Chapter 10

Reviewing the TIMSS 2007 Item Statistics

John F. Olson, Michael O. Martin, Ina V.S. Mullis, Pierre Foy, Ebru Erberber, and Corinna Preuschoff

10.1 Overview

For TIMSS 2007, similar to the process used in TIMSS 2003 and previous assessments, the TIMSS & PIRLS International Study Center conducted a review of a range of diagnostic statistics to examine and evaluate the psychometric characteristics of each achievement item in the 59 countries and 8 benchmarking participants that participated in TIMSS 2007. This review of item statistics was conducted before applying item response theory (IRT) scaling to the TIMSS 2007 achievement data to derive student mathematics and science achievement scores for analysis and reporting. The review of item statistics played a crucial role in the quality assurance of the TIMSS 2007 data, enabling the detection of unusual item properties that could signal a problem or error for a particular country. For example, an item that was uncharacteristically easy or difficult, or had an unusually low discriminating power, could indicate a potential problem with either translation or printing. Similarly, a constructed-response item with unusually low scoring reliability could indicate a problem with a scoring guide in a particular country. In the rare instances where such items were found, the country's translation verification documents and printed booklets were examined for flaws or inaccuracies and, if necessary, the item was removed from the international database for that country.

This chapter describes the basic item statistics that were calculated, the review criteria that were applied, statistics on the different types of reliability that were analyzed, and a summary of the reviews of the TIMSS 2007 item statistics. Examples from the TIMSS 2007 assessment are provided to illustrate the review process.

10.2 Statistics for Item Analysis

To begin the review process, the TIMSS & PIRLS International Study Center computed item analysis statistics for all 353 mathematics and science achievement items at the fourth grade and 429 items at the eighth grade that were administered in the TIMSS 2007 assessment. The properties of the items in each of the 59 countries and 8 benchmarking entities that participated were then carefully reviewed. Exhibits 10.1 and 10.2 show actual samples of the statistics calculated for a multiple-choice and a constructed-response item, respectively.

Exhibit 10.1 International Item Statistics for a Multiple-choice Item

```
Trends in International Mathematics and Science Study - TIMSS 2007 Assessment Results
International Item Statistics (Unweighted) - Review Version -  8th Grade
For Internal Review Only: DO NOT CITE OR CIRCULATE

Mathematics: Algebra / Knowing (M042077 - M10_06)
Label: Expression to equivalent to 4(3+x)
 Type: MC    Key: C

------------------------------------------------------------------------------------------------------------------------------------------------------------------
                         |                   |                      Percentages                   |                 Point Biserials                |         |
Country                  |    N  Diff  Disc  | Pct_A  Pct_B  Pct_C  Pct_D  Pct_E  Pct_OM Pct_NR  |  PB_A   PB_B   PB_C   PB_D   PB_E   PB_OM    |  RDIFF  |  Flags
------------------------------------------------------------------------------------------------------------------------------------------------------------------
Algeria                  |  752  16.2  0.11  |  40.2   19.8   16.2   19.3    .     4.5    2.0   |  0.03  -0.14   0.11   0.01    .    -0.06    |   0.36  | DCH___
Armenia                  |  683  73.2  0.46  |  15.4    5.0   73.2    3.7    .     2.8    0.4   | -0.29  -0.19   0.45  -0.19    .    -0.08    |  -1.01  | ____F_
Australia                |  578  31.1  0.40  |  32.0   19.0   31.1   15.6    .     2.2    0.3   | -0.11  -0.28   0.40  -0.01    .    -0.14    |   0.97  | __H___
Bahrain                  |  597  48.4  0.47  |  22.1   13.4   48.4   14.9    .     1.2    0.2   | -0.15  -0.31   0.47  -0.16    .    -0.10    |  -1.11  | __E___
Bosnia and Herzegovin    |  606  58.1  0.51  |  22.9    5.8   58.1   11.1    .     2.1    0.0   | -0.30  -0.28   0.51  -0.14    .    -0.13    |  -0.97  | ____F_
Botswana                 |  586  24.2  0.36  |  30.9   28.3   24.2   15.4    .     1.2    2.0   |  0.04  -0.30   0.36  -0.10    .    -0.05    |  -0.13  | _CH___
Bulgaria                 |  576  73.6  0.55  |  14.8    5.7   73.6    3.8    .     2.1    0.9   | -0.32  -0.29   0.55  -0.21    .    -0.12    |  -1.44  | __E_F_
Chinese Taipei           |  572  85.8  0.61  |   7.5    2.3   85.8    3.7    .     0.7    0.0   | -0.42  -0.30   0.61  -0.28    .    -0.06    |  -1.05  | ____F_
Colombia                 |  682  26.0  0.25  |  36.2   15.1   26.0   21.1    .     1.6    2.7   | -0.14  -0.17   0.25   0.06    .    -0.06    |  -0.29  | ______
Cyprus                   |  622  56.9  0.47  |  22.0    7.6   56.9   12.2    .     1.3    0.0   | -0.27  -0.28   0.47  -0.11    .    -0.11    |  -0.73  | ____F_
Czech Republic           |  693  59.2  0.49  |  20.3    5.2   59.2   15.2    .     0.1    0.0   | -0.33  -0.20   0.49  -0.18    .    -0.05    |  -0.38  | ____F_
Egypt                    |  945  56.9  0.42  |  19.3   11.9   56.9   11.3    .     0.6    0.0   | -0.16  -0.29   0.42  -0.15    .    -0.07    |  -1.28  | __E___
El Salvador              |  568  24.6  0.32  |  34.9   19.9   24.6   19.0    .     1.6    1.4   | -0.10  -0.22   0.32   0.01    .    -0.07    |  -0.60  | _C____
England                  |  571  46.1  0.59  |  20.8   18.4   46.1   13.7    .     1.1    0.0   | -0.16  -0.46   0.59  -0.13    .    -0.09    |   0.53  | __H___
Georgia                  |  616  55.7  0.51  |  23.9    7.1   55.7   11.2    .     2.1    1.1   | -0.30  -0.22   0.51  -0.16    .    -0.09    |  -1.19  | __E_F_
Ghana                    |  751  42.1  0.46  |  24.5   18.0   42.1   13.8    .     1.6    1.1   | -0.14  -0.27   0.46  -0.12    .    -0.08    |  -1.45  | __E___
Hong Kong SAR            |  502  87.5  0.51  |   5.6    2.8   87.5    3.8    .     0.4    0.0   | -0.37  -0.32   0.51  -0.13    .    -0.10    |  -1.39  | __E_F_
Hungary                  |  599  63.4  0.60  |  15.7    6.5   63.4   13.4    .     1.0    0.2   | -0.30  -0.30   0.60  -0.26    .    -0.13    |  -0.45  | ____F_
Indonesia                |  591  36.9  0.47  |  28.4   13.7   36.9   20.3    .     0.7    0.5   | -0.16  -0.26   0.47  -0.14    .     0.02    |  -0.50  | ______
Iran, Islamic Rep. of    |  574  40.9  0.45  |  30.7    7.1   40.9   20.7    .     0.5    0.2   | -0.18  -0.18   0.45  -0.21    .    -0.09    |  -0.61  | ____F_
Israel                   |  453  71.3  0.52  |  14.8    4.0   71.3    9.3    .     0.7    1.5   | -0.27  -0.26   0.52  -0.22    .    -0.14    |  -1.48  | __E_F_
Italy                    |  626  48.1  0.38  |  26.0    6.9   48.1   17.9    .     1.1    0.5   | -0.19  -0.23   0.38  -0.10    .    -0.07    |  -0.21  | __H_F_
Jordan                   |  761  53.9  0.58  |  18.3   14.5   53.9   12.5    .     0.9    0.1   | -0.19  -0.37   0.58  -0.23    .    -0.10    |  -1.02  | __E___
Korea, Rep. of           |  603  85.6  0.58  |   7.5    3.6   85.6    3.3    .     0.0    0.0   | -0.37  -0.32   0.58  -0.25    .     0.00    |  -0.89  | ____F_
Kuwait                   |  571  23.5  0.32  |  31.2   20.8   23.5   21.9    .     2.6    0.7   |  0.00  -0.23   0.32  -0.07    .    -0.10    |  -0.36  | _C____
Lebanon                  |  538  79.0  0.38  |  11.7    2.0   79.0    6.5    .     0.7    0.6   | -0.24  -0.13   0.38  -0.19    .    -0.13    |  -2.04  | __E_F_
Lithuania                |  571  62.7  0.59  |  16.1    7.7   62.7   12.6    .     0.9    0.0   | -0.31  -0.32   0.59  -0.24    .    -0.07    |  -0.52  | ____F_
Malaysia                 |  634  44.6  0.50  |  27.3   13.1   44.6   14.4    .     0.6    0.2   | -0.15  -0.35   0.50  -0.15    .    -0.12    |  -0.10  | __H___
Malta                    |  671  60.4  0.51  |  18.8    8.9   60.4   10.9    .     1.0    0.1   | -0.17  -0.35   0.51  -0.22    .    -0.14    |  -0.65  | ____F_
Mongolia                 |  610  57.4  0.49  |  18.7    7.9   57.4   14.4    .     1.6    0.3   | -0.28  -0.25   0.49  -0.20    .     0.00    |  -1.38  | __E_F_
Morocco                  |  614  40.9  0.42  |  27.9   15.5   40.9   14.3    .     1.5    0.5   | -0.17  -0.31   0.42  -0.05    .    -0.05    |  -0.94  | ______
Norway                   |  656  15.5  0.15  |  36.1   29.3   15.5   15.2    .     3.8    0.3   |  0.10  -0.18   0.15   0.01    .    -0.14    |   1.54  | _CH___
Oman                     |  678  41.7  0.49  |  20.5   20.8   41.7   16.1    .     0.9    0.3   | -0.09  -0.34   0.49  -0.17    .    -0.03    |  -1.08  | __E___
Palestinian Nat'l Aut    |  653  37.1  0.44  |  23.6   22.4   37.1   15.3    .     1.7    0.3   | -0.03  -0.34   0.44  -0.14    .    -0.08    |  -0.71  | ______
Qatar                    | 1017  31.7  0.26  |  27.7   23.1   31.7   15.7    .     1.8    0.1   | -0.02  -0.20   0.26  -0.06    .    -0.04    |  -1.07  | __E___
Romania                  |  601  65.6  0.53  |  20.6    5.8   65.6    7.5    .     0.5    0.3   | -0.33  -0.25   0.52  -0.17    .    -0.10    |  -1.17  | __E_F_
Russian Federation       |  633  77.3  0.52  |  12.6    2.8   77.3    6.2    .     1.1    0.8   | -0.32  -0.15   0.52  -0.29    .    -0.12    |  -1.27  | __E_F_
Saudi Arabia             |  612  22.2  0.31  |  30.2   23.0   22.2   23.4    .     1.1    0.2   |  0.03  -0.24   0.31  -0.07    .    -0.06    |  -0.47  | _C____
Scotland                 |  564  43.1  0.54  |  19.3   22.2   43.1   13.8    .     1.6    0.2   | -0.13  -0.38   0.54  -0.13    .    -0.10    |   0.20  | __H___
Serbia                   |  572  69.9  0.61  |  16.3    4.0   69.9    8.6    .     1.2    0.2   | -0.37  -0.26   0.61  -0.26    .    -0.15    |  -1.22  | __E_F_
Singapore                |  655  86.3  0.58  |   8.1    1.8   86.3    3.5    .     0.3    0.0   | -0.38  -0.28   0.58  -0.28    .    -0.12    |  -1.13  | ____F_
Slovenia                 |  570  37.5  0.41  |  36.5    4.4   37.5   21.1    .     0.5    0.0   | -0.25  -0.18   0.41  -0.09    .    -0.06    |   0.64  | __H_F_
Sweden                   |  725  24.6  0.20  |  36.6   24.1   24.6   11.6    .     3.2    0.5   |  0.08  -0.23   0.20   0.03    .    -0.18    |   1.36  | _CH___
Syrian Arab Republic     |  670  46.1  0.45  |  23.3   13.9   46.1   15.4    .     1.3    0.3   | -0.18  -0.24   0.45  -0.13    .    -0.11    |  -1.02  | __E___
Thailand                 |  763  37.0  0.55  |  27.0   18.1   37.0   16.9    .     1.0    0.4   | -0.14  -0.34   0.55  -0.16    .    -0.10    |   0.00  | __H___
Tunisia                  |  580  34.8  0.48  |  14.5   31.2   34.8   18.4    .     1.0    0.3   | -0.09  -0.35   0.48  -0.07    .    -0.07    |  -0.25  | ______
Turkey                   |  663  53.5  0.60  |  21.0   14.9   53.5    9.8    .     0.8    0.2   | -0.30  -0.34   0.60  -0.18    .    -0.04    |  -0.91  | ____F_
Ukraine                  |  640  66.4  0.51  |  16.7    6.7   66.4    9.4    .     0.8    0.2   | -0.26  -0.29   0.51  -0.22    .    -0.08    |  -1.35  | __E_F_
United States            | 1049  50.4  0.53  |  28.5    4.9   50.4   15.3    .     0.9    0.2   | -0.29  -0.24   0.53  -0.19    .    -0.08    |   0.14  | __H_F_
------------------------------------------------------------------------------------------------------------------------------------------------------------------
International Avg.       |    .  50.5  0.46  |  22.6   12.5   50.5   13.1    .     1.3    0.5   | -0.19  -0.27   0.46  -0.14    .    -0.09    |  -0.61  | ______
------------------------------------------------------------------------------------------------------------------------------------------------------------------
Basque Country, Spain    |  331  66.5  0.40  |  18.7    2.1   66.5   11.2    .     1.5    0.0   | -0.26  -0.19   0.40  -0.12    .    -0.16    |  -0.83  | ____F_
British Columbia, Can    |  600  34.8  0.42  |  41.7    8.0   34.8   13.7    .     1.8    0.3   | -0.21  -0.19   0.42  -0.10    .    -0.10    |   0.88  | __H_F_
Massachusetts, US        |  282  56.0  0.58  |  25.2    2.8   56.0   14.5    .     1.4    0.0   | -0.35  -0.20   0.58  -0.24    .    -0.14    |   0.39  | __H_F_
Minnesota, US            |  247  42.1  0.52  |  35.6    5.7   42.1   16.2    .     0.4    0.0   | -0.26  -0.23   0.52  -0.20    .    -0.11    |   0.79  | __H_F_
Ontario, Canada          |  496  21.4  0.26  |  44.6   11.3   21.4   20.2    .     2.6    0.4   | -0.02  -0.24   0.26  -0.01    .    -0.09    |   1.88  | _CH___
Quebec, Canada           |  590  56.1  0.44  |  24.6    5.4   56.1   12.7    .     1.2    0.7   | -0.20  -0.24   0.44  -0.19    .    -0.11    |   0.01  | __H_F_
------------------------------------------------------------------------------------------------------------------------------------------------------------------

Keys:  Diff: Percent correct score; Disc: Item discrimination; Pct_A...E: Percent choosing option; Pct_OM, NR: Percent Omitted and Not Reached;
       PB_A...E: Point Biserial for  option; PB_OM: Point Biserial for Omittted. RDIFF= Rasch difficulty.

Flags: A= Ability not ordered/Attractive distractor; C= Difficulty less than chance; D= Negative/low discrimination; E= Easier than average;
       F= Distractor chosen by less than 10%; H= Harder than average; R= Scoring reliability < 80%; V= Difficulty greater than 95.
```

Exhibit 10.2 **International Item Statistics for a Constructed-response Item**

```
Trends in International Mathematics and Science Study - TIMSS 2007 Assessment Results
International Item Statistics (Unweighted) - Review Version -  8th Grade
For Internal Review Only: DO NOT CITE OR CIRCULATE

Mathematics: Number / Applying (M042304B - M04_05B)
Label: Complete the table for cycling
 Type: CR   Key: X

-------------------------------------------------------------------------------------------------------------------------------------------------------
                        |                  |              Percentages              |       Point Biserials         |        |     Reliability     |
Country                 |    N  Diff  Disc |  Pct_0  Pct_1  Pct_2  Pct_OM Pct_NR   |  PB_0   PB_1   PB_2  PB_OM    |  RDIFF | Cases  Score   Code |  Flags
-------------------------------------------------------------------------------------------------------------------------------------------------------
Algeria                 |  680  12.8  0.26 |   57.6   22.1    1.8   18.5    1.6   | -0.18   0.10   0.32  -0.02    |   0.49 |  188   60.1   57.4 | ____FR
Armenia                 |  676  15.3  0.54 |   39.8    3.4   13.6   43.2    0.4   | -0.01   0.18   0.50  -0.39    |   1.63 |  219   98.2   96.3 | __H_F_
Australia               |  571  38.2  0.60 |   33.5   37.1   19.6    9.8    0.3   | -0.35   0.14   0.49  -0.32    |   0.35 |  190   96.8   96.8 | __E___
Bahrain                 |  600  15.4  0.50 |   58.2   24.2    3.3   14.3    0.5   | -0.32   0.30   0.37  -0.11    |   0.80 |  200   98.5   98.5 | ____F_
Bosnia and Herzegovin   |  600  19.8  0.48 |   32.2   27.7    6.0   34.2    0.0   | -0.12   0.24   0.38  -0.30    |   1.09 |  210   90.5   89.5 | __H_F_
Botswana                |  596   6.2  0.54 |   74.0    9.4    1.5   15.1    0.3   | -0.24   0.35   0.41  -0.13    |   1.12 |  207   96.6   96.6 | ____F_
Bulgaria                |  569  22.8  0.60 |   36.4   21.6   12.0   30.1    0.4   | -0.27   0.25   0.49  -0.29    |   1.25 |  182   80.2   79.1 | __H___
Chinese Taipei          |  581  57.7  0.65 |   20.8   31.5   42.0    5.7    0.0   | -0.44   0.00   0.53  -0.37    |   1.00 |  219   98.2   98.2 | ______
Colombia                |  701  24.6  0.58 |   46.1   32.4    8.4   13.1    2.0   | -0.39   0.25   0.46  -0.13    |  -0.22 |  197   98.5   98.0 | __E_F_
Cyprus                  |  635  34.8  0.53 |   21.1   39.7   15.0   24.3    0.0   | -0.17   0.11   0.45  -0.34    |   0.22 |    .      .      . | __E___
Czech Republic          |  693  43.3  0.57 |   23.2   42.1   22.2   12.4    0.1   | -0.31   0.03   0.49  -0.27    |   0.10 |  200   96.5   96.0 | __E___
Egypt                   |  955  15.7  0.64 |   65.3   18.8    6.3    9.5    0.2   | -0.46   0.38   0.46  -0.14    |   0.88 |  203  100.0  100.0 | ____F_
El Salvador             |  569  10.2  0.43 |   63.4   17.2    1.6   17.8    2.2   | -0.27   0.30   0.31  -0.08    |   0.56 |  203  100.0  100.0 | ____F_
England                 |  560  43.0  0.63 |   25.2   41.8   22.1   10.9    0.0   | -0.37   0.10   0.51  -0.32    |   0.49 |   84   98.8   98.8 | ______
Georgia                 |  642  15.1  0.49 |   32.7   19.3    5.5   42.5    0.8   | -0.12   0.33   0.32  -0.31    |   0.87 |  212   91.0   89.2 | ____F_
Ghana                   |  754   3.7  0.38 |   70.0    6.4    0.5   23.1    0.4   | -0.11   0.32   0.19  -0.10    |   1.35 |  269   99.3   99.3 | __H_F_
Hong Kong SAR           |  484  50.2  0.60 |   25.4   38.4   31.0    5.2    0.4   | -0.39   0.05   0.48  -0.34    |   1.07 |  164   95.1   93.9 | __H___
Hungary                 |  584  49.4  0.64 |   17.5   44.3   27.2   11.0    0.0   | -0.36  -0.05   0.56  -0.29    |   0.21 |  181   94.5   92.8 | __E___
Indonesia               |  596  11.1  0.55 |   63.8   13.4    4.4   18.5    0.2   | -0.27   0.34   0.40  -0.17    |   0.88 |  203   93.1   92.6 | ____F_
Iran, Islamic Rep. of   |  568  17.3  0.53 |   54.0   24.6    4.9   16.4    0.4   | -0.32   0.29   0.40  -0.16    |   0.71 |  182   98.4   97.8 | ____F_
Israel                  |  485  28.5  0.57 |   36.5   32.2   12.4   19.0    0.8   | -0.31   0.30   0.40  -0.31    |   0.63 |  239   90.8   85.4 | ______
Italy                   |  620  32.8  0.47 |   23.9   41.5   12.1   22.6    0.2   | -0.22   0.14   0.39  -0.25    |   0.58 |  216   99.5   99.5 | ______
Jordan                  |  752  14.4  0.52 |   68.8   18.8    5.1    7.4    0.1   | -0.37   0.33   0.37  -0.14    |   1.10 |  221   98.6   98.6 | __H_F_
Korea, Rep. of          |  601  53.7  0.71 |   24.5   33.6   36.9    5.0    0.0   | -0.47  -0.09   0.64  -0.29    |   1.22 |  212   97.2   97.2 | __H___
Kuwait                  |  587   8.7  0.54 |   70.2   11.2    3.1   15.5    0.5   | -0.22   0.31   0.43  -0.20    |   0.33 |  200   99.0   99.0 | ____F_
Lebanon                 |  549  23.8  0.59 |   39.0   34.8    6.4   19.9    0.2   | -0.37   0.38   0.36  -0.21    |   1.03 |  191   99.5   98.4 | ____F_
Lithuania               |  577  37.8  0.59 |   30.0   34.3   20.6   15.1    0.0   | -0.32   0.07   0.52  -0.27    |   0.58 |  174   96.6   95.4 | ______
Malaysia                |  627  32.5  0.61 |   43.1   33.3   15.8    7.8    0.0   | -0.46   0.21   0.48  -0.16    |   0.43 |  216   96.3   96.3 | ______
Malta                   |  660  42.7  0.63 |   29.7   34.2   25.6   10.5    0.2   | -0.39   0.10   0.52  -0.32    |   0.18 |  176   93.8   93.8 | __E___
Mongolia                |  631  15.3  0.35 |   42.2   24.2    3.2   30.4    0.6   |  0.05   0.17   0.32  -0.34    |   1.18 |  192   81.3   72.9 | __H_F_
Morocco                 |  615  20.7  0.22 |   42.3   33.7    3.9   20.2    0.5   |  0.00  -0.05   0.37  -0.13    |   0.39 |  176   93.2   93.2 | ___AF_
Norway                  |  656  34.0  0.54 |   28.7   39.9   14.0   17.4    0.5   | -0.24   0.18   0.42  -0.33    |   0.07 |  220   94.1   92.3 | __E___
Oman                    |  675   7.0  0.53 |   80.0   13.0    0.4    6.5    0.9   | -0.39   0.51   0.14  -0.09    |   1.93 |  226   99.1   99.1 | __H_F_
Palestinian Nat'l Aut   |  642   9.7  0.55 |   75.4   13.2    3.1    8.3    0.0   | -0.41   0.42   0.33  -0.09    |   0.83 |  231   94.4   94.4 | ____F_
Qatar                   | 1019   5.0  0.50 |   71.9    8.4    0.8   18.8    0.1   | -0.16   0.40   0.29  -0.16    |   0.98 |  197   98.0   97.5 | ____F_
Romania                 |  599  31.1  0.60 |   30.4   32.2   15.0   22.4    0.3   | -0.29   0.14   0.53  -0.28    |   0.44 |  209   98.6   97.6 | ______
Russian Federation      |  637  42.5  0.60 |   22.6   39.1   22.9   15.4    0.3   | -0.32   0.02   0.54  -0.28    |   0.58 |  212   99.1   99.1 | ______
Saudi Arabia            |  600   5.0  0.36 |   79.0    8.3    0.8   11.8    0.7   | -0.14   0.29   0.22  -0.14    |   0.91 |  211   99.1   99.1 | ____F_
Scotland                |  585  37.5  0.62 |   27.2   41.9   16.6   14.4    0.2   | -0.32   0.14   0.50  -0.34    |   0.37 |  116   99.1   98.3 | __E___
Serbia                  |  584  33.1  0.54 |   19.7   34.4   15.9   30.0    0.2   | -0.15   0.14   0.45  -0.38    |   0.65 |  206   97.6   97.1 | ______
Singapore               |  665  63.7  0.71 |   19.1   27.2   50.1    3.6    0.0   | -0.50  -0.14   0.64  -0.32    |   0.52 |  220   99.1   99.1 | ______
Slovenia                |  559  41.9  0.54 |   24.0   44.7   19.5   11.8    0.0   | -0.26  -0.04   0.52  -0.23    |   0.25 |  184  100.0  100.0 | __E___
Sweden                  |  751  41.2  0.64 |   22.6   34.0   24.2   19.2    0.3   | -0.25   0.04   0.56  -0.40    |  -0.03 |  253   96.0   94.9 | __E___
Syrian Arab Republic    |  668  17.0  0.38 |   58.1   25.6    4.2   12.1    0.0   | -0.24   0.14   0.36  -0.06    |   0.50 |  195   99.5   98.5 | ____F_
Thailand                |  768  32.9  0.59 |   39.1   32.6   16.7   11.7    0.1   | -0.37   0.16   0.49  -0.23    |   0.11 |  197   95.4   94.9 | __E___
Tunisia                 |  577  19.2  0.49 |   57.4   23.4    7.5   11.8    0.3   | -0.38   0.24   0.38  -0.05    |   0.30 |  205   87.3   87.3 | __E_F_
Turkey                  |  654  20.0  0.55 |   50.2   29.8    5.0   15.0    0.0   | -0.32   0.26   0.47  -0.17    |   1.18 |  184   97.8   97.8 | __H_F_
Ukraine                 |  646  18.7  0.43 |   46.1   20.4    8.5   24.9    0.0   | -0.12   0.18   0.36  -0.26    |   1.13 |  212   85.8   84.9 | __H_F_
United States           | 1066  41.3  0.62 |   31.8   40.8   20.9    6.5    0.6   | -0.40   0.09   0.52  -0.26    |   0.50 |  284   93.3   92.3 | ______
-------------------------------------------------------------------------------------------------------------------------------------------------------
International Avg.      |    .  26.9  0.54 |   42.7   27.7   13.1   16.5    0.4   | -0.28   0.18   0.43  -0.23    |   0.69 |    .   95.1   94.3 | ______
-------------------------------------------------------------------------------------------------------------------------------------------------------
Basque Country, Spain   |  326  38.5  0.50 |   23.9   45.1   16.0   15.0    0.0   | -0.20   0.00   0.48  -0.26    |   0.28 |  199   99.0   99.0 | __E___
British Columbia, Can   |  603  40.2  0.53 |   28.9   42.3   19.1    9.8    0.2   | -0.25   0.07   0.45  -0.32    |   0.49 |   85   95.3   94.1 | ______
Massachusetts, US       |  262  51.9  0.64 |   23.3   41.2   31.3    4.2    0.8   | -0.46   0.00   0.53  -0.24    |   0.54 |   68   95.6   95.6 | ______
Minnesota, US           |  257  42.4  0.59 |   31.5   45.9   19.5    3.1    0.0   | -0.47   0.08   0.48  -0.08    |   0.69 |   69   94.2   92.8 | ______
Ontario, Canada         |  492  43.2  0.57 |   28.5   40.4   23.0    8.1    1.4   | -0.33   0.07   0.47  -0.33    |   0.40 |  127   96.9   96.1 | ______
Quebec, Canada          |  569  51.0  0.61 |   19.2   43.9   29.0    7.9    0.9   | -0.31  -0.07   0.55  -0.31    |   0.21 |  183   95.1   95.1 | __E___
-------------------------------------------------------------------------------------------------------------------------------------------------------

Keys:  Diff: Percent correct score; Disc: Item discrimination; Pct_0...3: Percent obtaining score level; Pct_OM, NR: Percent Omitted and Not Reached;
       PB_0...3: Point Biserial for score level; PB_OM: Point Biserial for Omitted;  RDIFF= Rasch difficulty;
       Reliability: (Cases) Responses double scored; (Score) Percent agreement on score; (Code) Percent agreement on code.
Flags: A= Ability not ordered/Attractive distractor; C= Difficulty less than chance; D= Negative/low discrimination; E= Easier than average;
       F= Score obtained by less than 10%; H= Harder than average; R= Scoring reliability < 80%; V= Difficulty greater than 95.
```

For all items, regardless of item format, statistics included the number of students that responded in each country, the difficulty level (the percentage of students that answered the item correctly), and the discrimination index (the point-biserial correlation between success on the item and a total score).[1] Also provided was an estimate of the item's difficulty using a Rasch one-parameter IRT model. The international means of the item difficulties and item discriminations served as guides to the overall statistical properties of the items. Statistics for each item are displayed alphabetically by country, with the international average for each statistic in the bottom row. For those countries that tested in more than one language, statistics were calculated and examined separately by language group.

Statistics displayed for multiple-choice items included the percentage of students that chose each option, as well as the percentage of students that omitted or did not reach the item, and the point-biserial correlation between the response to each option and the total score. Statistics displayed for constructed-response items (which could have one or two score levels) included the difficulty and discrimination of each score level. Constructed-response item displays also provided information about the reliability with which the item was scored in each country, with the total number of double-scored cases and the percent exact agreement between the scorers.

10.2.1 Statistics used in Item Analysis

Definitions and detailed descriptions of the statistics that were calculated are provided below, with examples shown in Exhibits 10.1 and 10.2. The statistics were calculated separately, by grade, for mathematics and science. Statistics are listed in order of appearance in the item analysis output:

N: Number of students to whom the item was administered. If a student did not reach an item in the achievement booklet, the item was considered not administered for the purpose of the item analysis.[2]

Diff: Item difficulty is the average percent correct. For 1-point items, it is the percentage of students providing a fully correct response to the item. For the computation of this statistic, not reached items were treated as not administered.

Disc: Item discrimination was computed as the correlation between a correct response to the item and the overall score on all of the

[1] For the purpose of computing the discrimination index, the total score was the percentage of mathematics or science items a student answered correctly.

[2] In TIMSS, for the purposes of item analysis and item parameter estimation in scaling, items not reached by a student were treated as if they had not been administered. For purposes of estimating student proficiency, however, not reached items were treated as incorrectly answered.

mathematics or science items the student was administered.[3] Items exhibiting good measurement properties should have a moderately positive correlation.

PCT_A, **PCT_B**, **PCT_C**, **PCT_D**, and **PCT_E**: Used for multiple-choice items only (see Exhibit 10.1), each column indicates the percentage of students choosing the particular response option for the item (A, B, C, D, or E). Not reached items were excluded from the denominator for these calculations.

PCT_0, **PCT_1**, and **PCT_2**: Used for constructed-response items only (see Exhibit 10.2), each column indicates the percentage of students scoring at the particular score level, up to and including the maximum score level for the item. Not reached items were excluded from the denominator for these calculations.

PCT_OM: Percentage of students who, having reached the item, did not provide a response. Not reached items were excluded from the denominator when calculating this statistic.

PCT_NR: Percentage of students who did not reach the item. An item was coded as not reached when there was no evidence of a response to any subsequent items in the booklet and the response to the item preceding it was omitted.

PB_A, **PB_B**, **PB_C**, **PB_D**, and **PB_E**: Used for multiple-choice items only, these present the correlation between choosing each of the response options, A, B, C, D, or E, and the overall score on all of the mathematics or science items the student was administered. Items with good psychometric properties have near-zero or negative correlations for the distracter options (the incorrect options) and moderately positive correlations for the correct option.

PB_0, **PB_1**, and **PB_2**: Used for constructed-response items only, these present the correlation between the score levels on the item (0, 1, or 2) and the overall score on all of the mathematics or science items the student was administered. For items with good measurement properties, the correlation coefficients should change from negative to positive as the score on the item increases.

PB_OM: The correlation between a binary variable, indicating an omitted response to the item, and the overall score on all of the mathematics or science items the student was administered. This correlation should be negative or near zero.

[3] For constructed-response items, the discrimination is the correlation between the number of score points and total score.

RDIFF: An estimate of the item's difficulty based on a Rasch one-parameter IRT model applied to each country's sample. The difficulty estimate is expressed in the logit metric (with a positive logit indicating a difficult item) and was scaled so that the average Rasch item difficulty was zero within each country.

Reliability (Cases): To provide a measure of the reliability of the scoring of the constructed-response items, those items in approximately 25 percent of the test booklets in each country were scored by two independent scorers. This column indicates the number of times each item was double-scored in a country.

Reliability (Score): This column contains the percentage of exact agreement on the scores assigned by two independent scorers.

Reliability (Code): This column contains the percentage of exact agreement on the 2-digit scoring codes.

As an aid to reviewers, the item-analysis display includes a series of "flags" signaling the presence of one or more conditions that might indicate a problem with an item. The following conditions are flagged:

- Item difficulty exceeds 95 percent in the sample as a whole.
- Item difficulty is less than 25 percent for four-option multiple-choice items in the sample as a whole.
- One or more of the distracter percentages is less than 10 percent.
- One or more of the distracter percentages is greater than the percentage for the correct answer or the point-biserial correlation for one or more of the distracters exceeds zero.
- Item discrimination (i.e., the point-biserial for the correct answer) is less than 0.2.
- Item discrimination does not increase with each score level (for constructed-response items with more than one score level).
- The Rasch difficulty estimate is easier or harder than the average across countries.
- Scoring reliability for the score points is less than 80 percent (for constructed-response items only).

Although not all of these conditions necessarily indicate a problem, the flags are a useful way to draw attention to potential sources of concern.

In order to measure trends, TIMSS 2007 included items from TIMSS 2003 at the fourth grade and from TIMSS 2003 and 1999 (those items from 1999 that were administered again in 2003) at the eighth grade.[4] For these trend items, the review included an examination of changes in item statistics between the 2003 and 2007 administrations.

An example item statistics display for a fourth grade trend item is shown in Exhibit 10.3. The information in this exhibit is different from the item statistics presented in Exhibits 10.1 and 10.2, and presents countries' statistics from the TIMSS 2007 and 2003 assessments. In reviewing these item statistics, the aim was to detect any unusual changes in item properties between assessments, which might indicate a problem in using the item to measure change.

[4] For more information on trend items, see Chapter 2.

Exhibit 10.3 International Item Statistics for a Trend Item

Trends in International Mathematics and Science Study - TIMSS 2007 Bridge Assessment Results
Percent of Responses by Item Category (Science) - Trend Items - 4th Grade
For Internal Review Only: DO NOT CITE OR CIRCULATE

Science: Life Science / Factual Knowledge (S031233 - S11_03)
Label: Main features of four animals shown
Type: CR Key: X

COUNTRY	Year	N	10	70	71	79	99	V1	NOT REACHED	OMIT	1.GIRL % Right	2.BOY % Right
Armenia	2003	464	33.2	18.6	16.8	11.7	19.6	33.2	0.8	18.9	33.0	33.5
	2007	273	48.7	8.8	21.2	1.8	19.4	48.7	0.7	18.7	48.1	49.1
Australia	2003	375	64.0	17.5	11.0	5.5	2.1	64.0	0.0	2.1	65.3	62.7
	2007	291	64.6	17.2	14.1	3.4	0.7	64.6	0.0	0.7	63.5	65.6
Chinese Taipei	2003	380	76.5	3.3	18.7	0.3	1.3	76.5	0.0	1.3	76.5	76.4
	2007	300	64.0	6.7	26.0	3.0	0.3	64.0	0.0	0.3	63.3	64.6
England	2003	291	64.2	13.0	15.4	6.1	1.2	64.2	0.0	1.2	63.8	64.8
	2007	305	69.2	14.8	10.8	4.6	0.7	69.2	0.0	0.7	70.8	67.5
Hong Kong SAR	2003	373	69.8	7.9	18.9	3.0	0.4	69.8	0.0	0.4	72.5	67.7
	2007	268	69.8	5.2	22.0	1.9	1.1	69.8	0.0	1.1	69.9	69.6
Hungary	2003	268	70.5	15.4	9.9	3.6	0.6	70.5	0.0	0.6	75.9	65.1
	2007	288	72.2	11.1	11.8	3.8	1.0	72.2	0.0	1.0	73.5	71.1
Iran, Islamic Rep. of	2003	352	41.7	22.8	20.3	6.7	8.5	41.7	0.0	8.5	48.2	37.7
	2007	274	39.4	25.9	20.1	8.4	6.2	39.4	0.0	6.2	36.9	41.4
Italy	2003	353	72.4	9.2	14.3	3.7	0.5	72.4	0.0	0.5	75.4	69.8
	2007	323	66.3	10.5	15.2	3.1	5.0	66.3	0.0	5.0	65.2	67.0
Latvia	2003	295	61.0	19.5	13.5	4.8	1.1	61.0	0.0	1.1	62.0	60.1
	2007	277	68.2	15.2	11.6	3.6	1.4	68.2	0.0	1.4	71.1	65.5
Lithuania	2003	371	58.3	23.2	9.2	7.3	1.9	58.3	0.0	1.9	59.3	56.3
	2007	285	54.7	22.8	14.4	7.0	1.1	54.7	0.0	1.1	54.7	54.8
Morocco	2003	339	14.8	32.1	12.5	28.2	12.5	14.8	0.0	12.5	10.2	18.9
	2007	300	12.0	16.7	7.7	47.7	16.0	12.0	0.7	15.3	10.5	13.7
Netherlands	2003	242	54.7	15.4	19.9	10.0	0.0	54.7	0.0	0.0	57.3	51.8
	2007	237	63.3	15.2	14.8	5.5	1.3	63.3	0.0	1.3	62.5	64.1
New Zealand	2003	354	61.1	17.8	14.2	4.9	2.1	61.1	0.0	2.1	59.9	62.3
	2007	349	56.7	18.3	15.8	8.3	0.9	56.7	0.0	0.9	56.3	57.2
Norway	2003	361	58.6	20.5	12.8	6.3	1.8	58.6	0.0	1.8	61.7	55.9
	2007	290	55.9	16.9	20.7	4.1	2.4	55.9	0.0	2.4	57.4	54.5
Russian Federation	2003	325	53.9	16.6	8.3	17.1	4.2	53.9	0.0	4.2	51.7	56.1
	2007	323	62.8	15.5	11.5	7.4	2.8	62.8	0.0	2.8	60.1	65.0
Scotland	2003	330	54.5	19.1	17.3	6.8	2.3	54.5	0.4	1.9	54.2	54.9
	2007	286	59.1	21.7	10.1	7.7	1.4	59.1	0.0	1.4	50.0	66.3
Singapore	2003	562	78.1	9.7	9.1	2.4	0.7	78.1	0.0	0.7	80.7	75.5
	2007	360	78.3	10.3	9.4	1.7	0.3	78.3	0.0	0.3	81.4	75.1

V1 = Percent scoring 1 or better V2 = Percent scoring 2 or better
Percent right for boys and girls corresponds to the percent obtaining the maximum score on the item.
Because of missing gender information, some totals may appear inconsistent.

Exhibit 10.3 International Item Statistics for a Trend Item (Continued)

```
Trends in International Mathematics and Science Study - TIMSS 2007 Bridge Assessment Results
Percent of Responses by Item Category (Science) - Trend Items - 4th Grade
For Internal Review Only: DO NOT CITE OR CIRCULATE

Science: Life Science / Factual Knowledge (S031233 - S11_03)
Label: Main features of four animals shown
 Type: CR   Key: X


                                                                           NOT
                                                                         REACH         1.GIRL    2.BOY
 COUNTRY                    Year        N    10    70    71    79    99    V1    ED  OMIT  % Right  % Right

 -------------------------------------------------------------------------------------------------------
 Slovenia                   2003      280  58.9  18.4  13.6   6.4   2.7  58.9   0.0   2.7     63.2     56.4
                            2007      316  66.5  14.6   9.5   8.5   0.9  66.5   0.0   0.9     69.7     63.4

 Tunisia                    2003      354  18.5  26.2  20.3  27.2   7.8  18.5   0.0   7.8     16.4     20.6
                            2007      290  18.6  29.3  13.1  30.0   9.0  18.6   0.0   9.0     20.7     16.6

 United States              2003      809  60.7  20.0  11.5   6.7   1.0  60.7   0.0   1.0     62.9     58.6
                            2007      566  61.3  19.3  12.0   6.0   1.4  61.3   0.2   1.2     59.2     63.3


 International Avg.         2003        .  56.3  17.3  14.4   8.4   3.6  56.3   0.1   3.6     57.5     55.2
                            2007        .  57.6  15.8  14.6   8.4   3.7  57.6   0.1   3.6     57.2     57.8


 Ontario, Canada            2003      359  55.1  21.3  14.6   7.1   2.0  55.1   0.0   2.0     49.1     61.9
                            2007      255  50.6  22.7  12.9   8.6   5.1  50.6   0.0   5.1     49.3     52.1

 Quebec, Canada             2003      373  51.9  25.2  15.6   7.0   0.3  51.9   0.0   0.3     47.7     55.5
                            2007      276  61.2  15.2  13.4   6.9   3.3  61.2   0.0   3.3     64.5     58.0
```

```
V1 = Percent scoring 1 or better     V2 = Percent scoring 2 or better
Percent right for boys and girls corresponds to the percent obtaining the maximum score on the item.
Because of missing gender information, some totals may appear inconsistent.
```

10.2.2 Item-by-Country Interaction

Although countries are expected to exhibit some variation in performance across items, in general, as a whole, countries with high average performance on the assessment should perform relatively well on each of the items, and low-scoring countries should do less well on each of items. When this does not occur (i.e., when a high-scoring country has a low performance on an item on which other countries are doing well), there is said to be an item-by-country interaction. When large, such item-by-country interactions may be a sign of an item that is flawed in some way, and steps should be taken to address the problem.

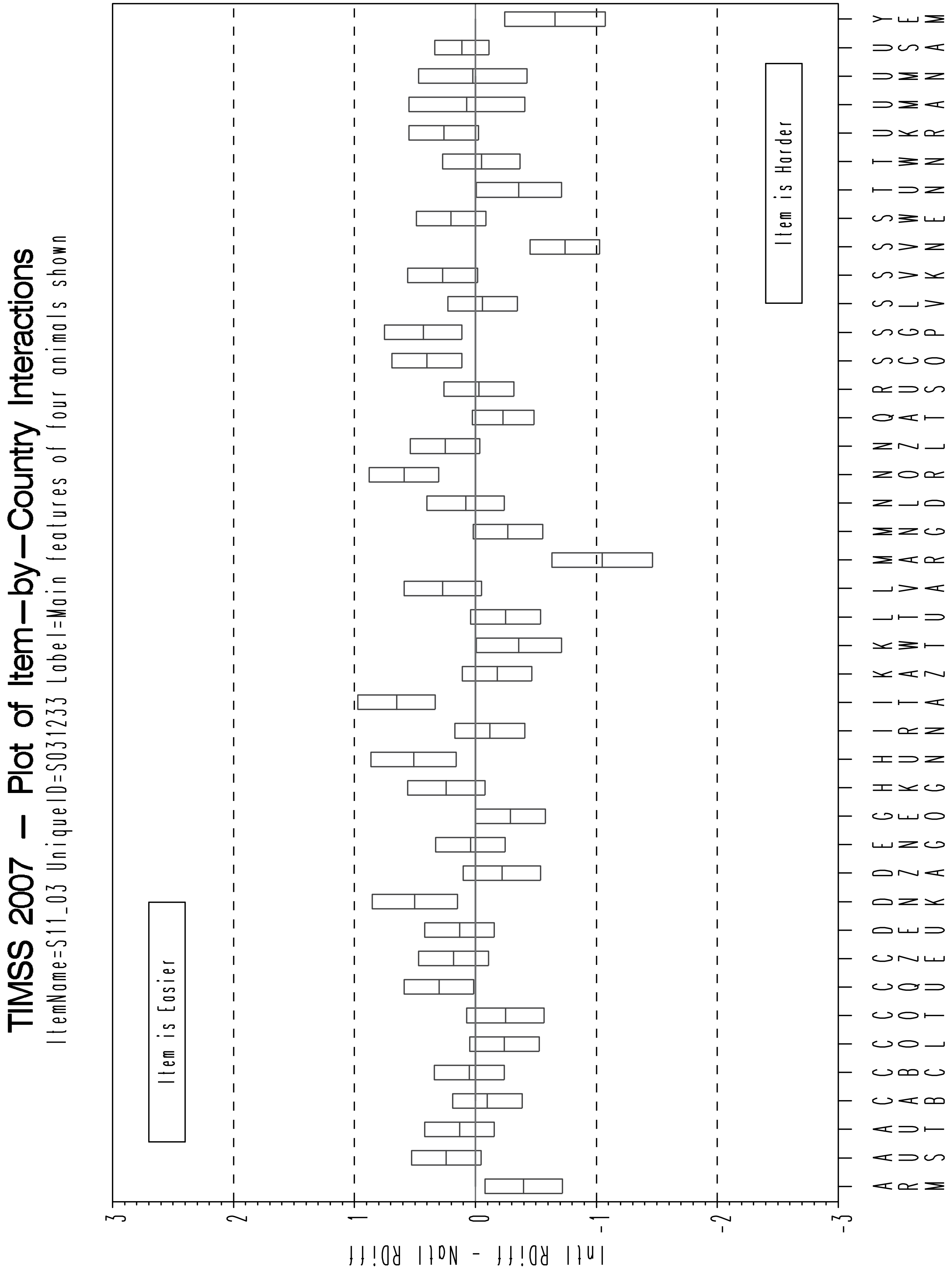
To assist in detecting sizeable item-by-country interactions, the TIMSS & PIRLS International Study Center produced a graphical display for each item showing the average probability across all countries of a correct response for a student of average international proficiency, compared with the probability of a correct response by a student of average proficiency in each country. Exhibit 10.4 provides an example of a TIMSS item-by-country interaction display. The probability for each country is presented as a 95 percent confidence interval, which includes a built-in Bonferroni correction for multiple comparisons. The limits for the confidence interval are computed as follows:

$$\text{Upper Limit} = 1 - \frac{e^{RDIFF_{ik} + SE_{RDIFF_{ik}} \text{x} Z_b}}{1 + e^{RDIFF_{ik} + SE_{RDIFF_{ik}} \text{x} Z_b}}$$

$$\text{Lower Limit} = 1 - \frac{e^{RDIFF_{ik} - SE_{RDIFF_{ik}} \text{x} Z_b}}{1 + e^{RDIFF_{ik} - SE_{RDIFF_{ik}} \text{x} Z_b}}$$

where $RDIFF_{ik}$ is the Rasch difficulty of item k within country i, $\text{SE}_{RDIFFik}$ is the standard error of the difficulty of item k in country i and Z_b is the critical value from the Z distribution, corrected for multiple comparisons using the Bonferroni procedure.

Exhibit 10.4 Sample Plot of Item-by-Country Interaction for a TIMSS 2007 Item



10.2.3 Trend Item Analysis

Because an important part of the TIMSS 2007 assessment was the measuring of trends across cycles, there was an additional stage of the review process to ensure that the trend items had similar characteristics in both cycles (i.e., an item that was relatively easy in 2003 should be relatively easy in 2007). The comparison between cycles was made in a number of ways. For each trend country, almanacs of item statistics displayed the percentage of students within each score category (or response option, for multiple-choice items) for each cycle, as well as the difficulty of the item and the percent correct by gender. While some changes were anticipated as countries' overall achievement may have improved or declined, items were noted if the trend difference was greater than 2 logits for a particular country.

In addition, TIMSS 2007 included a bridge study to examine the effect of changes to the assessment design and booklets. Countries measuring trend were required to participate in a bridge study where they administered a subset of TIMSS 2003 booklets in TIMSS 2007 under the TIMSS 2003 conditions. During the trend item analysis and review stage, comparisons then were made for these items to examine for any differences.

The TIMSS & PIRLS International Study Center used two different graphical displays to examine the differences between item difficulties in 2003 to 2007. The first of these, shown in Exhibit 10.5, displays the difference in Rasch difficulty estimates (in logits) for trend comparisons between 2003 and the 2007 bridge data. A positive difference indicates that the item was relatively easier in a country in 2007, and a negative difference indicates that an item was relatively more difficult. The second, Exhibit 10.6, shows a country's performance on all trend items simultaneously. Individually for each country, a scatterplot graphed the Rasch difficulty of each item in 2003 against the difficulty for that item in 2007. Where there are no differences between the difficulties in the 2003 and 2007 bridge data, the data points will align on or near the diagonal indicating a one-to-one correlation between cycles.

These graphs were used in conjunction with one another to detect items that performed differently in the two cycles. When such items were found, the source of the difference was investigated using booklets from both cycles, translation verifier's comments, national adaptation forms, and trend scoring reliability data.

Exhibit 10.5 Sample Plot of Difference in Rasch Difficulties for a TIMSS 2007 Item

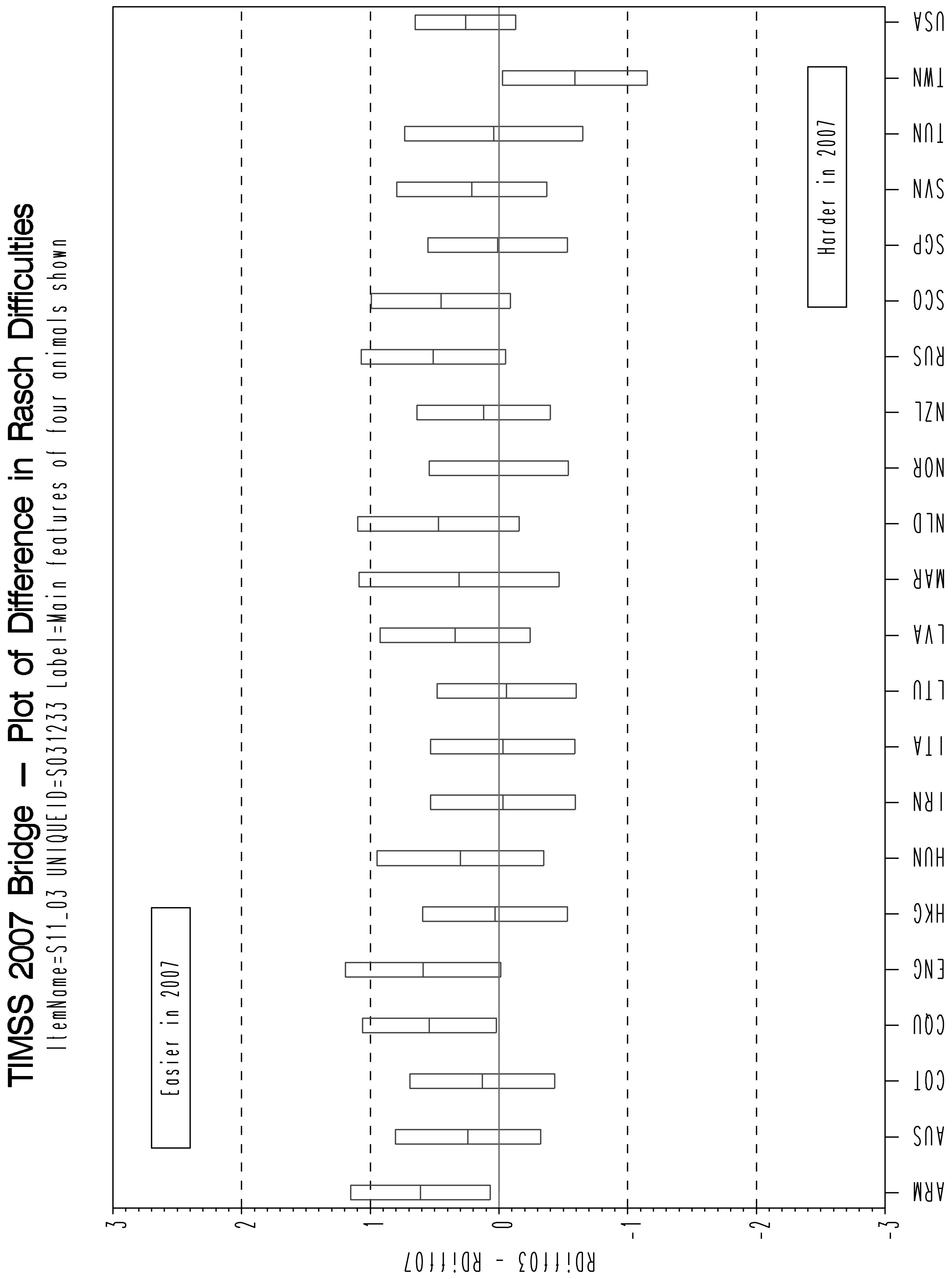
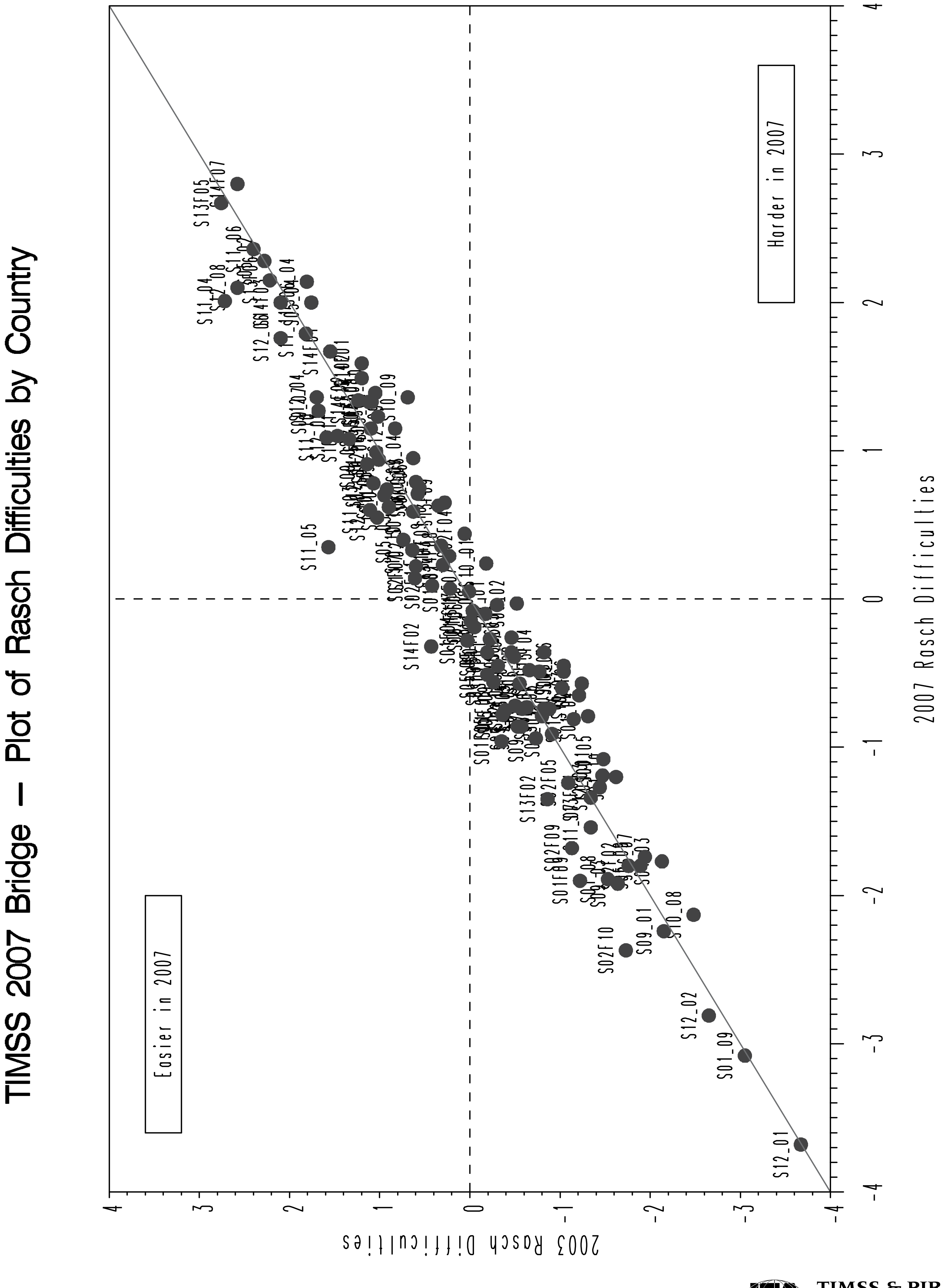


Exhibit 10.6 Sample Plot of Rasch Difficulties by Country



10.3 Reliability

10.3.1 Test Reliability

Exhibits 10.7 and 10.8 display the mathematics and science test reliability coefficients for each country. This coefficient is the median Cronbach's alpha reliability across the 14 test booklets. In general, at both grade levels in mathematics, median reliabilities were relatively high, with an international median (the median of the reliability coefficients for all countries) of 0.83 for fourth grade and 0.88 for eighth grade. In science, median reliabilities were 0.80 for fourth grade and 0.84 for eighth grade. Despite the generally high reliabilities, there were some countries with median reliabilities below 0.70 at one or both grades in mathematics, namely Algeria, Botswana, El Salvador, Ghana, Kuwait, Qatar, Saudi Arabia, and Yemen. Countries with median reliabilities below 0.70 at one or both grades in science were Algeria and Yemen.

Exhibit 10.7 Cronbach's Alpha Reliability Coefficient – TIMSS 2007 Mathematics Test

Country	Reliability Coefficient	
	4th Grade	8th Grade
Algeria	0.77	0.66
Armenia	0.87	0.88
Australia	0.86	0.89
Austria	0.82	
Bahrain		0.80
Bosnia and Herzegovina		0.84
Botswana		0.69
Bulgaria		0.90
Chinese Taipei	0.83	0.93
Colombia	0.77	0.77
Cyprus		0.88
Czech Republic	0.83	0.88
Denmark	0.84	
Egypt		0.84
El Salvador	0.70	0.63
England	0.88	0.90
Georgia	0.83	0.84
Germany	0.83	
Ghana		0.68
Hong Kong SAR	0.81	0.92
Hungary	0.88	0.90
Indonesia		0.83
Iran, Islamic Rep. of	0.81	0.84
Israel		0.90
Italy	0.85	0.87
Japan	0.85	0.91
Jordan		0.88
Kazakhstan	0.87	
Korea, Rep. of		0.92
Kuwait	0.69	0.69
Latvia	0.83	
Lebanon		0.84
Lithuania	0.85	0.89
Malaysia		0.88
Malta		0.89
Morocco	0.78	0.76
Netherlands	0.79	
New Zealand	0.87	
Norway	0.82	0.84
Oman		0.80
Palestinian Nat'l Auth.		0.83
Qatar	0.58	0.64
Romania		0.90
Russian Federation	0.86	0.90
Saudi Arabia		0.62
Scotland	0.85	0.89
Serbia		0.89
Singapore	0.87	0.92
Slovak Republic	0.86	
Slovenia	0.84	0.88
Sweden	0.82	0.87
Syrian Arab Republic		0.79
Thailand		0.88
Tunisia	0.78	0.78
Turkey		0.91
Ukraine	0.84	0.88
United States	0.85	0.89
Yemen	0.55	
International Median	0.83	0.88
Benchmarking Participants		
Alberta, Canada	0.82	
Basque Country, Spain		0.85
British Columbia, Canada	0.84	0.87
Dubai, UAE	0.84	0.89
Massachusetts, US	0.82	0.89
Minnesota, US	0.86	0.87
Ontario, Canada	0.82	0.87
Quebec, Canada	0.82	0.87

Exhibit 10.8 Cronbach's Alpha Reliability Coefficient - TIMSS 2007 Science Test

Country	Reliability Coefficient	
	4th Grade	8th Grade
Algeria	0.76	0.65
Armenia	0.88	0.88
Australia	0.81	0.85
Austria	0.81	
Bahrain		0.84
Bosnia and Herzegovina		0.82
Botswana		0.75
Bulgaria		0.87
Chinese Taipei	0.80	0.88
Colombia	0.81	0.78
Cyprus		0.82
Czech Republic	0.80	0.83
Denmark	0.80	
Egypt		0.82
El Salvador	0.79	0.71
England	0.82	0.87
Georgia	0.77	0.79
Germany	0.80	
Ghana		0.72
Hong Kong SAR	0.78	0.86
Hungary	0.83	0.84
Indonesia		0.76
Iran, Islamic Rep. of	0.83	0.83
Israel		0.88
Italy	0.82	0.83
Japan	0.78	0.85
Jordan		0.87
Kazakhstan	0.80	
Korea, Rep. of		0.85
Kuwait	0.82	0.82
Latvia	0.76	
Lebanon		0.83
Lithuania	0.76	0.85
Malaysia		0.85
Malta		0.88
Morocco	0.79	0.73
Netherlands	0.73	
New Zealand	0.83	
Norway	0.79	0.82
Oman		0.82
Palestinian Nat'l Auth.		0.85
Qatar	0.77	0.78
Romania		0.84
Russian Federation	0.82	0.85
Saudi Arabia		0.73
Scotland	0.80	0.85
Serbia		0.83
Singapore	0.86	0.91
Slovak Republic	0.82	
Slovenia	0.79	0.83
Sweden	0.79	0.85
Syrian Arab Republic		0.80
Thailand		0.84
Tunisia	0.85	0.73
Turkey		0.85
Ukraine	0.80	0.84
United States	0.82	0.86
Yemen	0.69	
International Median	0.80	0.84
Benchmark Participants		
Alberta, Canada	0.79	
Basque Country, Spain		0.81
British Columbia, Canada	0.79	0.83
Dubai, UAE	0.85	0.86
Massachusetts, US	0.78	0.86
Minnesota, US	0.81	0.85
Ontario, Canada	0.81	0.82
Quebec, Canada	0.76	0.81

10.3.2 Scoring Reliability for Constructed-response Items

About one-third of the items in the TIMSS 2007 assessment were constructed-response items, comprising nearly half of the score points for the assessment.[5] An essential requirement for use of such items is that they be reliably scored by all participants. That is, a particular student response should receive the same score, regardless of the scorer. In conducting TIMSS 2007, measures taken to ensure that the constructed-response items were scored reliably in all countries included developing scoring guides for each constructed-response question (which provided descriptions of acceptable responses for each score point value)[6] and providing extensive training in the application of the scoring guides. Procedures for organizing and monitoring the scoring sessions are outlined in the *TIMSS 2007 Survey Operations Procedures Unit 5: Scoring the TIMSS 2007 Assessment* (TIMSS, 2006).

10.3.2.1 Within-Country Scoring Reliability

To gather and document information about the within-country agreement among scorers, a random sample of at least 200 student responses to each item was selected to be scored independently by two scorers.[7] The inter-rater agreement for each item in each country was examined as part of the item review process. The average and range of the within-country exact percent of agreement across all items for both grades is presented in Exhibit 10.9 for mathematics and Exhibit 10.10 for science.

Agreement across items was high on average across countries. The exact percent agreement was 98 percent at both grades in mathematics and 96 percent at both grades in science. All countries had an average exact percent agreement above 92 percent at the fourth grade and 95 percent at the eighth grade in mathematics and above 85 percent at the fourth grade and 90 at the eighth grade in science.

[5] For details on the development of the TIMSS 2007 assessment items, see Chapter 2.

[6] A discussion of the development of the scoring guides for constructed-response items is provided in Chapter 2.

[7] Since individual items appear in two booklets, 100 of each of the 14 booklets were chosen randomly for double-scoring. For a sample of 4,500, this amounts to almost 25 percent of the total sample.

Exhibit 10.9 TIMSS 2007 Within-country Scoring Reliability for the Fourth Grade Constructed-response Mathematics Items

Countries	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Algeria	92	58	99	85	54	98
Armenia	99	94	100	97	91	100
Australia	100	98	100	99	95	100
Austria	99	95	100	99	94	100
Chinese Taipei	98	84	100	97	83	100
Colombia	99	93	100	97	89	100
Czech Republic	98	90	100	96	77	100
Denmark	97	83	100	93	74	99
El Salvador	99	96	100	98	85	100
England	99	91	100	98	89	100
Georgia	97	88	100	94	68	100
Germany	97	75	100	95	71	100
Hong Kong SAR	100	98	100	100	98	100
Hungary	100	97	100	99	95	100
Iran, Islamic Rep. of	99	96	100	96	84	100
Italy	99	94	100	99	79	100
Japan	99	94	100	98	84	100
Kazakhstan	99	96	100	99	94	100
Kuwait	100	98	100	98	95	100
Latvia	95	41	100	92	39	100
Lithuania	98	88	100	97	50	100
Morocco	95	33	100	88	29	98
Netherlands	97	86	100	95	72	100
New Zealand	99	95	100	97	90	100
Norway	99	92	100	97	88	100
Qatar	99	91	100	95	78	100
Russian Federation	100	98	100	99	96	100
Scotland	99	91	100	98	87	100
Singapore	99	93	100	97	90	100
Slovak Republic	99	92	100	98	90	100
Slovenia	100	99	100	99	94	100
Sweden	98	89	100	97	87	100
Tunisia	98	86	100	93	77	99
Ukraine	100	98	100	100	98	100
United States	98	83	100	96	72	100
Yemen	98	83	100	93	80	99
International Avg.	98	88	100	96	81	100
Benchmark Participants						
Alberta, Canada	99	93	100	98	90	100
British Columbia, Canada	99	96	100	99	91	100
Dubai, UAE	97	87	100	94	78	100
Massachusetts, US	98	82	100	96	72	100
Minnesota, US	98	79	100	96	68	100
Ontario, Canada	99	88	100	98	88	100
Quebec, Canada	98	90	100	97	86	100

Exhibit 10.9 TIMSS 2007 Within-country Scoring Reliability for the Eighth Grade Constructed-response Mathematics Items (Continued)

Countries	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Algeria	95	60	100	90	57	97
Armenia	99	94	100	97	75	100
Australia	99	93	100	97	86	100
Bahrain	100	97	100	99	96	100
Bosnia and Herzegovina	98	90	100	96	83	100
Botswana	98	84	100	96	76	100
Bulgaria	96	70	100	94	68	100
Chinese Taipei	98	47	100	97	43	100
Colombia	99	92	100	97	89	100
Czech Republic	98	86	100	96	81	100
Egypt	99	94	100	97	89	100
El Salvador	100	98	100	100	96	100
England	99	94	100	98	85	100
Georgia	97	76	100	95	75	100
Ghana	100	98	100	99	92	100
Hong Kong SAR	99	95	100	99	94	100
Hungary	98	84	100	97	80	100
Indonesia	98	90	100	95	82	100
Iran, Islamic Rep. of	99	93	100	97	86	100
Israel	96	82	100	92	69	99
Italy	99	85	100	98	68	100
Japan	97	84	100	94	71	100
Jordan	100	97	100	98	93	100
Korea, Rep. of	99	96	100	99	93	100
Kuwait	99	96	100	98	93	100
Lebanon	100	97	100	98	94	100
Lithuania	98	94	100	97	91	100
Malaysia	99	96	100	99	96	100
Malta	97	81	100	95	73	100
Norway	99	94	100	97	86	100
Oman	99	95	100	97	93	100
Palestinian Nat'l Auth.	98	89	100	96	83	100
Qatar	99	91	100	98	86	100
Romania	99	96	100	99	95	100
Russian Federation	100	98	100	99	96	100
Saudi Arabia	100	97	100	99	92	100
Scotland	99	95	100	98	89	100
Serbia	99	94	100	98	93	100
Singapore	98	93	100	97	91	100
Slovenia	100	98	100	99	96	100
Sweden	98	86	100	96	84	100
Syrian Arab Republic	99	95	100	98	93	100
Thailand	98	89	100	97	82	100
Tunisia	97	87	100	95	74	100
Turkey	100	95	100	99	92	100
Ukraine	98	80	100	97	79	100
United States	97	86	100	95	77	100
Morocco	95	75	100	89	57	99
International Avg.	98	89	100	97	83	100
Benchmark Participants						
Basque Country, Spain	99	89	100	98	85	100
British Columbia, Canada	98	89	100	97	85	100
Dubai, UAE	97	87	100	95	83	100
Massachusetts, US	97	78	100	95	74	100
Minnesota, US	97	81	100	95	76	100
Ontario, Canada	98	87	100	97	85	100
Quebec, Canada	99	94	100	98	91	100

Exhibit 10.10 TIMSS 2007 Within-country Scoring Reliability for the Fourth Grade Constructed-response Science Items

Countries	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Algeria	88	69	98	78	50	96
Armenia	98	93	100	95	77	100
Australia	99	95	100	98	92	100
Austria	98	90	100	96	89	100
Chinese Taipei	97	74	100	96	74	100
Colombia	98	92	100	97	89	100
Czech Republic	94	78	100	91	74	100
Denmark	91	72	100	86	68	99
El Salvador	99	78	100	98	72	100
England	98	88	100	95	84	100
Georgia	92	68	100	86	68	98
Germany	93	73	100	91	69	100
Hong Kong SAR	99	98	100	99	97	100
Hungary	99	96	100	99	94	100
Iran, Islamic Rep. of	97	83	100	96	78	100
Italy	98	85	100	97	82	100
Japan	97	88	100	95	82	100
Kazakhstan	99	97	100	99	97	100
Kuwait	99	94	100	96	89	99
Latvia	85	42	100	80	36	99
Lithuania	95	80	100	92	78	100
Morocco	93	75	100	85	43	98
Netherlands	92	71	100	88	61	99
New Zealand	97	90	100	95	86	100
Norway	97	88	100	95	87	99
Qatar	99	94	100	96	88	100
Russian Federation	100	99	100	100	98	100
Scotland	97	87	100	95	80	100
Singapore	96	90	100	95	90	100
Slovak Republic	99	97	100	98	93	100
Slovenia	99	93	100	99	93	100
Sweden	93	65	100	89	62	100
Tunisia	93	77	100	88	67	99
Ukraine	100	98	100	100	98	100
United States	94	68	100	90	66	100
Yemen	96	85	100	89	67	98
International Avg.	96	83	100	93	78	100
Benchmark Participants						
Alberta, Canada	98	86	100	97	86	100
British Columbia, Canada	99	89	100	96	84	100
Dubai, UAE	93	73	100	89	71	99
Massachusetts, US	94	72	100	91	65	100
Minnesota, US	94	74	100	91	55	100
Ontario, Canada	98	90	100	97	88	100
Quebec, Canada	99	91	100	97	88	100

Exhibit 10.10 TIMSS 2007 Within-country Scoring Reliability for the Eighth Grade Constructed-response Science Items (Continued)

Countries	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Algeria	94	75	100	89	70	99
Armenia	98	89	100	95	75	100
Australia	97	88	100	95	86	100
Bahrain	94	78	100	90	67	100
Bosnia and Herzegovina	95	74	100	91	72	99
Botswana	95	79	100	89	73	100
Bulgaria	91	69	100	86	59	100
Chinese Taipei	94	66	100	90	63	100
Colombia	98	88	100	96	84	100
Czech Republic	93	75	100	90	64	100
Egypt	97	88	100	94	80	99
El Salvador	100	98	100	99	92	100
England	97	88	100	95	80	100
Georgia	92	67	100	85	53	100
Ghana	99	96	100	98	94	100
Hong Kong SAR	99	96	100	98	94	100
Hungary	95	86	100	92	80	100
Indonesia	97	81	100	92	75	100
Iran, Islamic Rep. of	97	86	100	95	79	100
Israel	92	73	100	84	66	99
Italy	96	63	100	94	60	100
Japan	91	54	100	85	54	100
Jordan	99	93	100	96	74	100
Korea, Rep. of	99	95	100	98	87	100
Kuwait	99	88	100	97	87	100
Lebanon	100	97	100	98	95	100
Lithuania	97	90	100	96	84	100
Malaysia	99	96	100	98	93	100
Malta	93	81	100	89	75	99
Norway	97	88	100	95	85	100
Oman	99	95	100	94	81	100
Palestinian Nat'l Auth.	94	82	100	88	69	99
Qatar	99	95	100	98	91	100
Romania	99	89	100	98	89	100
Russian Federation	99	93	100	98	92	100
Saudi Arabia	99	90	100	98	88	100
Scotland	97	84	100	95	77	100
Serbia	97	74	100	94	74	100
Singapore	96	90	100	94	90	100
Slovenia	100	95	100	99	93	100
Sweden	92	70	100	88	64	100
Syrian Arab Republic	99	92	100	98	91	100
Thailand	90	73	100	83	63	100
Tunisia	91	61	100	85	61	100
Turkey	97	81	100	94	63	100
Ukraine	92	68	100	86	52	100
United States	93	73	100	88	61	100
Morocco	90	58	99	81	49	98
International Avg.	96	82	100	93	76	100
Benchmark Participants						
Basque Country, Spain	97	86	100	96	77	100
British Columbia, Canada	96	81	100	92	77	100
Dubai, UAE	96	88	100	94	83	100
Massachusetts, US	92	76	100	88	62	100
Minnesota, US	93	77	100	89	61	100
Ontario, Canada	96	84	100	93	82	100
Quebec, Canada	95	84	100	92	82	100

10.3.2.2 Trend Item Scoring Reliability

TIMSS 2007 also took steps to show that the constructed-response items from 2003 that were used in 2007 were scored in the same way in both assessments. In anticipation of this, countries that participated in TIMSS 2003 sent samples of scored student booklets from the 2003 data collection to the IEA Data Processing and Research Center, where they were digitally scanned and stored in presentation software for later use. As a check on scoring consistency from 2003 to 2007, staff members working in each country on scoring the 2007 fourth- and eighth-grade data were asked also to score these 2003 responses using the DPC software.

As shown in Exhibit 10.11 for mathematics and Exhibit 10.12 for science, there was a very high degree of scoring consistency, with 97 percent exact agreement for both grades in mathematics, on average internationally, between the scores awarded in 2003 and those given by the 2007 scorers. The average exact percent agreement in science was 93 percent for fourth grade and 94 percent for eighth grade. There also was high agreement in mathematics at the diagnostic score level, with 96 and 94 percent exact agreement, on average, for grades four and eight, respectively. It was somewhat less in science, with 86 percent at grade four and 88 percent at grade eight, on average.

Exhibit 10.11 TIMSS 2007 Trend Scoring Reliability (2003–2007) for the Fourth Grade Constructed-response Mathematics Items

Countries	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Armenia	96	90	100	93	65	99
Australia	97	84	100	96	83	100
Chinese Taipei	97	93	100	96	88	100
England	98	92	100	97	87	100
Hong Kong SAR	99	93	100	98	87	100
Hungary	99	96	100	97	92	100
Iran, Islamic Rep. of	98	95	100	96	86	100
Japan	98	93	100	96	88	100
Lithuania	97	88	100	94	74	100
Netherlands	97	90	99	95	88	99
New Zealand	98	95	100	97	90	100
Norway	98	96	100	97	93	100
Russian Federation	99	95	100	98	92	100
Scotland	96	91	100	95	90	100
Singapore	95	86	100	93	83	100
Slovenia	96	68	99	93	47	99
Tunisia	98	97	100	95	81	100
United States	98	92	100	96	88	100
International Avg.	97	91	100	96	83	100
Benchmark Participants						
Alberta, Canada	98	91	99	96	85	99
British Columbia, Canada	98	91	99	96	85	99
Ontario, Canada	98	91	99	96	85	99
Quebec, Canada	98	91	99	96	85	99

Exhibit 10.11 TIMSS 2007 Trend Scoring Reliability (2003–2007) for the Eighth Grade Constructed-response Mathematics Items (Continued)

Countries	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Armenia	96	80	100	94	74	100
Bahrain	98	79	100	96	77	100
Botswana	95	87	99	93	81	98
Bulgaria	95	80	100	92	76	100
Chinese Taipei	96	83	100	94	70	100
Egypt	97	82	100	92	75	100
England	97	92	100	95	83	100
Ghana	99	96	100	97	93	100
Hong Kong SAR	98	94	100	97	90	100
Hungary	96	88	100	94	80	100
Indonesia	98	88	100	95	88	100
Iran, Islamic Rep. of	98	92	100	95	88	99
Israel	95	86	99	91	75	98
Japan	97	91	100	95	80	100
Jordan	97	63	100	96	45	100
Korea, Rep. of	96	86	100	94	81	100
Lithuania	97	82	100	93	70	100
Malaysia	97	89	100	95	82	99
Norway	97	87	100	94	79	100
Palestinian Nat'l Auth.	95	83	100	93	80	100
Russian Federation	98	94	100	95	84	100
Scotland	94	84	100	92	77	100
Serbia	96	87	100	94	85	99
Singapore	96	80	100	94	78	100
Slovenia	96	86	100	94	75	100
Sweden	97	89	100	94	82	100
Tunisia	98	90	100	95	82	100
United States	97	88	100	94	74	100
International Avg.	97	86	100	94	79	100
Benchmark Participants						
Basque Country, Spain	97	89	100	95	80	100
British Columbia, Canada	96	83	100	92	68	99
Ontario, Canada	96	83	100	92	68	99
Quebec, Canada	96	83	100	92	68	99

Exhibit 10.12 TIMSS 2007 Trend Scoring Reliability (2003–2007) for the Fourth Grade Constructed-response Science Items

Countries	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Armenia	91	75	99	80	57	91
Australia	93	88	100	88	77	99
Chinese Taipei	91	33	99	85	33	97
England	95	86	99	90	79	99
Hong Kong SAR	93	86	100	89	73	99
Hungary	94	85	100	88	72	100
Iran, Islamic Rep. of	92	80	99	84	77	98
Japan	92	85	99	87	70	98
Lithuania	94	87	100	85	71	99
Netherlands	92	84	97	85	75	97
New Zealand	94	85	100	87	67	100
Norway	95	88	99	91	81	99
Russian Federation	95	85	100	91	72	97
Scotland	92	80	100	88	69	100
Singapore	92	84	99	88	77	95
Slovenia	89	75	100	65	40	88
Tunisia	94	76	99	86	74	97
United States	92	84	99	84	64	98
International Avg.	93	80	99	86	68	97
Benchmark Participants						
Alberta, Canada	91	80	100	84	65	99
British Columbia, Canada	91	80	100	84	65	99
Ontario, Canada	91	80	100	84	65	99
Quebec, Canada	91	80	100	84	65	99

Exhibit 10.12 TIMSS 2007 Trend Scoring Reliability (2003–2007) for the Eighth Grade Constructed-response Science Items (Continued)

Countries	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Armenia	93	75	99	87	56	99
Bahrain	96	91	99	90	81	97
Botswana	92	79	99	86	67	98
Bulgaria	94	85	100	88	70	100
Chinese Taipei	91	67	100	81	36	100
Egypt	91	74	98	82	65	98
England	91	67	100	87	59	100
Ghana	99	95	100	96	87	99
Hong Kong SAR	95	87	100	91	74	100
Hungary	94	88	99	89	73	98
Indonesia	96	91	100	91	80	100
Iran, Islamic Rep. of	94	86	100	87	72	100
Israel	94	85	100	86	61	100
Japan	94	78	100	85	57	100
Jordan	99	96	100	98	85	100
Korea, Rep. of	94	80	100	88	68	99
Lithuania	94	82	100	87	74	100
Malaysia	95	86	100	91	75	99
Norway	93	84	100	87	72	100
Palestinian Nat'l Auth.	94	87	100	87	76	99
Russian Federation	97	92	100	93	86	99
Scotland	94	83	100	90	68	100
Serbia	95	86	99	90	74	99
Singapore	93	80	100	87	69	100
Slovenia	91	77	99	83	64	99
Sweden	93	83	100	87	76	99
Tunisia	97	84	100	90	75	100
United States	92	79	99	85	71	99
International Avg.	94	83	100	88	70	99
Benchmark Participants						
Basque Country, Spain	95	87	100	90	75	99
British Columbia, Canada	91	79	99	84	65	98
Ontario, Canada	91	79	99	84	65	98
Quebec, Canada	91	79	99	84	65	98

10.3.2.3 Cross-Country Scoring Reliability Study

Because of the many different languages in use in TIMSS 2007, establishing the reliability of constructed-response scoring across all countries was not feasible. However, TIMSS 2007 did conduct a cross-country study of scoring reliability among Northern Hemisphere countries that had scorers who were proficient in English.[8] A sample of student responses was provided by the English-speaking Southern Hemisphere countries. It included 200 student responses for each of 18 fourth-grade and 20 eighth-grade mathematics items and 23 fourth-grade and 20 eighth-grade science items (81 in total, representing about one-quarter of constructed-response items at the two grades) collected from Australia, Botswana, New Zealand, and Singapore. This set of 16,200 student responses in English was then scored independently in each country that had two scorers proficient in English. In all, 52 scorers from 30 countries at fourth grade and 67 scorers from 38 countries at eighth grade participated in the study. Scoring for this study took place shortly after the other scoring reliability activities were completed. Making all possible comparisons among scorers gave 1,225 comparisons at fourth grade and 2,211 comparisons at eighth grade for each student response to each item. This resulted in 265,200 total comparisons at fourth grade and 442,200 total comparisons at eighth grade when aggregated across all 200 student responses to that item. Agreement across countries was defined in terms of the percentage of these comparisons that were in exact agreement.

Exhibits 10.13 and 10.14 show that scorer reliability across countries was high for mathematics, with the percent exact agreement averaging 95 percent across the 18 items for the correctness score and 93 percent for the diagnostic score at fourth grade, and 91 percent across the 20 mathematics items for the correctness score and 90 percent for the diagnostic score at eighth grade. For science, the percent exact agreement averaged 91 percent across the 23 items for the correctness score and 86 percent for the diagnostic score at fourth grade, and 83 percent across the 20 science items for the correctness score and 76 percent for the diagnostic score at eighth grade.

[8] See Chapter 6 for further details.

Exhibit 10.13 TIMSS 2007 Cross-country Scoring Reliability for Constructed-response Mathematics Items – Fourth Grade

Item Label	Total Valid Comparisons	Exact Percent Agreement	
		Correctness Score Agreement	Diagnostic Score Agreement
M04_02 - M041056	265200	98	96
M04_04 - M041076	265200	99	98
M04_07 - M041146	265200	92	92
M04_09 - M041258A	265200	96	94
M04_09 - M041258B	265200	86	74
M04_11 - M041275	265200	85	85
M05_02 - M031309	265200	99	99
M05_04 - M031242A	265200	98	97
M05_04 - M031242B	265200	97	96
M05_05 - M031247	265200	94	91
M11_02 - M031009	265200	100	99
M11_04 - M031316	265200	99	99
M11_06 - M031079B	261579	99	99
M11_06 - M031079C	261579	97	97
M11_09 - M031325	265200	97	92
M12_04 - M041059	265200	99	95
M12_13 - M041276A	265200	98	98
M12_13 - M041276B	265200	83	79
Average Percent Agreement		95	93

TIMSS 2007 Cross-country Scoring Reliability for Constructed-response Mathematics Items – Eighth Grade

Item Label	Total Valid Comparisons	Exact Percent Agreement	
		Correctness Score Agreement	Diagnostic Score Agreement
M04_05 - M042304A	442200	94	93
M04_05 - M042304B	442200	86	85
M04_05 - M042304C	442200	93	93
M04_05 - M042304D	442200	79	77
M04_11 - M042130	442200	92	87
M04_12 - M042303A	442200	93	90
M04_12 - M042303B	442200	88	88
M05_03 - M032640	442200	91	91
M05_04 - M032344	442002	94	94
M05_05 - M032754	442200	92	92
M05_06 - M032755	442200	89	84
M11_02 - M032725	442200	94	92
M11_03 - M032683	442200	89	83
M11_13 - M032681A	442200	93	91
M11_13 - M032681B	442200	93	92
M11_13 - M032681C	442200	94	94
M12_03 - M042194	442200	95	95
M12_04 - M042114A	442200	93	91
M12_04 - M042114B	442200	94	94
M12_07 - M042050	442200	95	95
Average Percent Agreement		91	90

Exhibit 10.14 TIMSS 2007 Cross-country Scoring Reliability for Constructed-response Science Items – Fourth Grade

Item Label	Total Valid Comparisons	Exact Percent Agreement	
		Correctness Score Agreement	Diagnostic Score Agreement
S04_02 - S041023	265200	90	88
S04_04 - S041001	265200	74	74
S04_05 - S041029	265200	91	86
S04_08 - S041179	265200	99	99
S04_11 - S041216	265200	94	93
S04_12 - S041061	265200	99	99
S04_13 - S041202	265200	84	81
S05_02 - S031240A	265200	83	74
S05_02 - S031240B	265200	84	76
S05_04 - S031235A	265200	95	87
S05_04 - S031235B	265200	92	83
S05_06 - S031399A	265200	92	85
S05_06 - S031399B	265200	97	86
S05_07 - S031393	265200	93	84
S05_08 - S031278	265200	93	85
S11_03 - S031233	265200	99	97
S11_04 - S031204	265200	88	87
S11_06 - S031299	265200	92	90
S11_10 - S031088A	265200	95	93
S11_10 - S031088B	265200	83	68
S12_01 - S041027	265200	99	99
S12_02 - S041043	265200	89	85
S12_05 - S041006	265200	79	79
Average Percent Agreement		91	86

TIMSS 2007 Cross-country Scoring Reliability for Constructed-response Science Items – Eighth Grade

Item Label	Total Valid Comparisons	Exact Percent Agreement	
		Correctness Score Agreement	Diagnostic Score Agreement
S04_04 - S042052	442200	65	65
S04_06 - S042043	442200	89	66
S04_07 - S042196	416000	90	90
S04_09 - S042292	415874	78	71
S04_11 - S042232A	409600	86	86
S04_11 - S042232B	409600	86	86
S04_13 - S042149	442200	75	75
S04_14 - S042155	442200	83	83
S05_02 - S022292	416000	89	74
S05_06 - S022078	416000	89	79
S05_08 - S022281	416000	88	82
S05_11 - S032519	442200	75	63
S05_14 - S032120A	442200	77	63
S05_14 - S032120B	442200	86	69
S11_03 - S032306	442200	82	78
S11_04 - S032640	442200	81	68
S11_06 - S032570	416000	80	75
S11_08 - S032272	416000	92	82
S11_10 - S032060	442200	93	93
Average Percent Agreement		83	76

10.4 Summary of Review of TIMSS 2007 Item Statistics

Based on the information from the comprehensive collection of item analyses and reliability data that were computed and summarized for TIMSS 2007, as described in this chapter, the TIMSS & PIRLS International Study Center thoroughly reviewed all item statistics for every participating country to ensure that the items were performing comparably across countries. In particular, items with the following problems were considered for possible deletion from the international database:

- An error was detected during TIMSS 2007 translation verification but was not corrected before test administration.
- Data checking revealed a multiple-choice item with more or fewer options than in the international version.
- The item analysis showed the item to have a negative biserial, or, for an item with more than 1 score point, a nonmonotonic relationship between score level and total score.
- The item-by-country interaction results showed a very large negative interaction for a particular country.
- For constructed-response items, the within-country scoring reliability data showed an agreement of less than 70 percent.
- For trend items, an item performed substantially differently in 2007 compared to 2003, or an item was not included in the 2003 assessment for a particular country.

When the item statistics indicated a problem with an item, the documentation from the translation verification[9] was used as an aid in checking the test booklets. If a question remained about potential translation or cultural issues, however, then the National Research Coordinator was consulted before deciding how the item should be treated. If a problem could be detected by the TIMSS & PIRLS International Study Center (such as a negative point-biserial for a correct answer or too few options for a multiple-choice item), the item was deleted from the international scaling.

The checking of the TIMSS 2007 achievement data involved review of 782 items for 59 countries and 8 benchmarking participants at both grades (total of more than 52,000 item-country combinations), and resulted in the detection of very few items that were inappropriate for international comparisons. Among the few items singled out in the review process

[9] See Chapter 4 for a description of the process for translating and verifying the TIMSS 2007 data-collection instruments.

were mostly items with differences attributable to either translation or printing problems. Appendix C, Country Adaptations to Items and Item Scoring, provides a list of deleted items, as well as a list of recodes made to constructed-response item codes.

References

TIMSS & PIRLS International Study Center. (2006). *TIMSS 2007 survey operations procedures unit 5: Scoring the TIMSS 2007 assessment*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

TIMSS & PIRLS
International Study Center
Lynch School of Education, Boston College