CHAPTER 9

Reviewing the PIRLS 2021 Achievement Item Statistics

Jessie Bristol
Ina V.S. Mullis
Bethany Fishbein
Pierre Foy

Overview

Conducting a review of achievement item statistics is an essential step in assuring the quality of the achievement data before applying item response theory (IRT) methods to derive student achievement estimates for analysis and reporting. The TIMSS & PIRLS International Study Center conducts an item-by-item, country-by-country review of key diagnostic statistics to detect items with unusual psychometric properties or reveal anomalous patterns in the data for a particular country. An uncharacteristically difficult item or one with unusually low discriminating power in a particular country can indicate a potential problem with translation or other technical errors. Similarly, a human-scored constructed-response item with low scoring reliability can point to a problem in applying the scoring guide. In rare instances where an item is found to be faulty for a particular country, the research staff at the TIMSS & PIRLS International Study Center examine the country's translation verification records and digital instrument archives for flaws or inaccuracies. In some cases, the data may be removed from the international database.

In addition to evaluating the performance of each individual PIRLS 2021 item, the TIMSS & PIRLS International Study Center conducted analyses to detect and evaluate any possible differences in the measurement properties of paper trend items between the PIRLS 2021 and PIRLS 2016 assessments. Aggregate-level analyses were conducted for additional quality assurance of the PIRLS 2021 data. Item position effects were evaluated to ensure student performance was not affected substantially by the position of the texts in the assessment booklets. Analysis by booklet difficulty—more difficult or less difficult—allowed for detecting anomalous data patterns in particular countries and for evaluating the effectiveness of the group adaptive design (Martin et al., 2019). Extensive analyses of each country's item statistics allowed for detecting any irregular patterns relative to previous cycles or to the international distribution across countries.

TIMSS & PIRLS
International Study Center
Lynch School of Education
BOSTON COLLEGE

https://doi.org/10.6017/lse.tpisc.tr2103.kb5892

The PIRLS 2021 Achievement Item Review

The TIMSS & PIRLS International Study Center computed item statistics for all achievement items in the 2021 assessment, including digitalPIRLS, paperPIRLS, and the paper "bridge" booklets. Altogether, the PIRLS 2021 item review included statistics for 769 items (including item parts) across 23 unique texts. This included 297 paperPIRLS items (18 texts), 349 digitalPIRLS items (18 texts), and 123 ePIRLS items (5 texts). The bridge booklets consisted only of the 8 paper-based trend texts with 124 trend items that were also in paperPIRLS, and their item statistics were reviewed alongside paperPIRLS item statistics.

As data collection coincided with the height of the COVID-19 pandemic, many schools necessarily delayed administering the PIRLS 2021 assessment. For this reason, PIRLS 2021 data collection occurred over a span of two years instead of the typical range of a few months. The TIMSS & PIRLS International Study Center reviewed PIRLS 2021 item statistics over the course of two years and met three times to conduct formal adjudications of the achievement data in preparation for IRT scaling. Executive Directors, along with the PIRLS Coordinator and Analysis Unit staff, met in March 2022 for three consecutive working days, in May 2022 for one working day, and in December 2022 for one working day. During the meetings, the Executive Directors, together with staff, made decisions about necessary modifications to the data and about areas requiring further analyses. The review was conducted item-by-item simultaneously for digitalPIRLS and paperPIRLS countries. During the review, members viewed both versions of an item and its scoring guide alongside item statistics and graphical displays of item statistics. Reviewers also referenced country reports about translation errors, printing issues, or other technical problems. Graphical displays of item statistics helped reviewers detect inconsistent or systematic patterns in a particular country's data that warranted further investigation.

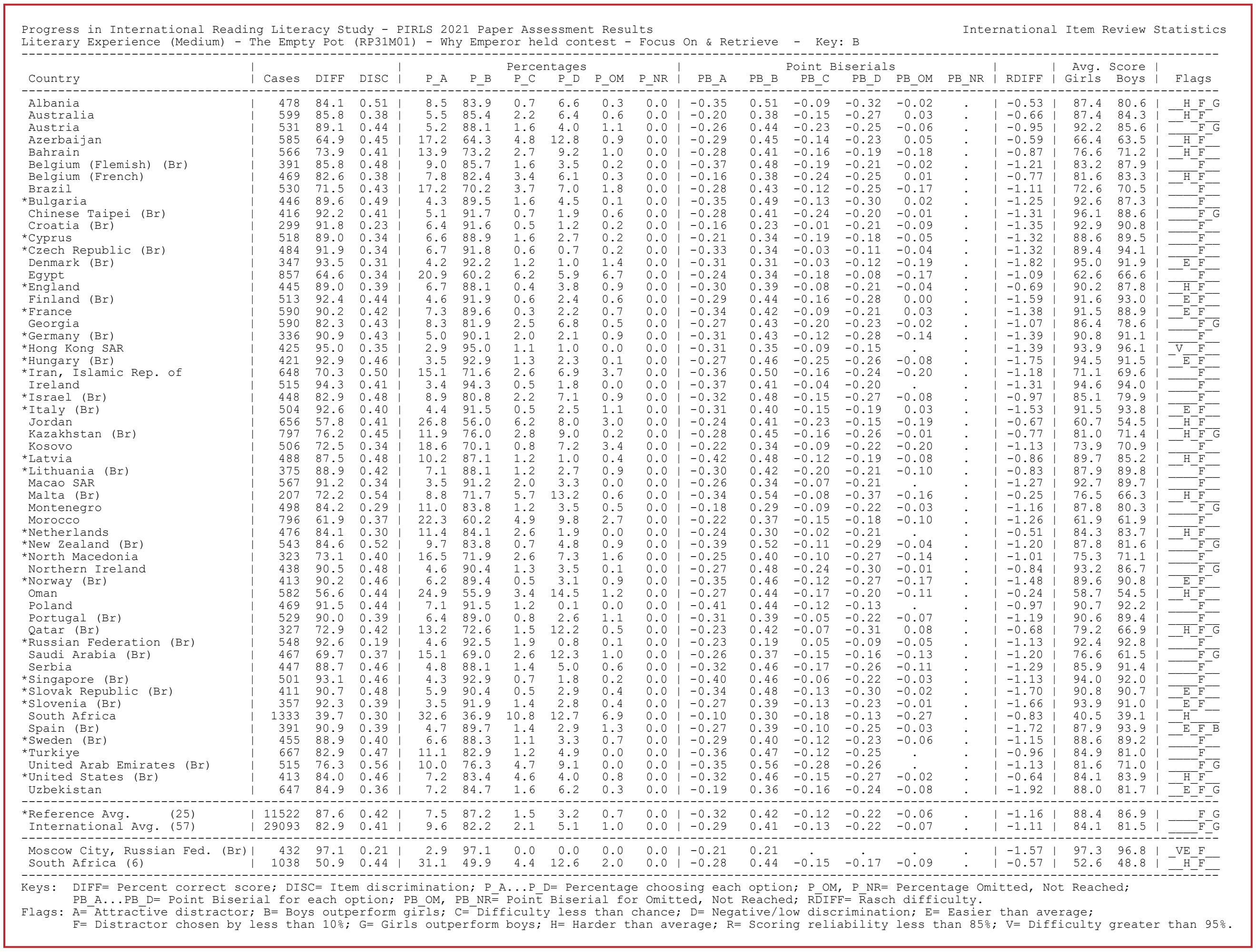
Following each item review meeting, the TIMSS & PIRLS International Study Center contacted National Research Coordinators from participating countries and benchmarking entities to inquire about concerns or anomalies that were detected in the data. Analysis Unit staff communicated decisions about item deletions or recodes to IEA Hamburg so that they could edit the international data files.

Item Review Statistics

The TIMSS & PIRLS International Study Center computed item statistics for each of the PIRLS 2021 participating countries. These data were combined for internal and external review in item almanacs for paper data and digital data, respectively. Each item almanac page included statistics for all countries that administered that particular item. The paperPIRLS item almanacs included data from paperPIRLS countries plus the bridge samples from digitalPIRLS trend countries.

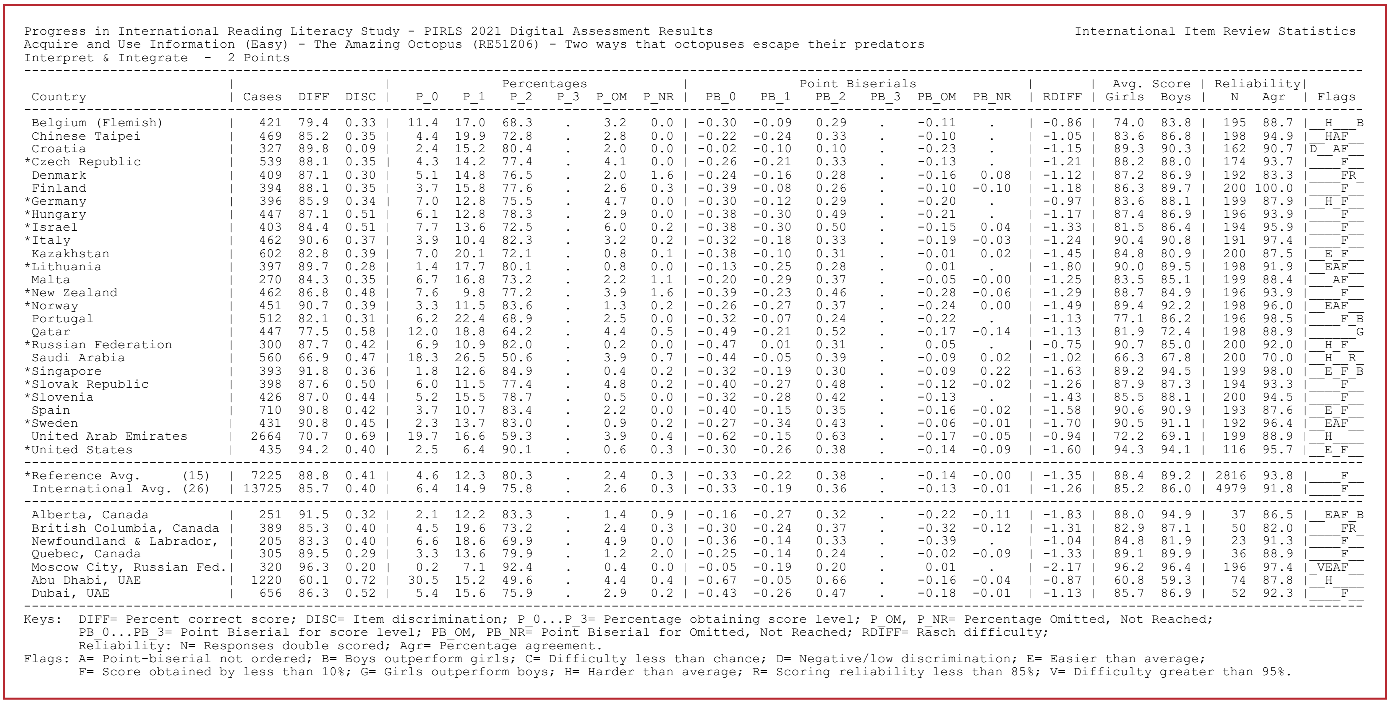
Exhibits 9.1 and 9.2 show examples of the statistics calculated for a selected-response item and a constructed-response item, respectively.

Exhibit 9.1: Example International Item Statistics for a PIRLS 2021 Selected-Response Item



```
Progress in International Reading Literacy Study - PIRLS 2021 Paper Assessment Results                    International Item Review Statistics
Literary Experience (Medium) - The Empty Pot (RP31M01) - Why Emperor held contest - Focus On & Retrieve  -  Key: B
------------------------------------------------------------------------------------------------------------------------------------------------------------
                               |                   |            Percentages              |             Point Biserials                |       | Avg. Score |
 Country                       | Cases  DIFF  DISC |   P_A   P_B   P_C   P_D  P_OM  P_NR |   PB_A   PB_B   PB_C   PB_D  PB_OM  PB_NR | RDIFF | Girls  Boys |  Flags
------------------------------------------------------------------------------------------------------------------------------------------------------------
 Albania                       |   478  84.1  0.51 |   8.5  83.9   0.7   6.6   0.3   0.0 |  -0.35   0.51  -0.09  -0.32  -0.02      . | -0.53 |  87.4  80.6 | __H_F_G
 Australia                     |   599  85.8  0.38 |   5.5  85.4   2.2   6.4   0.6   0.0 |  -0.20   0.38  -0.15  -0.27   0.03      . | -0.66 |  87.4  84.3 | __H_F__
 Austria                       |   531  89.1  0.44 |   5.2  88.1   1.6   4.0   1.1   0.0 |  -0.26   0.44  -0.23  -0.25  -0.06      . | -0.95 |  92.2  85.6 | ____F_G
 Azerbaijan                    |   585  64.9  0.45 |  17.2  64.3   4.8  12.8   0.9   0.0 |  -0.29   0.45  -0.14  -0.23   0.05      . | -0.59 |  66.4  63.5 | __H_F__
 Bahrain                       |   566  73.9  0.41 |  13.9  73.2   2.7   9.2   1.0   0.0 |  -0.28   0.41  -0.16  -0.19  -0.18      . | -0.87 |  76.6  71.2 | __H_F__
 Belgium (Flemish) (Br)        |   391  85.8  0.48 |   9.0  85.7   1.6   3.5   0.2   0.0 |  -0.37   0.48  -0.19  -0.21  -0.02      . | -1.21 |  83.2  87.9 | ____F__
 Belgium (French)              |   469  82.6  0.38 |   7.8  82.4   3.4   6.1   0.3   0.0 |  -0.16   0.38  -0.24  -0.25   0.01      . | -0.77 |  81.6  83.3 | __H_F__
 Brazil                        |   530  71.5  0.43 |  17.2  70.2   3.7   7.0   1.8   0.0 |  -0.28   0.43  -0.12  -0.25  -0.17      . | -1.11 |  72.6  70.5 | ____F__
*Bulgaria                      |   446  89.6  0.49 |   4.3  89.5   1.6   4.5   0.1   0.0 |  -0.35   0.49  -0.13  -0.30   0.02      . | -1.25 |  92.6  87.3 | ____F__
 Chinese Taipei (Br)           |   416  92.2  0.41 |   5.1  91.7   0.7   1.9   0.6   0.0 |  -0.28   0.41  -0.24  -0.20  -0.01      . | -1.31 |  96.1  88.6 | ____F_G
 Croatia (Br)                  |   299  91.8  0.23 |   6.4  91.6   0.5   1.2   0.2   0.0 |  -0.16   0.23  -0.01  -0.21  -0.09      . | -1.35 |  92.9  90.8 | ____F__
*Cyprus                        |   518  89.0  0.34 |   6.6  88.9   1.6   2.7   0.2   0.0 |  -0.21   0.34  -0.19  -0.18  -0.05      . | -1.32 |  88.6  89.5 | ____F__
*Czech Republic (Br)           |   484  91.9  0.34 |   6.7  91.8   0.6   0.7   0.2   0.0 |  -0.33   0.34  -0.03  -0.11  -0.04      . | -1.32 |  89.4  94.1 | ____F__
 Denmark (Br)                  |   347  93.5  0.31 |   4.2  92.2   1.2   1.0   1.4   0.0 |  -0.31   0.31  -0.03  -0.12  -0.19      . | -1.82 |  95.0  91.9 | __E_F__
 Egypt                         |   857  64.6  0.34 |  20.9  60.2   6.2   5.9   6.7   0.0 |  -0.24   0.34  -0.18  -0.08  -0.17      . | -1.09 |  62.6  66.6 | ____F__
*England                       |   445  89.0  0.39 |   6.7  88.1   0.4   3.8   0.9   0.0 |  -0.30   0.39  -0.08  -0.21  -0.04      . | -0.69 |  90.2  87.8 | __H_F__
 Finland (Br)                  |   513  92.4  0.44 |   4.6  91.9   0.6   2.4   0.6   0.0 |  -0.29   0.44  -0.16  -0.28   0.00      . | -1.59 |  91.6  93.0 | __E_F__
*France                        |   590  90.2  0.42 |   7.3  89.6   0.3   2.2   0.7   0.0 |  -0.34   0.42  -0.09  -0.21   0.03      . | -1.38 |  91.5  88.9 | __E_F__
 Georgia                       |   590  82.3  0.43 |   8.3  81.9   2.5   6.8   0.5   0.0 |  -0.27   0.43  -0.20  -0.23  -0.02      . | -1.07 |  86.4  78.6 | ____F_G
*Germany (Br)                  |   336  90.9  0.43 |   5.0  90.1   2.0   2.1   0.9   0.0 |  -0.31   0.43  -0.12  -0.28  -0.14      . | -1.39 |  90.8  91.1 | ____F__
*Hong Kong SAR                 |   425  95.0  0.35 |   2.9  95.0   1.1   1.0   0.0   0.0 |  -0.31   0.35  -0.09  -0.15      .      . | -1.39 |  93.9  96.1 | _V__F__
*Hungary (Br)                  |   421  92.9  0.46 |   3.5  92.9   1.3   2.3   0.1   0.0 |  -0.27   0.46  -0.25  -0.26  -0.08      . | -1.75 |  94.5  91.5 | __E_F__
*Iran, Islamic Rep. of         |   648  70.3  0.50 |  15.1  71.6   2.6   6.9   3.7   0.0 |  -0.36   0.50  -0.16  -0.24  -0.20      . | -1.18 |  71.1  69.6 | ____F__
 Ireland                       |   515  94.3  0.41 |   3.4  94.3   0.5   1.8   0.0   0.0 |  -0.37   0.41  -0.04  -0.20      .      . | -1.31 |  94.6  94.0 | ____F__
*Israel (Br)                   |   448  82.9  0.48 |   8.9  80.8   2.2   7.1   0.9   0.0 |  -0.32   0.48  -0.15  -0.27  -0.08      . | -0.97 |  85.1  79.9 | ____F__
*Italy (Br)                    |   504  92.6  0.40 |   4.4  91.5   0.5   2.5   1.1   0.0 |  -0.31   0.40  -0.15  -0.19   0.03      . | -1.53 |  91.5  93.8 | __E_F__
 Jordan                        |   656  57.8  0.41 |  26.8  56.0   6.2   8.0   3.0   0.0 |  -0.24   0.41  -0.23  -0.15  -0.19      . | -0.67 |  60.7  54.5 | __H_F__
 Kazakhstan (Br)               |   797  76.2  0.45 |  11.9  76.0   2.8   9.0   0.2   0.0 |  -0.28   0.45  -0.16  -0.26  -0.01      . | -0.77 |  81.0  71.4 | __H_F_G
 Kosovo                        |   506  72.5  0.34 |  18.6  70.1   0.8   7.2   3.4   0.0 |  -0.22   0.34  -0.09  -0.22  -0.20      . | -1.13 |  73.9  70.9 | ____F__
*Latvia                        |   488  87.5  0.48 |  10.2  87.1   1.2   1.0   0.4   0.0 |  -0.42   0.48  -0.12  -0.19  -0.08      . | -0.86 |  89.7  85.2 | __H_F__
*Lithuania (Br)                |   375  88.9  0.42 |   7.1  88.1   1.2   2.7   0.9   0.0 |  -0.30   0.42  -0.20  -0.21  -0.10      . | -0.83 |  87.9  89.8 | ____F__
 Macao SAR                     |   567  91.2  0.34 |   3.5  91.2   2.0   3.3   0.0   0.0 |  -0.26   0.34  -0.07  -0.21      .      . | -1.27 |  92.7  89.7 | ____F__
 Malta (Br)                    |   207  72.2  0.54 |   8.8  71.7   5.7  13.2   0.6   0.0 |  -0.34   0.54  -0.08  -0.37  -0.16      . | -0.25 |  76.5  66.3 | __H_F__
 Montenegro                    |   498  84.2  0.29 |  11.0  83.8   1.2   3.5   0.5   0.0 |  -0.18   0.29  -0.09  -0.22  -0.03      . | -1.16 |  87.8  80.3 | ____F_G
 Morocco                       |   796  61.9  0.37 |  22.3  60.2   4.9   9.8   2.7   0.0 |  -0.22   0.37  -0.15  -0.18  -0.10      . | -1.26 |  61.9  61.9 | ____F__
*Netherlands                   |   476  84.1  0.30 |  11.4  84.1   2.6   1.9   0.0   0.0 |  -0.24   0.30  -0.02  -0.21      .      . | -0.51 |  84.3  83.7 | __H_F__
*New Zealand (Br)              |   543  84.6  0.52 |   9.7  83.8   0.7   4.8   0.9   0.0 |  -0.39   0.52  -0.11  -0.29  -0.04      . | -1.20 |  87.8  81.6 | ____F_G
*North Macedonia               |   323  73.1  0.40 |  16.5  71.9   2.6   7.3   1.6   0.0 |  -0.25   0.40  -0.10  -0.27  -0.14      . | -1.01 |  75.3  71.1 | ____F__
 Northern Ireland              |   438  90.5  0.48 |   4.6  90.4   1.3   3.5   0.1   0.0 |  -0.27   0.48  -0.24  -0.30  -0.01      . | -0.84 |  93.2  86.7 | ____F_G
*Norway (Br)                   |   413  90.2  0.46 |   6.2  89.4   0.5   3.1   0.9   0.0 |  -0.35   0.46  -0.12  -0.27  -0.17      . | -1.48 |  89.6  90.8 | __E_F__
 Oman                          |   582  56.6  0.44 |  24.9  55.9   3.4  14.5   1.2   0.0 |  -0.27   0.44  -0.17  -0.20  -0.11      . | -0.24 |  58.7  54.5 | __H_F__
 Poland                        |   469  91.5  0.44 |   7.1  91.5   1.2   0.1   0.0   0.0 |  -0.41   0.44  -0.12  -0.13      .      . | -0.97 |  90.7  92.2 | ____F__
 Portugal (Br)                 |   529  90.0  0.39 |   6.4  89.0   0.8   2.6   1.1   0.0 |  -0.31   0.39  -0.05  -0.22  -0.07      . | -1.19 |  90.6  89.4 | ____F__
 Qatar (Br)                    |   327  72.9  0.42 |  13.2  72.6   1.5  12.2   0.5   0.0 |  -0.23   0.42  -0.07  -0.31   0.08      . | -0.68 |  79.2  66.9 | __H_F_G
*Russian Federation (Br)       |   548  92.6  0.19 |   4.6  92.5   1.9   0.8   0.1   0.0 |  -0.23   0.19   0.05  -0.09  -0.05      . | -1.13 |  92.4  92.8 | ____F__
 Saudi Arabia (Br)             |   467  69.7  0.37 |  15.1  69.0   2.6  12.3   1.0   0.0 |  -0.26   0.37  -0.15  -0.16  -0.13      . | -1.20 |  76.6  61.5 | ____F_G
 Serbia                        |   447  88.7  0.46 |   4.8  88.1   1.4   5.0   0.6   0.0 |  -0.32   0.46  -0.17  -0.26  -0.11      . | -1.29 |  85.9  91.4 | ____F__
*Singapore (Br)                |   501  93.1  0.46 |   4.3  92.9   0.7   1.8   0.2   0.0 |  -0.40   0.46  -0.06  -0.22  -0.03      . | -1.13 |  94.0  92.0 | ____F__
*Slovak Republic (Br)          |   411  90.7  0.48 |   5.9  90.4   0.5   2.9   0.4   0.0 |  -0.34   0.48  -0.13  -0.30  -0.02      . | -1.70 |  90.8  90.7 | __E_F__
*Slovenia (Br)                 |   357  92.3  0.39 |   3.5  91.9   1.4   2.8   0.4   0.0 |  -0.27   0.39  -0.13  -0.23  -0.01      . | -1.66 |  93.9  91.0 | __E_F__
 South Africa                  |  1333  39.7  0.30 |  32.6  36.9  10.8  12.7   6.9   0.0 |  -0.10   0.30  -0.18  -0.13  -0.27      . | -0.83 |  40.5  39.1 | __H____
 Spain (Br)                    |   391  90.9  0.39 |   4.7  89.7   1.4   2.9   1.3   0.0 |  -0.27   0.39  -0.10  -0.25  -0.03      . | -1.72 |  87.9  93.9 | __E_F_B
*Sweden (Br)                   |   455  88.9  0.40 |   6.6  88.3   1.1   3.3   0.7   0.0 |  -0.29   0.40  -0.12  -0.23  -0.06      . | -1.15 |  88.6  89.2 | ____F__
*Turkiye                       |   667  82.9  0.47 |  11.1  82.9   1.2   4.9   0.0   0.0 |  -0.36   0.47  -0.12  -0.25      .      . | -0.96 |  84.9  81.0 | ____F__
 United Arab Emirates (Br)     |   515  76.3  0.56 |  10.0  76.3   4.7   9.1   0.0   0.0 |  -0.35   0.56  -0.28  -0.26      .      . | -1.13 |  81.6  71.0 | ____F_G
*United States (Br)            |   413  84.0  0.46 |   7.2  83.4   4.6   4.0   0.8   0.0 |  -0.32   0.46  -0.15  -0.27  -0.02      . | -0.64 |  84.1  83.9 | __H_F__
 Uzbekistan                    |   647  84.9  0.36 |   7.2  84.7   1.6   6.2   0.3   0.0 |  -0.19   0.36  -0.16  -0.24  -0.08      . | -1.92 |  88.0  81.7 | __E_F_G
------------------------------------------------------------------------------------------------------------------------------------------------------------
*Reference Avg.     (25)       | 11522  87.6  0.42 |   7.5  87.2   1.5   3.2   0.7   0.0 |  -0.32   0.42  -0.12  -0.22  -0.06      . | -1.16 |  88.4  86.9 | ____F_G
 International Avg. (57)       | 29093  82.9  0.41 |   9.6  82.2   2.1   5.1   1.0   0.0 |  -0.29   0.41  -0.13  -0.22  -0.07      . | -1.11 |  84.1  81.5 | ____F_G
------------------------------------------------------------------------------------------------------------------------------------------------------------
 Moscow City, Russian Fed. (Br)|   432  97.1  0.21 |   2.9  97.1   0.0   0.0   0.0   0.0 |  -0.21   0.21      .      .      .      . | -1.57 |  97.3  96.8 | _VE_F__
 South Africa (6)              |  1038  50.9  0.44 |  31.1  49.9   4.4  12.6   2.0   0.0 |  -0.28   0.44  -0.15  -0.17  -0.09      . | -0.57 |  52.6  48.8 | __H_F__
------------------------------------------------------------------------------------------------------------------------------------------------------------
Keys:  DIFF= Percent correct score; DISC= Item discrimination; P_A...P_D= Percentage choosing each option; P_OM, P_NR= Percentage Omitted, Not Reached;
       PB_A...PB_D= Point Biserial for each option; PB_OM, PB_NR= Point Biserial for Omitted, Not Reached; RDIFF= Rasch difficulty.
Flags: A= 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 less than 85%; V= Difficulty greater than 95%.
```

Exhibit 9.2: Example International Item Statistics for a PIRLS 2021 Constructed-Response Item



```
Progress in International Reading Literacy Study - PIRLS 2021 Digital Assessment Results                            International Item Review Statistics
Acquire and Use Information (Easy) - The Amazing Octopus (RE51Z06) - Two ways that octopuses escape their predators
Interpret & Integrate  -  2 Points
---------------------------------------------------------------------------------------------------------------------------------------------------------------
                          |                   |              Percentages              |              Point Biserials                |       |  Avg. Score | Reliability|
 Country                  | Cases  DIFF  DISC |   P_0   P_1   P_2   P_3  P_OM  P_NR |  PB_0   PB_1   PB_2   PB_3  PB_OM  PB_NR  | RDIFF | Girls  Boys |    N   Agr | Flags
---------------------------------------------------------------------------------------------------------------------------------------------------------------
 Belgium (Flemish)        |   421  79.4  0.33 |  11.4  17.0  68.3     .   3.2   0.0 | -0.30  -0.09   0.29      .  -0.11      .  | -0.86 |  74.0  83.8 |  195  88.7 |__H___B
 Chinese Taipei           |   469  85.2  0.35 |   4.4  19.9  72.8     .   2.8   0.0 | -0.22  -0.24   0.33      .  -0.10      .  | -1.05 |  83.6  86.8 |  198  94.9 |__HAF__
 Croatia                  |   327  89.8  0.09 |   2.4  15.2  80.4     .   2.0   0.0 | -0.02  -0.10   0.10      .  -0.23      .  | -1.15 |  89.3  90.3 |  162  90.7 |D__AF__
*Czech Republic           |   539  88.1  0.35 |   4.3  14.2  77.4     .   4.1   0.0 | -0.26  -0.21   0.33      .  -0.13      .  | -1.21 |  88.2  88.0 |  174  93.7 |____F__
 Denmark                  |   409  87.1  0.30 |   5.1  14.8  76.5     .   2.0   1.6 | -0.24  -0.16   0.28      .  -0.16   0.08  | -1.12 |  87.2  86.9 |  192  83.3 |____FR_
 Finland                  |   394  88.1  0.35 |   3.7  15.8  77.6     .   2.6   0.3 | -0.39  -0.08   0.26      .  -0.10  -0.10  | -1.18 |  86.3  89.7 |  200 100.0 |____F__
*Germany                  |   396  85.9  0.34 |   7.0  12.8  75.5     .   4.7   0.0 | -0.30  -0.12   0.29      .  -0.20      .  | -0.97 |  83.6  88.1 |  199  87.9 |__H_F__
*Hungary                  |   447  87.1  0.51 |   6.1  12.8  78.3     .   2.9   0.0 | -0.38  -0.30   0.49      .  -0.21      .  | -1.17 |  87.4  86.9 |  196  93.9 |____F__
*Israel                   |   403  84.4  0.51 |   7.7  13.6  72.5     .   6.0   0.2 | -0.38  -0.30   0.50      .  -0.15   0.04  | -1.33 |  81.5  86.4 |  194  95.9 |____F__
*Italy                    |   462  90.6  0.37 |   3.9  10.4  82.3     .   3.2   0.2 | -0.32  -0.18   0.33      .  -0.19  -0.03  | -1.24 |  90.4  90.8 |  191  97.4 |____F__
 Kazakhstan               |   602  82.8  0.39 |   7.0  20.1  72.1     .   0.8   0.1 | -0.38  -0.10   0.31      .  -0.01   0.02  | -1.45 |  84.8  80.9 |  200  87.5 |__E_F__
*Lithuania                |   397  89.7  0.28 |   1.4  17.7  80.1     .   0.8   0.0 | -0.13  -0.25   0.28      .   0.01      .  | -1.80 |  90.0  89.5 |  198  91.9 |__EAF__
 Malta                    |   270  84.3  0.35 |   6.7  16.8  73.2     .   2.2   1.1 | -0.20  -0.29   0.37      .  -0.05  -0.00  | -1.25 |  83.5  85.1 |  199  88.4 |___AF__
*New Zealand              |   462  86.8  0.48 |   7.6   9.8  77.2     .   3.9   1.6 | -0.39  -0.23   0.46      .  -0.28  -0.06  | -1.29 |  88.7  84.9 |  196  93.9 |____F__
*Norway                   |   451  90.7  0.39 |   3.3  11.5  83.6     .   1.3   0.2 | -0.26  -0.27   0.37      .  -0.24   0.00  | -1.49 |  89.4  92.2 |  198  96.0 |__EAF__
 Portugal                 |   512  82.1  0.31 |   6.2  22.4  68.9     .   2.5   0.0 | -0.32  -0.07   0.24      .  -0.22      .  | -1.13 |  77.1  86.2 |  196  98.5 |____F_B
 Qatar                    |   447  77.5  0.58 |  12.0  18.8  64.2     .   4.4   0.5 | -0.49  -0.21   0.52      .  -0.17  -0.14  | -1.13 |  81.9  72.4 |  198  88.9 |______G
*Russian Federation       |   300  87.7  0.42 |   6.9  10.9  82.0     .   0.2   0.0 | -0.47   0.01   0.31      .   0.05      .  | -0.75 |  90.7  85.0 |  200  92.0 |__H_F__
 Saudi Arabia             |   560  66.9  0.47 |  18.3  26.5  50.6     .   3.9   0.7 | -0.44  -0.05   0.39      .  -0.09   0.02  | -1.02 |  66.3  67.8 |  200  70.0 |__H__R_
*Singapore                |   393  91.8  0.36 |   1.8  12.6  84.9     .   0.4   0.2 | -0.32  -0.19   0.30      .  -0.09   0.22  | -1.63 |  89.2  94.5 |  199  98.0 |__E_F_B
*Slovak Republic          |   398  87.6  0.50 |   6.0  11.5  77.4     .   4.8   0.2 | -0.40  -0.27   0.48      .  -0.12  -0.02  | -1.26 |  87.9  87.3 |  194  93.3 |____F__
*Slovenia                 |   426  87.0  0.44 |   5.2  15.5  78.7     .   0.5   0.0 | -0.32  -0.28   0.42      .  -0.13      .  | -1.43 |  85.5  88.1 |  200  94.5 |____F__
 Spain                    |   710  90.8  0.42 |   3.7  10.7  83.4     .   2.2   0.0 | -0.40  -0.15   0.35      .  -0.16  -0.02  | -1.58 |  90.6  90.9 |  193  87.6 |__E_F__
*Sweden                   |   431  90.8  0.45 |   2.3  13.7  83.0     .   0.9   0.2 | -0.27  -0.34   0.43      .  -0.06  -0.01  | -1.70 |  90.5  91.1 |  192  96.4 |__EAF__
 United Arab Emirates     |  2664  70.7  0.69 |  19.7  16.6  59.3     .   3.9   0.4 | -0.62  -0.15   0.63      .  -0.17  -0.05  | -0.94 |  72.2  69.1 |  199  88.9 |__H____
*United States            |   435  94.2  0.40 |   2.5   6.4  90.1     .   0.6   0.3 | -0.30  -0.26   0.38      .  -0.14  -0.09  | -1.60 |  94.3  94.1 |  116  95.7 |__E_F__
---------------------------------------------------------------------------------------------------------------------------------------------------------------
*Reference Avg.      (15) |  7225  88.8  0.41 |   4.6  12.3  80.3     .   2.4   0.3 | -0.33  -0.22   0.38      .  -0.14  -0.00  | -1.35 |  88.4  89.2 | 2816  93.8 |____F__
 International Avg. (26)  | 13725  85.7  0.40 |   6.4  14.9  75.8     .   2.6   0.3 | -0.33  -0.19   0.36      .  -0.13  -0.01  | -1.26 |  85.2  86.0 | 4979  91.8 |____F__
---------------------------------------------------------------------------------------------------------------------------------------------------------------
 Alberta, Canada          |   251  91.5  0.32 |   2.1  12.2  83.3     .   1.4   0.9 | -0.16  -0.27   0.32      .  -0.22  -0.11  | -1.83 |  88.0  94.9 |   37  86.5 |__EAF_B
 British Columbia, Canada |   389  85.3  0.40 |   4.5  19.6  73.2     .   2.4   0.3 | -0.30  -0.24   0.37      .  -0.32  -0.12  | -1.31 |  82.9  87.1 |   50  82.0 |____FR_
 Newfoundland & Labrador, |   205  83.3  0.40 |   6.6  18.6  69.9     .   4.9   0.0 | -0.36  -0.14   0.33      .  -0.39      .  | -1.04 |  84.8  81.9 |   23  91.3 |____F__
 Quebec, Canada           |   305  89.5  0.29 |   3.3  13.6  79.9     .   1.2   2.0 | -0.25  -0.14   0.24      .  -0.02  -0.09  | -1.33 |  89.1  89.9 |   36  88.9 |____F__
 Moscow City, Russian Fed.|   320  96.3  0.20 |   0.2   7.1  92.4     .   0.4   0.0 | -0.05  -0.19   0.20      .   0.01      .  | -2.17 |  96.2  96.4 |  196  97.4 |_VEAF__
 Abu Dhabi, UAE           |  1220  60.1  0.72 |  30.5  15.2  49.6     .   4.4   0.4 | -0.67  -0.05   0.66      .  -0.16  -0.04  | -0.87 |  60.8  59.3 |   74  87.8 |__H____
 Dubai, UAE               |   656  86.3  0.52 |   5.4  15.6  75.9     .   2.9   0.2 | -0.43  -0.26   0.47      .  -0.18  -0.01  | -1.13 |  85.7  86.9 |   52  92.3 |____F__
---------------------------------------------------------------------------------------------------------------------------------------------------------------
Keys:  DIFF= Percent correct score; DISC= Item discrimination; P_0...P_3= Percentage obtaining score level; P_OM, P_NR= Percentage Omitted, Not Reached;
       PB_0...PB_3= Point Biserial for score level; PB_OM, PB_NR= Point Biserial for Omitted, Not Reached; RDIFF= Rasch difficulty;
       Reliability: N= Responses double scored; Agr= Percentage agreement.
Flags: A= Point-biserial not ordered; B= Boys outperform girls; C= Difficulty less than chance; D= Negative/low discrimination; E= Easier than average;
       F= Score obtained by less than 10%; G= Girls outperform boys; H= Harder than average; R= Scoring reliability less than 85%; V= Difficulty greater than 95%.
```

Item statistics for all items included the number of students who responded in each country, an estimated item difficulty (the percentage of students that answered the item correctly), and the discrimination index (the point-biserial correlation between success on the item and total score).[1] Also provided was an estimate of the item difficulty parameter of the Rasch IRT model. Statistics for each item were displayed alphabetically by country, together with an international average (based on all participating countries) and a reference average (based on a pool of countries that have participated regularly in the PIRLS assessments) for each statistic. The reference countries are shown with an asterisk next to their names. The international and reference averages of the item difficulties and item discriminations guided the evaluation of the overall statistical properties of the items. The item almanacs also listed the benchmarking participants.

Statistics displayed for selected-response items included the percentage of students that chose each response option, the percentage of students that omitted or did not reach the item, and the point-biserial correlations for each of these categories. Statistics displayed for constructed-response items (which could have 1, 2, or 3 score points) included the percentage of students and point-biserial for each score level. Constructed-response item tables also provided information about the reliability with which each item was scored in each country, showing the total number

[1] For computing point biserial correlations, the total score is the percentage of points a student has scored on the items they were administered. Not reached and omitted responses are not included in the total score.

of responses that were scored twice and the percentage of score agreement between the scorers on these responses.

Included with these statistics are percentages for the categories *Omitted* and *Not Reached*. Omitted responses are defined as missing responses that occur between valid item responses, while Not Reached are those occurring at the end of a booklet. Both types of item non-response did not contribute to calculating the proportion correct and discrimination statistics, but the percentage of students is reported.

The definitions and detailed descriptions of the item statistics are given below. The statistics are listed in order of their appearance in the item review outputs:

CASES: This is the number of students to whom the item was administered. Students with Omitted (OM) or Not Reached (NR) codes are included in this figure. The number of students differs across items depending on the number of booklets where the item appears and the group-adaptive booklet rotation in each country.

DIFF: The item difficulty is estimated as the percentage correct on an item. For a 1-point item, including all selected-response items, it is the percent of students who provided a fully correct response. For items worth 2 or 3 points, it is the expected score divided by the maximum score achieved on the item as follows:

$$\frac{\sum_{k=1}^{K} k \cdot p_k}{K},$$

where K is the maximum number of score points, k is the score point category, and p_k is the proportion of students who received k points on the item. For example, for a 2-point item, if 25 percent of students scored 2 points, 50 percent scored 1 point, and the other 25 percent scored 0 points, then the average percent correct for such an item would be 50 percent. For this statistic, Omitted (OM) and Not Reached (NR) responses were excluded.

DISC: Item discrimination is computed as the correlation between the response to an item and the total score on all items administered to a student. Items exhibiting good measurement properties should have a moderately positive correlation, indicating that more able students are likely to get the item right and that less able ones are likely to get it wrong. For this statistic, Omitted (OM) and Not Reached (NR) responses were excluded.

Percentages (0/A, 1/B, 2/C, 3/D, E, F, OM, NR): These statistics represent the percentage of respondents choosing the different response options (for selected-response items) or the percentage of respondents by score points awarded (for constructed-response items), along with the percentage of responses coded as Omitted (OM) and Not Reached (NR). The percentages are computed based on "CASES" summed across all response categories, including Omitted (OM) and Not Reached (NR). The percentages sum to 100 percent.

Point-Biserials (0/A, 1/B, 2/C, 3/D, E, F, OM, NR): These statistics represent the point-biserial correlation between each response option (for selected-response items) or score level (for constructed-response items) and the total score on all items administered to a student. The point-biserial is also calculated for the categories Omitted (OM) and Not Reached (NR).

RDIFF: An estimate of the item difficulty based on the Rasch IRT model applied to the achievement data of a given country. The difficulty estimate is expressed in the logit metric (with a positive logit indicating a difficult item) and is scaled so that the average Rasch item difficulty across all items within each country is zero.

Average Score (Girls/Boys): The average DIFF across girls and across boys.

Reliability (N): Available for human-scored constructed-response items, the number of responses that were scored independently by two raters (double scored) for a given item in a country.

Reliability (Agr): Available for human-scored constructed-response items, the percent that the two scorers agreed on the score point value assigned to the item response.

Flags: As an aid to the reviewers, the item review displays included a series of flags signaling the presence of one or more conditions that might indicate an item requires further review. The flags seldom indicate an actual problem but serve to draw attention to potential sources of concern. The following conditions were flagged:

- The item discrimination (DISC) was less than 0.10 (flag D)
- The item difficulty (DIFF) was less than 0.25 for selected-response items (flag C)
- The item difficulty (DIFF) exceeded 0.95 (flag V)
- The Rasch difficulty estimate (RDIFF) for a given country showed the item was either easier (flag E) or more difficult (flag H) than the international average for that item
- The point-biserial correlation for at least one distracter in a selected-response item was positive, or the point-biserial correlations across the score levels of a constructed-response item were not ordered (flag A)
- The percentage of students selecting one of the response options for a selected-response item, or one of the score values for a constructed-response item, was less than 10% (flag F)
- Scoring reliability for agreement on the score value of a constructed-response item was less than 0.85 (flag R)

Scoring Reliability for Human-Scored Items

Constructed-response items made up roughly half of the score points in the PIRLS 2021 assessment. For many of these constructed-response items, scoring required human judgment to assign appropriate score points to the student responses. To ensure that the items requiring human scoring were scored reliably in all countries, the TIMSS & PIRLS International Study Center developed detailed scoring guides for each constructed-response item. The scoring guides provided descriptions and examples of acceptable responses for each score point value. The TIMSS & PIRLS International Study Center also provided extensive training in applying the scoring guides. See Chapter 1 for more information on developing the scoring guides, and see Chapter 4 for information on the human-scoring process.

The following sections describe how PIRLS 2021 assessed and documented human-scoring reliability within-country, over time (trend), and across countries.

Within-Country Scoring Reliability

To gather and document information about the within-country agreement among scorers for PIRLS 2021, a random sample of approximately 200 student responses per item was scored independently by two scorers. The TIMSS & PIRLS International Study Center examined the inter-scorer agreement for each item in each country as part of the item review process, flagging any countries where an item's scoring agreement was below 75 percent for further review. Appendix 9A shows the average and range of the within-country percentages of score point agreement across all human-scored items for PIRLS 2021 (paperPIRLS and digitalPIRLS countries), and for the PIRLS 2021 bridge samples. The average within-country score point agreement for PIRLS was 95 percent, ranging from an average minimum of 84 percent to a maximum of 100 percent. For the bridge data, the average within-country score point agreement was also 95 percent.

Trend Item Scoring Reliability Study

The TIMSS & PIRLS International Study Center also took steps to show that the 2021 human-scored constructed-response items used in PIRLS 2016 were scored in the same way in both assessments. In anticipation of this, countries participating in PIRLS 2016 sent samples of scored student booklets from the 2016 data collection to IEA Hamburg, where they were digitally scanned and stored for later use. As a check on scoring consistency from one administration to the next, staff members working in each country on scoring the 2021 data also were asked to score these 2016 responses using IEA Hamburg's CodingExpert Software. Each country scored 200 responses for 18 PIRLS reading items from three passages.

There was a very high degree of scoring consistency in PIRLS 2021. The exact agreement between the scores awarded in 2016 and those given by the 2021 scorers was 94 percent on average internationally. The average and range of scoring consistency over time can be found in Appendix 9B.

Cross-Country Scoring Reliability Study

The TIMSS & PIRLS International Study Center documented the consistency of scoring across countries. Since participating PIRLS countries use many different languages, it was not possible to establish the reliability of constructed-response scoring across all countries; however, a cross-country study of scoring reliability was conducted among all countries that had scorers who were proficient in English. Cross-country scoring included 200 student responses for 18 PIRLS reading items from three passages. This common set of student responses was then scored independently by each country using IEA Hamburg's CodingExpert Software.

In all, scorers from 53 countries and one benchmarking entity participated in the process. Having 54 independent scorers gave a total of 1,431 possible comparisons for each student response to each item. With 200 responses per item expected to be scored by each country, a maximum of 286,200 total comparisons were available to obtain the cross-country scoring reliability agreement for any given item.

Agreement across countries was defined as the percentage of these comparisons that were in exact agreement. On average, internationally, scorer reliability across countries in PIRLS 2021 was high at 92.5 percent. See Appendix 9C for the results of the cross-country scoring reliability study.

Item-by-Country Interactions

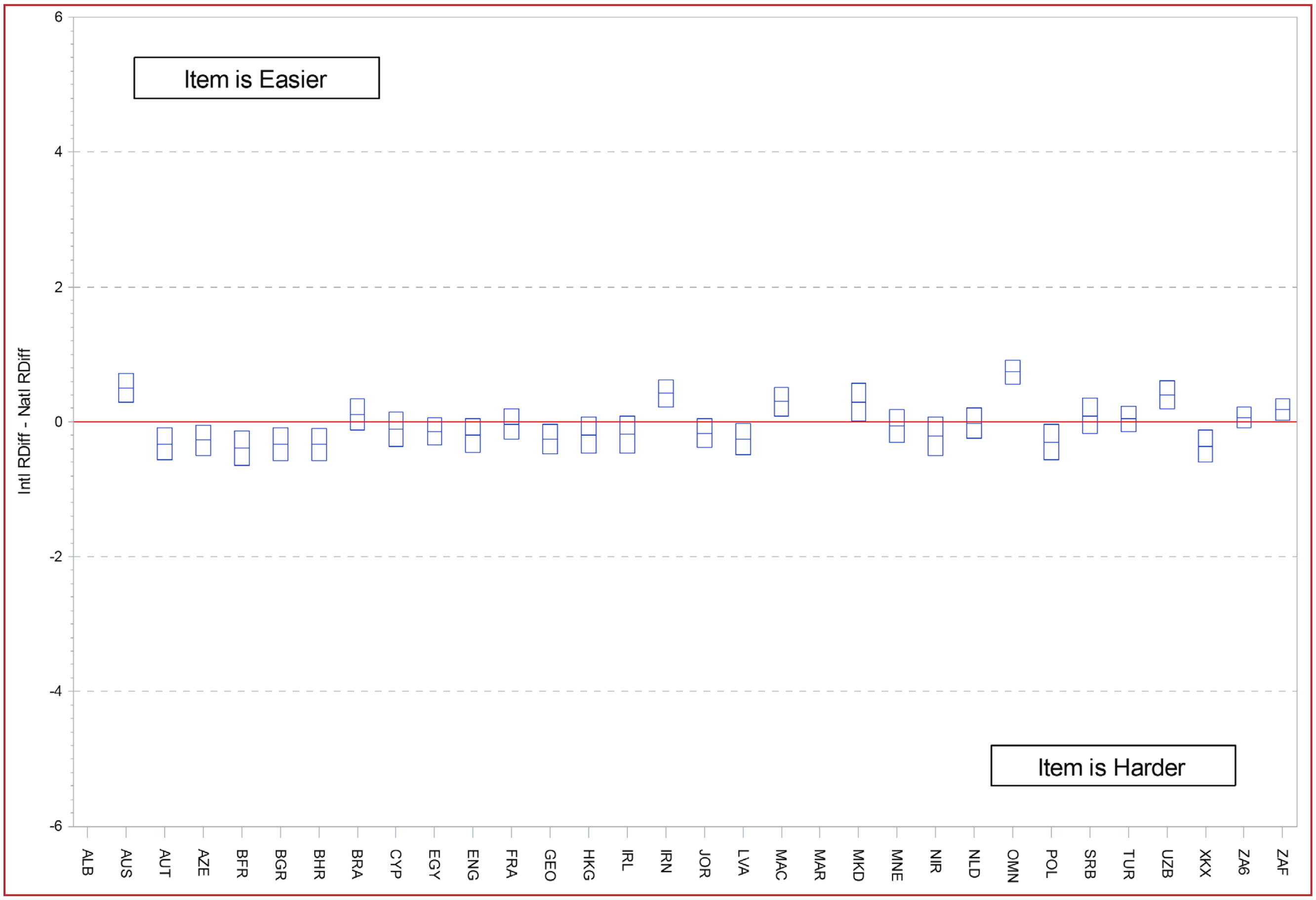
Although countries are expected to exhibit some variation in performance across items, in general, countries with high average performance on the assessment should perform relatively well on each item, and low-scoring countries should perform less well. When the opposite happens (e.g., when a high-performing country has low performance on an item on which other countries did well), it is called an "item-by-country interaction" or country-level "differential item functioning" (DIF). The presence of relatively large item-by-country interactions may indicate that an item is flawed for that particular country. This can cause misfit of the IRT measurement model to the achievement data, which could negatively impact achievement estimates (see Chapter 11).

The TIMSS & PIRLS International Study Center used two types of statistics with graphical displays to detect instances of country DIF. The first method was conducted based on within-country statistics using Rasch item difficulties reported in the item data almanacs. The second was conducted for each country's data but based on international item parameters estimated using two- and three-parameter generalized partial credit IRT models (see Chapter 10).

Rasch Method

The first graphical display for a particular item, shown in Exhibit 9.3, shows the difference between each country's Rasch item difficulty and the international average Rasch item difficulty. When this difference is greater than 2 logits or less than –2 logits, it is considered an item-by-country interaction and is flagged for the Analysis Unit staff to address a potential problem.

Exhibit 9.3: Example Item-by-Country Rasch Plot for a PIRLS 2021 Item



In each of these item-by-country interaction displays, the difference for each country is presented as a 95 percent confidence interval, which includes a Bonferroni correction for multiple comparisons across the participating countries. The limits for this confidence interval were computed as follows:

$$\begin{aligned}\text{Upper Limit} &= RDIFF_{i.} - RDIFF_{ik} + SE(RDIFF_{ik}) \cdot Z_b \\ \text{Lower Limit} &= RDIFF_{i.} - RDIFF_{ik} - SE(RDIFF_{ik}) \cdot Z_b\end{aligned} \qquad (9.1)$$

where $RDIFF_{ik}$ is the Rasch difficulty of item i in country k, $RDIFF_{i.}$ is the international average Rasch difficulty of item i, $SE(RDIFF_{ik})$ is the standard error of the Rasch difficulty of item i in country k, and Z_b is the critical value.

IRT Method

As an additional criterion to detect country DIF, the TIMSS & PIRLS International Study Center used international IRT parameters to generate theoretical item response functions for each item. Given a student's latent ability θ, the function gives a probability of answering a given item correctly. Graphs of these functions are known as item characteristic curves (ICCs). For each country, empirical ICCs were calculated for each item from the latent abilities estimated for each student that responded to the item. These country-level empirical ICCs were plotted alongside the international theoretical ICCs (see example in Exhibit 9.4). The country-level empirical functions themselves are based on an estimated latent ability distribution that uses the IRT model, and they are therefore also referred to as item functions based on pseudo counts. When the empirical results for an item fall near the fitted international curves, the IRT model for that item fits the country's data well and provides an accurate and reliable measurement of the underlying proficiency scale.

Exhibit 9.4: Example Country-Level ICC Plot for a PIRLS 2021 Item

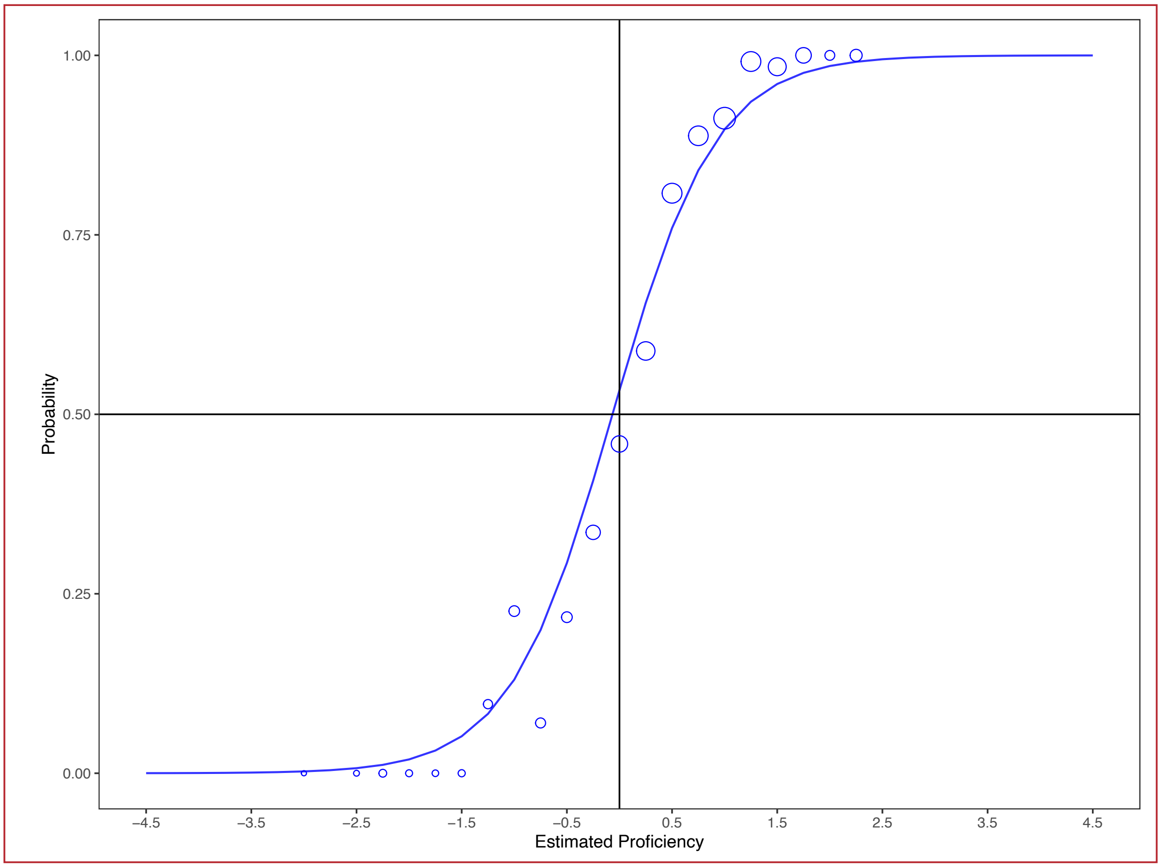


Exhibit 9.4 shows an example of a country's empirical ICC plotted with the fitted international curve. The horizontal x-axis represents the proficiency scale on the logit metric, and the vertical y-axis represents the probability of a correct response. The fitted curve based on the estimated international item parameters is shown as a solid line. Country-level empirical results based on pseudo counts are represented by circles. The center of each circle represents the empirical percentage of correct responses, and the size of each circle is proportional to the estimated number of students contributing to the empirical percent correct in its corresponding interval. Visual inspection of the country ICC plots can help detect country-level DIF, also known as "misfit," when the circles do not align well to the solid line.

The fit of a country's data to the IRT model, or the level of fit between a country's empirical ICC and the fitted international curve, is quantified by the root mean square difference (RMSD) statistic. The RMSD is the square root of the average of squared differences (i.e., the area) between the country-level empirical curve, shown as bubbles, and the international fitted curve, shown as the straight line, weighted by the size of the bubbles. The RMSD statistic is sensitive to country-specific deviations from the international parameters in both item difficulty and item discrimination. When the RMSD value is close to zero, it signifies a good fit of the items, implying that the model with international item parameters is an accurate representation of the responses within that specific country.

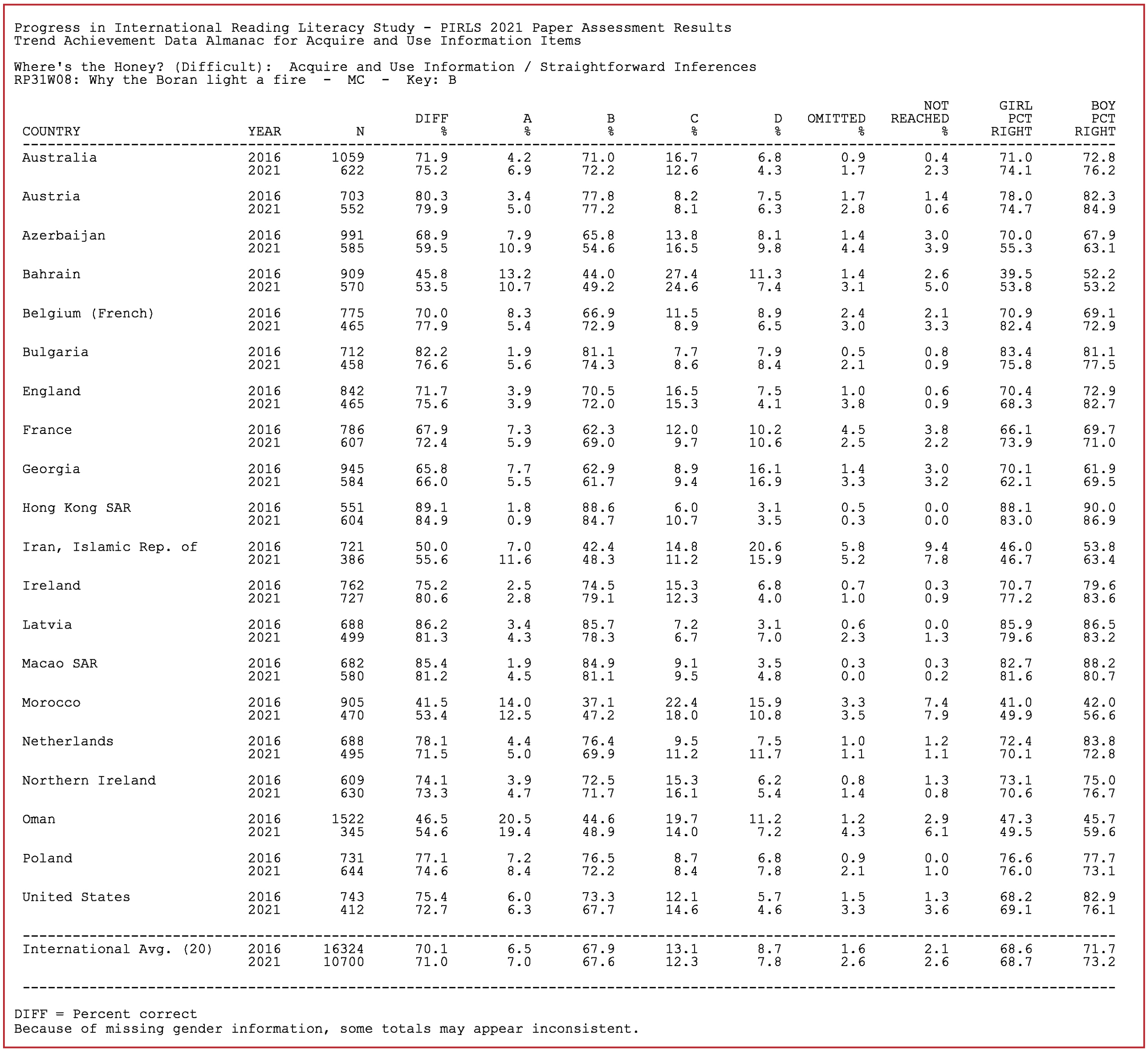
The median absolute deviation (MAD) outlier detection method applied to the RMSD values calculated for each country and item (von Davier & Bezirhan, 2022) was used as a diagnostic tool to help identify potential country-level item misfit. MAD is a robust measure of dispersion that was employed as a flagging rule rather than an arbitrary cut-off value. This method flags an item as a possible misfit for a country if its distance from the median of the absolute distances of all other observations exceeds a predetermined threshold. For PIRLS 2021 item review, a conservative threshold of 4.5 was used.

Review of Item Statistics for Measuring Trends

Successive PIRLS assessments include achievement items from previous assessments in order to measure trends, as well as new items developed for each successive assessment. Accordingly, the PIRLS 2021 assessment included texts and items used in 2006, 2011, and 2016 with ones developed specifically for 2021. Therefore, an important review step included checking that these "trend items" had statistical properties in 2021 similar to those they had in the previous 2016 assessment (e.g., a PIRLS item that was relatively easy in 2016 should still be relatively easy in 2021).

As shown in the example in Exhibit 9.5, the trend item review focused on statistics for paper trend items from countries that participated in both the current and previous assessments (2021 and 2016). This included statistics for the digitalPIRLS bridge samples. For each country, trend item statistics included the percentage of students in each score category (or response option for selected-response items) for each assessment, the difficulty of the item, and the percent correct by gender. The primary aim of reviewing these item statistics was to detect any unusual international-level changes in item difficulties between administrations, which might indicate a problem in using the item to measure trends. At the country level, it is typical for sampling variance to cause small differences in statistics between assessments. However, larger differences can indicate a more systematic change or data problem.

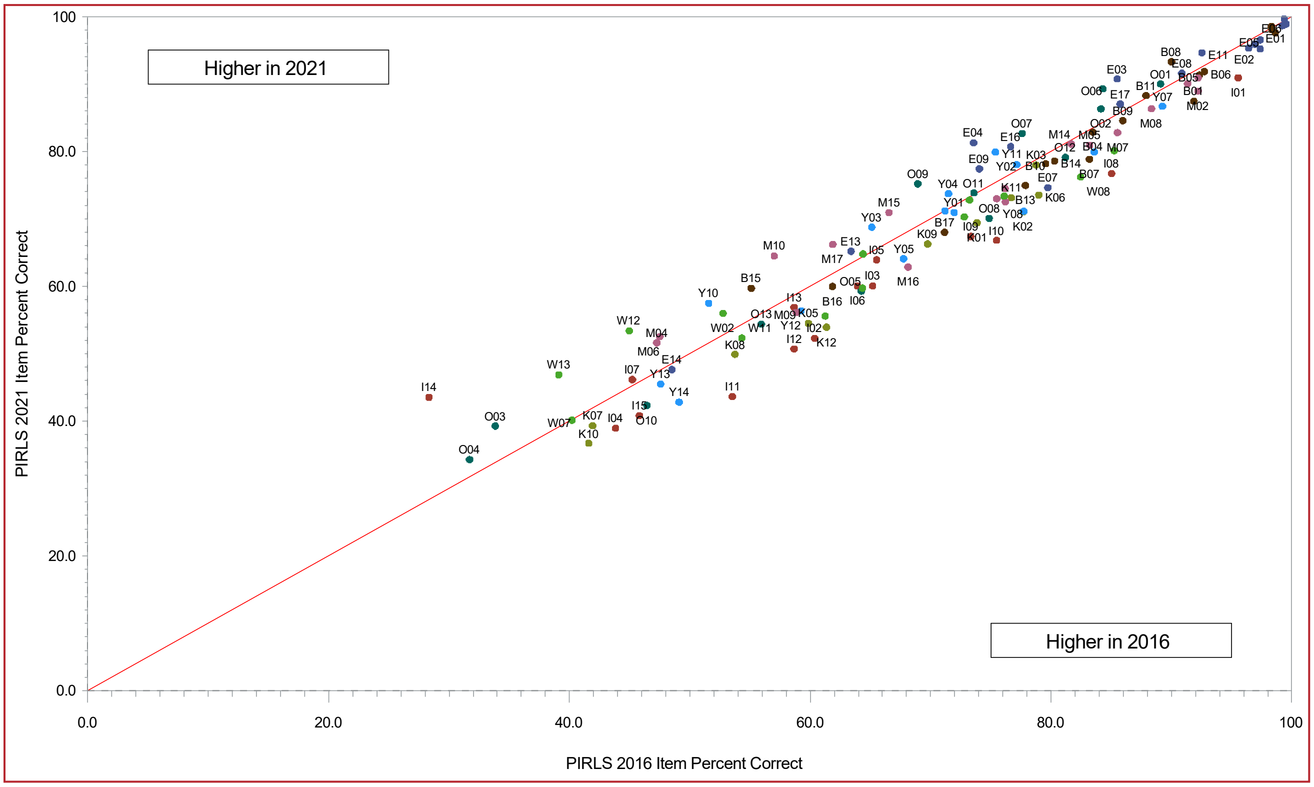
Exhibit 9.5: Example Item Statistics in 2021 and 2016 for a PIRLS 2021 Trend Item



```
Progress in International Reading Literacy Study - PIRLS 2021 Paper Assessment Results
Trend Achievement Data Almanac for Acquire and Use Information Items

Where's the Honey? (Difficult):  Acquire and Use Information / Straightforward Inferences
RP31W08: Why the Boran light a fire  -  MC  -  Key: B

                                                                                                         NOT      GIRL       BOY
                                         DIFF         A         B         C         D   OMITTED   REACHED       PCT       PCT
 COUNTRY                   YEAR         N         %         %         %         %         %         %         %     RIGHT     RIGHT
 ----------------------------------------------------------------------------------------------------------------------------------
 Australia                 2016      1059      71.9       4.2      71.0      16.7       6.8       0.9       0.4      71.0      72.8
                           2021       622      75.2       6.9      72.2      12.6       4.3       1.7       2.3      74.1      76.2

 Austria                   2016       703      80.3       3.4      77.8       8.2       7.5       1.7       1.4      78.0      82.3
                           2021       552      79.9       5.0      77.2       8.1       6.3       2.8       0.6      74.7      84.9

 Azerbaijan                2016       991      68.9       7.9      65.8      13.8       8.1       1.4       3.0      70.0      67.9
                           2021       585      59.5      10.9      54.6      16.5       9.8       4.4       3.9      55.3      63.1

 Bahrain                   2016       909      45.8      13.2      44.0      27.4      11.3       1.4       2.6      39.5      52.2
                           2021       570      53.5      10.7      49.2      24.6       7.4       3.1       5.0      53.8      53.2

 Belgium (French)          2016       775      70.0       8.3      66.9      11.5       8.9       2.4       2.1      70.9      69.1
                           2021       465      77.9       5.4      72.9       8.9       6.5       3.0       3.3      82.4      72.9

 Bulgaria                  2016       712      82.2       1.9      81.1       7.7       7.9       0.5       0.8      83.4      81.1
                           2021       458      76.6       5.6      74.3       8.6       8.4       2.1       0.9      75.8      77.5

 England                   2016       842      71.7       3.9      70.5      16.5       7.5       1.0       0.6      70.4      72.9
                           2021       465      75.6       3.9      72.0      15.3       4.1       3.8       0.9      68.3      82.7

 France                    2016       786      67.9       7.3      62.3      12.0      10.2       4.5       3.8      66.1      69.7
                           2021       607      72.4       5.9      69.0       9.7      10.6       2.5       2.2      73.9      71.0

 Georgia                   2016       945      65.8       7.7      62.9       8.9      16.1       1.4       3.0      70.1      61.9
                           2021       584      66.0       5.5      61.7       9.4      16.9       3.3       3.2      62.1      69.5

 Hong Kong SAR             2016       551      89.1       1.8      88.6       6.0       3.1       0.5       0.0      88.1      90.0
                           2021       604      84.9       0.9      84.7      10.7       3.5       0.3       0.0      83.0      86.9

 Iran, Islamic Rep. of     2016       721      50.0       7.0      42.4      14.8      20.6       5.8       9.4      46.0      53.8
                           2021       386      55.6      11.6      48.3      11.2      15.9       5.2       7.8      46.7      63.4

 Ireland                   2016       762      75.2       2.5      74.5      15.3       6.8       0.7       0.3      70.7      79.6
                           2021       727      80.6       2.8      79.1      12.3       4.0       1.0       0.9      77.2      83.6

 Latvia                    2016       688      86.2       3.4      85.7       7.2       3.1       0.6       0.0      85.9      86.5
                           2021       499      81.3       4.3      78.3       6.7       7.0       2.3       1.3      79.6      83.2

 Macao SAR                 2016       682      85.4       1.9      84.9       9.1       3.5       0.3       0.3      82.7      88.2
                           2021       580      81.2       4.5      81.1       9.5       4.8       0.0       0.2      81.6      80.7

 Morocco                   2016       905      41.5      14.0      37.1      22.4      15.9       3.3       7.4      41.0      42.0
                           2021       470      53.4      12.5      47.2      18.0      10.8       3.5       7.9      49.9      56.6

 Netherlands               2016       688      78.1       4.4      76.4       9.5       7.5       1.0       1.2      72.4      83.8
                           2021       495      71.5       5.0      69.9      11.2      11.7       1.1       1.1      70.1      72.8

 Northern Ireland          2016       609      74.1       3.9      72.5      15.3       6.2       0.8       1.3      73.1      75.0
                           2021       630      73.3       4.7      71.7      16.1       5.4       1.4       0.8      70.6      76.7

 Oman                      2016      1522      46.5      20.5      44.6      19.7      11.2       1.2       2.9      47.3      45.7
                           2021       345      54.6      19.4      48.9      14.0       7.2       4.3       6.1      49.5      59.6

 Poland                    2016       731      77.1       7.2      76.5       8.7       6.8       0.9       0.0      76.6      77.7
                           2021       644      74.6       8.4      72.2       8.4       7.8       2.1       1.0      76.0      73.1

 United States             2016       743      75.4       6.0      73.3      12.1       5.7       1.5       1.3      68.2      82.9
                           2021       412      72.7       6.3      67.7      14.6       4.6       3.3       3.6      69.1      76.1

 ----------------------------------------------------------------------------------------------------------------------------------
 International Avg. (20)   2016     16324      70.1       6.5      67.9      13.1       8.7       1.6       2.1      68.6      71.7
                           2021     10700      71.0       7.0      67.6      12.3       7.8       2.6       2.6      68.7      73.2

 ----------------------------------------------------------------------------------------------------------------------------------

DIFF = Percent correct
Because of missing gender information, some totals may appear inconsistent.
```

The TIMSS & PIRLS International Study Center plotted trends in item percent correct by country. Exhibit 9.6 shows the item percent correct for a sample country. Countries with data points clustered tightly along the diagonal performed similarly across 2016 and 2021. Countries where the item percent correct was higher in 2021 (clustered above the diagonal) or higher in 2016 (clustered below the diagonal) were flagged for further review in extreme cases.

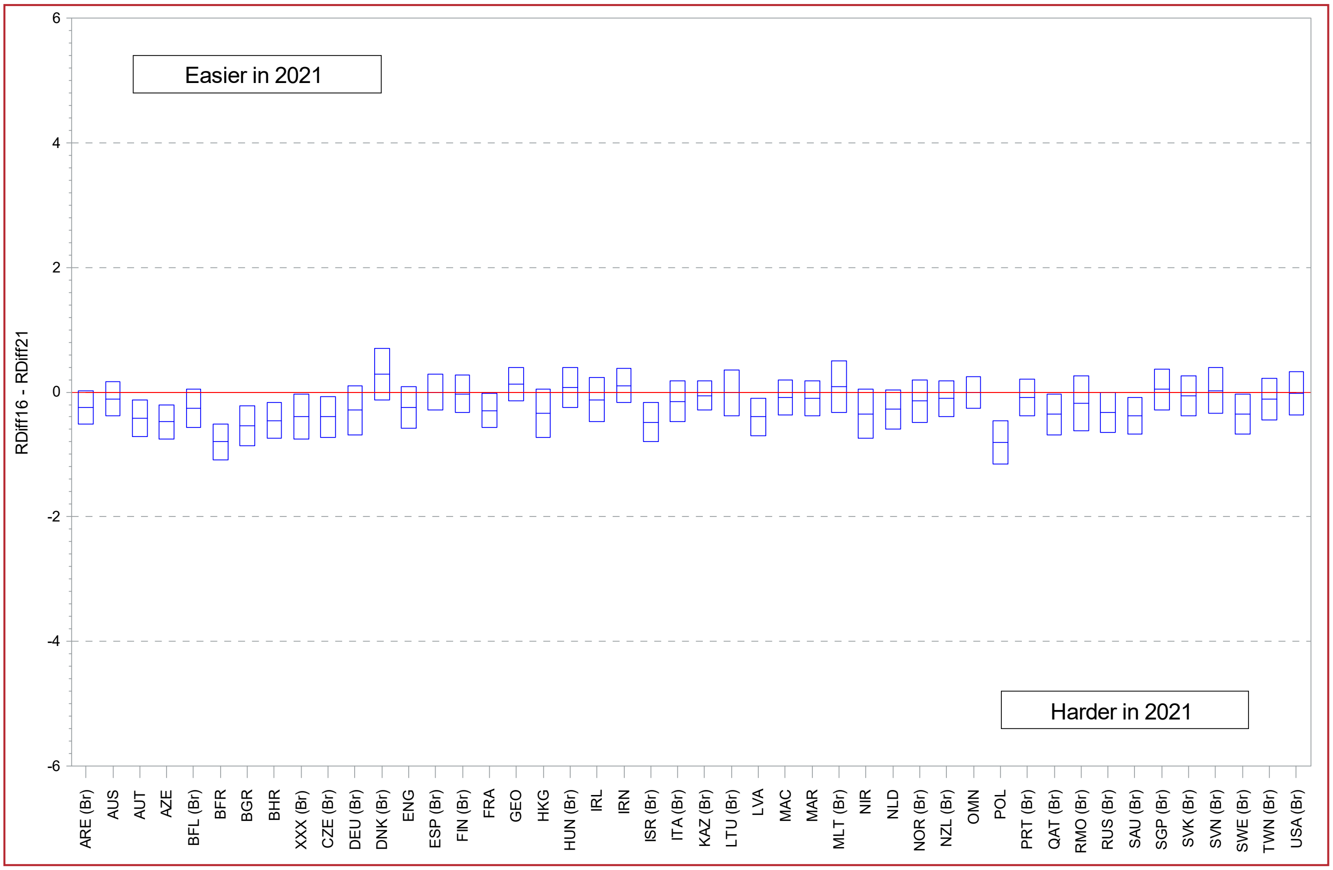
Exhibit 9.6: Example Plot of Trends in Item Percent Correct by Country



The TIMSS & PIRLS International Study Center used two different graphical displays to examine the differences in countries' Rasch item difficulties between 2021 and 2016. While some variability of country-level difficulties is due to sampling error, systematic changes in item difficulties could be due to overall achievement that may have improved or declined. Items were flagged for review if the difference between the Rasch difficulties across the two assessments for a particular country was greater than 2 logits.

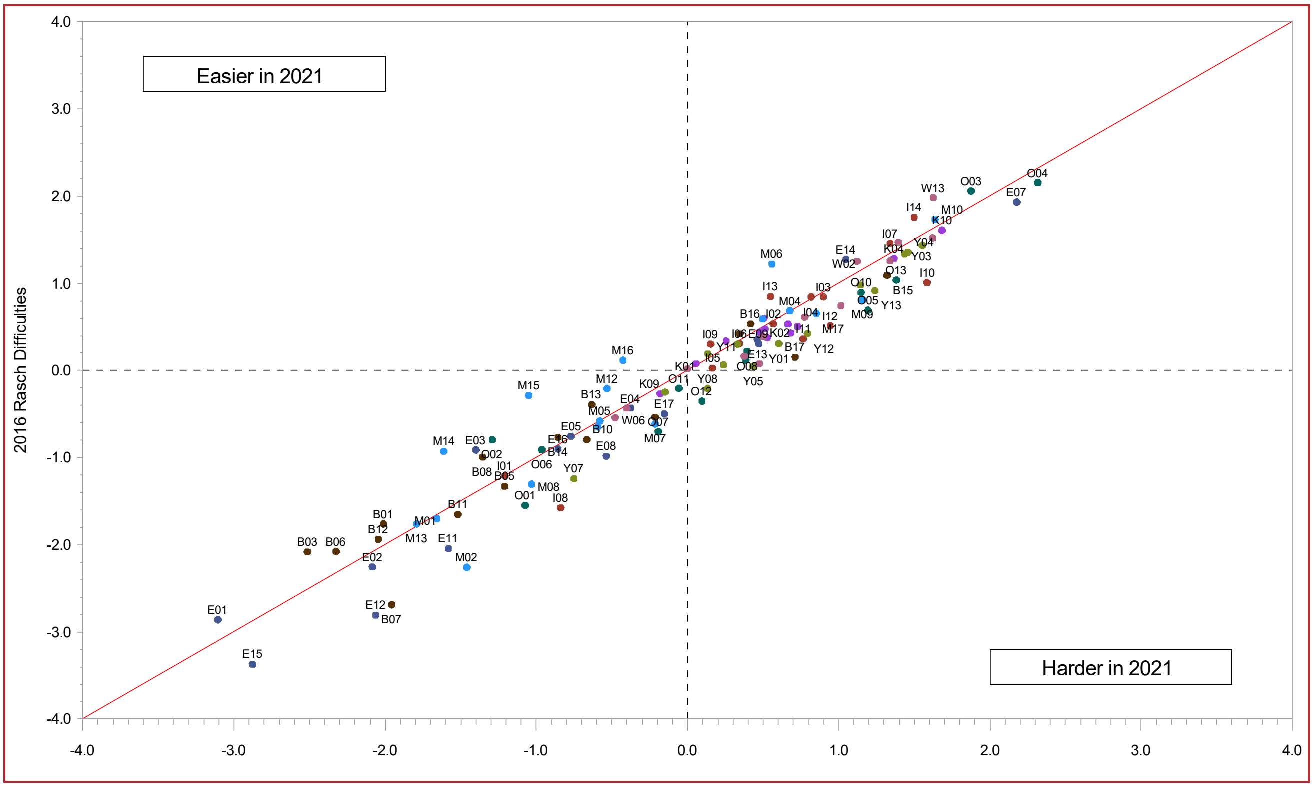
The first of these displays (Exhibit 9.7) shows each country's difference in Rasch item difficulty between 2021 and 2016 for a given item. The difference in Rasch item difficulty for each country is displayed as a confidence interval, calculated using equation (9.1) but using each country's 2021 and 2016 Rasch difficulties in place of the national and international difficulties. A positive difference for a country indicates that the item estimate was relatively easier in 2021, and a negative difference indicates that the item estimate was relatively more difficult.

Exhibit 9.7: Example Plot of Differences in Rasch Item Difficulties Between 2021 and 2016 for a PIRLS 2021 Trend Item



The second graphical display, presented in Exhibit 9.8, shows the performance of a given country on all trend items simultaneously. For each country, the graph plots the 2021 Rasch difficulty of every trend item against its Rasch difficulty in 2016. When there are no differences between the difficulties in the two successive administrations, the data points align on or near the diagonal. Some deviations are expected due to the limited sample size per country, but large deviations from the diagonal were noted for further investigation.

Exhibit 9.8: Example Plot of Rasch Trend Item Difficulties Across PIRLS 2021 and 2016



Text Position and Booklet Effects

As described in the PIRLS 2021 Assessment Design (Martin et al., 2019), assessment items are arranged in blocks according to the text to which they pertain. These text and item blocks are assembled into achievement booklets, paired with a second text of the other purpose (literary or informational). The group adaptive assessment design in PIRLS 2021 divided the 18 texts that were in both paper and digital versions of the assessment into three levels of text difficulty—difficult, medium, and easy—and combined these into two levels of booklet difficulty. More difficult booklets (9) comprised two difficult texts or one medium and one difficult text. Less difficult booklets (9) consisted of easy and medium texts or two easy texts. All countries administered all 18 texts and all 18 booklets, but the balance of more difficult and less difficult booklets varied with the reading achievement level of the students in the country (Martin et al., 2019).

Each text appeared in two booklets, as the first half of one booklet and the second half of another, with a few exceptions. The first exception came because the group adaptive design dictated that when texts of two different difficulties are combined in a booklet, the easier text should always come first. This required three "easy" texts—*Learning a New Language*, *The Amazing*

Octopus, and *The Summer My Father was 10*—to only ever appear in the first half of a booklet, and three "difficult" texts—*Icelandic Horses*, *Oliver and the Griffin*, and *World's Bank for Seeds*—to only ever appear in the second half of a booklet. The second exception came because all informational texts appeared in more than two booklets for digitalPIRLS countries. The digitalPIRLS assessment included 65 additional booklets to incorporate ePIRLS into the design. Booklets 19 through 38 included all possible combinations of pairs of the five ePIRLS tasks. The rest, booklets 39 through 83, included all possible combinations of pairs of one digital informational text and one ePIRLS task. In these combinations, the ePIRLS tasks were always placed in the second half of the booklet.

To examine whether the particular position in which a text was administered affected student performance, the TIMSS & PIRLS International Study Center computed item statistics for each of the two positions in which each text appeared in the booklet design—either position 1 or position 2. For text and item sets that always appeared either first or second in a booklet, the positions were defined by whether they are paired with an easier text (position 1) or a more difficult text (position 2). The results are reported in Appendix 9D for each assessment averaged across countries, as well as for each country across texts. A summary of results with the average differences in item statistics between the booklet positions is provided in Exhibit 9.9.

Exhibit 9.9: Summary of International Average Item Statistics by Booklet Position (Weighted)

	Average Percent Correct Across Items			Average Percent Omitted Responses Across Items			Average Percent Not Reached Across Items		
	Position 1	Position 2	Difference	Position 1	Position 2	Difference	Position 1	Position 2	Difference
paperPIRLS	61.7	60.6	–1.1	6.7	6.9	0.2	3.5	3.1	–0.4
digitalPIRLS	62.6	62.7	0.2	4.0	3.7	–0.4	1.6	1.0	–0.6
ePIRLS	58.0	56.0	–2.0	5.2	4.5	–0.7	3.0	1.2	–1.8
Bridge	65.1	64.3	–0.9	5.0	5.1	0.1	1.8	1.6	–0.2

The results indicate a minimal impact of text position on the PIRLS 2021 item statistics. Passages appearing in the first and second half of the booklet were similarly difficult in positions 1 and 2 across paperPIRLS, digitalPIRLS, ePIRLS, and bridge assessments. Text and item sets in positions 1 and 2 also had similar non-response rates. Across countries, differences in average item percent correct between position 1 and position 2 ranged from –2.0 for ePIRLS items to 0.2 for digitalPIRLS items. Differences in average percent omitted ranged from –0.7 for ePIRLS items to 0.2 for paperPIRLS items. Differences in average percent not reached ranged from –0.2 for bridge items to –1.8 for ePIRLS items.

As an additional validation for the PIRLS 2021 group adaptive design, the TIMSS & PIRLS International Study Center reviewed rates of non-response for all countries by booklet type, as well as distributions of ability estimates from the IRT model by booklet type. Appendix 9E reports these results for each country, along with the proportion of more difficult and less difficult booklets administered in each country. Across participating countries, the overall non-response rate was 7.2 percent. The average non-response rate for more difficult booklets was 9.5 percent, compared to 5.2 percent for less difficult booklets. Across countries on average, standard deviations of ability estimates differed by 0.02 logits between more difficult booklets and less difficult booklets. On average, higher performing countries showed smaller standard deviations for more difficult booklets, while lower performing countries showed smaller standard deviations for less difficult booklets.

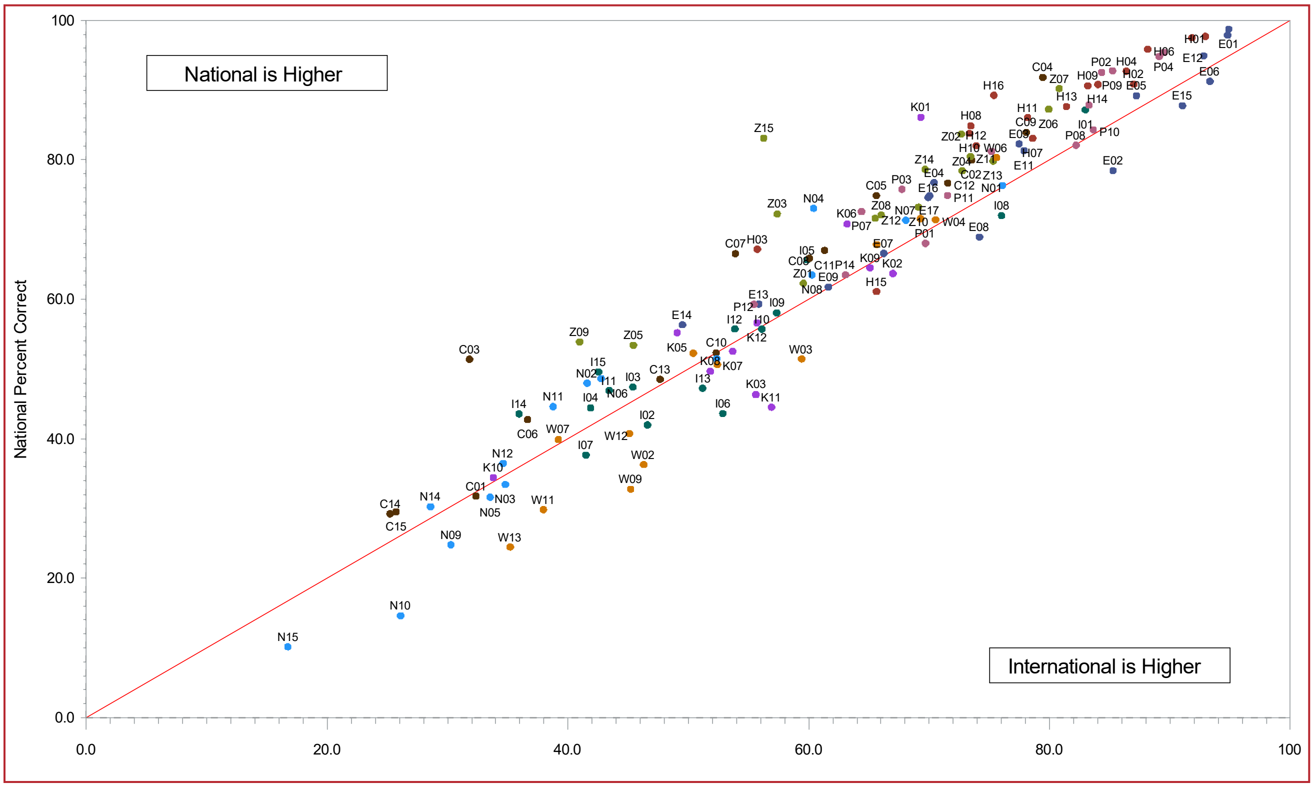
Detecting Anomalies in the PIRLS 2021 Achievement Data

To ensure that each participating country and benchmarking entity had data adhering to PIRLS' quality standards, the TIMSS & PIRLS International Study Center conducted analyses of item statistics at the country level. Several graphical displays were produced for each PIRLS participant: item percent correct, item point-biserial correlations, and item non-response (Omitted or Not Reached). The graphs were analyzed to detect any anomalous patterns in any particular country's data relative to the international average or to their previous PIRLS performance. Irregular patterns might indicate systematic errors occurring in a country's data, which may be due to errors in collecting and processing the data. For any anomalous patterns detected in the item statistics for a particular country, the National Research Coordinator was contacted to discuss the nature of the anomalies and resolve any issues.

The first set of graphical displays compared each country's performance to the international average for all items simultaneously where item performance was defined as item percent correct, item discrimination (point-biserial correlation), and item percent non-response. An example is shown in Exhibit 9.10 for item percent correct. For each country, the graph plots the 2021 item percent correct of all items against their 2021 international averages. Typical patterns show data points falling across the range of the x- and y-axis, with small and random deviations from the diagonal. There will be more points above the diagonal for higher-performing countries and more points below for lower-performing countries. Otherwise, the points should align closely with the diagonal. The best-fit line should be approximately linear and parallel to the diagonal. Any patterns largely deviating from this were noted for further investigation. Plots comparing national and international item discrimination (point-biserial correlation) and national and international percent non-response have similar patterns, but with data points more tightly clustered together since their range is smaller.

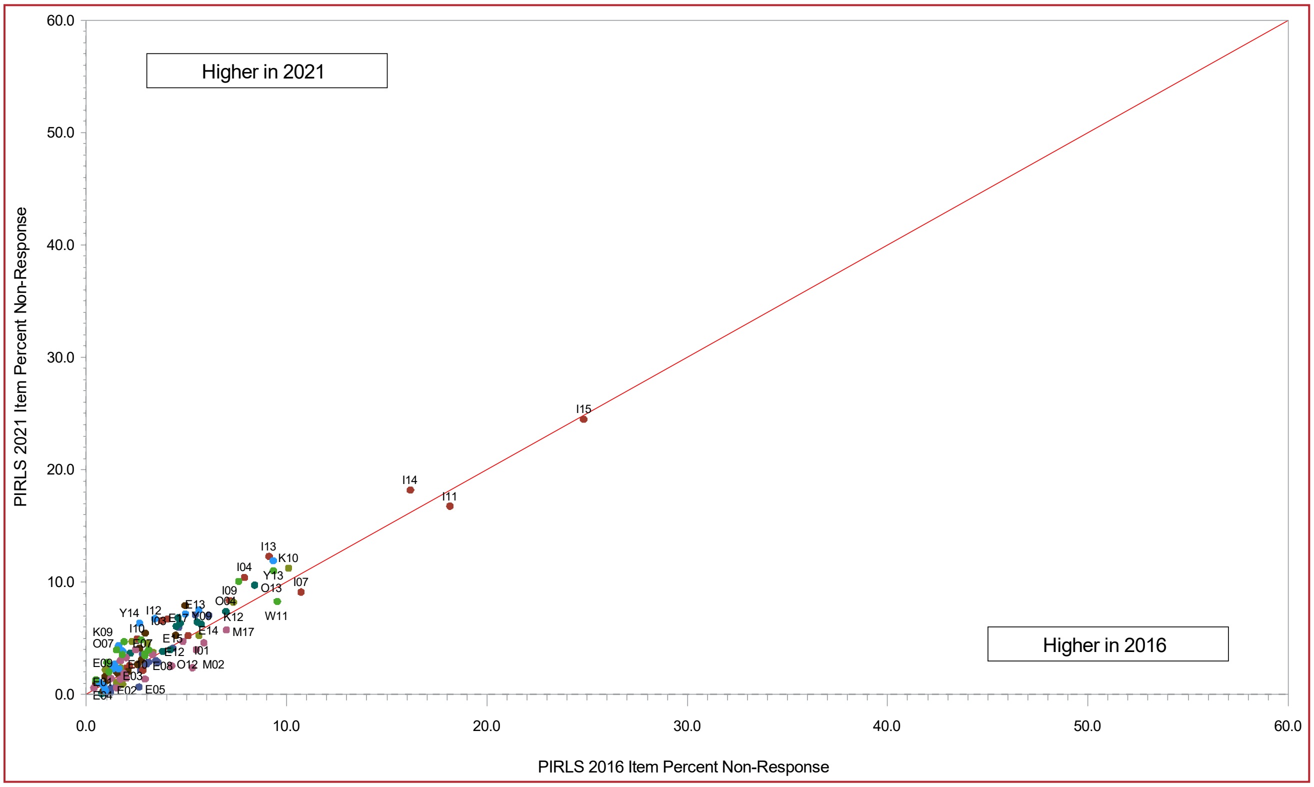
These plots of national versus international item statistics were also compared against the same plots produced in PIRLS 2016. If the patterns for the assessments were unusually different, it might have indicated a problem in the 2021 data. The plots were also examined separately for selected-response and constructed-response formatted items to ensure similar patterns. It was expected that the relationship between national and international statistics for both item types would also match that from PIRLS 2016.

Exhibit 9.10: Example Plot of Item Percent Correct Across National and International by Country



The second set of graphical displays compared each country's PIRLS 2021 trend item performance with their PIRLS 2016 item performance for all items simultaneously, where item performance was defined in terms of percent correct (similar to Exhibit 9.6), item discrimination (point-biserial correlation), and item percent non-response (Omitted or Not Reached). For this step in the item review process, the Analysis Unit was looking across items for unusual patterns and not at individual items. An example is shown in Exhibit 9.11 for item percent non-response, displaying a typical pattern. For each country, the graphs plot the 2021 item percent non-response of every trend item against its item percent non-response in 2016, with points colored according to the text to which they pertain. When there were few or no differences between the non-response rates in the two successive administrations, the data points aligned on or near the diagonal from the graph origin. While some changes were anticipated due to sampling variation or because countries' overall achievement may have improved or declined, unusually large deviations from the diagonal were noted for further investigation. For all statistics plotted, it was expected that comparisons would show similar patterns for both selected-response and constructed-response item types, and any differences would not relate to the difficulty of the item.

Exhibit 9.11: Example Plot of Item Percent Non-Response Across PIRLS 2021 and 2016 by Country



An additional set of plots were produced comparing each country's PIRLS 2021 item performance with their item performance from the field test conducted one year earlier. These plots were similar to the example in Exhibit 9.11 above comparing 2021 and 2016 performance, with the expectation of smaller differences. Large differences in item performance compared to the field test would be considered an implausible change in performance, warranting further review.

Item Review Outcomes

Using all the information from the comprehensive collection of item analyses and reliability data that were computed and summarized for PIRLS 2021, the TIMSS & PIRLS International Study Center thoroughly reviewed all item statistics for every participating country and benchmarking participant to ensure that the items were performing comparably across countries and modes. In particular, the following observations led to items being considered for possible deletion from the international database:

- An error was detected for a particular country during translation verification but was not corrected before test administration
- Data checking revealed a selected-response item with more or fewer options than in the international version for a particular country
- The item analysis showed the item to have a negative biserial, or, for an item with more than 1 score point, point-biserial correlations that did not increase with each score level
- For selected-response items, the item review revealed a faulty distracter influencing the item statistics for all countries
- The item-by-country interaction results showed a very large negative or positive interaction for a particular country
- For constructed-response items, the within-country scoring reliability data showed an agreement of less than 75 percent
- For paper trend items, an item performed substantially differently in 2021 compared to the PIRLS 2016 administration, or an item was not included in the previous assessment for a particular country

When the item statistics indicated a problem with an item, the documentation from the translation verification 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.

Checking the PIRLS 2021 achievement data involved a review of 769 items 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, most items were removed due to differences attributable to translation problems, which were first detected with RMSDs using the MAD outlier method. Score codes for some constructed-response items were recoded if the point-biserial correlations did not behave as expected. Decisions about deleting items for all countries were most often implemented for both digitalPIRLS and paperPIRLS versions, with a few exceptions.

Appendix 9E includes a list of deleted items, as well as a list of recodes made to constructed-response items. For Albania, a larger number of items were deleted due to data problems that caused severe differential item functioning: 104 items were deleted from all booklets, and 58 items were deleted from only one booklet.

A number of PIRLS 2021 items were used to derive scores for analysis and IRT scaling purposes. Appendix 9F includes details about how score points were calculated for each derived item.

Review of Item Statistics Between digitalPIRLS and paperPIRLS

The PIRLS 2021 item review included steps to compare the statistical properties of the digital and paper versions of achievement items (e.g., an item that was relatively easy on paper should also be easy in digital format). The review focused on comparing item statistics for trend items administered in digital format (to the regular sample of students) to the statistics of trend items administered in paper format to the bridge samples. The review followed procedures similar to those conducted for trend analysis but with greater emphasis on IRT methods. The analyses of differences between statistics of the digitalPIRLS and bridge samples are provided in Chapter 12 of this volume.

References

Martin, M. O., von Davier, M., Foy, P., & Mullis, I. V. S. (2019). PIRLS 2021 assessment design. In I. V. S. Mullis & M.O. Martin's (Eds.), *PIRLS 2021 Assessment Frameworks*. Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/pirls2021/frameworks/

von Davier, M., & Bezirhan, U. (2022). A robust method for detecting item misfit in large-scale assessments. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644221105819

Appendix 9A: PIRLS 2021 Within-Country Scoring Reliability for Human-Scored Items

PIRLS 2021 Within-Country Scoring Reliability for Human-Scored Items

Country	PIRLS			Bridge		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement Across Items		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement Across Items	
		Minimum	Maximum		Minimum	Maximum
Australia	93	78	100	–	–	–
Austria	95	75	100	–	–	–
Azerbaijan	99	95	100	–	–	–
Bahrain	100	98	100	–	–	–
Belgium (Flemish)	93	78	100	93	75	100
Belgium (French)	94	81	100	–	–	–
Bulgaria	98	92	100	–	–	–
Chinese Taipei	99	94	100	100	99	100
Croatia	94	84	100	98	93	100
Cyprus	93	84	100	–	–	–
Czech Republic	96	90	100	97	90	100
Denmark	91	73	100	89	74	100
Egypt	98	91	100	–	–	–
England	95	88	100	–	–	–
Finland	99	96	100	100	99	100
France	94	69	100	–	–	–
Georgia	99	95	100	–	–	–
Germany	92	80	100	93	80	100
Hong Kong SAR	100	96	100	–	–	–
Hungary	92	76	100	94	86	100
Iran, Islamic Rep. of	95	87	100	–	–	–
Ireland	99	93	100	–	–	–
Israel	95	88	100	100	100	100
Italy	96	86	100	95	87	100
Jordan	99	94	100	–	–	–

PIRLS 2021 Within-Country Scoring Reliability for Human-Scored Items (Continued)

Country	PIRLS			Bridge		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement Across Items		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement Across Items	
		Minimum	Maximum		Minimum	Maximum
Kazakhstan	90	75	100	90	76	100
Latvia	96	83	100	–	–	–
Lithuania	97	91	100	97	91	100
Macao SAR	98	92	100	–	–	–
Malta	88	73	100	89	63	100
Montenegro	99	95	100	–	–	–
Morocco	90	67	100	–	–	–
Netherlands	94	76	100	–	–	–
New Zealand	95	87	100	94	83	100
North Macedonia	88	69	100	–	–	–
Norway (5)	96	86	100	95	87	100
Oman	93	77	100	–	–	–
Poland	98	88	100	–	–	–
Portugal	97	90	100	98	95	100
Qatar	92	81	100	93	80	100
Russian Federation	93	85	100	94	85	100
Saudi Arabia	88	62	100	92	79	100
Serbia	98	87	100	–	–	–
Singapore	98	91	100	100	98	100
Slovak Republic	92	77	100	99	92	100
Slovenia	95	90	100	94	85	100
South Africa	92	76	99	–	–	–
Spain	93	86	100	99	95	100
Sweden	94	83	100	93	85	100
Turkiye	96	86	100	–	–	–
United Arab Emirates	93	82	100	90	72	99
United States*	–	–	–	96	84	100
Uzbekistan	95	83	100	–	–	–
International Average	**95**	**84**	**100**	**95**	**86**	**100**

PIRLS 2021 Within-Country Scoring Reliability for Human-Scored Items (Continued)

Country	PIRLS			Bridge		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement Across Items		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement Across Items	
		Minimum	Maximum		Minimum	Maximum
Benchmarking Participants						
Alberta, Canada	91	69	100	–	–	–
British Columbia, Canada	91	74	100	–	–	–
Newfoundland & Labrador, Can.	91	68	100	–	–	–
Quebec, Canada	91	67	100	–	–	–
Moscow City, Russian Fed.	96	89	100	99	96	100
South Africa (6)	90	66	99	–	–	–

A dash (–) indicates comparable data not available.

Albania, Brazil, Kosovo, and Northern Ireland were excluded from this analysis due to concerns with their reliability data.

* The United States administered the PIRLS 2021 digital assessment and the PIRLS 2021 paper bridge assessment. The United States opted to report the paper bridge results.

Appendix 9B: PIRLS 2021 Trend Scoring Reliability for Human-Scored Items

PIRLS 2021 Trend Scoring Reliability for Human-Scored Items

Country	PIRLS		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement Across Items	
		Minimum	Maximum
Australia	93	69	100
Austria	96	86	100
Azerbaijan	90	68	100
Bahrain	93	75	100
Belgium (Flemish)	94	76	100
Belgium (French)	93	76	100
Bulgaria	97	90	100
Chinese Taipei	96	83	100
Czech Republic	95	79	100
Denmark	93	80	100
England	94	76	100
Finland	97	89	100
France	94	75	100
Georgia	92	68	100
Germany	96	82	100
Hong Kong SAR	97	82	100
Hungary	95	82	100
Iran, Islamic Rep. of	92	74	100
Ireland	96	85	100
Israel	93	76	99
Italy	93	83	100
Kazakhstan	84	49	100
Latvia	94	79	100
Lithuania	95	78	100
Macao SAR	96	84	100
Netherlands	94	76	100
New Zealand	94	83	100

PIRLS 2021 Trend Scoring Reliability for Human-Scored Items (Continued)

Country	PIRLS		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement Across Items	
		Minimum	Maximum
Northern Ireland	97	89	100
Norway	95	83	100
Oman	91	65	100
Poland	93	73	100
Portugal	93	70	100
Qatar	89	59	100
Russian Federation	96	84	100
Saudi Arabia	89	64	99
Singapore	97	84	100
Slovak Republic	95	76	100
Spain	91	72	100
Sweden	96	79	100
United Arab Emirates	93	80	100
United States	94	84	100
International Average	**94**	**77**	**100**
Benchmarking Participant			
Moscow City, Russian Fed.	98	91	100

Morocco and Slovenia did not participate in the trend reliability scoring procedures.

Appendix 9C: PIRLS 2021 Cross-Country Scoring Reliability for Human-Scored Items

PIRLS 2021 Cross-Country Scoring Reliability for Human-Scored Items

Item	Total Valid Comparisons	Percent Exact Agreeement
The Empty Pot – RP31M02	286,147	96.5
The Empty Pot – RP31M04	286,041	85.0
The Empty Pot – RP31M09	286,094	88.5
The Empty Pot – RP31M10	286,094	92.9
Where's the Honey? – RP31M16	286,094	93.8
Where's the Honey? – RP31W01	286,147	94.8
Where's the Honey? – RP31W02	286,147	80.6
Where's the Honey? – RP31W04	286,147	98.1
Where's the Honey? – RP31W11	286,041	97.8
Where's the Honey? – RP31W13	286,094	83.9
How Did We Learn to Fly? – RP41E01	286,200	99.5
How Did We Learn to Fly? – RP41E02	286,200	99.6
How Did We Learn to Fly? – RP41E07	286,147	86.8
How Did We Learn to Fly? – RP41E10	286,200	96.9
How Did We Learn to Fly? – RP41E12	286,147	99.7
How Did We Learn to Fly? – RP41E13	285,938	93.8
How Did We Learn to Fly? – RP41E14	286,147	76.8
How Did We Learn to Fly? – RP41E15	286,200	99.5
Average Percent Agreement		**92.5**

Egypt, Morocco, Serbia, and Kosovo did not participate in the cross-country reliability scoring procedures.

Appendix 9D: PIRLS 2021 Item Statistics by Booklet Position

PIRLS 2021 International Average Text Statistics by Booklet Position— paperPIRLS, Part 1/2

Text		Sample Sizes		Percent Correct		
		Position 1	Position 2	Position 1	Position 2	Difference
Literary Texts						
Difficult	Shiny Straw	7,696	7,567	53.3	51.2	-2.1
	[2] Oliver and the Griffin	7,615	7,612	49.7	49.9	0.2
	The Ink Drinker	7,667	7,570	50.5	48.8	-1.7
Medium	Pemba Sherpa	7,768	8,305	70.5	69.4	-1.1
	The Ostrich and the Hat	7,653	8,412	57.1	56.0	-1.1
	The Empty Pot	7,687	8,438	62.7	61.8	-0.9
Easy	[1] Learning a New Language	8,481	8,460	68.8	67.8	-1.0
	[1] The Summer My Father Was 10	8,483	8,544	78.8	78.6	-0.2
	Library Mouse	8,455	8,383	78.2	76.8	-1.4
Informational Texts						
Difficult	Where's the Honey?	7,701	7,624	50.0	47.6	-2.4
	[2] Icelandic Horses	7,536	7,515	46.2	44.9	-1.3
	[2] The World's Bank of Seeds	7,551	7,561	43.0	44.4	1.4
Medium	How Did We Learn to Fly?	7,677	8,291	73.3	70.8	-2.5
	Marie Curie	7,757	8,414	51.9	50.8	-1.0
	Sharks	7,711	8,186	53.7	51.9	-1.8
Easy	[1] The Amazing Octopus	8,498	8,452	66.4	67.5	1.1
	Training a Deaf Polar Bear	8,577	8,351	76.9	74.3	-2.7
	Hungry Plant	8,533	8,418	80.2	79.2	-1.0
Overall		**143,046**	**146,103**	**61.7**	**60.6**	**-1.1**

[1] Easy texts that always appear in position 1. Position 1 statistics refer to the text when paired with another easy text; position 2 statistics refer to the text when paired with a medium text.

[2] Difficult texts that always appear in position 2. Position 1 statistics refer to the text when paired with a medium text; position 2 statistics refer to the text when paired with another difficult text.

PIRLS 2021 International Average Text Statistics by Booklet Position— paperPIRLS Part 2/2

Text		Percent Omitted Responses			Percent Not-Reached Responses		
		Position 1	Position 2	Difference	Position 1	Position 2	Difference
Literary Texts							
Difficult	Shiny Straw	6.5	7.1	0.6	4.0	3.4	-0.6
	[2] Oliver and the Griffin	12.2	11.3	-0.9	3.3	3.0	-0.3
	The Ink Drinker	10.2	10.9	0.7	5.1	3.9	-1.1
Medium	Pemba Sherpa	3.9	4.8	0.9	3.4	3.5	0.2
	The Ostrich and the Hat	7.3	6.9	-0.4	4.5	4.0	-0.5
	The Empty Pot	5.3	5.1	-0.2	3.6	2.8	-0.8
Easy	[1] Learning a New Language	3.9	4.3	0.4	2.7	3.0	0.2
	[1] The Summer My Father Was 10	2.9	3.1	0.2	0.9	1.0	0.1
	Library Mouse	3.1	4.0	0.9	1.6	1.7	0.1
Informational Texts							
Difficult	Where's the Honey?	9.1	9.9	0.9	3.3	2.4	-0.9
	[2] Icelandic Horses	12.8	12.3	-0.5	5.7	5.8	0.1
	[2] The World's Bank of Seeds	8.3	8.1	-0.2	7.1	6.9	-0.2
Medium	How Did We Learn to Fly?	4.6	5.3	0.7	1.8	1.7	-0.1
	Marie Curie	8.6	8.1	-0.5	5.7	4.2	-1.5
	Sharks	9.4	9.6	0.2	4.4	2.7	-1.7
Easy	[1] The Amazing Octopus	7.2	7.3	0.0	4.3	4.4	0.1
	Training a Deaf Polar Bear	3.1	3.4	0.3	1.5	1.4	-0.1
	Hungry Plant	2.6	2.5	-0.1	0.7	0.6	-0.1
Overall		**6.7**	**6.9**	**0.2**	**3.5**	**3.1**	**-0.4**

[1] Easy texts that always appear in position 1. Position 1 statistics refer to the text when paired with another easy text; position 2 statistics refer to the text when paired with a medium text.

[2] Difficult texts that always appear in position 2. Position 1 statistics refer to the text when paired with a medium text; position 2 statistics refer to the text when paired with another difficult text.

PIRLS 2021 International Average Text Statistics by Booklet Position—digitalPIRLS with ePIRLS, Part 1/2

Text		Sample Sizes		Percent Correct		
		Position 1	Position 2	Position 1	Position 2	Difference
Literary Texts						
Difficult	Shiny Straw	6,004	6,019	58.8	58.3	-0.5
	[2] Oliver and the Griffin	5,982	6,024	52.7	53.0	0.2
	The Ink Drinker	6,037	6,007	54.1	53.8	-0.3
Medium	Pemba Sherpa	6,084	5,670	68.9	67.5	-1.4
	The Ostrich and the Hat	6,003	5,617	61.0	62.1	1.1
	The Empty Pot	6,072	5,636	66.0	66.1	0.1
Easy	[1] Learning a New Language	5,669	5,651	71.9	71.3	-0.6
	[1] The Summer My Father Was 10	5,724	5,666	75.9	75.6	-0.2
	Library Mouse	5,713	5,683	77.8	77.9	0.2
Informational Texts						
Difficult	Where's the Honey?	8,014	6,064	49.8	50.1	0.3
	[2] Icelandic Horses	7,929	6,029	51.6	51.7	0.1
	[2] The World's Bank of Seeds	7,982	5,997	47.7	47.5	-0.2
Medium	How Did We Learn to Fly?	7,959	5,646	68.3	67.1	-1.1
	Marie Curie	7,910	5,654	53.1	54.1	1.1
	Sharks	8,009	5,696	54.3	55.1	0.8
Easy	[1] The Amazing Octopus	7,616	5,674	69.1	69.4	0.2
	Training a Deaf Polar Bear	7,577	5,660	69.0	70.2	1.3
	Hungry Plant	7,579	5,719	76.1	78.5	2.4
Overall (digitalPIRLS)		**123,863**	**104,112**	**62.6**	**62.7**	**0.2**
ePIRLS Tasks						
Rainforests		7,046	10,540	57.1	55.8	-1.3
The Legend of Troy		7,026	10,538	62.6	60.5	-2.1
Zebra and Wildebeest Migration		7,031	10,471	60.5	58.5	-2.0
Oceans		6,997	10,505	60.9	59.0	-1.9
Voyages of Discovery		7,025	10,436	48.7	46.3	-2.4
Overall (ePIRLS)		**35,125**	**52,490**	**58.0**	**56.0**	**-2.0**

[1] Easy texts that always appear in position 1. Position 1 statistics refer to the text when paired with another easy text; position 2 statistics refer to the text when paired with a medium text.

[2] Difficult texts that always appear in position 2. Position 1 statistics refer to the text when paired with a medium text; position 2 statistics refer to the text when paired with another difficult text.

PIRLS 2021 International Average Text Statistics by Booklet Position—digitalPIRLS with ePIRLS, Part 2/2

Text		Percent Omitted Responses			Percent Not-Reached Responses		
		Position 1	Position 2	Difference	Position 1	Position 2	Difference
Literary Texts							
Difficult	Shiny Straw	3.8	3.5	-0.4	1.2	0.6	-0.7
	[2] Oliver and the Griffin	7.0	6.0	-0.9	1.2	0.9	-0.3
	The Ink Drinker	6.9	6.8	-0.1	3.0	1.1	-1.9
Medium	Pemba Sherpa	2.5	2.5	0.0	1.9	1.3	-0.5
	The Ostrich and the Hat	3.7	3.5	-0.1	1.7	1.1	-0.6
	The Empty Pot	2.5	2.0	-0.5	1.6	1.0	-0.6
Easy	[1] Learning a New Language	1.8	1.8	0.0	0.9	0.9	-0.1
	[1] The Summer My Father Was 10	1.6	1.8	0.2	0.5	0.5	0.1
	Library Mouse	1.3	1.3	0.0	0.7	0.3	-0.4
Informational Texts							
Difficult	Where's the Honey?	5.8	5.7	-0.1	1.9	0.9	-1.0
	[2] Icelandic Horses	7.1	6.6	-0.4	2.3	1.8	-0.5
	[2] The World's Bank of Seeds	5.6	5.4	-0.2	3.1	2.8	-0.4
Medium	How Did We Learn to Fly?	3.2	2.7	-0.5	1.1	0.4	-0.6
	Marie Curie	6.8	5.5	-1.4	3.2	1.7	-1.4
	Sharks	5.9	4.7	-1.2	2.2	0.7	-1.5
Easy	[1] The Amazing Octopus	3.8	3.7	-0.1	2.1	2.0	-0.1
	Training a Deaf Polar Bear	2.3	1.7	-0.6	0.8	0.3	-0.5
	Hungry Plant	1.0	0.8	-0.1	0.3	0.2	-0.2
Overall (digitalPIRLS)		**4.0**	**3.7**	**-0.4**	**1.6**	**1.0**	**-0.6**
ePIRLS Tasks							
Rainforests		4.1	3.5	-0.6	1.9	0.7	-1.1
The Legend of Troy		5.2	4.3	-0.9	3.2	1.2	-2.0
Zebra and Wildebeest Migration		5.1	4.4	-0.7	3.2	1.3	-1.9
Oceans		4.7	4.1	-0.6	2.7	0.9	-1.8
Voyages of Discovery		6.9	6.1	-0.8	4.1	1.8	-2.3
Overall (ePIRLS)		**5.2**	**4.5**	**-0.7**	**3.0**	**1.2**	**-1.8**

[1] Easy texts that always appear in position 1. Position 1 statistics refer to the text when paired with another easy text; position 2 statistics refer to the text when paired with a medium text.

[2] Difficult texts that always appear in position 2. Position 1 statistics refer to the text when paired with a medium text; position 2 statistics refer to the text when paired with another difficult text.

PIRLS 2021 International Average Text Statistics by Booklet Position— Bridge, Part 1/2

Text	Sample Sizes		Percent Correct		
	Position 1	Position 2	Position 1	Position 2	Difference
Literary Texts					
Shiny Straw	5,700	5,687	61.3	60.2	-1.2
Oliver and the Griffin	5,689	5,702	60.1	58.5	-1.5
[1] The Ink Drinker	5,756	5,687	77.6	77.6	0.0
Pemba Sherpa	5,716	5,693	69.6	69.1	-0.5
Informational Texts					
Where's the Honey?	5,706	5,656	57.7	55.3	-2.4
[2] Icelandic Horses	5,663	5,724	55.6	54.8	-0.8
[1] The World's Bank of Seeds	5,709	5,729	79.5	79.5	0.0
[2] How Did We Learn to Fly?	5,689	5,670	59.6	59.2	-0.5
Overall	**45,628**	**45,548**	**65.1**	**64.3**	**-0.9**

[1] Texts that always appear in position 1. Position 1 statistics refer to the booklet in which the text has the higher percent correct statistic.

[2] Texts that always appear in position 2. Position 1 statistics refer to the booklet in which the text has the higher percent correct statistic.

PIRLS 2021 International Average Text Statistics by Booklet Position— Bridge, Part 2/2

Text	Percent Omitted Responses			Percent Not-Reached Responses		
	Position 1	Position 2	Difference	Position 1	Position 2	Difference
Literary Texts						
Shiny Straw	4.1	4.6	0.5	1.8	1.3	-0.5
Oliver and the Griffin	6.5	6.5	0.0	2.7	2.0	-0.7
[1] The Ink Drinker	2.3	2.1	-0.2	1.6	1.3	-0.2
Pemba Sherpa	3.0	3.2	0.3	1.7	1.5	-0.2
Informational Texts						
Where's the Honey?	5.8	6.2	0.4	1.8	1.5	-0.3
[2] Icelandic Horses	9.0	8.5	-0.6	2.5	2.8	0.3
[1] The World's Bank of Seeds	2.5	2.5	-0.1	0.7	0.9	0.3
[2] How Did We Learn to Fly?	6.6	6.9	0.3	1.4	1.4	0.0
Overall	**5.0**	**5.1**	**0.1**	**1.8**	**1.6**	**-0.2**

[1] Texts that always appear in position 1. Position 1 statistics refer to the booklet in which the text has the higher percent correct statistic.

[2] Texts that always appear in position 2. Position 1 statistics refer to the booklet in which the text has the higher percent correct statistic.

PIRLS 2021 Country Average Item Statistics by Booklet Position— paperPIRLS

Country	Sample Sizes		Average Percent Correct Across Items		Average Percent Omitted Responses Across Items		Average Percent Not Reached Responses Across Items	
	Position 1	Position 2	Position 1	Position 2	Position 1	Position 2	Position 1	Position 2
Albania	4,197	4,216	73.8	71.2	6.3	6.1	0.5	0.8
Australia	5,503	5,398	72.0	70.5	2.5	2.9	1.2	1.3
Austria	4,854	4,743	70.2	69.4	5.6	6.1	0.9	0.9
Azerbaijan	5,188	5,099	52.2	50.2	11.1	12.3	3.7	4.8
Bahrain	5,186	5,108	56.3	55.3	7.6	7.7	4.4	3.9
Belgium (French)	4,250	4,283	63.6	61.3	7.6	7.6	3.4	2.6
Brazil	4,870	4,787	50.8	49.0	11.1	12.0	7.1	6.5
Bulgaria	4,063	4,002	72.0	71.1	4.5	4.3	1.1	1.1
Cyprus	4,629	4,509	67.4	65.2	6.1	6.5	3.5	2.8
Egypt	6,704	8,165	43.7	43.8	19.1	21.0	11.6	11.5
England	4,128	4,122	75.0	74.1	2.7	3.3	0.7	0.8
France	5,376	5,267	67.8	66.7	7.6	8.1	2.6	2.1
Georgia	5,276	5,093	63.4	62.5	6.8	7.1	2.9	2.9
Hong Kong SAR	4,340	3,298	78.5	78.0	3.6	3.9	0.7	0.7
Iran, Islamic Rep. of	5,129	6,627	47.2	46.9	13.2	13.5	8.1	7.1
Ireland	5,283	4,026	78.1	77.8	1.6	1.6	0.6	0.8
Jordan	5,225	6,648	43.9	43.2	15.9	15.2	11.9	10.4
Kosovo	4,580	4,472	45.9	45.1	7.0	7.2	3.7	3.4
Latvia	4,377	4,322	69.5	69.0	4.5	4.5	1.4	1.5
Macao SAR	5,127	5,057	71.0	69.4	3.4	3.4	0.8	0.6
Montenegro	4,517	4,401	62.8	61.4	9.4	9.7	5.0	4.0
Morocco	6,070	7,834	37.5	37.3	9.6	9.3	7.5	4.7
Netherlands	4,345	4,244	69.7	67.5	3.3	3.8	1.2	1.3
North Macedonia	2,930	2,861	52.6	50.8	9.4	9.6	4.3	4.6
Northern Ireland	4,586	3,498	76.2	75.7	2.2	2.3	0.6	0.4
Oman	4,587	5,908	51.8	50.8	8.1	8.0	7.4	6.3
Poland	4,741	3,588	74.6	74.6	5.7	6.1	0.8	0.9
Serbia	4,053	4,003	67.1	66.8	6.9	6.5	2.1	1.7
South Africa	10,578	13,631	27.3	26.2	9.5	9.2	10.2	8.3

PIRLS 2021 Country Average Item Statistics by Booklet Position— paperPIRLS (Continued)

Country	Sample Sizes		Average Percent Correct Across Items		Average Percent Omitted Responses Across Items		Average Percent Not Reached Responses Across Items	
	Position 1	Position 2	Position 1	Position 2	Position 1	Position 2	Position 1	Position 2
Turkiye	6,015	6,034	63.0	60.4	3.7	3.7	1.3	1.1
Uzbekistan	5,851	5,781	49.3	48.5	6.3	5.4	6.1	4.1
International Average	**156,553**	**161,020**	**60.8**	**59.8**	**7.2**	**7.4**	**3.8**	**3.4**
Benchmarking Participant								
South Africa (6)	8,082	10,380	40.0	38.7	3.2	3.1	2.0	2.3

PIRLS 2021 Country Average Item Statistics by Booklet Position — digitalPIRLS and ePIRLS

Country	Sample Sizes		Average Percent Correct Across Items		Average Percent Omitted Responses Across Items		Average Percent Not Reached Responses Across Items	
	Position 1	Position 2	Position 1	Position 2	Position 1	Position 2	Position 1	Position 2
Belgium (Flemish)	5,138	5,061	56.5	56.2	4.9	4.8	1.5	1.0
Chinese Taipei	5,578	5,521	64.1	63.9	4.1	3.8	0.8	0.5
Croatia	3,954	3,909	68.4	68.0	3.1	3.2	0.7	0.4
Czech Republic	6,673	6,535	64.1	64.7	5.5	4.7	1.5	0.8
Denmark	4,840	4,739	64.1	64.3	4.4	3.4	2.7	1.4
Finland	7,639	6,369	66.4	67.2	4.8	4.2	1.0	0.6
Germany	4,620	4,570	61.3	60.8	7.0	6.9	2.3	1.4
Hungary	5,344	5,276	63.4	63.6	3.1	2.6	1.0	0.5
Israel	4,916	4,845	58.4	57.5	6.0	5.7	4.6	2.6
Italy	5,484	5,390	63.6	63.9	5.3	4.6	2.1	0.8
Kazakhstan	7,056	6,984	53.8	53.9	2.5	1.8	1.8	0.7
Lithuania	4,656	4,583	66.8	66.1	2.7	3.1	0.5	0.4
Malta	3,056	2,990	57.7	57.6	3.6	3.5	1.6	0.9
New Zealand	5,591	5,454	60.4	59.3	4.8	5.2	1.9	1.3
Norway (5)	5,407	5,305	64.4	63.2	4.1	3.7	1.5	0.9
Portugal	6,099	6,097	59.2	58.8	4.6	3.9	2.5	1.2
Qatar	5,260	5,220	51.9	50.8	5.5	4.5	4.7	2.4
Russian Federation	5,681	4,750	70.8	69.7	2.6	1.8	1.4	0.5
Saudi Arabia	4,357	5,154	42.7	42.7	5.6	3.9	6.2	2.7
Singapore	7,286	6,151	73.8	73.1	0.7	0.8	0.1	0.2
Slovak Republic	4,863	4,784	61.5	62.1	4.7	3.7	1.3	0.8
Slovenia	5,135	5,065	59.5	58.8	5.8	5.4	1.7	1.1
Spain	8,599	8,471	59.8	59.1	5.7	5.4	2.7	1.3
Sweden	5,187	5,130	65.3	65.0	4.6	4.3	1.7	1.2
United Arab Emirates	26,569	28,250	52.8	51.3	3.9	3.4	2.8	1.4
International Average	**158,988**	**156,603**	**61.2**	**60.9**	**4.4**	**3.9**	**2.0**	**1.1**
Benchmarking Particpants								
Alberta, Canada	3,027	2,991	63.7	63.5	3.1	3.3	1.5	1.0
British Columbia, Canada	4,730	4,582	63.4	62.2	3.7	3.4	1.6	1.1
Newfoundland & Labrador, Can.	2,445	2,419	61.9	60.4	5.5	4.9	3.8	2.2

PIRLS 2021 Country Average Item Statistics by Booklet Position— digitalPIRLS and ePIRLS (Continued)

Country	Sample Sizes		Average Percent Correct Across Items		Average Percent Omitted Responses Across Items		Average Percent Not Reached Responses Across Items	
	Position 1	Position 2	Position 1	Position 2	Position 1	Position 2	Position 1	Position 2
Quebec, Canada	3,762	3,695	67.8	66.2	3.3	3.0	1.8	1.1
Moscow City, Russian Fed.	6,267	5,221	77.4	76.8	1.3	1.1	0.4	0.2
Abu Dhabi, UAE	9,429	11,296	44.3	43.8	4.8	4.3	2.9	1.6
Dubai, UAE	7,727	7,676	66.4	66.1	2.3	2.0	1.6	0.9

PIRLS 2021 Country Average Item Statistics by Booklet Position— Bridge

Country	Sample Sizes		Average Percent Correct Across Items		Average Percent Omitted Responses Across Items		Average Percent Not Reached Responses Across Items	
	Position 1	Position 2	Position 1	Position 2	Position 1	Position 2	Position 1	Position 2
Belgium (Flemish)	1,607	1,628	59.8	58.6	4.5	4.9	1.9	1.1
Chinese Taipei	1,671	1,666	72.0	71.1	2.9	3.8	0.5	0.5
Croatia	1,231	1,212	71.5	71.3	4.1	3.9	1.0	0.4
Czech Republic	1,886	1,918	70.7	69.7	6.6	5.8	1.9	1.3
Denmark	1,411	1,388	68.9	67.5	4.5	5.1	2.1	1.9
Finland	2,055	2,063	69.4	69.1	4.5	4.4	1.1	1.0
Germany	1,339	1,337	67.6	67.3	7.8	8.1	1.6	1.9
Hungary	1,681	1,698	70.0	68.1	4.5	4.1	1.8	1.8
Israel	1,783	1,731	64.9	62.8	7.1	7.1	4.9	4.7
Italy	1,971	1,978	69.4	68.8	3.7	4.0	0.7	0.9
Kazakhstan	3,190	3,209	58.8	58.9	2.7	2.1	1.1	0.9
Lithuania	1,516	1,507	68.7	67.4	4.3	4.6	0.9	0.7
Malta	826	838	59.5	58.1	5.2	5.8	1.3	1.3
New Zealand	2,204	2,213	63.0	61.7	3.7	3.8	1.5	1.6
Norway (5)	1,655	1,673	65.1	64.9	5.8	5.5	2.1	1.9
Portugal	2,111	2,075	66.1	66.2	5.5	5.1	2.3	1.7
Qatar	1,336	1,331	57.1	54.9	5.6	5.7	2.8	2.9
Russian Federation	2,195	2,178	75.3	75.1	2.3	2.4	0.2	0.4
Saudi Arabia	1,864	1,856	41.7	41.4	8.3	7.8	5.9	5.4
Singapore	1,991	1,983	76.2	75.0	1.3	1.6	0.2	0.2
Slovak Republic	1,645	1,628	65.7	66.0	5.2	5.3	1.3	1.0
Slovenia	1,411	1,400	64.7	63.4	7.1	7.1	1.2	1.3
Spain	1,566	1,563	62.7	62.3	6.8	7.1	1.9	2.1
Sweden	1,833	1,864	67.8	67.2	5.6	5.6	2.1	2.1
United Arab Emirates	2,002	1,959	55.0	54.6	5.0	5.3	2.3	2.0
United States	1,648	1,652	70.4	68.7	3.3	3.2	1.8	1.1
International Average	**45,628**	**45,548**	**65.5**	**64.6**	**4.9**	**5.0**	**1.8**	**1.6**
Benchmarking Participant								
Moscow City, Russian Fed.	1,701	1,688	81.6	82.1	1.8	1.7	0.4	0.3

Appendix 9E: PIRLS 2021 Group Adaptive Design Outcomes

PIRLS 2021 Average Percent of Item Non-Response and Posterior Standard Deviations by Booklet Difficulty

Country	Booklet Design	Average Percent of Item Non–Response				Posterior Standard Deviation			
		Overall	More Difficult Booklet Average	Less Difficult Booklet Average	Difference	Overall	More Difficult Booklet Average	Less Difficult Booklet Average	Difference
Albania	50/50	5.9	7.7	3.8	–3.9	0.34	0.32	0.36	+0.04
Australia	50/50	3.5	4.9	2.2	–2.7	0.28	0.26	0.31	+0.05
Austria	50/50	6.1	8.2	4.0	–4.2	0.24	0.22	0.26	+0.03
Azerbaijan	50/50	13.8	16.8	10.8	–6.0	0.29	0.29	0.28	–0.01
Bahrain	50/50	10.3	13.2	7.4	–5.8	0.30	0.30	0.30	+0.00
Belgium (Flemish)	50/50	4.8	6.8	2.8	–4.0	0.23	0.23	0.23	+0.01
Belgium (French)	50/50	9.2	12.9	5.7	–7.1	0.25	0.24	0.26	+0.02
Brazil	50/50	16.3	20.1	12.7	–7.4	0.33	0.34	0.33	–0.01
Bulgaria	50/50	4.8	6.3	3.2	–3.1	0.28	0.26	0.31	+0.05
Chinese Taipei	50/50	3.7	5.2	2.2	–3.0	0.24	0.23	0.25	+0.02
Croatia	50/50	3.0	4.5	1.4	–3.2	0.24	0.22	0.25	+0.03
Cyprus	50/50	8.1	11.7	4.5	–7.1	0.26	0.25	0.28	+0.03
Czech Republic	50/50	5.1	7.1	3.1	–4.0	0.24	0.23	0.25	+0.02
Denmark	50/50	4.8	6.7	2.9	–3.8	0.24	0.23	0.25	+0.02
Egypt	30/70	26.5	32.2	24.3	–8.0	0.38	0.42	0.36	–0.05
England	50/50	3.5	4.5	2.5	–2.0	0.28	0.25	0.31	+0.06
Finland	70/30	5.1	6.1	2.8	–3.3	0.24	0.23	0.26	+0.02
France	50/50	9.1	12.3	5.9	–6.3	0.25	0.23	0.26	+0.03
Georgia	50/50	8.8	11.2	6.5	–4.7	0.27	0.25	0.28	+0.03
Germany	50/50	7.0	9.5	4.5	–5.0	0.25	0.24	0.25	+0.01
Hong Kong SAR	70/30	4.3	5.5	1.9	–3.6	0.25	0.24	0.29	+0.05
Hungary	50/50	2.7	3.8	1.6	–2.3	0.25	0.23	0.26	+0.02
Iran, Islamic Rep. of	30/70	17.4	22.8	15.3	–7.5	0.30	0.32	0.30	–0.02
Ireland	70/30	2.3	2.7	1.4	–1.3	0.27	0.25	0.31	+0.06
Israel	50/50	7.6	10.2	5.0	–5.1	0.26	0.26	0.26	+0.01
Italy	50/50	5.1	7.1	3.0	–4.0	0.23	0.22	0.24	+0.02
Jordan	30/70	21.9	29.2	18.9	–10.3	0.36	0.39	0.34	–0.05

PIRLS 2021 Average Percent of Item Non-Response and Posterior Standard Deviations by Booklet Difficulty (Continued)

Country	Booklet Design	Average Percent of Item Non–Response				Posterior Standard Deviation			
		Overall	More Difficult Booklet Average	Less Difficult Booklet Average	Difference	Overall	More Difficult Booklet Average	Less Difficult Booklet Average	Difference
Kazakhstan	50/50	2.4	3.4	1.5	–1.9	0.24	0.23	0.24	+0.01
Kosovo	50/50	9.6	11.9	7.3	–4.6	0.28	0.29	0.27	–0.02
Latvia	50/50	5.4	7.0	3.8	–3.2	0.26	0.24	0.28	+0.04
Lithuania	50/50	2.6	3.4	1.7	–1.7	0.24	0.23	0.25	+0.03
Macao SAR	50/50	3.4	5.0	1.9	–3.1	0.26	0.24	0.28	+0.04
Malta	50/50	3.9	5.2	2.6	–2.6	0.25	0.25	0.26	+0.01
Montenegro	50/50	12.6	16.6	8.4	–8.2	0.26	0.25	0.27	+0.02
Morocco	30/70	13.0	16.5	11.5	–4.9	0.33	0.37	0.32	–0.05
Netherlands	50/50	4.2	5.9	2.5	–3.4	0.25	0.23	0.27	+0.04
New Zealand*	50/50, 30/70	5.6	6.9	4.4	–2.5	0.26	0.25	0.27	+0.01
North Macedonia	50/50	12.3	15.4	9.2	–6.2	0.28	0.28	0.28	–0.00
Northern Ireland	70/30	2.8	3.4	1.5	–1.9	0.27	0.25	0.31	+0.06
Norway (5)	50/50	4.1	5.7	2.4	–3.3	0.25	0.24	0.26	+0.02
Oman	30/70	11.8	16.9	9.7	–7.2	0.32	0.32	0.31	–0.01
Poland	70/30	6.9	8.5	3.4	–5.1	0.25	0.24	0.29	+0.05
Portugal	50/50	4.9	6.8	3.0	–3.8	0.23	0.23	0.24	+0.01
Qatar	50/50	6.6	8.4	4.8	–3.7	0.27	0.27	0.27	+0.00
Russian Federation	70/30	2.9	3.7	1.2	–2.4	0.24	0.23	0.26	+0.03
Saudi Arabia	30/70	6.1	8.3	5.2	–3.1	0.27	0.28	0.27	–0.01
Serbia	50/50	7.6	9.9	5.4	–4.5	0.25	0.23	0.27	+0.04
Singapore	70/30	0.8	1.0	0.5	–0.5	0.26	0.25	0.29	+0.04
Slovak Republic	50/50	4.3	5.8	2.7	–3.1	0.24	0.23	0.25	+0.02
Slovenia	50/50	5.6	7.6	3.4	–4.2	0.23	0.23	0.24	+0.01
South Africa	30/70	16.8	20.1	15.5	–4.6	0.42	0.47	0.39	–0.08
Spain	50/50	6.2	8.8	3.6	–5.2	0.23	0.23	0.24	+0.01
Sweden	50/50	4.5	6.3	2.7	–3.6	0.25	0.24	0.26	+0.02
Turkiye	50/50	4.2	6.0	2.4	–3.6	0.27	0.25	0.28	+0.03
United Arab Emirates	50/50	4.5	5.9	3.5	–2.4	0.30	0.29	0.31	+0.02

PIRLS 2021 Average Percent of Item Non-Response and Posterior Standard Deviations by Booklet Difficulty (Continued)

Country	Booklet Design	Average Percent of Item Non–Response				Posterior Standard Deviation			
		Overall	More Difficult Booklet Average	Less Difficult Booklet Average	Difference	Overall	More Difficult Booklet Average	Less Difficult Booklet Average	Difference
Uzbekistan	50/50	9.5	12.7	6.4	–6.3	0.28	0.28	0.28	–0.01
International Average		**7.2**	**9.5**	**5.2**	**–4.3**	**0.27**	**0.26**	**0.28**	**+0.02**
Benchmarking Participants									
Alberta, Canada	50/50	3.6	4.6	2.5	–2.1	0.25	0.24	0.26	+0.02
British Columbia, Canada	50/50	3.8	5.2	2.4	–2.8	0.25	0.24	0.26	+0.02
Newfoundland & Labrador, Canada	50/50	6.6	9.0	4.2	–4.8	0.25	0.25	0.26	+0.01
Quebec, Canada	50/50	3.6	5.1	2.1	–3.0	0.23	0.22	0.25	+0.02
Moscow City, Russian Federation	70/30	1.4	1.8	0.5	–1.4	0.24	0.23	0.27	+0.04
South Africa (6)	30/70	4.4	6.2	3.6	–2.6	0.33	0.35	0.32	–0.03
Abu Dhabi, UAE	30/70	5.0	6.8	4.2	–2.6	0.34	0.34	0.34	–0.01
Dubai, UAE	50/50	2.7	3.6	1.8	–1.8	0.27	0.25	0.28	+0.03

* New Zealand used a 30/70 rotation for the Maori-medium stratum.

The United States was excluded from this analysis.

Appendix 9F: Modifications to the PIRLS 2021 Achievement Data

PIRLS 2021
Items Excluded from Scaling for All Countries*
Sharks Item 4 – RE21K04, RP21K04 (attractive distracter)
Items Deleted for All Countries
Learning a New Language Item 12 – RE51R12, RP51R12 (attractive distracter)
The Legend of Troy Item 7 – E041T07 (attractive distracter)
Voyages of Discovery Item 17C – E051V17C (attractive distracter)
Voyages of Discovery Item 19B – E051V19B (severe differential item functioning)
Items Recoded for All Countries
Ostrich and the Hat Item 12 – RE51T12, RP51T12 (2 to 1)
Icelandic Horses Item 15 – RE41I15, RP41I15 (2 to 1)
Zebra and Wildebeest Migration Item 12 – E041Z12 (2 to 1)

Items Deleted by Country
Belgium (Flemish) Shiny Straw Item 1 – RE21Y01 (translation error)
Belgium (French) Learning a New Language Item 16 – RP51R16 (low discrimination)
Bulgaria Sharks Item 11 – RP21K11 (low discrimination) Training a Deaf Polar Bear Item 14 – RP31P14 (severe differential item functioning)
Chinese Taipei Shiny Straw Item 3 – RP21Y03 (low discrimination) Voyages of Discovery Item 5 – E051V05 (negative discrimination)
Croatia Pemba Sherpa Item 9 – RP41B09 (negative discrimination) Pemba Sherpa Item12 – RP41B12 (negative discrimination)
Denmark Ink Drinker Item 3 – RE51D03 (translation error) The Summer My Father Was 10 Item 4 – RE31U04 (low discrimination)
Egypt Ostrich and the Hat Item 2 – RP51T02 (negative discrimination) World's Bank for Seeds Item 3 – RP51N03 (negative discrimination)

PIRLS 2021

Items Deleted by Country

Finland

Rainforests Item 16 – E041R16 (translation error)
Voyages of Discovery Item 12– E051V12 (translation error)

France

Sharks Item 12 – RP21K12 (low reliability)

Georgia

Sharks Item 3 – RP21K04 (negative discrimination)

Hong Kong SAR

Ostrich and the Hat Item 2 – RP51T02 (negative discrimination)
Learning a New Language Item 13 – RP51R13 (severe differential item functioning)

Jordan

Ostrich and the Hat Item 2 – RP51T02 (negative discrimination)
World's Bank for Seeds Item 3 – RP51N03 (negative discrimination)

Kosovo

Pemba Sherpa Item 6 – RP41B06 (low discrimination)
World's Bank for Seeds Item 3 – RP51N03 (negative discrimination)
Sharks Item 3 – RP21K04 (negative discrimination)
Hungry Plant Item 4 – RP41H04 (negative discrimination)

Latvia (Russian language only)

Learning a New Language Item 16 – RP51R16 (printing error)

Lithuania

Ostrich and the Hat Item 2 – RE51T02 (translation error)

Macao SAR

Shiny Straw Item 10 – RP21Y10 (severe differential item functioning)

Montenegro

Pemba Sherpa Item 6 – RP41B06 (low discrimination)
Library Mouse Item 11 – RP41M11 (translation error)
Hungry Plant Item 3 – RP41H04 (negative discrimination)

Morocco

Empty Pot Item 9 – RP31M09 (poor reliability)
World's Bank for Seeds Item 3 – RP51N03 (negative discrimination)
Sharks Item 11 – RP21K11 (low discrimination)

Netherlands

Empty Pot Item 4 – RP31M04 (low discrimination)
Sharks Item 4 – RP21K04 (translation error)

North Macedonia

Pemba Sherpa Item 6 – RP41B06 (low discrimination)
World's Bank for Seeds Item 3 – RP51N03 (negative discrimination)
Sharks Item 4 – RP21K04 (translation error)

PIRLS 2021

Items Deleted by Country

Norway

Ink Drinker Item 6 – RE51D06 (translation error)

Oman

Ostrich and the Hat Item 2 – RP51T02 (negative discrimination)
World's Bank for Seeds Item 3 – RP51N03 (negative discrimination)
Sharks Item 3 – RP21K03 (low discrimination)

Russian Federation

Legend of Troy Item 9 – E041T09 (translation error)

Saudi Arabia

Shiny Straw Item 10 – RE21Y10 (translation error)
Shiny Straw Item 14 – RE21Y14 (poor reliability)
Oliver and the Griffin Item 6 – RE41O06 (translation error)

Slovak Republic

Empty Pot Item 3 – RE31M04 (translation error)
Hungry Plant Item 9 – RE41H09 (low discrimination)

South Africa

World's Bank for Seeds Item 3 – RP51N03 (negative discrimination)
Sharks Item 3 – RP21K03 (negative discrimination)

Spain

The Summer My Father Was 10 Item 4 – RE31U04 (translation error)
Library Mouse Item 15 – RE41M15 (low discrimination)

Sweden

Ostrich and the Hat Item 1 – RE51T01 (negative discrimination)
Zebra and Wildebeest Migration Item 2 – E041Z02 (translation error)

Uzbekistan

World's Bank for Seeds Item 3 – RP51N03 (negative discrimination)

Moscow City, Russian Federation

Legend of Troy Item 9 – E041T09 (translation error)

Items beginning with "RE" are digitalPIRLS items and beginning with "E0" are ePIRLS items. Items beginning with "RP" are paperPIRLS items, or bridge items. paperPIRLS trend items deleted or recoded for all countries were also modified for digitalPIRLS bridge samples.

Appendix 9G: Derived Items in PIRLS 2021

PIRLS 2021
The Sumer My Father Was 10 Item 12 – RE31U12: Item parts A, B, and D are combined to create a 1-point item, where 1 score point is awarded if all parts are correct
Learning a New Language Item 5 – RE51R05: Item parts A, B, C, D, and E are combined to create a 2-point item, where 2 score points are awarded if all 5 parts are correct and 1 score point is awarded if 4 parts are correct
Learning a New Language Item 15 – RE51R15: Item parts A, C, D, and E are combined to create a 1-point item, where 1 score point is awarded if all parts are correct
Learning a New Language Item 17 – RE51R17, RP51R17: Item parts A and B are combined to create a 2-point item, where 2 score points are awarded if both parts are correct and 1 score point is awarded if 1 part is correct
The Empty Pot Item 17 – RE31M17, RP31M17: Item parts A, B, and C are combined to create a 3-point item, where 3 score points are awarded if all parts are correct, 2 score points are awarded if 2 parts are correct, and 1 score point is awarded if 1 part is correct
Ostrich and the Hat Item 5 – RE51T05: Item parts A, B, C, D, and E are combined to create a 2-point item, where 2 score points are awarded if all 5 parts are correct and 1 score point is awarded if 4 parts are correct
Ostrich and the Hat Item 14 – RE51T14, RP51T14: Item parts A, B, C, D, and E are combined to create a 1-point item, where 1 score point is awarded if all 5 parts are correct
The Ink Drinker Item 11 – RE51D11, RP51D11: Item parts A and B are combined to create a 2-point item, where 2 score points are awarded if both parts are correct and 1 score point is awarded if 1 part is correct
The Ink Drinker Item 12 – RE51D12, RP51D12: Item parts A and B are combined to create a 3-point item, where 3 score points are awarded if both parts are correct, 2 score points are awarded if item part B is answered correctly, and 1 score point is awarded if item part A is answered correctly
Training a Deaf Polar Bear Item 14 – RE31P14: Item parts A, C, and D are combined to create a 1-point item, where 1 score point is awarded if all parts are correct
The Amazing Octopus Item 1 – RE51Z01: Item parts A, B, C, D, and E are combined to create a 2-point item, where 2 score points are awarded if all 5 parts are correct and 1 score point is awarded if 4 parts are correct
How Did We Learn to Fly? Item 16 – RE41E16: Item parts A, C, D, and E are combined to create a 1-point item, where 1 score point is awarded if all parts are correct
Marie Curie Prize-Winning Scientist Item 1 – RE51C01: Item parts A, B, C, D, and E are combined to create a 2-point item, where 2 score points are awarded if all 5 parts are correct and 1 score point is awarded if 4 parts are correct
Marie Curie Prize-Winning Scientist Item 7 – RE51C07: Item parts A, B, C, and D are combined to create a 1-point item, where 1 score point is awarded if all parts are correct
Marie Curie Prize-Winning Scientist Item 13 – RE51C13, RP51C13: Item parts A and B are combined to create a 2-point item, where 2 score points are awarded if both parts are correct and 1 score point is awarded if 1 part is correct
Where's the Honey? Item 7 – RE31W07, RP31W07: Item parts A, B, and C are combined to create a 3-point item, where 3 score points are awarded if all parts are correct, 2 score points are awarded if 2 parts are correct, and 1 score point is awarded if 1 part is correct

PIRLS 2021
The World's Bank for Seeds Item 2 – RE51N02: Item parts A, B, C, D, and E are combined to create a 2-point item, where 2 score points are awarded if all 5 parts are correct and 1 score point is awarded if 4 parts are correct
The World's Bank for Seeds Item 6 – RE51N06: Item parts A, B, C, D, and E are combined to create a 2-point item, where 2 score points are awarded if all 5 parts are correct and 1 score point is awarded if 4 parts are correct
The World's Bank for Seeds Item 9 – RE51N09, RP51N09: Item parts A and B are combined to create a 2-point item, where 2 score points are awarded if both parts are correct and 1 score point is awarded if 1 part is correct
The World's Bank for Seeds Item 10 – RE51N10: Item parts A, B, D, and E are combined to create a 1-point item, where 1 score point is awarded if all parts are correct
The World's Bank for Seeds Item 13 – RE51N13, RP51N13: Item parts A and B are combined to create a 2-point item, where 2 score points are awarded if both parts are correct and 1 score point is awarded if 1 part is correct
Rainforests Item 3 – E041R03: Item parts A, B, C, and D are combined to create a 2-point item, where 2 score points are awarded if all 4 parts are correct and 1 score point is awarded if 3 parts are correct
Rainforests Item 7 – E041R07: Item parts A, B, C, and D are combined to create a 2-point item, where 2 score points are awarded if all 4 parts are correct and 1 score point is awarded if 3 parts are correct
The Legend of Troy Item 18 – E041T18: Item parts A, B, and D are combined to create a 1-point item, where 1 score point is awarded if all parts are correct
Zebra and Wildebeest Migration Item 20 – E041Z20: Item parts A, B, C, and D are combined to create a 2-point item, where 2 score points are awarded if all 4 parts are correct and 1 score point is awarded if 3 parts are correct
Voyages of Discovery Item 9 – E051V09: Item parts A, B, C, D, and E are combined to create a 2-point item, where 2 score points are awarded if all 5 parts are correct and 1 score point is awarded if 4 parts are correct
Voyages of Discovery Item 17 – E051V17: Item parts A, B, D, and E are combined to create a 2-point item, where 2 score points are awarded if all 4 parts are correct and 1 score point is awarded if 3 parts are correct
Voyages of Discovery Item 18 – E051V18: Item parts A, B, C, and D are combined to create a 1-point item, where 1 score point is awarded if all parts are correct

Items beginning with "RE" are digitalPIRLS items and item beginning with "E0" are ePIRLS items. Items beginning with "RP" are paperPIRLS items, or bridge items. Derived paperPIRLS trend items were also derived for digitalPIRLS bridge samples.