum for more than 50% of the students. The NRCs had considerable flexibility in selecting items and may have considered items inappropriate for other reasons. All participating countries except Thailand returned the information for analysis.

Tables B.1 and B.2 present the TCMA results for the eighth and seventh grades, respectively. The first row of each table indicates that at both grades the countries varied substantially in the number of items considered appropriate. At the eighth

¹ See Appendix A for more information on the test development.

² Because there also may be curriculum areas covered in some countries that are not covered by the TIMSS tests, the TCMA does not provide complete information about how well the TIMSS tests cover the curricula of the countries.

grade, more than half of the countries indicated that items representing three-quarters or more of the score points (110 out of a possible 146) were appropriate,³ with the percent ranging from 100% in Spain, Iceland, and the United States to approximately 40% in Korea (59 score points) and French-speaking Belgium (58 score points). Fewer items were selected at the seventh grade, but nearly half of the countries selected at least 60%, with several selecting at least three-quarters of the score points. All items were selected at the seventh grade as well as the eighth grade in both the United States and Iceland. At the seventh grade there were also several countries, including Korea and Japan, which retained about 30% or less. That lower percentages of items were selected for the TCMA at the seventh grade is consistent with the instrument-development process, which put more emphasis on the upper-grade curriculum.

Since most countries indicated that some items were not included in their intended curricula at the two grades tested, the question becomes whether the inclusion of these items had any effect on the international performance comparisons.⁴ The TCMA results provide a method for answering this question, providing evidence that it is reasonable to make cross-national comparisons on the basis of the TIMSS science test.

Each of the first columns in Tables B.1 and B.2 shows the overall average percent correct for each country (as discussed in Chapter 2 and reproduced here for convenience in making comparisons). The countries are presented in the order of their overall performance, from highest to lowest. To interpret these tables, reading across a row provides the average percent correct for the students in the country identified by that row on the items selected by each of the countries named across the top of the table. For example, at the eighth grade, Singapore, where the average percent correct was 72% on its own set of items, also had 72% for the items selected by Korea, 73% for those selected by Japan, 69% for those selected by the Czech Republic, and so forth. The column for a country shows how each of the other countries performed on the subset of items selected for its own students. Using the set of items selected by Hong Kong as an example, on average, 71% of these items were answered correctly by the Singaporean students, 65% by the Korean students, 66% by the Japanese, and so forth. The shaded diagonal elements in each table show how each country performed on the subset of items that it selected based on its own curriculum. Thus, the Hong Kong students themselves averaged 59% correct responses on the items identified by Hong Kong for the analysis.

³ Of the 135 items in the test, some items were assigned more score points than others. In particular, some items had two parts, and some extended-response items were scored on a two-point scale and others on a three-point scale. The total number of score points available for analysis was 146. The TCMA uses the score points in order to give the same weight to items that they received in the test scoring.

⁴ It should be noted that the performance levels presented in Tables B.1 and B.2 are based on average percents correct as was done in Chapter 2, which is different from the average scale scores that were presented in Chapter 1. The cost and delay of scaling would have been prohibitive for the TCMA analyses.

Table B.2 Test-Curriculum Matching Analysis Results - Science - Lower Grade (Seventh Grade*) Average Percent Correct Based on Subsets of Items Specially Identified by Each Country as Addressing Its Curriculum (See Table B.4 for corresponding standard errors)

Instructions: 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.

Table with columns for Country and various countries (Singapore, Korea, Japan, Czech Republic, Slovenia, Belgium (Fl), Bulgaria, Netherlands, England, Hungary, Austria, Slovak Republic, United States, Canada, Australia, Hong Kong, Germany, Ireland, Sweden, New Zealand, Norway, Switzerland, Russian Federation, Spain, Scotland, Iceland, France, Belgium (Fr), Romania, Greece, Denmark, Iran, Islamic Rep., Latvia (LSS), Portugal, Cyprus, Colombia, Lithuania, South Africa). Rows list countries: Singapore, Korea, Japan, Czech Republic, Slovenia, Belgium (Fl), Bulgaria, Netherlands, England, Hungary, Austria, Slovak Republic, United States, Canada, Australia, Hong Kong, Germany, Ireland, Sweden, New Zealand, Norway, Switzerland, Russian Federation, Spain, Scotland, Iceland, France, Belgium (Fr), Romania, Greece, Denmark, Iran, Islamic Rep., Latvia (LSS), Portugal, Cyprus, Colombia, Lithuania, South Africa, and International Average.

*Seventh grade in most countries; see Table 2 for more information about the grades tested in each country.

**Of the 135 items in the science test, some items had two parts and some extended-response items were scored on a two- or three-point scale, resulting in 146 total score points.

() Standard errors for the average percent of correct responses on all items appear in parentheses. Standard errors for scores based on subsets of items are provided in Table B.4.

Because results are rounded to the nearest whole number, some totals may appear inconsistent.

Countries shown in italics did not satisfy one or more guidelines for sample participation rates, age/grade specifications, or classroom sampling procedures (see Figure A.3 for details).

Because population coverage falls below 65% Latvia is annotated LSS for Latvian Speaking Schools only.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

The international averages of each country's selected items presented across the last row of the tables show that the selection of items for the participating countries varied somewhat in average difficulty, ranging from 55% to 59% at the eighth grade and from 49% to 56% at seventh grade. Despite these differences, the overall picture provided by both Tables B.1 and B.2 reveals that different item selections do not make a major difference in how well countries perform relative to each other. The items selected by some countries were more difficult than those selected by others. The relative performance of countries on the various item selections did vary somewhat, but generally not in a statistically significant manner.⁵

Comparing the diagonal element for a country with the overall average percentage correct shows the difference between performance on this subset of items and performance on the test as a whole. In general, there were only small increases in each country's performance on its own subset of items. To illustrate, the average percent correct for eighth-grade students in Singapore was 70%. The diagonal element shows that Singaporean students had about the same average percent correct (72%) based on the smaller set of items selected as relevant to the curriculum in Singapore as they did overall. In the eighth grade, most countries had a difference of less than 5 percentage points between the two performance measures, with the largest difference of 7% for the Russian Federation (65% compared to 58%). Performance differences between the entire TIMSS test and the subset of items selected for the TCMA were, in general, somewhat larger for seventh-grade students, including a few countries with an average performance that was about 10 percentage points higher on the subsets of items selected for the TCMA for their own students – Switzerland, France, and the Russian Federation. Even these increases are not particularly large, however, considering that France and Switzerland both selected less than one-quarter of the items at the seventh grade.

It is clear that the selection of items does not have a major effect on the general relationship among countries. Countries that had substantially higher or lower performance on the overall test in comparison to each other also had higher or lower relative performance on the different sets of items selected for the TCMA. For example, at the eighth grade, Singapore had the highest average percent correct on the test as a whole and on all of the different item selections, with Japan, Korea, and the Czech Republic among the four highest-performing countries in all cases. 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. As the most extreme example, consider the 49 score points selected by the Russian Federation for the seventh grade. The Russian students did substantially better on these items than on the test as a whole, with 61% correct responses to these items, on average, compared to 50% average correct on the items on the test as a whole.

⁵ Small differences in performance in these tables are not statistically significant. The standard errors for the estimated average percent correct statistics can be found in Tables B.3 and B.4. We can say with 95% confidence that the value for the entire population will fall between the sample estimate plus or minus two standard errors.

However, all other countries also did better on these particular items, with an international average of 54% for the items selected by the Russian Federation compared with 50% on the test as a whole. Only 8 of the 22 countries that performed better than the Russian students on the overall test also did so on the items selected by the Russian Federation. However, 10 countries with the same or higher overall performance were within 5 percentage points of the Russian students on these items.

The TCMA results provide evidence that the TIMSS science test provides a reasonable basis for comparing achievement for the participating countries. This result is not unexpected, since making the test as fair as possible was a major consideration in test 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 substantially on the same items. Insofar as countries rejected items that would be difficult for their own 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 results is largely unaffected.

Table B.4 Standard Errors for the Test-Curriculum Matching Analysis Results - Science - Lower Grade (Seventh Grade*)

See Table B.3 for the Test-Curriculum Matching Analysis Results

Instructions: Read across the row for the standard error for the score based on the test items included by each of the countries across the top. Read down the column under a country name for the standard error for the score of the country down the left on the items included by the country listed on the top. Read along the diagonal for the standard error for the score for each different country based on its own decisions about the test items to include.

Table with 24 columns (Country, Average Percent Correct on All Items, Number of Score Points Included, and 135 countries) and 24 rows (Singapore, Korea, Japan, Czech Republic, Slovenia, Belgium (Fl), Bulgaria, Netherlands, England, Hungary, Austria, Slovak Republic, United States, Canada, Australia, Hong Kong, Germany, Ireland, Sweden, New Zealand, Norway, Switzerland, Russian Federation, Spain, Scotland, France, Iceland, Belgium (Fr), Romania, Denmark, Latvia (LSS), Portugal, Cyprus, Lithuania, Colombia, South Africa). Values are standard errors, with some cells containing bolded values and parentheses indicating standard error.

*Seventh grade in most countries; see Table 2 for more information about the grades tested in each country.

**Of the 135 items in the science test, some items had two parts and some extended-response items were scored on a two- or three-point scale, resulting in 146 total score points.

() Standard errors for the average percent of correct responses on all items appear in parentheses. The matrix contains standard errors corresponding to the average percent of correct responses based on TCMA subsets of items, as displayed in Table B.2. Because results are rounded to the nearest whole number, some totals may appear inconsistent.

Countries shown in italics did not satisfy one or more guidelines for sample participation rates, age/grade specifications, or classroom sampling procedures (see Figure A.3 for details). Because population coverage falls below 65% Latvia is annotated LSS for Latvian Speaking Schools only.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.