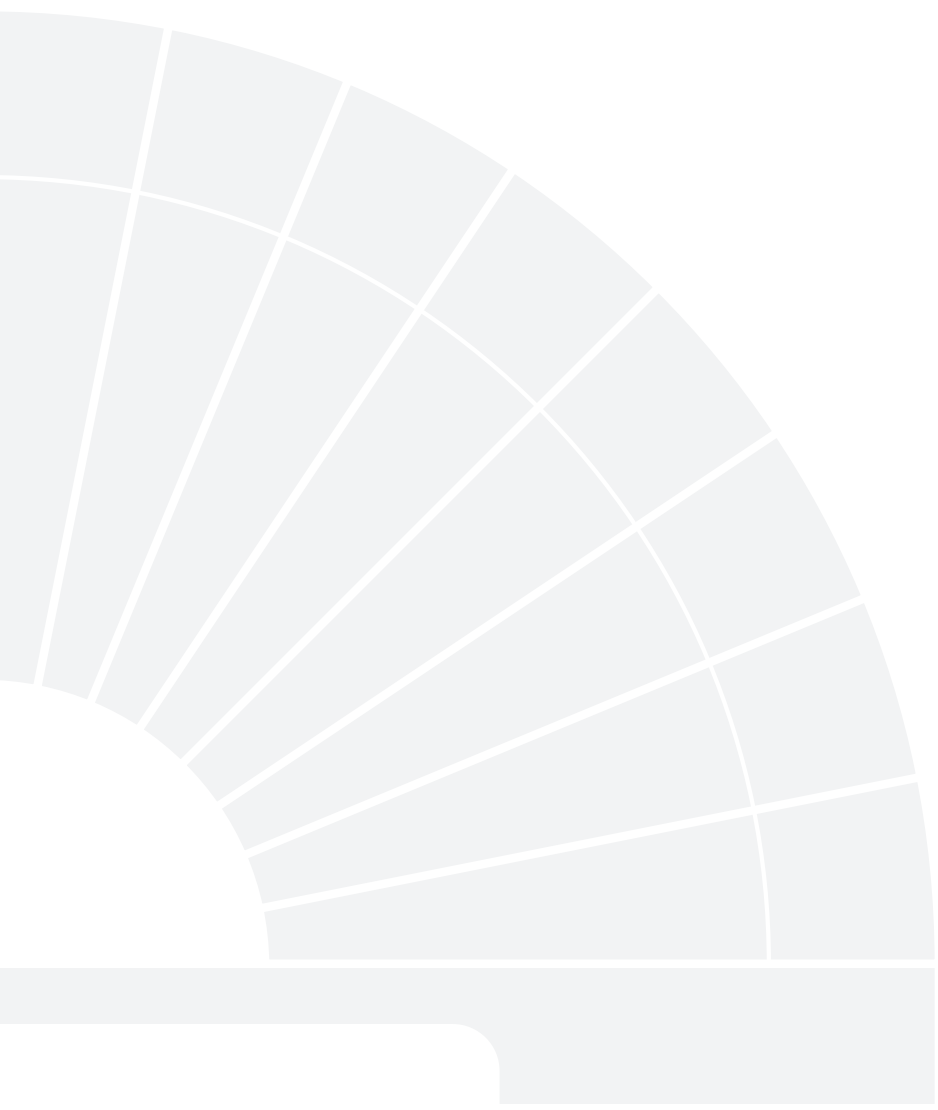


The Test-Curriculum
Matching Analysis:
Mathematics





When comparing student achievement across countries, it is important that the comparisons be as fair as possible. LSS has worked toward 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. Similar to the procedures used for developing the original TIMSS instruments, developing the TIMSS 1999 tests involved a series of reviews by representatives of the participating countries, experts in mathematics, and testing specialists.¹ The National Research Coordinators (NRCS) from each country formally approved the TIMSS 1999 tests, thus accepting them as being sufficiently fair to compare their students' mathematics achievement with that of students from other countries.

Although the tests were developed to represent a set of agreed-upon mathematics content, differences among the curricula of participating countries result in various topics being taught at different grades. To restrict test items to 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 appropriateness of the TIMSS 1999 mathematics test for the eighth-grade students in the participating countries. TCMA also shows how student performance for individual countries varies when based only on the test questions that are judged to be relevant to their own curricula.²

To gather data about the extent to which the TIMSS 1999 tests were relevant to the curricula of the participating countries, each NRC reported whether each item was in that country's intended curriculum at the grade tested. The NRC was asked to choose a person or persons who were very familiar with the curriculum at the grade tested to make this 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 percent of the students. The NRCS had considerable flexibility in selecting items and may have considered items inappropriate for other reasons. All participating countries returned the information for analysis.

Exhibits C.1 and C.2 present the TCMA results for the TIMSS 1999 tests. Exhibit C.1 shows the average percent correct for each country on items selected as appropriate and on the test as a whole. Exhibit C.2 shows the standard errors corresponding to the percentages presented in Exhibit C.1.



¹ See Appendix A for more information on test development.

² Because there may also be curriculum areas covered in some countries that are not covered by the TIMSS 1999 tests, the TCMA does not provide complete information about how well the tests cover the curricula of the countries.

In Exhibit C.1, the last row of the exhibit indicates that the countries varied substantially in the number of items (score points) identified as appropriate.³ The percentages ranged from 100 percent (169 score points) in Chinese Taipei, the Slovak Republic, Latvia (LSS), the United States, Lithuania, Moldova, and Indonesia to 58 percent (98 score points) in Chile. Thirty-four of the 38 countries indicated that the items representing three-quarters or more of the score points (127 out of a possible 169) were appropriate.

Since most countries indicated that some items were not included in their intended curriculum at the grade tested, the data were analyzed to determine whether the inclusion of these items had any effect on the international performance comparisons.⁴


The first column in Exhibit C.1 shows the average percent correct on all test items for each country. The countries are presented in order of their overall performance based on overall percent correct, from highest to lowest. To interpret this exhibit, reading across a row provides the average percent correct for the students in that country on the items selected by each of the countries listed across the top of the exhibit. For example, Singapore, where the average percent correct was 77 percent on its own set of items, also had 78 percent correct for the items selected by Korea, 77 percent for the items selected by Hong Kong, and so forth. The column for a country listed across the top shows how each of the other countries performed on the subset of items selected as appropriate for its own students. Using the set of items selected by Finland as an example, on average 77 percent of these items were answered correctly by students in Singapore, 73 percent by students in Korea, 72 percent by those in Hong Kong, and so forth. The shaded diagonal element in the exhibit shows how each country performed on the subset of items that it selected based on its own curriculum. Thus, Finnish students averaged 56 percent correct on the set of items identified by Finland for the analysis.

The international averages of each country's selected items are presented across the second to the last row of the exhibit. They show that the selection of items for the participating countries varied somewhat in average difficulty, ranging from 48 to 54 percent. Despite these differences, the overall picture presented by Exhibit C.1 reveals that different item selections do not make a major difference in how well countries perform relative to one another. The items selected by some countries were more difficult than those selected by others. The relative performance of countries on various item selections did vary somewhat, but generally not in a statistically significant manner.⁵

³ Of the 162 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. The total number of score points available for analysis was 169. The TCMA uses score points in order to give the same weight to items given them in test scoring.

⁴ It should be noted that the performance levels presented in Exhibit C.1 are based on average percents, which are different from the average scale scores that are presented in Chapter 1.

⁵ Small differences in performance shown in this exhibit are not statistically significant. The standard errors for the estimated average percent correct statistics are in Exhibit C.2. It can be said with 95 percent confidence that the value for the entire population falls between the sample estimate plus or minus two standard errors.



Comparing the diagonal element for a country with the overall average percent correct shows the difference between performance on the subset of items chosen as appropriate 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 Singapore was 77 percent. The diagonal element shows that Singaporean students had the same percent correct (77 percent) based on the smaller set of items selected as they did overall. All countries had a difference of less than five percentage points between the two performance measures, with the largest difference four percent for the Netherlands (65 percent compared with 61 percent).

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 relative performance on all items also had higher or lower relative performance on the different sets of items selected for the TCMA. For example, Singapore had the highest average percent correct on the test as a whole and on most of the different item selections, with Korea, Hong Kong, and Chinese Taipei 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 an example, consider the 149 score points selected by Jordan. The Jordanian students did better on these items than on the test as a whole, with 41 percent correct on these items, on average, compared with 38 percent correct on all items. However, most other countries also did better on these particular items, with an international average of 54 percent correct on the items selected by Jordan. All 30 countries that performed better than Jordan on the overall test also performed better on the items selected by Jordan.

The TCMA results provide evidence that the TIMSS 1999 mathematics test provides a reasonable basis for comparing achievement of 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 on essentially 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 results is largely unaffected.

