Chapter 10

Item Analysis and Review

Michael O. Martin, Ann M. Kennedy, and Kathleen L. Trong

10.1 Overview

An important stage in creating the PIRLS 2006 achievement scale was an extensive review of the item statistics prior to item response theory (IRT) scaling. This review was conducted by the TIMSS & PIRLS International Study Center and involved evaluating the psychometric characteristics of each item within and across the participating countries. The purpose of this review was to ensure the quality of PIRLS achievement data by screening items for unusual item characteristics that could be attributed to an error, identifying the source, and rectifying the problem. For example, an item with low discrimination in one country atypical of the item's discrimination power in general may indicate a translation or printing problem. Also, for the trend items, item statistics were compared between 2001 and 2006.

In the few cases where country-level problems were identified, the TIMSS & PIRLS International Study Center consulted translation verification materials, checked printed booklets, and contacted National Research Coordinators (NRCs) to determine the source of the problem. When necessary, the item was removed from the international database for that country. This chapter describes the review process and the basic item statistics that were employed, using examples from the assessment.

10.2 Statistics for Item Analysis

As a first step, the TIMSS & PIRLS International Study Center created data almanacs containing the basic statistics for each achievement item. Exhibits 10.1 and 10.2 show examples of these statistics for a multiple-choice and constructed-response item, respectively. As these exhibits show, statistics were computed for each country individually, as well as on average internationally. The five Canadian provinces, listed below the international average row, were not included in the calculation of the international average.

For each item, almanacs include the number of students who were administered the item, item difficulty, item discrimination, and the percentage of boys and girls who responded correctly. For multiple-choice items, the percentage of students who chose each option and the percentage of students who omitted or did not reach the item was computed. In addition, the point-biserial correlation between each option and the total score was calculated. For constructed-response items, the percentage of students at each score level, the difficulty and discrimination for each score level (items could have up to 3 points), the number of responses that were double-scored, and the reliability between the two scorers were calculated. More detailed descriptions of these statistics are provided below.

- N: The number of students who were administered the item. If a student did not reach the item, it was considered not administered during item analysis.[1]
- Diff: Item difficulty, calculated as the percentage of students providing a correct response to the item. For constructed-response items worth more than one point, this is the students' average score as a percentage of the maximum score points for the item. Items that were not reached by the students were treated as not administered when computing this statistic.
- Disc: Item discrimination, calculated as the correlation between a correct response to the item and the total score on all items in the test booklet.[2] For constructed-response items worth more than one point, the correlation between the number of score points and total score was used. Items exhibiting good measurement properties should have a moderately positive correlation.

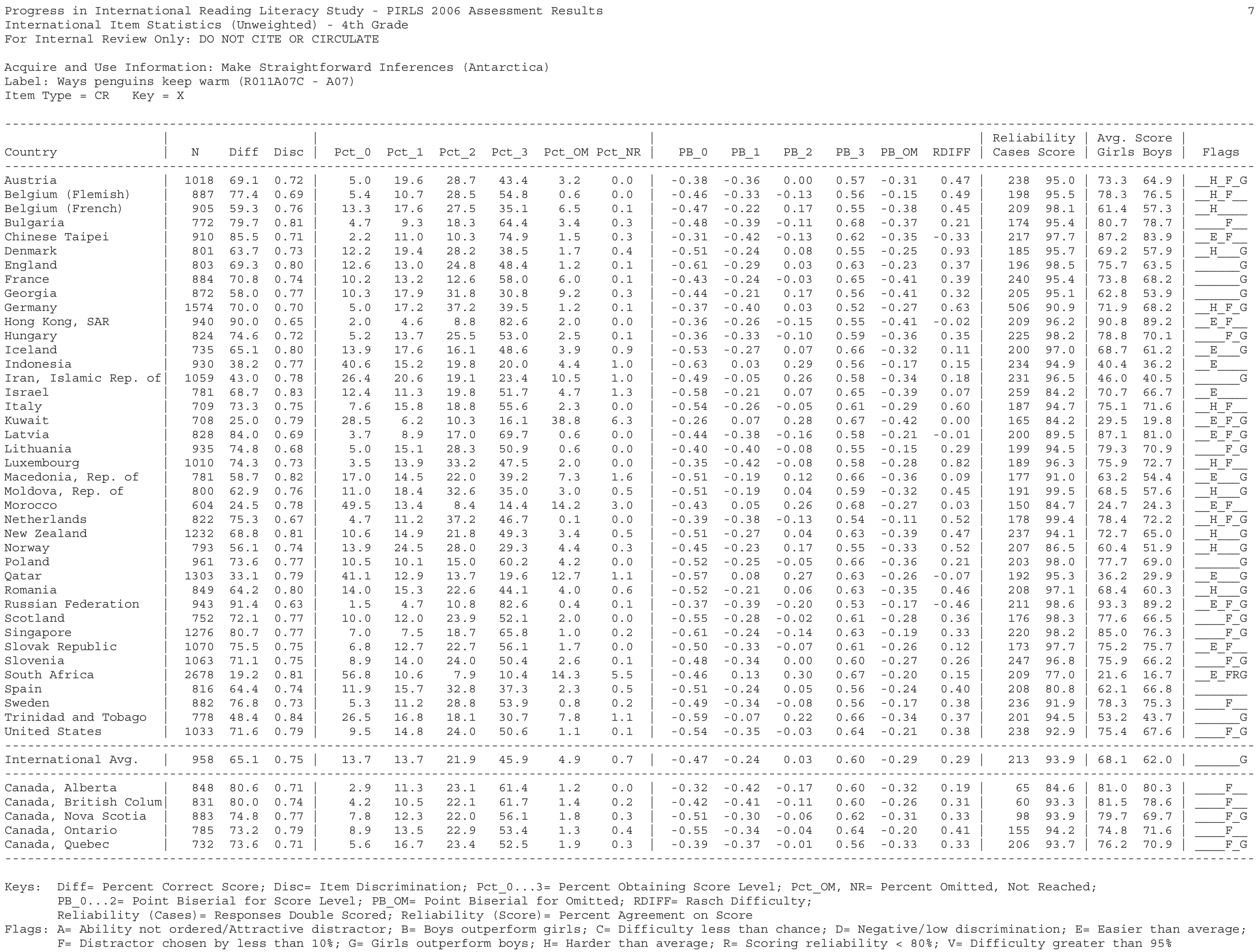
[1] For the purpose of item analysis and item parameter estimation for scaling, items not reached by the student were considered not administered. However, these items were treated as incorrect when estimating student proficiency.

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

Exhibit 10.1 International Item Statistics for a Multiple-choice Item

```
Progress in International Reading Literacy Study - PIRLS 2006 Assessment Results
International Item Statistics (Unweighted) - 4th Grade
For Internal Review Only: DO NOT CITE OR CIRCULATE

Acquire and Use Information: Focus on and Retrieve Explicitly Stated Information and Ideas (Searching for Food)
Label: How do other ants from nest find food too (R021S04M - S04)
Item Type = MC   Key = C

--------------------------------------------------------------------------------------------------------------------------------------------------------------------
                         |                  |                                                      |                                             |     Avg. Score |
Country                  |    N   Diff  Disc|  Pct_A  Pct_B  Pct_C  Pct_D  Pct_In Pct_OM Pct_NR |   PB_A   PB_B   PB_C   PB_D  PB_In  PB_OM | RDIFF  Girls Boys |  Flags
--------------------------------------------------------------------------------------------------------------------------------------------------------------------
Austria                  | 1008  69.3  0.51 |   13.1    5.0   69.3   10.1    0.0    2.5    0.0 |  -0.22  -0.25   0.51  -0.25   0.00  -0.18 | -0.40  65.8  72.6 | ____F_B
Belgium (Flemish)        |  859  65.0  0.42 |   19.7    2.1   65.0   12.3    0.0    0.9    0.0 |  -0.27  -0.10   0.42  -0.23   0.00  -0.05 | -0.31  63.0  67.0 | __H_F__
Belgium (French)         |  907  60.5  0.49 |   17.4    6.7   60.5   13.6    0.0    1.8    0.0 |  -0.25  -0.22   0.49  -0.20   0.00  -0.15 | -0.70  57.9  63.1 | ____F__
Bulgaria                 |  756  69.7  0.54 |   17.2    4.5   69.7    8.3    0.0    0.3    0.0 |  -0.29  -0.28   0.54  -0.29   0.00  -0.05 | -0.44  68.9  70.5 | ____F__
Chinese Taipei           |  912  82.1  0.49 |   10.0    3.4   82.1    3.7    0.0    0.8    0.0 |  -0.34  -0.27   0.49  -0.13   0.00  -0.12 | -1.32  81.1  83.1 | __E_F__
Denmark                  |  770  78.7  0.46 |   10.4    3.0   78.7    7.3    0.0    0.6    0.1 |  -0.26  -0.27   0.46  -0.24   0.00  -0.07 | -0.88  80.7  76.5 | ____F__
England                  |  808  69.4  0.53 |   15.6    4.1   69.4    9.8    0.0    1.1    0.1 |  -0.22  -0.29   0.53  -0.30   0.00  -0.15 | -0.58  69.9  69.0 | ____F__
France                   |  875  80.5  0.42 |    8.9    3.4   80.5    6.1    0.0    1.1    0.0 |  -0.20  -0.25   0.42  -0.22   0.00  -0.10 | -1.50  83.4  77.6 | __E_F_G
Georgia                  |  845  43.1  0.53 |   21.7   17.0   43.1   16.6    0.0    1.7    0.1 |  -0.17  -0.26   0.53  -0.21   0.00  -0.12 | -0.32  39.8  46.4 | __H____
Germany                  | 1572  78.1  0.51 |    8.5    4.2   78.1    8.0    0.0    1.3    0.1 |  -0.28  -0.25   0.51  -0.25   0.00  -0.14 | -0.86  78.1  78.0 | __E_F__
Hong Kong, SAR           |  947  78.8  0.44 |   14.7    2.4   78.8    4.0    0.0    0.1    0.0 |  -0.28  -0.31   0.44  -0.16   0.00  -0.05 | -0.99  79.5  78.1 | __E_F__
Hungary                  |  791  67.6  0.54 |   11.9    7.6   67.6   11.3    0.0    1.6    0.0 |  -0.31  -0.20   0.54  -0.25   0.00  -0.15 | -0.40  62.2  72.9 | ____F_B
Iceland                  |  730  65.6  0.49 |   13.4    5.5   65.6   14.4    0.0    1.1    0.1 |  -0.25  -0.25   0.49  -0.24   0.00  -0.05 | -0.73  64.9  66.3 | ____F__
Indonesia                |  915  38.3  0.41 |   21.6   14.4   38.3   24.3    0.0    1.4    0.1 |  -0.09  -0.22   0.41  -0.17   0.00  -0.09 | -0.51  40.0  36.4 | _______
Iran, Islamic Rep. of    | 1070  52.4  0.50 |   22.2   12.0   52.4   11.4    0.0    2.0    0.3 |  -0.17  -0.29   0.50  -0.21   0.00  -0.13 | -1.01  51.7  53.0 | __E____
Israel                   |  676  65.7  0.59 |   14.6   10.1   65.7    8.4    0.0    1.2    0.3 |  -0.29  -0.30   0.59  -0.24   0.00  -0.17 | -1.07  61.8  69.5 | __E_F_B
Italy                    |  720  67.9  0.49 |   14.6    6.5   67.9   10.6    0.0    0.4    0.0 |  -0.31  -0.26   0.49  -0.17   0.00  -0.06 | -0.48  67.9  68.0 | ____F__
Kuwait                   |  669  28.6  0.43 |   14.8   27.8   28.6   20.9    0.0    7.9    1.5 |  -0.08  -0.11   0.43  -0.15   0.00  -0.15 | -0.63  32.8  23.1 | ______G
Latvia                   |  817  66.3  0.45 |   17.9    4.4   66.3   10.6    0.0    0.7    0.1 |  -0.22  -0.24   0.45  -0.22   0.00  -0.08 | -0.25  67.5  65.2 | __H_F__
Lithuania                |  946  58.8  0.48 |   20.1    5.5   58.8   13.6    0.0    2.0    0.0 |  -0.20  -0.19   0.48  -0.29   0.00  -0.10 | -0.12  56.8  60.7 | __H_F__
Luxembourg               |  991  71.2  0.52 |   13.9    4.5   71.2    9.3    0.0    1.0    0.0 |  -0.29  -0.26   0.52  -0.25   0.00  -0.09 | -0.40  68.2  74.4 | ____F_B
Macedonia, Rep. of       |  787  42.7  0.51 |   27.4   13.6   42.7   12.3    0.0    3.9    0.1 |  -0.12  -0.30   0.51  -0.20   0.00  -0.16 | -0.50  43.6  41.8 | _______
Moldova, Rep. of         |  776  56.2  0.55 |   19.6   11.3   56.2   12.4    0.0    0.5    0.0 |  -0.27  -0.29   0.55  -0.22   0.00  -0.03 | -0.40  55.4  57.0 | _______
Morocco                  |  613  32.6  0.43 |   17.0   27.4   32.6   16.5    0.0    6.5    0.5 |  -0.09  -0.17   0.43  -0.13   0.00  -0.16 | -1.12  32.7  32.5 | __E____
Netherlands              |  795  70.8  0.45 |   16.6    2.8   70.8    9.7    0.0    0.1    0.0 |  -0.31  -0.17   0.45  -0.20   0.00  -0.01 | -0.56  69.9  71.8 | ____F__
New Zealand              | 1204  63.5  0.55 |   18.7    5.1   63.5   11.3    0.0    1.4    0.2 |  -0.25  -0.33   0.55  -0.24   0.00  -0.14 | -0.40  65.0  62.1 | ____F__
Norway                   |  744  58.1  0.45 |   21.2    3.0   58.1   15.6    0.0    2.2    0.3 |  -0.23  -0.21   0.45  -0.18   0.00  -0.18 | -0.41  56.1  60.0 | ____F__
Poland                   |  943  64.6  0.50 |   11.6    6.5   64.6   15.8    0.0    1.6    0.1 |  -0.22  -0.25   0.50  -0.23   0.00  -0.15 | -0.68  63.6  65.5 | ____F__
Qatar                    | 1287  31.9  0.50 |   17.1   25.2   31.9   24.0    0.0    1.9    0.4 |  -0.15  -0.22   0.50  -0.17   0.00  -0.04 | -0.89  34.3  29.4 | __E____
Romania                  |  831  53.3  0.50 |   28.3    7.9   53.3    9.3    0.0    1.2    0.4 |  -0.24  -0.28   0.50  -0.18   0.00  -0.13 | -0.42  51.6  55.0 | ____F__
Russian Federation       |  915  73.7  0.47 |   12.6    2.8   73.7   10.6    0.0    0.3    0.0 |  -0.23  -0.18   0.47  -0.30   0.00  -0.11 | -0.36  72.7  74.7 | ____F__
Scotland                 |  752  64.0  0.52 |   17.3    4.3   64.0   13.4    0.0    1.1    0.0 |  -0.29  -0.21   0.52  -0.25   0.00  -0.09 | -0.49  67.1  60.8 | ____F__
Singapore                | 1260  73.7  0.57 |   17.1    3.5   73.7    5.4    0.0    0.3    0.1 |  -0.36  -0.27   0.57  -0.27   0.00  -0.05 | -0.75  74.8  72.6 | ____F__
Slovak Republic          | 1073  68.2  0.52 |   12.5    8.5   68.2   10.3    0.0    0.6    0.1 |  -0.21  -0.25   0.52  -0.31   0.00  -0.08 | -0.58  69.7  66.8 | ____F__
Slovenia                 | 1044  59.4  0.43 |   24.0    5.3   59.4   10.2    0.0    1.1    0.1 |  -0.23  -0.20   0.43  -0.20   0.00  -0.06 | -0.22  58.7  60.0 | __H_F__
South Africa             | 2771  24.5  0.47 |   25.8   22.3   24.5   16.0    0.0   11.4    0.5 |  -0.07  -0.16   0.47  -0.07   0.00  -0.24 | -0.63  24.2  24.9 | _C_____
Spain                    |  793  51.2  0.49 |   26.4    7.3   51.2   14.0    0.0    1.1    0.0 |  -0.25  -0.21   0.49  -0.21   0.00  -0.08 | -0.05  46.6  55.2 | __H_F_B
Sweden                   |  854  62.3  0.43 |   19.3    4.7   62.3   11.7    0.0    2.0    0.0 |  -0.21  -0.17   0.43  -0.18   0.00  -0.22 |  0.01  58.8  65.8 | __H_F_B
Trinidad and Tobago      |  780  50.1  0.55 |   25.5    8.3   50.1   14.5    0.0    1.5    0.3 |  -0.28  -0.25   0.55  -0.20   0.00  -0.10 | -0.63  49.2  51.0 | ____F__
United States            | 1041  60.7  0.46 |   26.5    4.2   60.7    8.5    0.0    0.1    0.1 |  -0.29  -0.26   0.46  -0.16   0.00  -0.01 | -0.20  60.4  61.0 | __H_F__
--------------------------------------------------------------------------------------------------------------------------------------------------------------------
International Avg.       |  946  60.5  0.49 |   17.7    8.2   60.5   11.9    0.0    1.8    0.1 |  -0.23  -0.24   0.49  -0.21   0.00  -0.11 | -0.58  59.9  61.0 | ____F_B
--------------------------------------------------------------------------------------------------------------------------------------------------------------------
Canada, Alberta          |  855  71.1  0.41 |   17.8    2.5   71.1    8.2    0.0    0.5    0.0 |  -0.26  -0.11   0.41  -0.23   0.00  -0.07 | -0.40  70.1  72.1 | ____F__
Canada, British Colum    |  809  69.0  0.44 |   16.9    2.6   69.0   10.6    0.0    0.9    0.0 |  -0.25  -0.20   0.44  -0.23   0.00  -0.10 | -0.31  69.1  68.9 | ____F__
Canada, Nova Scotia      |  863  63.0  0.48 |   19.7    2.8   63.0   13.3    0.0    1.2    0.0 |  -0.25  -0.26   0.48  -0.22   0.00  -0.12 | -0.33  61.4  64.6 | ____F__
Canada, Ontario          |  781  65.7  0.44 |   19.3    4.1   65.7   10.4    0.0    0.5    0.0 |  -0.23  -0.24   0.44  -0.21   0.00  -0.06 | -0.38  62.5  68.8 | ____F__
Canada, Quebec           |  762  62.9  0.48 |   19.4    4.2   62.9   12.5    0.0    1.0    0.0 |  -0.23  -0.22   0.48  -0.24   0.00  -0.12 | -0.42  60.2  65.5 | ____F__
--------------------------------------------------------------------------------------------------------------------------------------------------------------------

Keys:  Diff= Percent Correct Score; Disc= Item Discrimination; Pct_A...D= Percent Choosing Each Option; Pct_In, OM, NR= Percent Invalid, Omitted, Not Reached;
       PB_A...D= Point Biserial for Each Option; PB_OM= Point Biserial for Omitted; RDIFF= Rasch Difficulty

Flags: A= Ability not ordered/Attractive distractor; B= Boys outperform girls; C= Difficulty less than chance; D= Negative/low discrimination; E= Easier than average;
       F= Distractor chosen by less than 10%; G= Girls outperform boys; H= Harder than average; V= Difficulty greater than 95%
```

Exhibit 10.2 International Item Statistics for a Constructed-response Item



```
Progress in International Reading Literacy Study - PIRLS 2006 Assessment Results                                                                                        7
International Item Statistics (Unweighted) - 4th Grade
For Internal Review Only: DO NOT CITE OR CIRCULATE

Acquire and Use Information: Make Straightforward Inferences (Antarctica)
Label: Ways penguins keep warm (R011A07C - A07)
Item Type = CR   Key = X

-----------------------------------------------------------------------------------------------------------------------------------------------------------------------
                     |                   |                                           |                                           |  Reliability | Avg. Score |
Country              |     N  Diff  Disc |  Pct_0  Pct_1  Pct_2  Pct_3 Pct_OM Pct_NR |   PB_0   PB_1   PB_2   PB_3  PB_OM  RDIFF |  Cases Score | Girls  Boys |   Flags
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------
Austria              |  1018  69.1  0.72 |    5.0   19.6   28.7   43.4    3.2    0.0 |  -0.38  -0.36   0.00   0.57  -0.31   0.47 |    238  95.0 |  73.3  64.9 | __H_F_G
Belgium (Flemish)    |   887  77.4  0.69 |    5.4   10.7   28.5   54.8    0.6    0.0 |  -0.46  -0.33  -0.13   0.56  -0.15   0.49 |    198  95.5 |  78.3  76.5 | __H_F__
Belgium (French)     |   905  59.3  0.76 |   13.3   17.6   27.5   35.1    6.5    0.1 |  -0.47  -0.22   0.17   0.55  -0.38   0.45 |    209  98.1 |  61.4  57.3 | __H____
Bulgaria             |   772  79.7  0.81 |    4.7    9.3   18.3   64.4    3.4    0.3 |  -0.48  -0.39  -0.11   0.68  -0.37   0.21 |    174  95.4 |  80.7  78.7 | ____F__
Chinese Taipei       |   910  85.5  0.71 |    2.2   11.0   10.3   74.9    1.5    0.3 |  -0.31  -0.42  -0.13   0.62  -0.35  -0.33 |    217  97.7 |  87.2  83.9 | __E_F__
Denmark              |   801  63.7  0.73 |   12.2   19.4   28.2   38.5    1.7    0.4 |  -0.51  -0.24   0.08   0.55  -0.25   0.93 |    185  95.7 |  69.2  57.9 | __H___G
England              |   803  69.3  0.80 |   12.6   13.0   24.8   48.4    1.2    0.1 |  -0.61  -0.29   0.03   0.63  -0.23   0.37 |    196  98.5 |  75.7  63.5 | ______G
France               |   884  70.8  0.74 |   10.2   13.2   12.6   58.0    6.0    0.1 |  -0.43  -0.24  -0.03   0.65  -0.41   0.39 |    240  95.4 |  73.8  68.2 | ______G
Georgia              |   872  58.0  0.77 |   10.3   17.9   31.8   30.8    9.2    0.3 |  -0.44  -0.21   0.17   0.56  -0.41   0.32 |    205  95.1 |  62.8  53.9 | ______G
Germany              |  1574  70.0  0.70 |    5.0   17.2   37.2   39.5    1.2    0.1 |  -0.37  -0.40   0.03   0.52  -0.27   0.63 |    506  90.9 |  71.9  68.2 | __H_F_G
Hong Kong, SAR       |   940  90.0  0.65 |    2.0    4.6    8.8   82.6    2.0    0.0 |  -0.36  -0.26  -0.15   0.55  -0.41  -0.02 |    209  96.2 |  90.8  89.2 | __E_F__
Hungary              |   824  74.6  0.72 |    5.2   13.7   25.5   53.0    2.5    0.1 |  -0.36  -0.33  -0.10   0.59  -0.36   0.35 |    225  98.2 |  78.8  70.1 | ____F_G
Iceland              |   735  65.1  0.80 |   13.9   17.6   16.1   48.6    3.9    0.9 |  -0.53  -0.27   0.07   0.66  -0.32   0.11 |    200  97.0 |  68.7  61.2 | __E___G
Indonesia            |   930  38.2  0.77 |   40.6   15.2   19.8   20.0    4.4    1.0 |  -0.63   0.03   0.29   0.56  -0.17   0.15 |    234  94.9 |  40.4  36.2 | __E____
Iran, Islamic Rep. of|  1059  43.0  0.78 |   26.4   20.6   19.1   23.4   10.5    1.0 |  -0.49  -0.05   0.26   0.58  -0.34   0.18 |    231  96.5 |  46.0  40.5 | ______G
Israel               |   781  68.7  0.83 |   12.4   11.3   19.8   51.7    4.7    1.3 |  -0.58  -0.21   0.07   0.65  -0.39   0.07 |    259  84.2 |  70.7  66.7 | __E____
Italy                |   709  73.3  0.75 |    7.6   15.8   18.8   55.6    2.3    0.0 |  -0.54  -0.26  -0.05   0.61  -0.29   0.60 |    187  94.7 |  75.1  71.6 | __H_F__
Kuwait               |   708  25.0  0.79 |   28.5    6.2   10.3   16.1   38.8    6.3 |  -0.26   0.07   0.28   0.67  -0.42   0.00 |    165  84.2 |  29.5  19.8 | __E_F_G
Latvia               |   828  84.0  0.69 |    3.7    8.9   17.0   69.7    0.6    0.0 |  -0.44  -0.38  -0.16   0.58  -0.21  -0.01 |    200  89.5 |  87.1  81.0 | __E_F_G
Lithuania            |   935  74.8  0.68 |    5.0   15.1   28.3   50.9    0.6    0.0 |  -0.40  -0.40  -0.08   0.55  -0.15   0.29 |    199  94.5 |  79.3  70.9 | ____F_G
Luxembourg           |  1010  74.3  0.73 |    3.5   13.9   33.2   47.5    2.0    0.0 |  -0.35  -0.42  -0.08   0.58  -0.28   0.82 |    189  96.3 |  75.9  72.7 | __H_F__
Macedonia, Rep. of   |   781  58.7  0.82 |   17.0   14.5   22.0   39.2    7.3    1.6 |  -0.51  -0.19   0.12   0.66  -0.36   0.09 |    177  91.0 |  63.2  54.4 | __E___G
Moldova, Rep. of     |   800  62.9  0.76 |   11.0   18.4   32.6   35.0    3.0    0.5 |  -0.51  -0.19   0.04   0.59  -0.32   0.45 |    191  99.5 |  68.5  57.6 | __H___G
Morocco              |   604  24.5  0.78 |   49.5   13.4    8.4   14.4   14.2    3.0 |  -0.43   0.05   0.26   0.68  -0.27   0.03 |    150  84.7 |  24.7  24.3 | __E_F__
Netherlands          |   822  75.3  0.67 |    4.7   11.2   37.2   46.7    0.1    0.0 |  -0.39  -0.38  -0.13   0.54  -0.11   0.52 |    178  99.4 |  78.4  72.2 | __H_F_G
New Zealand          |  1232  68.8  0.81 |   10.6   14.9   21.8   49.3    3.4    0.5 |  -0.51  -0.27   0.04   0.63  -0.39   0.47 |    237  94.1 |  72.7  65.0 | __H___G
Norway               |   793  56.1  0.74 |   13.9   24.5   28.0   29.3    4.4    0.3 |  -0.45  -0.23   0.17   0.55  -0.33   0.52 |    207  86.5 |  60.4  51.9 | __H___G
Poland               |   961  73.6  0.77 |   10.5   10.1   15.0   60.2    4.2    0.0 |  -0.52  -0.25  -0.05   0.66  -0.36   0.21 |    203  98.0 |  77.7  69.0 | ______G
Qatar                |  1303  33.1  0.79 |   41.1   12.9   13.7   19.6   12.7    1.1 |  -0.57   0.08   0.27   0.63  -0.26  -0.07 |    192  95.3 |  36.2  29.9 | __E___G
Romania              |   849  64.2  0.80 |   14.0   15.3   22.6   44.1    4.0    0.6 |  -0.52  -0.21   0.06   0.63  -0.35   0.46 |    208  97.1 |  68.4  60.3 | __H___G
Russian Federation   |   943  91.4  0.63 |    1.5    4.7   10.8   82.6    0.4    0.1 |  -0.37  -0.39  -0.20   0.53  -0.17  -0.46 |    211  98.6 |  93.3  89.2 | __E_F_G
Scotland             |   752  72.1  0.77 |   10.0   12.0   23.9   52.1    2.0    0.0 |  -0.55  -0.28  -0.02   0.61  -0.28   0.36 |    176  98.3 |  77.6  66.5 | ____F_G
Singapore            |  1276  80.7  0.77 |    7.0    7.5   18.7   65.8    1.0    0.2 |  -0.61  -0.24  -0.14   0.63  -0.19   0.33 |    220  98.2 |  85.0  76.3 | ____F_G
Slovak Republic      |  1070  75.5  0.75 |    6.8   12.7   22.7   56.1    1.7    0.0 |  -0.50  -0.33  -0.07   0.61  -0.26   0.12 |    173  97.7 |  75.2  75.7 | __E_F__
Slovenia             |  1063  71.1  0.75 |    8.9   14.0   24.0   50.4    2.6    0.1 |  -0.48  -0.34   0.00   0.60  -0.27   0.26 |    247  96.8 |  75.9  66.2 | ____F_G
South Africa         |  2678  19.2  0.81 |   56.8   10.6    7.9   10.4   14.3    5.5 |  -0.46   0.13   0.30   0.67  -0.20   0.15 |    209  77.0 |  21.6  16.7 | __E_FRG
Spain                |   816  64.4  0.74 |   11.9   15.7   32.8   37.3    2.3    0.5 |  -0.51  -0.24   0.05   0.56  -0.24   0.40 |    208  80.8 |  62.1  66.8 | _______
Sweden               |   882  76.8  0.73 |    5.3   11.2   28.8   53.9    0.8    0.2 |  -0.49  -0.34  -0.08   0.56  -0.17   0.38 |    236  91.9 |  78.3  75.3 | ____F__
Trinidad and Tobago  |   778  48.4  0.84 |   26.5   16.8   18.1   30.7    7.8    1.1 |  -0.59  -0.07   0.22   0.66  -0.34   0.37 |    201  94.5 |  53.2  43.7 | ______G
United States        |  1033  71.6  0.79 |    9.5   14.8   24.0   50.6    1.1    0.1 |  -0.54  -0.35  -0.03   0.64  -0.21   0.38 |    238  92.9 |  75.4  67.6 | ____F_G
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------
International Avg.   |   958  65.1  0.75 |   13.7   13.7   21.9   45.9    4.9    0.7 |  -0.47  -0.24   0.03   0.60  -0.29   0.29 |    213  93.9 |  68.1  62.0 | ______G
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------
Canada, Alberta      |   848  80.6  0.71 |    2.9   11.3   23.1   61.4    1.2    0.0 |  -0.32  -0.42  -0.17   0.60  -0.32   0.19 |     65  84.6 |  81.0  80.3 | ____F__
Canada, British Colum|   831  80.0  0.74 |    4.2   10.5   22.1   61.7    1.4    0.2 |  -0.42  -0.41  -0.11   0.60  -0.26   0.31 |     60  93.3 |  81.5  78.6 | ____F__
Canada, Nova Scotia  |   883  74.8  0.77 |    7.8   12.3   22.0   56.1    1.8    0.3 |  -0.51  -0.30  -0.06   0.62  -0.31   0.33 |     98  93.9 |  79.7  69.7 | ____F_G
Canada, Ontario      |   785  73.2  0.79 |    8.9   13.5   22.9   53.4    1.3    0.4 |  -0.55  -0.34  -0.04   0.64  -0.20   0.41 |    155  94.2 |  74.8  71.6 | ____F__
Canada, Quebec       |   732  73.6  0.71 |    5.6   16.7   23.4   52.5    1.9    0.3 |  -0.39  -0.37  -0.01   0.56  -0.33   0.33 |    206  93.7 |  76.2  70.9 | ____F_G
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------

Keys:  Diff= Percent Correct Score; Disc= Item Discrimination; Pct_0...3= Percent Obtaining Score Level; Pct_OM, NR= Percent Omitted, Not Reached;
       PB_0...2= Point Biserial for Score Level; PB_OM= Point Biserial for Omitted; RDIFF= Rasch Difficulty;
       Reliability (Cases)= Responses Double Scored; Reliability (Score)= Percent Agreement on Score
Flags: A= Ability not ordered/Attractive distractor; B= Boys outperform girls; C= Difficulty less than chance; D= Negative/low discrimination; E= Easier than average;
       F= Distractor chosen by less than 10%; G= Girls outperform boys; H= Harder than average; R= Scoring reliability < 80%; V= Difficulty greater than 95%
```

- Pct_A, Pct_B, Pct_C, Pct_D: Used for multiple-choice items only, each column indicates the percentage of students choosing the particular response option (A, B, C, or D). Students who did not reach the item were excluded from the denominator for these calculations.
- Pct_0, Pct_1, Pct_2, Pct_3: Used for constructed-response items only, each column indicates the percentage of students earning the particular number of score points (0, 1, 2, or 3) for that item. Students who did not reach the item were excluded from the denominator for these calculations.
- Pct_In: Used for multiple-choice items only, this column indicates the percentage of students who provided an invalid response to the item. A typical invalid response was a student selecting multiple response options for an item.
- Pct_OM: This column indicates the percentage of students who reached the item but did not provide a response. Students who did not reach the item were excluded from the denominator when calculating this statistic.
- Pct_NR: This column indicates the percentage of students who did not reach the item. An item was considered not reached if the student did not respond to any subsequent items in the test booklet, and the previous item was omitted.
- PB_A, PB_B, PB_C, PB_D: Used for multiple-choice items only, each column indicates the point-biserial correlation between the particular option (A, B, C, or D) and the total score. Items with good psychometric properties have near-zero or negative correlation coefficients for the incorrect options and a moderately positive correlation coefficient for the correct option.
- PB_0, PB_1, PB_2, PB_3: Used for constructed-response items only, each column indicates the correlation between the particular score level (0, 1, 2, or 3) and the total score. Items with good psychometric properties should increase with each score level.
- PB_In: Used for multiple-choice items only, this column indicates the correlation between an invalid response to the item (i.e., selecting multiple response options) and total score. For an item with good psychometric properties, this should be negative or near zero.

- PB_OM: This column indicates the correlation between a binary variable indicating an omitted response to the item and the total score. For an item with good psychometric properties, this should be negative or near zero.
- RDIFF: This is an estimate of the item's difficulty based on a Rasch one-parameter IRT model. The difficulty estimate is expressed in logits (with a positive logit indicating a difficult item) and was scaled so that the average Rasch item difficulty was zero within each country.
- Avg. Score Girls/Boys: These columns indicate the average difficulty for the item separately for boys and girls.
- Reliability Cases: To provide a measure of the scoring reliability of constructed-response items, those items in approximately one quarter of the test booklets were scored by two independent scorers in each country. This column indicates the number of responses that were double-scored for an item.
- Reliability Score: Used for constructed-response items only, this column indicates the percentage of exact agreement between two independent scorers.

As an aid during the review process, the almanacs also include a series of "flags" to highlight conditions that may warrant a closer look. While not all flags necessarily signify a faulty item, they draw attention to potential problems.

The following conditions are flagged:

- Difficulty levels for the item are significantly different for boys and girls;
- Item difficulty is less than chance (e.g., 25% for multiple-choice items);
- Item difficulty exceeds 95 percent;
- Item discrimination (i.e., the point-biserial for the correct option) is less than 0.2;
- Rasch difficulty estimate is below the average of all items;
- Rasch difficulty estimate is above the average of all items;
- For multiple choice items, a greater percentage of students chose the incorrect response than the percentage of students who chose the correct option, or the point-biserial correlation for one or more of the distracters exceeds zero;

- Less than 10 percent of students chose one or more of the distracters (for multiple-choice items) or earned one or more of the score levels (for constructed-response items);
- Scoring reliability is less than 80 percent; and
- Students with lower total scores are more likely to answer an item correctly than students with higher total scores.

In addition to item-level statistics, the TIMSS & PIRLS International Study Center also examined descriptive statistics for each test block as a whole to gain a sense of the overall performance of the items associated with each passage. These statistics included the number of students who were administered the block, the minimum, maximum, and average number of items students answered correctly, the standard deviation, the percent correct overall and, for boys and girls separately, the minimum, maximum, and average point-biserial correlation across the items, and the Cronbach's alpha reliability coefficient for the block. Each of these was calculated for individual countries and on average internationally.

Eleven countries and the five Canadian provinces administered the assessment in more than one language.[3] In these cases, an additional step was taken to examine item statistics for each language group. Due to the fact that some language groups make up a small proportion of the overall sample in a country, problems with a particular language within a country may not have been evident when the languages are combined, and may indicate a translation error for that language. The same process and statistics that were described above were used for this step.

10.3 Examining Item-by-Country Interactions

While it is reasonable to expect country performance to vary somewhat across items, countries with high average performance should generally perform well on each of the items, and countries with lower overall performance should do less well on individual items. When this pattern is not followed (i.e., a high-performing country does poorly on a particular item), this is called an item-by-country interaction. If this interaction is large, it could indicate a problem with the item that should be investigated and addressed.

To easily detect item-by-country interactions, the TIMSS & PIRLS International Study Center created plots that graphically display the Rasch difficulties for each item. Exhibit 10.3 displays the 95 percent confidence interval

[3] See Chapter 5 for details about translation procedures.

for the national average Rasch difficulty estimate. The limits for the confidence intervals were computed as follows:

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

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

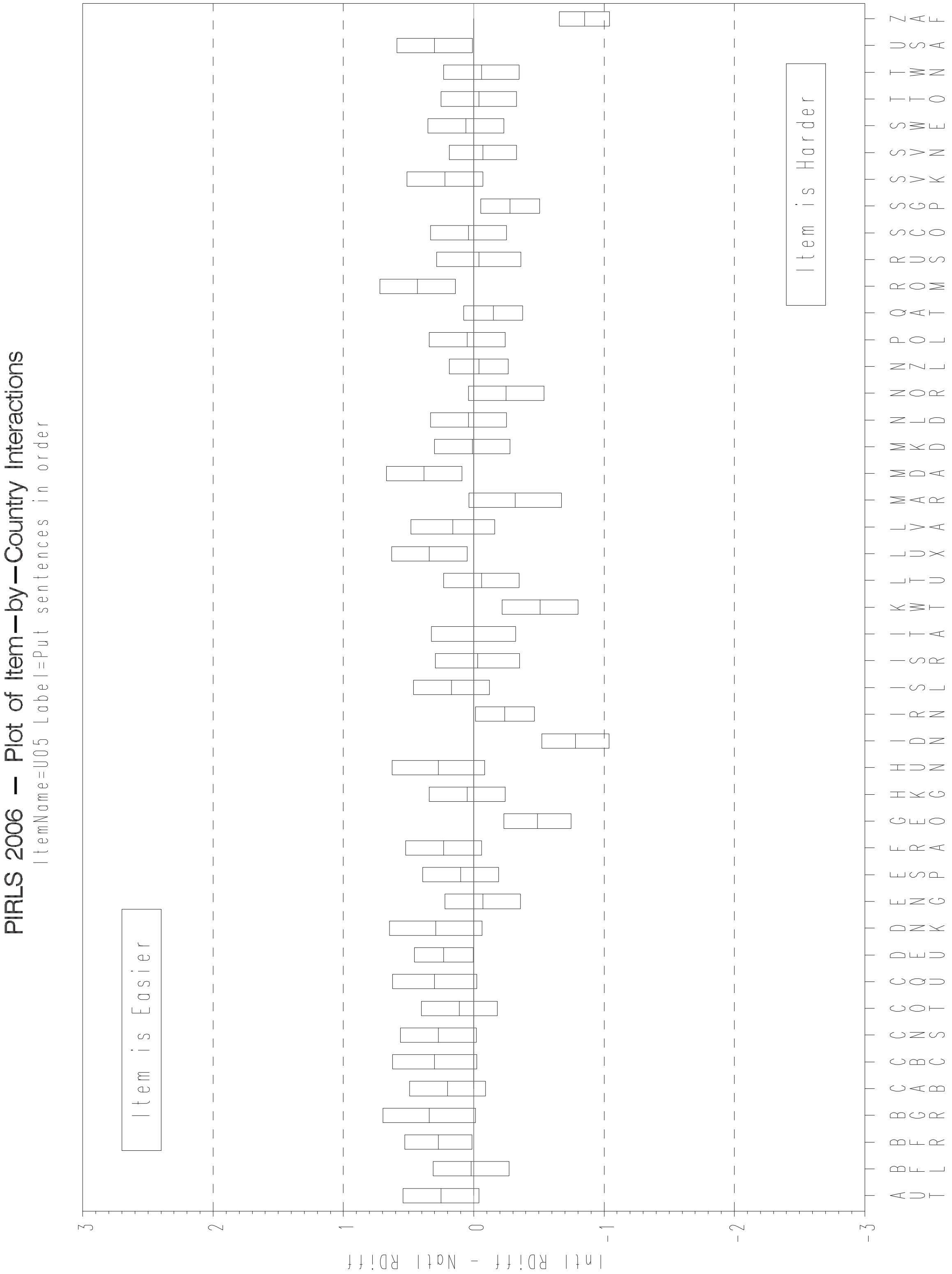
where $RDIFF_{ik}$ is the Rasch difficulty of item *k* within country *i*, $SE_{RDIFFik}$ is the standard error of the difficulty of item *k* in country *i*, and Z_b is the critical value from the *Z* distribution, corrected for multiple comparisons using the Bonferroni procedure. With the international average Rasch difficulty scaled to zero as a reference point, countries with a positive difference between the national and international Rasch difficulty found the item easier, and countries with a negative difference found the item more difficult.

10.4 Trend Item Analysis

Because an important part of the PIRLS 2006 assessment was the measuring 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 2001 should be relatively easy in 2006). 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 reading achievement may have improved or declined, items were noted if the percent correct changed by more than 15 percent for a particular country.

The TIMSS and PIRLS International Study Center used two different graphical displays to examine the differences between item difficulties in 2001 to 2006. The first of these, shown in Exhibit 10.4, displayed the difference in Rasch difficulty estimates (in logits) between the two assessment cycles. A positive difference indicates that the item was relatively easier in a country in 2006, and a negative difference indicates that an item was relatively more difficult. The second shows a country's performance on all trend items simultaneously.

Exhibit 10.3 Sample Plot of Item-by-Country Interactions for a PIRLS 2006 Item



Individually for each country, a scatterplot graphed the Rasch difficulty of each item in 2001 against the difficulty for that item in 2006. Where there are no differences between the difficulties in 2001 and 2006, 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.

10.5 Scoring Reliability for Constructed-response Items

Almost two thirds of the score points in the PIRLS 2006 assessment were from constructed-response items. In order to include these types of items in the assessment, it is essential that they are scored reliably within and across countries. In other words, a particular student response should receive the same score, regardless of the scorer. To ensure that this was the case, specific scoring guides were created for each constructed-response item that provided descriptions of each score level and sample student responses. Countries received extensive training in the application of each of these guides, using genuine student responses as examples and practice materials. Procedures for organizing and monitoring the scoring sessions were provided in the *PIRLS 2006 Survey Operations Procedures Unit 4* (TIMSS & PIRLS International Study Center, 2005). In addition to this training, countries were required to provide several types of scoring reliability data to document the consistency with which the scoring guides were applied. These are described in the following sections.

10.5.1 Within-country Scoring Reliability

This first type of scoring reliability data documented the extent to which items were scored consistently within each country. For each constructed-response item, 200 randomly selected student responses were independently marked by two scorers in the country. The percent agreement between these scorers was included in the item almanacs described above and examined as part of the review process. During this review, items were noted for further examination if agreement for a particular country fell below 70 percent. On average, the exact percent agreement across items was very high at 93 percent. All countries had an average percent of exact agreement above 81 percent. The average and range of these percentages is presented in Exhibit 10.6.

Exhibit 10.4 Sample Plot of Difference in Rasch Difficulties for a PIRLS 2006 Item

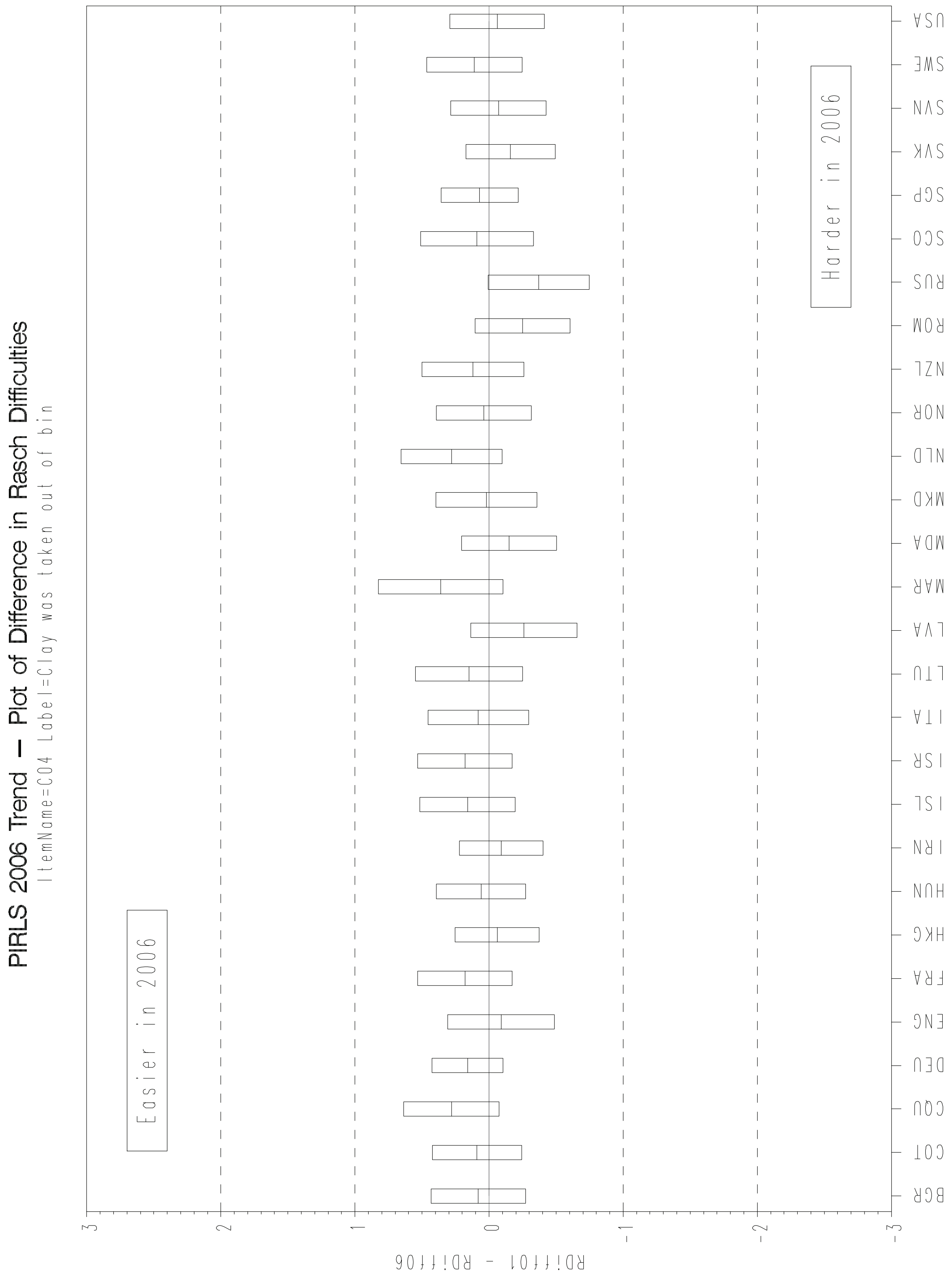


Exhibit 10.5 Sample Plot of Rasch Difficulties by Country for a PIRLS 2006 Item

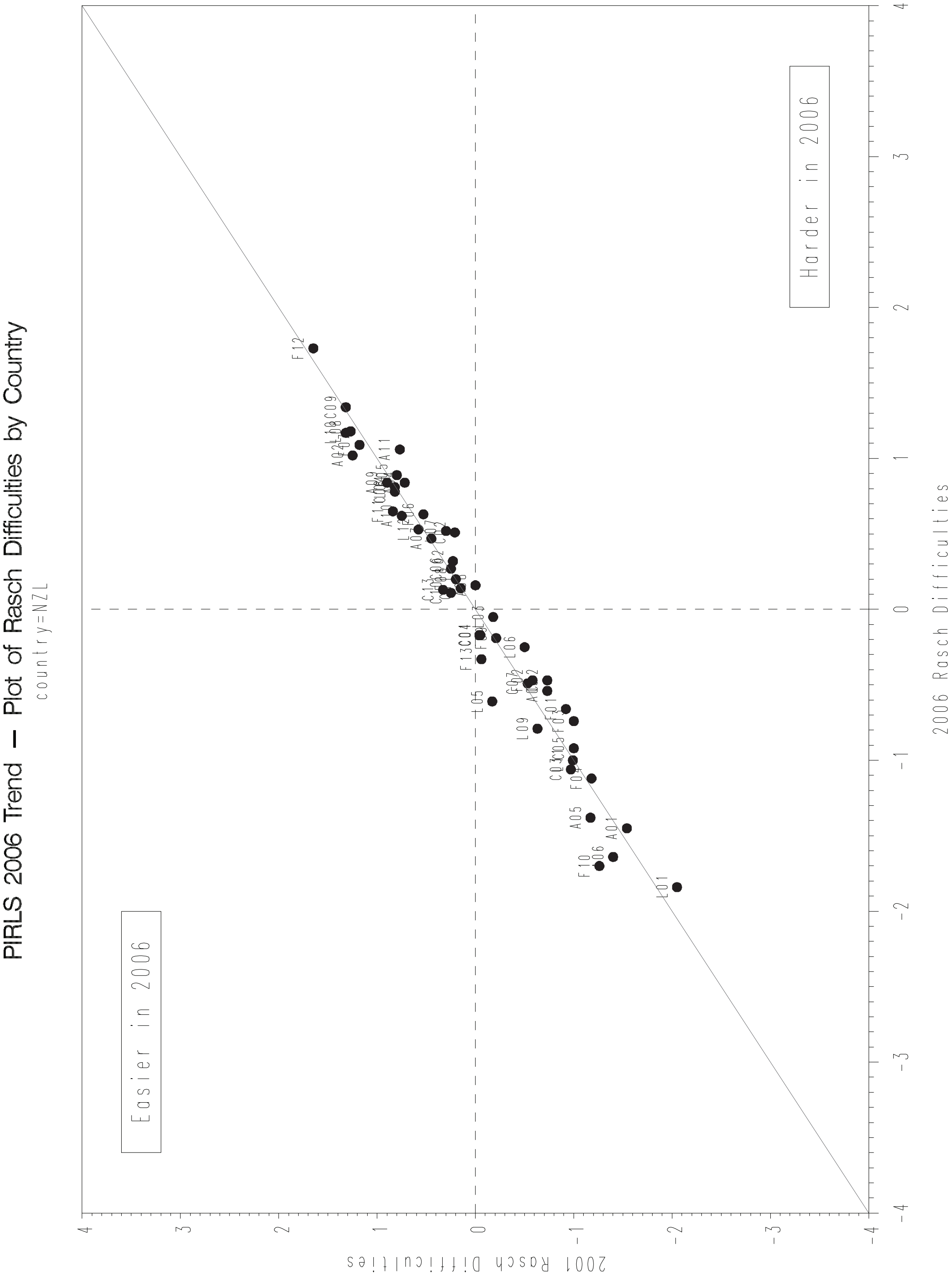


Exhibit 10.6 PIRLS 2006 Within-country Constructed-response Scoring Reliability Data

Countries	Average of Percent Exact Agreement Across Items	Range of Percent Exact Agreement	
		Minimum	Maximum
Austria	95	80	100
Belgium (Flemish)	90	73	99
Belgium (French)	97	90	100
Bulgaria	98	94	100
Canada, Alberta	91	67	100
Canada, British Columbia	92	70	100
Canada, Nova Scotia	93	84	100
Canada, Ontario	94	80	100
Canada, Quebec	95	87	100
Chinese Taipei	95	78	100
Denmark	97	90	100
England	98	93	100
France	89	69	100
Georgia	85	65	98
Germany	89	76	99
Hong Kong SAR	96	85	100
Hungary	98	89	100
Iceland	95	88	99
Indonesia	95	76	100
Iran, Islamic Rep. of	93	83	99
Israel	91	80	98
Italy	95	85	100
Kuwait	86	80	95
Latvia	90	78	100
Lithuania	97	91	100
Luxembourg	94	82	100
Macedonia, Rep. of	88	78	96
Moldova, Rep. of	99	97	100
Morocco	89	71	97
Netherlands	99	93	100
New Zealand	93	80	98
Norway	83	66	97
Poland	97	93	100
Qatar	97	93	99
Romania	99	96	100
Russian Federation	99	97	100
Scotland	97	89	100
Singapore	98	94	100
Slovak Republic	96	88	100
Slovenia	98	92	100
South Africa	82	63	92
Spain	81	61	96
Sweden	92	72	100
Trinidad and Tobago	93	71	100
United States	93	82	100
International Avg.	93	82	99

10.5.2 Cross-country Scoring Reliability

It also was important to document the consistency of scoring across countries. To accomplish this goal, the TIMSS & PIRLS International Study Center collected 200 student responses to each of the 23 constructed-response items from four assessment blocks, for a total of 4,600 student responses. Due to the wide range of languages used in PIRLS 2006 and the logistic issues this presents for cross-country scoring, responses were only collected from participants that administered the assessment in English (the Canadian province of Ontario, England, New Zealand, Scotland, Singapore, South Africa, and the United States). These responses were scanned by the IEA Data Processing and Research Center (DPC) and provided to each country in a software program that facilitated the scoring process.

Countries were asked to provide at least two scorers who were proficient in English to score this set of responses, following the main scoring activities. Each student response was scored by 62 scorers from across the countries, giving a total of 1,891 comparisons for each student response to each item, and 378,200 total comparisons across the set of 200 responses per item.[4] Exhibit 10.7 shows the percentage of exact agreement for each item. On average, there was 87 percent agreement, with variation across the individual items.

[4] The number of comparisons varies across items because not all scorers scored all items.

Exhibit 10.7 PIRLS 2006 Cross-country Constructed-response Scoring Reliability

Purpose	Item Label	Total Valid Comparisons	Exact Percent Agreement
Literary Experience	Flowers F06C	377504	91%
	Flowers F07C	377957	80%
	Flowers F08C	375960	92%
	Flowers F09C	378078	93%
	Flowers F10C	376869	97%
	Flowers F12C	375684	63%
	Unbelievable Night U05C	377224	99%
	Unbelievable Night U06C	377385	93%
	Unbelievable Night U08C	378078	76%
	Unbelievable Night U10C	377453	96%
	Unbelievable Night U12C	377302	87%
Acquire and Use Information	Antartica A01C	378200	95%
	Antartica A03C	378139	98%
	Antartica A04C	377542	89%
	Antartica A07C	378139	88%
	Antartica A08C	377722	80%
	Antartica A09C	377370	83%
	Antartica A11C	377363	81%
	Day Hiking N02C	377897	91%
	Day Hiking N03C	378139	94%
	Day Hiking N08C	376927	92%
	Day Hiking N11C	377773	77%
	Day Hiking N12C	330146	76%
	Average Percent Agreement		87%

10.5.3 Trend Scoring Reliability

Extensive efforts were made to ensure that constructed-response items were scored consistently across testing cycles. In preparation for this, the IEA DPC scanned 200 student responses to each of the 26 trend constructed-response items from the PIRLS 2001 reliability booklet samples and provided the responses, along with the original scores from 2001, in a scoring software program.[5] As part of the scoring training activities, at least two scorers in each trend country were asked to score the set of student responses for each item from the four trend blocks, for a total of 5,200 student responses. After scoring half of these responses, scorers used the software to compare their scores to one

[5] A number of participants were unable to complete the trend scoring reliability task because of software difficulties or because it was not possible to scan their 2001 students booklets.

another, as well as to the original score given in 2001. If agreement for any item fell below 85 percent, retraining was required and the responses for that item were rescored. Exhibit 10.8 shows the results of the trend scoring reliability task. Overall, the percent agreement across the trend constructed-response items was high—90 percent on average across countries.

Exhibit 10.8 PIRLS 2006 Trend Scoring Reliability (2001–2006) for the Constructed-response Items

Countries	Average Percent Exact Agreement Across Items
Canada, Ontario	–
Canada, Quebec	–
England	89
France	90
Germany	88
Hong Kong SAR	93
Hungary	91
Iceland	–
Iran, Islamic Rep. of	92
Israel	96
Italy	91
Latvia	84
Lithuania	92
Macedonia, Rep. of	81
Moldova, Rep. of	–
Morocco	–
Netherlands	93
New Zealand	90
Norway	90
Romania	–
Russian Federation	–
Scotland	88
Singapore	88
Slovak Republic	92
Slovenia	–
Sweden	89
United States	93
International Avg.	90

A dash (–) indicates data are not available.

10.6 Item Review Procedures

Using the range of techniques described in the previous sections, the TIMSS & PIRLS International Study Center reviewed the performance of each item within every participating country to ensure that the results were comparable internationally. In particular, items with the following problems were considered for possible deletion from the international database:

- An error was detected in the translation of the item that was not fixed before test administration;
- Data checking revealed a multiple-choice item with more or fewer options than the international version;
- The item analysis showed the item to have a negative biserial, or, for items with more than one score point, point biserials that did not increase with each score level;
- The item-by-country interaction results showed a very large interaction for a particular country;
- For constructed-response items, the within-country scoring reliability data showed an agreement of less than 70 percent; and
- For trend items, an item performed substantially differently in 2006 compared to 2001.

In cases where a potential problem was detected, test booklets (including PIRLS 2001 booklets, if necessary), and documentation from the translation verification and national adaptation process were reviewed. Additionally, the NRC was consulted for clarification or confirmation of translation, printing, or scoring problems. Of the 126 items used in the assessment, only one item was removed from the international database for all countries. In only three countries were one or two additional items problematic for international comparisons. The main cause of errors in items was translation or printing errors. A list of deleted and recoded items is provided in Appendix C.

References

TIMSS & PIRLS International Study Center. (2005). *Survey operations procedures unit 4: Scoring the PIRLS 2006 assessment*. Chestnut Hill, MA: Boston College.