

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 field of mathematics 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’ mathematics achievement with that of students from other countries.

Although the TIMSS test was developed to represent a set of agreed-upon mathematics content areas, there are differences among the curricula of participating countries that result in various mathematics 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 mathematics 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 the 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.

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.

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 grade, half of the countries indicated that items representing 90% or more of the score points (145 out of a possible 162) were appropriate,³ with the percent ranging from 100% in Hungary and the United States to 47% (76 score points) in Greece. Although, in general, fewer items were selected at the seventh grade than at the eighth grade, nearly half of the countries selected items representing at least three-quarters of the score points (121), and several countries selected items representing 90% or more. The number of score points represented by the selected items for the seventh grade ranged from 59 (36%) in Denmark to 162 (100%) in the United States. That somewhat 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 mathematics 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, eighth-grade Korean students had an average of 71% correct on the items that Singapore selected as appropriate for the Singaporean students, an average of 72% percent correct on the items selected for the Japanese students, 73% correct for its own items, 72% on the items selected by Hong Kong, 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 Switzerland as an example, on average, 80% of these items were answered correctly by the Singaporean students, 75% by the Japanese students, 72% by the students from Hong Kong, 71% by the Belgian (Flemish) students, and so forth. The shaded diagonal elements in

³ Of the 151 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 162. The TCMA uses the score points in order to give the same importance 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.

each table show how each country performed on the subset of items that it selected based on its own curriculum. Thus, the Swiss students themselves averaged 64% correct responses on the items identified by Switzerland for the analysis.

The international averages 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 54% to 58% at the eighth grade and from 48% to 61% at the 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 small increases in each country's performance on its own subset of items. To illustrate, the average percent correct for eighth-grade students in the Russian Federation is 60%. The diagonal element shows that Russian students had about the same average percent correct (62%) based on the smaller set of items selected as relevant to the curriculum in the Russian Federation as they did overall. In the eighth grade, the differences were extremely small (2 average percentage points or less) for most countries. Only a few countries had an average percent correct on their own selected items more than 3 percentage points higher than their average on the test as a whole. 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 several countries with average performance that was 5 to 10 percentage points higher on the items selected for the TCMA for their own students. The largest increase (16 average percentage points) was for the seventh-grade students in Denmark.

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. At the eighth grade, Singapore, Japan, Korea, and Hong Kong were the highest-performing countries and in the same order of performance, both on the test as a whole and on all the different sets of item selections. At the seventh 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, Hong Kong, and Belgium (Flemish) among the top five highest performing countries in all cases. Although there were some changes in

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.

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 59 score points selected by Denmark for the seventh grade. Denmark did substantially better on these items than on the test as a whole, with 60% correct responses to these items, on average, compared to only 44% average correct on the test as a whole. However, all other countries also did better on these particular items, with an international average of 61% for the items selected by Denmark compared with 49% on the test as a whole. Also, for example, Scotland, Norway, and Latvia (LSS), which also averaged 44% correct over all items at the seventh grade, performed similarly to Denmark on the set of items selected by Denmark – 58%, 59%, and 56%, respectively.

The TCMA results provide evidence that the TIMSS mathematics 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 improves 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.

