Appendix C

The Test-Curriculum Matching Analysis: Science

To ensure that comparisons of student achievement across countries would be as fair and equitable as possible, TIMSS developed extensive assessment frameworks and specifications that addressed the important aspects of science in countries' curricula and instructional programs, and went to great lengths to develop assessment items that faithfully represented those specifications. Similar to the procedures used for developing the original TIMSS instruments, developing the TIMSS 2003 tests involved a series of reviews by representatives of the participating countries, experts in science, and testing specialists.[1] The National Research Coordinators (NRCs) from each country formally approved the TIMSS 2003 tests, thus accepting them as being sufficiently fair to compare their students' science achievement with that of students from other countries.

Although the tests were developed to represent an agreed-upon framework and were intended to have as much in common across

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 2003 science test for the eighth- and fourth-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 2003 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 (eighth or fourth grade in most countries). The NRC was asked to choose a person or persons who were very familiar with the curriculum at these grades to make this determination. Since an item might be in the curriculum for some but not all students in a country, an item was to be 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 participants returned the information for analysis except Syria at eighth grade and Yemen at fourth grade.

Exhibits C.1 and C.2 present the TCMA results for the TIMSS 2003 tests at eighth and fourth grades. Exhibit C.1 shows the average percent correct on the science items selected as appropriate by each country. Exhibit C.2 shows the standard errors corresponding to the percentages presented in Exhibit C.1.

In Exhibit C.1, the last row of the exhibit shows the number of items (score points) identified as appropriate in each country.[3] At the eighth grade, the percentage of score points ranged from 100 percent (206 score points) in Israel and Saudi Arabia to 31 percent (63 score points) in Belgium (Flemish). Generally, the proportion of items judged

2 2003 ,

3 . , , - .

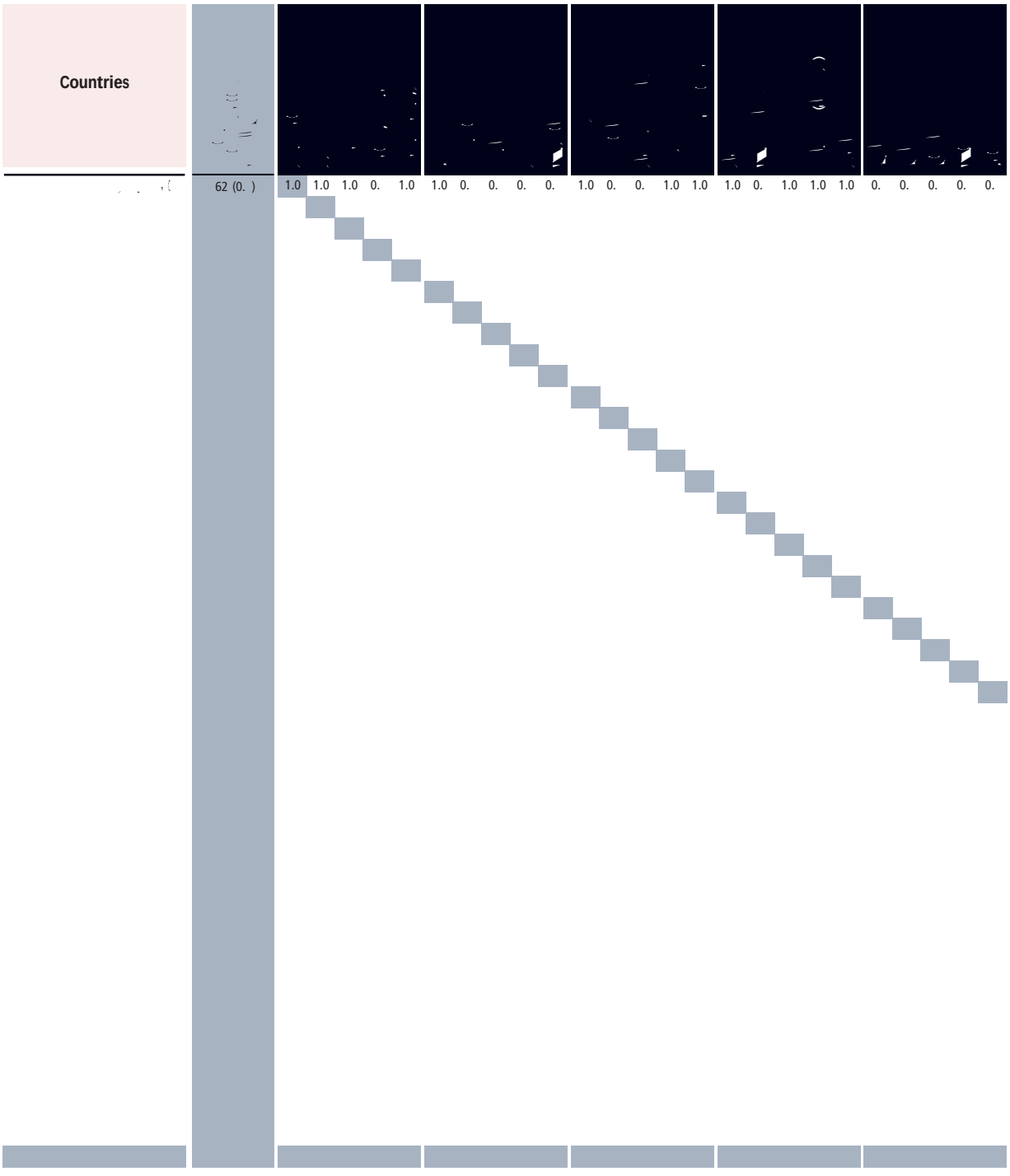
appropriate was high, with 40 of the 50 participants indicating that items representing three-quarters or more of the score points (154 out of a possible 206) were appropriate. Only Belgium (Flemish) and Chile considered less than 50 percent of the score points appropriate. At the fourth grade, the percentage of score points ranged from 98 percent (161 score points) in Hungary, the United States, Latvia, Lithuania, Moldova, and Armenia to 27 percent (44 score points) in Japan. Eighteen of the 28 fourth-grade participants indicated that items representing three-quarters or more of the score points (124 out of a possible 165) 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]

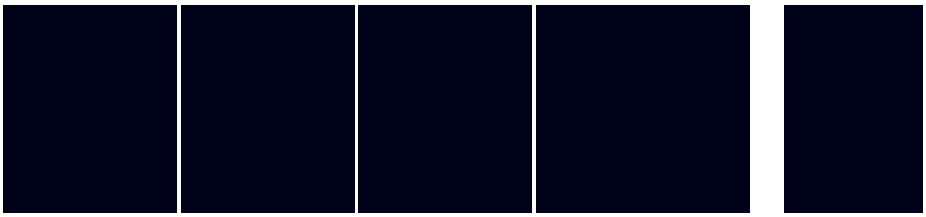
The first column in Exhibit C.1 shows the average percent correct on all test items for each participant. Subsequent columns show the performance of each participant on those items judged appropriate by the participant listed at the head of the column. Participants are presented in order of their performance based on average percent correct on all items, from highest to lowest. To interpret 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, at the eighth-grade, Singapore, where the average percent correct was 65 percent on its own set of items, had 63 percent correct for the items selected by Chinese Taipei, 63 percent for the items selected by Estonia, 62 percent for the items selected by Korea, and so forth. The column for a country listed across the top shows how each of the other participants performed on the subset of items selected as appropriate for that country's students. Using the set of items selected by Bulgaria as an example, on average, 63 percent of these items were answered correctly by students in Singapore, 62 percent by students in Chinese Taipei, 58 percent by students in Estonia, 58 percent by those in Korea, and so forth. The shaded

4 .1 1.

across down diagonal



t acros s



It is clear that the selection of items does not have a major effect on the general relationship among countries. Countries 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 eighth grade, Singapore had the highest average percent correct on the test as a whole and on all of the different item selections, with Chinese Taipei, Estonia, and Korea 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. As an example, consider the 195 score points selected by Armenia. The students in Armenia did a little better on these items than on the test as a whole, with 39 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 44 percent correct on the items selected by Armenia. All 29 participants that performed better than Armenia on the overall test also performed better on the items selected by Armenia.

The TCMA results provide evidence that the TIMSS 2003 science test provides a reasonable basis for comparing achievement of the participating countries and benchmarking entities. 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 . As ejected ageas .w