CHAPTER 11

Reviewing the TIMSS Advanced 2015 Achievement Item Statistics

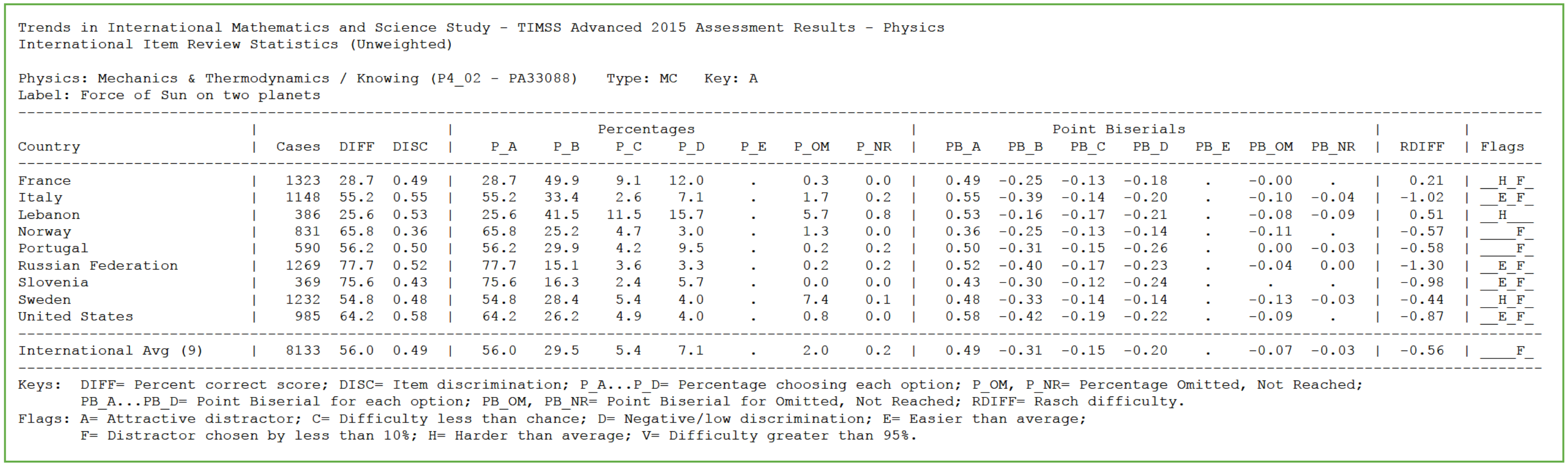
Pierre Foy
Michael O. Martin
Ina V.S. Mullis
Liqun Yin
Kerry Cotter
Jenny Liu

The TIMSS & PIRLS International Study Center conducted a review of a range of diagnostic statistics to examine and evaluate the psychometric characteristics of each achievement item across the countries that participated in the TIMSS Advanced 2015 assessments. This review of item statistics is essential to the successful application of item response theory (IRT) scaling to derive student achievement scores for analysis and reporting. This review played a crucial role in the quality assurance of the TIMSS Advanced 2015 achievement data prior to scaling, making it possible to detect unusual item properties that could signal a problem or error for a particular country. For example, an item that was uncharacteristically easy or difficult, or had an unusually low discriminating power, could indicate a potential problem with either translation or printing. Similarly, a constructed response item with unusually low scoring reliability could indicate a problem with a scoring guide in a particular country. In the rare instances where such items were found, the country's translation verification documents and printed booklets were examined for flaws or inaccuracies and, if necessary, the item was removed from the international database for that country.

Statistics for Item Review

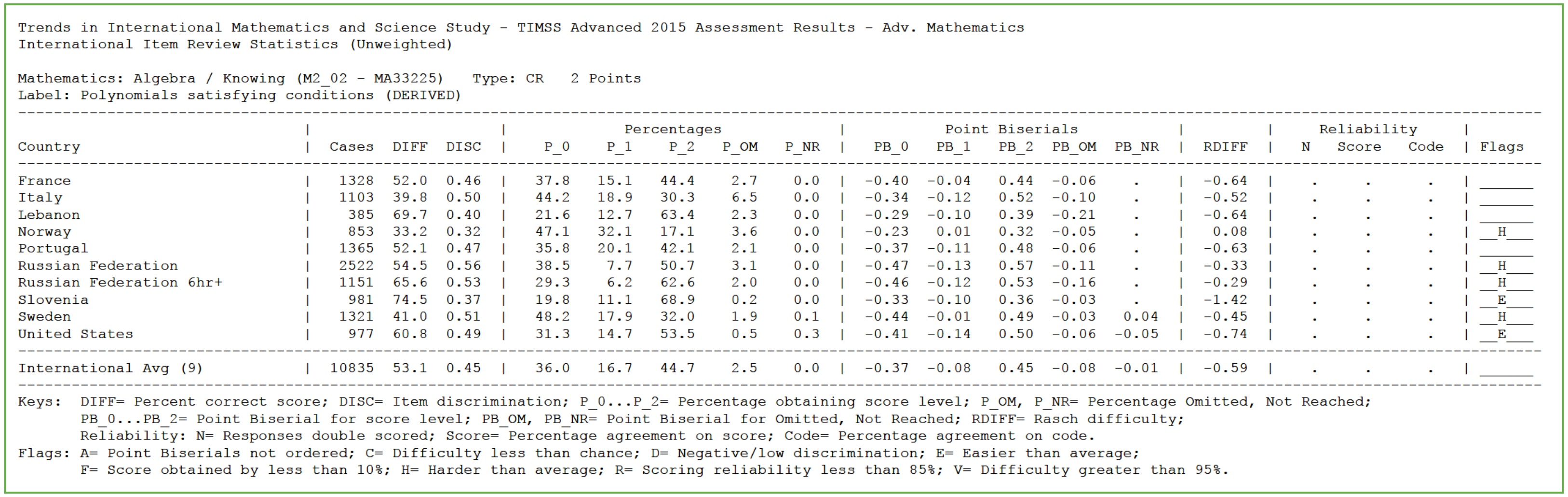
The TIMSS & PIRLS International Study Center computed item statistics for all achievement items in the 2015 assessments, including advanced mathematics (102 items) and physics (103 items). The item statistics for each of the participating countries were then carefully reviewed. Exhibits 11.1 and 11.2 show actual samples of the statistics calculated for a multiple-choice and a constructed response item, respectively.

Exhibit 11.1: Example International Item Statistics for a TIMSS Advanced 2015 Multiple-Choice Item



```
Trends in International Mathematics and Science Study - TIMSS Advanced 2015 Assessment Results - Physics
International Item Review Statistics (Unweighted)

Physics: Mechanics & Thermodynamics / Knowing (P4_02 - PA33088)    Type: MC    Key: A
Label: Force of Sun on two planets
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
                                    |                        |                           Percentages                      |                         Point Biserials                        |          |
Country                             |  Cases  DIFF  DISC     |     P_A     P_B     P_C     P_D     P_E    P_OM    P_NR     |   PB_A    PB_B    PB_C    PB_D    PB_E   PB_OM   PB_NR     |  RDIFF   | Flags
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
France                              |   1323  28.7  0.49     |    28.7    49.9     9.1    12.0      .      0.3     0.0     |   0.49   -0.25   -0.13   -0.18      .    -0.00      .       |   0.21   | __H_F_
Italy                               |   1148  55.2  0.55     |    55.2    33.4     2.6     7.1      .      1.7     0.2     |   0.55   -0.39   -0.14   -0.20      .    -0.10   -0.04     |  -1.02   | __E_F_
Lebanon                             |    386  25.6  0.53     |    25.6    41.5    11.5    15.7      .      5.7     0.8     |   0.53   -0.16   -0.17   -0.21      .    -0.08   -0.09     |   0.51   | __H___
Norway                              |    831  65.8  0.36     |    65.8    25.2     4.7     3.0      .      1.3     0.0     |   0.36   -0.25   -0.13   -0.14      .    -0.11      .       |  -0.57   | ____F_
Portugal                            |    590  56.2  0.50     |    56.2    29.9     4.2     9.5      .      0.2     0.2     |   0.50   -0.31   -0.15   -0.26      .     0.00   -0.03     |  -0.58   | ____F_
Russian Federation                  |   1269  77.7  0.52     |    77.7    15.1     3.6     3.3      .      0.2     0.2     |   0.52   -0.40   -0.17   -0.23      .    -0.04    0.00     |  -1.30   | __E_F_
Slovenia                            |    369  75.6  0.43     |    75.6    16.3     2.4     5.7      .      0.0     0.0     |   0.43   -0.30   -0.12   -0.24      .       .       .       |  -0.98   | __E_F_
Sweden                              |   1232  54.8  0.48     |    54.8    28.4     5.4     4.0      .      7.4     0.1     |   0.48   -0.33   -0.14   -0.14      .    -0.13   -0.03     |  -0.44   | __H_F_
United States                       |    985  64.2  0.58     |    64.2    26.2     4.9     4.0      .      0.8     0.0     |   0.58   -0.42   -0.19   -0.22      .    -0.09      .       |  -0.87   | __E_F_
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
International Avg (9)               |   8133  56.0  0.49     |    56.0    29.5     5.4     7.1      .      2.0     0.2     |   0.49   -0.31   -0.15   -0.20      .    -0.07   -0.03     |  -0.56   | ____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; C= Difficulty less than chance; D= Negative/low discrimination; E= Easier than average;
       F= Distractor chosen by less than 10%; H= Harder than average; V= Difficulty greater than 95%.
```

Exhibit 11.2: Example International Item Statistics for a TIMSS Advanced 2015 Constructed Response Item



```
Trends in International Mathematics and Science Study - TIMSS Advanced 2015 Assessment Results - Adv. Mathematics
International Item Review Statistics (Unweighted)

Mathematics: Algebra / Knowing (M2_02 - MA33225)   Type: CR   2 Points
Label: Polynomials satisfying conditions (DERIVED)
----------------------------------------------------------------------------------------------------------------------------------------------------------------------
                              |                   |              Percentages            |         Point Biserials             |        |      Reliability     |
Country                       | Cases  DIFF  DISC |   P_0   P_1   P_2   P_OM   P_NR     |  PB_0   PB_1   PB_2  PB_OM  PB_NR   |  RDIFF |  N    Score    Code  | Flags
----------------------------------------------------------------------------------------------------------------------------------------------------------------------
France                        |  1328  52.0  0.46 |  37.8  15.1  44.4    2.7    0.0     | -0.40  -0.04   0.44  -0.06    .     |  -0.64 |  .      .        .   | ______
Italy                         |  1103  39.8  0.50 |  44.2  18.9  30.3    6.5    0.0     | -0.34  -0.12   0.52  -0.10    .     |  -0.52 |  .      .        .   | ______
Lebanon                       |   385  69.7  0.40 |  21.6  12.7  63.4    2.3    0.0     | -0.29  -0.10   0.39  -0.21    .     |  -0.64 |  .      .        .   | ______
Norway                        |   853  33.2  0.32 |  47.1  32.1  17.1    3.6    0.0     | -0.23   0.01   0.32  -0.05    .     |   0.08 |  .      .        .   | __H___
Portugal                      |  1365  52.1  0.47 |  35.8  20.1  42.1    2.1    0.0     | -0.37  -0.11   0.48  -0.06    .     |  -0.63 |  .      .        .   | ______
Russian Federation            |  2522  54.5  0.56 |  38.5   7.7  50.7    3.1    0.0     | -0.47  -0.13   0.57  -0.11    .     |  -0.33 |  .      .        .   | __H___
Russian Federation 6hr+       |  1151  65.6  0.53 |  29.3   6.2  62.6    2.0    0.0     | -0.46  -0.12   0.53  -0.16    .     |  -0.29 |  .      .        .   | __H___
Slovenia                      |   981  74.5  0.37 |  19.8  11.1  68.9    0.2    0.0     | -0.33  -0.10   0.36  -0.03    .     |  -1.42 |  .      .        .   | __E___
Sweden                        |  1321  41.0  0.51 |  48.2  17.9  32.0    1.9    0.1     | -0.44  -0.01   0.49  -0.03   0.04   |  -0.45 |  .      .        .   | __H___
United States                 |   977  60.8  0.49 |  31.3  14.7  53.5    0.5    0.3     | -0.41  -0.14   0.50  -0.06  -0.05   |  -0.74 |  .      .        .   | __E___
----------------------------------------------------------------------------------------------------------------------------------------------------------------------
International Avg (9)         | 10835  53.1  0.45 |  36.0  16.7  44.7    2.5    0.0     | -0.37  -0.08   0.45  -0.08  -0.01   |  -0.59 |  .      .        .   | ______
----------------------------------------------------------------------------------------------------------------------------------------------------------------------
Keys:  DIFF= Percent correct score; DISC= Item discrimination; P_0...P_2= Percentage obtaining score level; P_OM, P_NR= Percentage Omitted, Not Reached;
       PB_0...PB_2= Point Biserial for score level; PB_OM, PB_NR= Point Biserial for Omitted, Not Reached; RDIFF= Rasch difficulty;
       Reliability: N= Responses double scored; Score= Percentage agreement on score; Code= Percentage agreement on code.
Flags: A= Point Biserials not ordered; C= Difficulty less than chance; D= Negative/low discrimination; E= Easier than average;
       F= Score obtained by less than 10%; H= Harder than average; R= Scoring reliability less than 85%; V= Difficulty greater than 95%.
```

For all items, regardless of format (i.e., multiple-choice or constructed response), statistics included the number of students that responded in each country, the difficulty level (the percentage of students that answered the item correctly), and the discrimination index (the point-biserial correlation between success on the item and total score).[1] Also provided was an estimate of the difficulty of the item using a Rasch one-parameter IRT model. Statistics for each item were displayed alphabetically by country, together with an international average—i.e., based on all participating countries listed above the international average for each statistic. The international averages of the item difficulties and item discriminations served as guides to the overall statistical properties of the items.

Statistics displayed for multiple-choice items included the percentage of students that chose each response option—as well as the percentage of students that omitted or did not reach the item—and the point-biserial correlations for each response option. Statistics displayed for constructed response items (which could have 1 or 2 score points) included the percent correct and point-biserial of 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 of double-scored responses, the percentage of score agreement between the scorers, and—because TIMSS Advanced has a 2-digit scoring scheme—the percentage of code agreement between scorers.

During item review, "not-reached" responses (i.e., items toward the end of the booklet that the student did not attempt)[2] were treated as "not administered" and thus did not contribute to the calculation of the item statistics. However, the percentage of students not reaching each item was reported. Omitted responses, although treated as incorrect, were tabulated separately from incorrect responses for the sake of distinguishing students who provided no form of response from students who attempted a response.

The definitions and detailed descriptions of the statistics that were calculated are given below. The statistics were calculated separately by subject, and within each table 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. Not-reached responses were not included in this count.

DIFF: The item difficulty is the average percent correct on an item. For a 1-point item, including all multiple-choice items, it is the percentage of students providing a fully correct response to the item. For 2-point items, it is the average percentage of points. For example, if 25 percent of students scored 2 points, 50 percent scored 1 point on a 2-point item, and the other 25 percent score 0 points, then the average percent correct for such an item would be 50 percent. For this statistic, not-reached responses were not included.

[1] For computing point-biserial correlations, the total score is the percentage of points a student has scored on the items (s)he was administered. Not-reached responses are not included in the total score.

[2] An item was considered "not-reached" if the item itself and the item immediately preceding it were not answered and no subsequent items had been attempted.

DISC: The 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 the more able students get the item right, the less able get it wrong. For this statistic, not-reached items were not included.

PCT_A, PCT_B, PCT_C, PCT_D, and PCT_E: Available for multiple-choice items. Each column indicates the percentage of students choosing the particular response option for the item (A, B, C, D, or E). Not-reached responses were excluded from the denominator.

PCT_0, PCT_1, and PCT_2: Available for constructed response items. Each column indicates the percentage of students responding at that particular score level, up to and including the maximum score level for the item. Not-reached items were excluded from the denominator.

PCT_OM: Percentage of students who, having reached the item, did not provide a response. Not reached responses were excluded from the denominator.

PCT_NR: Percentage of students who did not reach the item. This statistic is the number of students who did not reach an item as a percentage of all students who were administered that item, including those who omitted or did not reach that item.

PB_A, PB_B, PB_C, PB_D, or PB_E: Available for multiple-choice items. These columns show the point-biserial correlations between choosing each of the response options (A, B, C, D, or E) and the total score on all of the items administered to a student. Items with good psychometric properties have moderately positive correlations for the correct option and negative correlations for the distracters (the incorrect options). Not-reached responses were not included in these calculations.

PB_0, PB_1, and PB_2: Available for constructed response items. These columns present the point-biserial correlations between the score levels on the item (0, 1, or 2) and the overall score on all of the items the student was administered. For items with good measurement properties, the correlation coefficients should monotonically increase from negative to positive as the score on the item increases. Not-reached responses were not included in these calculations.

PB_OM: The point-biserial correlation between a binary variable indicating an omitted response to the item, and the total score on all items administered to a student. This correlation should be negative or near zero. Not-reached responses were not included in this statistic.

PB_NR: The point-biserial correlation between a binary variable indicating a not-reached response to the item, and the total score on all items administered to a student. This correlation should be negative or near zero.

RDIFF: An estimate of the difficulty of an item based on a Rasch one-parameter 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 was scaled so that the average Rasch item difficulty across all items within each country was zero.

Reliability (N): To provide a measure of the reliability of the scoring of the constructed response items, items in approximately 25 percent of the test booklets in each country were independently scored by two scorers. This column indicates the number of responses that were double-scored for a given item in a country.

Reliability (Score): This column contains the percentage of agreement on the score value of the two-digit diagnostic codes assigned by the two independent TIMSS Advanced scorers.

Reliability (Code): This column contains the percentage of agreement on the two-digit diagnostic codes assigned by the two independent TIMSS Advanced scorers.

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 a problem with an item. The following conditions were flagged:

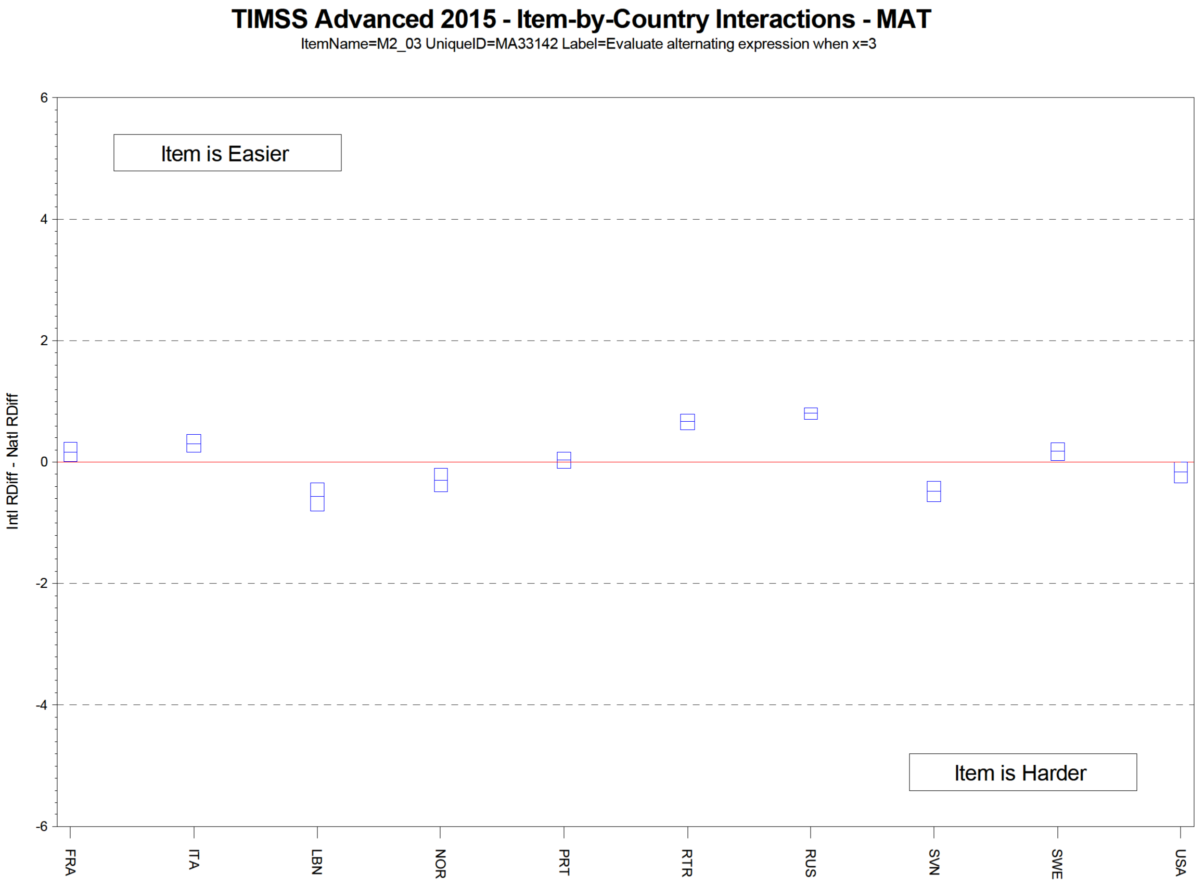
- The item discrimination (DISC) was less than 0.10 (flag D).
- The item difficulty (DIFF) was less than .25 for multiple-choice items (flag C).
- The item difficulty (DIFF) exceeded .95 (flag V).
- The Rasch difficulty estimate (RDIFF) for a given country made the item either easier (flag E) or more difficult (flag H) relative to the international average for that item.
- The point-biserial correlation for at least one distracter in a multiple-choice 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 multiple-choice item, or one of the score values for a constructed response item, was less than 10 percent (flag F).
- Scoring reliability for agreement on the score value of a constructed response item was less than 85 percent (flag R).

Although not all of these conditions necessarily indicated a problem, the flags were a useful tool to draw attention to potential sources of concern.

Item–by–Country Interaction

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 of the items, and low-scoring countries should do less well on each of the items. When this does not occur (e.g., when a high-performing country has low performance on an item on which other countries are doing well), there is said to be an item-by-country interaction. When large, such item-by-country interactions may be a sign that an item is flawed in some way and that steps should be taken to address the problem. To assist in detecting sizeable item-by-country interactions, the TIMSS & PIRLS International Study Center produced a graphical display for each item showing the difference between each country's Rasch item difficulty and the international average Rasch item difficulty across all countries. An example of the graphical displays is provided in Exhibit 11.3.

Exhibit 11.3: Example Plot of Item-by-Country Interaction for a TIMSS Advanced 2015 Item



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

$$\text{Upper Limit} = RDIFF_{i.} - RDIFF_{ik} + SE(RDIFFik) \cdot Z_b$$
$$\text{Lower Limit} = RDIFF_{i.} - RDIFF_{ik} - SE(RDIFFik) \cdot Z_b$$

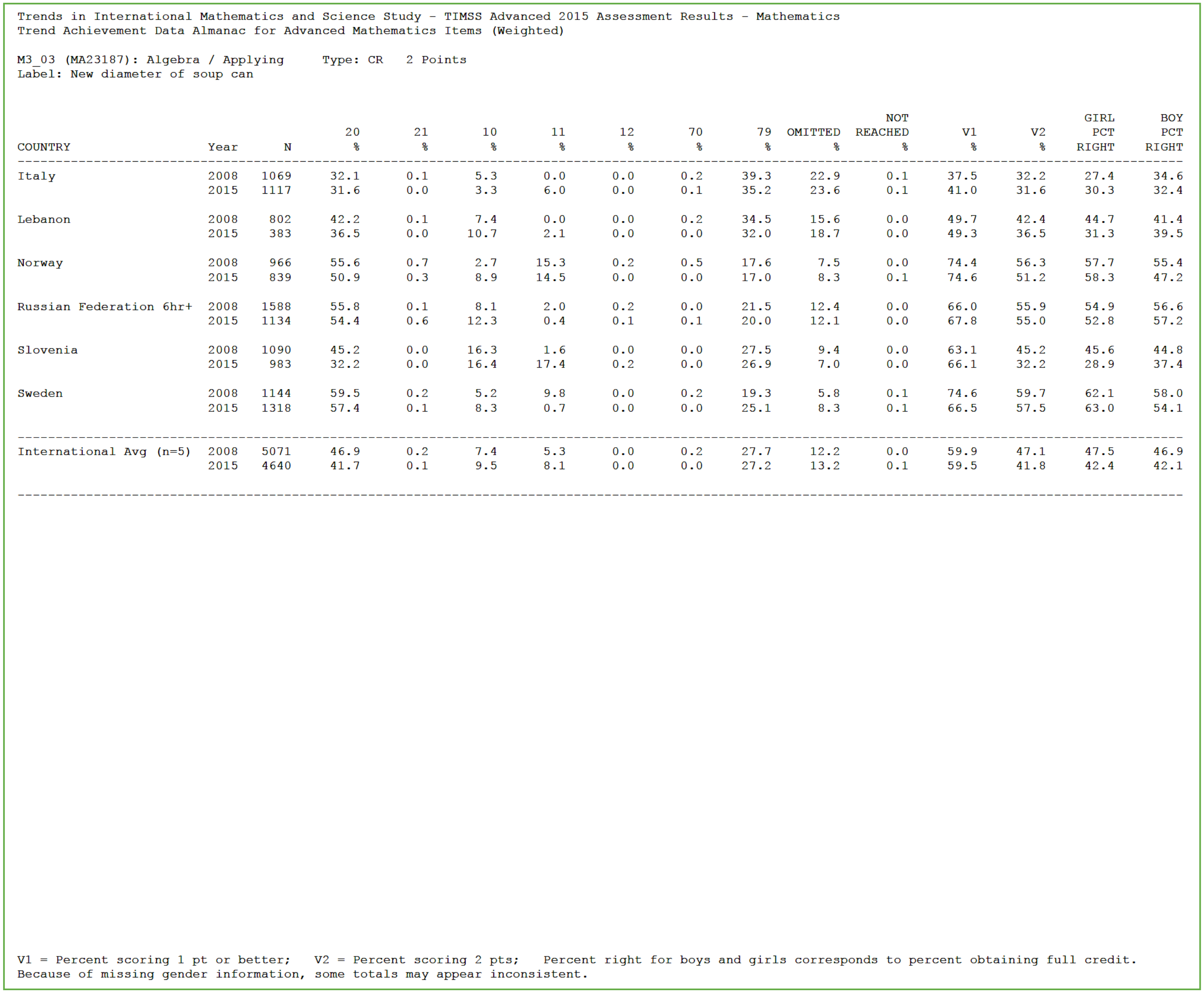
where $RDIFF_{\mathrm{ik}}$ is the Rasch difficulty of item i in country k, $RDIFF_{i.}$ is the international average Rasch difficulty of item i, $SE(RDIFF_{ik})$ is the standard error of the Rasch difficulty of item i in country k, and Z_b is the 95% critical value from the Z distribution corrected for multiple comparisons using the Bonferroni procedure.

Trend Item Review

In order to measure trends, TIMSS Advanced 2015 included achievement items from previous assessments as well as items developed for use for the first time in 2015. Accordingly, the TIMSS Advanced 2015 assessments included items from 1995, 2008, and 2015. An important review step, therefore, was to check that these "trend items" had statistical properties in 2015 similar to those they had in the previous assessments (e.g., a TIMSS Advanced item that was relatively easy in 2008 should still be relatively easy in 2015).

As can be seen in the example in Exhibit 11.4, the trend item review focused on statistics for trend items from the current and previous assessments (2015 and 2008) for countries that participated in both. For each country, trend item statistics included the percentage of students in each score category (or response option for multiple-choice items) for each assessment, as well as the difficulty of the item and the percent correct by gender. In reviewing these item statistics, the aim was to detect any unusual changes in item difficulties between administrations, which might indicate a problem in using the item to measure trends.

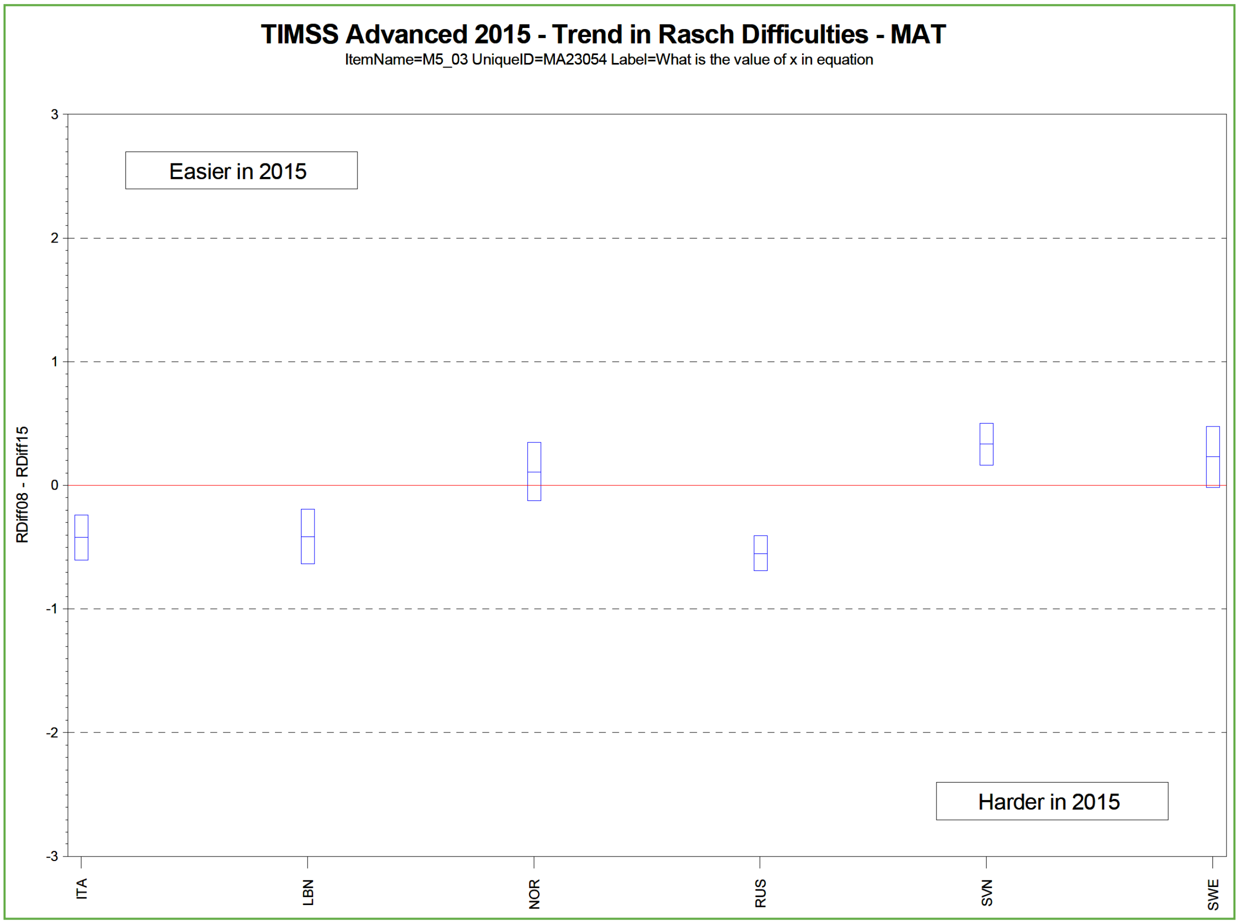
Exhibit 11.4: Example Item Statistics for a TIMSS Advanced 2015 Trend Item



```
Trends in International Mathematics and Science Study - TIMSS Advanced 2015 Assessment Results - Mathematics
Trend Achievement Data Almanac for Advanced Mathematics Items (Weighted)

M3_03 (MA23187): Algebra / Applying     Type: CR   2 Points
Label: New diameter of soup can


                                                                                                 NOT                        GIRL     BOY
                                          20      21      10      11      12      70      79  OMITTED  REACHED     V1      V2      PCT     PCT
COUNTRY                    Year     N      %       %       %       %       %       %       %       %        %       %       %    RIGHT   RIGHT
-------------------------------------------------------------------------------------------------------------------------------------------
Italy                      2008   1069   32.1     0.1     5.3     0.0     0.0     0.2    39.3    22.9      0.1    37.5    32.2    27.4    34.6
                           2015   1117   31.6     0.0     3.3     6.0     0.0     0.1    35.2    23.6      0.1    41.0    31.6    30.3    32.4

Lebanon                    2008    802   42.2     0.1     7.4     0.0     0.0     0.2    34.5    15.6      0.0    49.7    42.4    44.7    41.4
                           2015    383   36.5     0.0    10.7     2.1     0.0     0.0    32.0    18.7      0.0    49.3    36.5    31.3    39.5

Norway                     2008    966   55.6     0.7     2.7    15.3     0.2     0.5    17.6     7.5      0.0    74.4    56.3    57.7    55.4
                           2015    839   50.9     0.3     8.9    14.5     0.0     0.0    17.0     8.3      0.1    74.6    51.2    58.3    47.2

Russian Federation 6hr+    2008   1588   55.8     0.1     8.1     2.0     0.2     0.0    21.5    12.4      0.0    66.0    55.9    54.9    56.6
                           2015   1134   54.4     0.6    12.3     0.4     0.1     0.1    20.0    12.1      0.0    67.8    55.0    52.8    57.2

Slovenia                   2008   1090   45.2     0.0    16.3     1.6     0.0     0.0    27.5     9.4      0.0    63.1    45.2    45.6    44.8
                           2015    983   32.2     0.0    16.4    17.4     0.2     0.0    26.9     7.0      0.0    66.1    32.2    28.9    37.4

Sweden                     2008   1144   59.5     0.2     5.2     9.8     0.0     0.2    19.3     5.8      0.1    74.6    59.7    62.1    58.0
                           2015   1318   57.4     0.1     8.3     0.7     0.0     0.0    25.1     8.3      0.1    66.5    57.5    63.0    54.1

-------------------------------------------------------------------------------------------------------------------------------------------
International Avg (n=5)    2008   5071   46.9     0.2     7.4     5.3     0.0     0.2    27.7    12.2      0.0    59.9    47.1    47.5    46.9
                           2015   4640   41.7     0.1     9.5     8.1     0.0     0.0    27.2    13.2      0.1    59.5    41.8    42.4    42.1

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

V1 = Percent scoring 1 pt or better;    V2 = Percent scoring 2 pts;    Percent right for boys and girls corresponds to percent obtaining full credit.
Because of missing gender information, some totals may appear inconsistent.
```

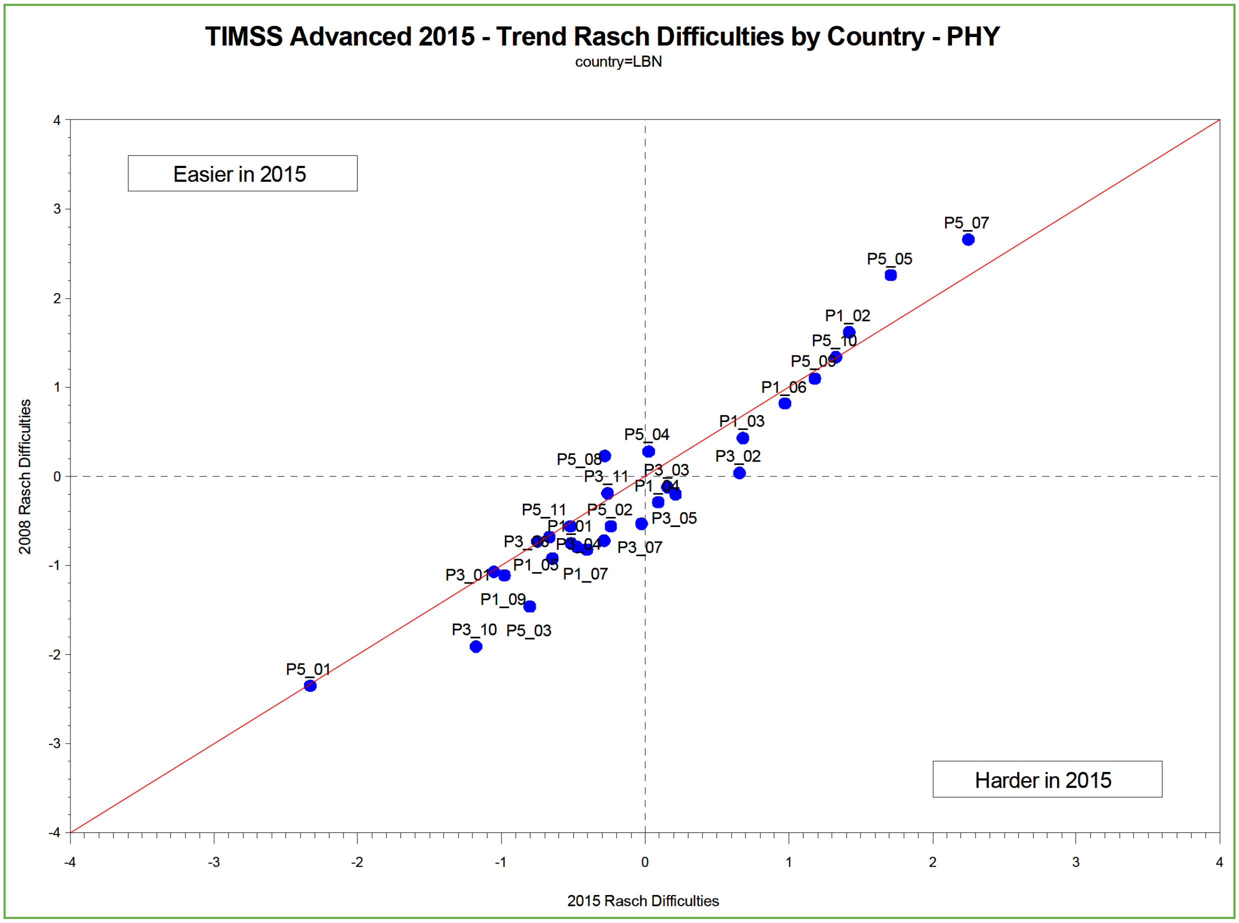
While some changes in item difficulties were anticipated as countries' overall achievement may have improved or declined, items were noted if the difference between the Rasch difficulties across the two assessments for a particular country was greater than 2 logits. The TIMSS & PIRLS International Study Center used two different graphical displays to examine the differences in item difficulties. The first of these, shown for an example item in Exhibit 11.5, displays the difference in Rasch item difficulty of the item between 2015 and 2008 for each country. A positive difference for a country indicates that the item was relatively easier in 2015, and a negative difference indicates that the item was relatively more difficult.

Exhibit 11.5: Example Plot of Differences in Rasch Item Difficulties Between 2015 and 2008 for a TIMSS Advanced Trend Item



The second graphical display, presented in Exhibit 11.6, shows the performance of a given country on all trend items simultaneously. For each country, the graph plots the 2015 Rasch difficulty of every trend item against its Rasch difficulty in 2008. Where there were no differences between the difficulties in the two successive administrations, the data points aligned on or near the diagonal.

Exhibit 11.6: Example Plot of Rasch Item Difficulties Across TIMSS Advanced Trend Items by Country



Reliability

Documenting the reliability of the TIMSS Advanced 2015 assessments was a critical quality control step in reviewing the items. As one indicator of reliability, the review considered Cronbach's Alpha coefficient of reliability calculated at the assessment booklet level. Secondly, the scoring of the constructed response items had to meet specific reliability criteria in terms of consistent within-country scoring and across assessment or trend–scoring.

Test Reliability

Exhibit 11.7 displays the TIMSS Advanced 2015 advanced mathematics and physics test reliability coefficients for every country, respectively. These coefficients are the median Cronbach's alpha reliability across all TIMSS Advanced 2015 assessment booklets. In general, reliabilities were relatively high, with international median reliabilities (the median of the reliability coefficients for all countries) of 0.88 for advanced mathematics and 0.84 for physics.

Exhibit 11.7: **Cronbach's Alpha Reliability Coefficient—TIMSS Advanced 2015**

Country	Reliability Coefficient	
	Advanced Mathematics	Physics
France	0.86	0.70
Italy	0.90	0.82
Lebanon	0.88	0.77
Norway	0.84	0.84
Portugal	0.86	0.78
Russian Federation	0.93	0.88
Russian Federation 6hr+[3]	0.93	—
Slovenia	0.88	0.85
Sweden	0.89	0.85
United States	0.91	0.84
International Median	**0.88**	**0.84**

Scoring Reliability for Constructed Response Items

A sizeable proportion of the items in the TIMSS Advanced 2015 assessments were constructed response items, comprising about half of the assessment score points. An essential requirement for use of such items is that they be reliably scored by all participants. That is, a particular student response should receive the same score, regardless of the scorer. In conducting TIMSS Advanced 2015, measures taken to ensure that the constructed response items were scored reliably in all countries included developing scoring guides for each constructed response question (that provided descriptions of acceptable responses for each score point value) and providing extensive training in the application of the scoring guides. See Chapter 1: Developing the TIMSS Advanced 2015 Achievement Items for more information on the scoring guides and see Chapter 6: Survey Operations Procedures for information on the scoring process.

Within-Country Scoring Reliability

To gather and document information about the within-country agreement among scorers for TIMSS Advanced 2015, a random sample of approximately 25 percent of the assessment booklets was selected to be scored independently by two scorers. The inter-scorer agreement for each item in each country was examined as part of the item review process. Exact percent agreement across items was high on average across countries, with 97 percent agreement in advanced mathematics and 96 percent agreement in physics, on average internationally. In TIMSS Advanced 2015 there also was high agreement at the diagnostic score level, with 96 percent agreement in advanced mathematics and 94 percent agreement in physics, on average internationally. See Appendix 11A

[3] For advanced mathematics, the Russian Federation participated in 2015 with an expanded population that included the more specialized students assessed in 1995 and 2008.

for the average and range of the within-country percentage of correctness score agreement across all items. The TIMSS Advanced Within-Country Scoring Reliability documents also provide the average and range of the within-country percentage of diagnostic score agreement.

Trend Item Scoring Reliability

The TIMSS & PIRLS International Study Center also took steps to show that the 2015 constructed response items used in TIMSS Advanced 2008 were scored in the same way in both assessments. In anticipation of this, countries that participated in TIMSS Advanced 2008 sent samples of scored student booklets from the 2008 data collections to the IEA Data Processing and Research Center (IEA DPC), 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 2015 data were asked also to score these 2008 responses using the Trend Reliability Scoring Software developed by the IEA DPC. Each country scored 200 responses for each of 11 advanced mathematics and 11 physics items.

There was a very high degree of scoring consistency in TIMSS Advanced 2015. The exact agreement between the scores awarded in 2008 and those given by the 2015 scorers was 93 percent in advanced mathematics and 92 percent in physics, on average internationally. There also was high agreement in TIMSS Advanced at the diagnostic score level, although somewhat less in physics than in advanced mathematics, on average. The average and range of scoring consistency over time can be found in Appendix 11B.

Item Review Procedures

Using the information from the comprehensive collection of item analyses and reliability data that were computed and summarized for TIMSS Advanced 2015, the TIMSS & PIRLS International Study Center thoroughly reviewed all item statistics for every participating country to ensure that the items were performing comparably across countries. In particular, items with the following problems were considered for possible deletion from the international database:

- An error was detected during translation verification but was not corrected before test administration
- Data checking revealed a multiple-choice item with more or fewer options than in the international version
- The item analysis showed the item to have a negative biserial, or, for an item with more than 1 score point, point biserials that did not increase with each score level
- The item-by-country interaction results showed a very large negative interaction for a particular country

- For constructed response items, the within-country scoring reliability data showed an agreement of less than 70 percent
- For trend items, an item performed substantially differently in 2015 compared to the TIMSS Advanced 2008 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.

The checking of the TIMSS Advanced 2015 achievement data involved review of more than 200 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 were mostly items with differences attributable to either translation or printing problems. See Appendix 11C: Country Adaptations to Items and Item Scoring for a list of deleted items, as well as a list of recodes made to constructed response item codes. There also were a number of items in each study that were combined, or derived, for scoring purposes. See Appendix 11D for details about how score points were awarded for each derived item.

Appendix 11A: TIMSS Advanced 2015 Within-Country Scoring Reliability for the Constructed Response Items

TIMSS Advanced 2015 Within-Country Scoring Reliability for the Advanced Mathematics Constructed Response Items

Country	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Minimum	Maximum		Minimum	Maximum
France	97	84	100	96	81	100
Italy	97	90	100	96	86	100
Lebanon	92	78	99	88	77	98
Norway	98	87	100	97	83	100
Portugal	100	99	100	100	99	100
Russian Federation	99	96	100	99	96	100
Russian Federation 6hr+[1]	99	96	100	99	96	100
Slovenia	99	92	100	99	92	100
Sweden	96	88	100	95	82	100
United States	96	83	100	95	83	100
International Avg.	**97**	**89**	**100**	**96**	**87**	**100**

[1] For advanced mathematics, the Russian Federation participated in 2015 with an expanded population that included the more specialized students assessed in 1995 and 2008.

TIMSS Advanced 2015 Within-Country Scoring Reliability for the Physics Constructed Response Items

Country	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Minimum	Maximum		Minimum	Maximum
France	96	83	100	94	83	100
Italy	95	85	100	94	85	100
Lebanon	90	59	100	83	50	99
Norway	97	91	100	96	90	100
Portugal	100	98	100	99	97	100
Russian Federation	97	88	100	97	86	100
Slovenia	96	86	100	95	77	100
Sweden	96	90	100	94	85	100
United States	96	84	100	94	79	100
International Avg.	**96**	**85**	**100**	**94**	**81**	**100**

Appendix 11B: Trend Scoring Reliability for the Constructed Response Items

TIMSS Advanced 2015 Trend Scoring Reliability for the Advanced Mathematics Constructed Response Items

Country	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Minimum	Maximum		Minimum	Maximum
Italy	93	73	100	92	73	99
Norway	96	91	99	94	85	99
Russian Federation 6hr+[1]	92	72	100	91	72	100
Slovenia	92	77	99	90	76	98
Sweden	94	87	100	90	82	97
International Avg.	**93**	**80**	**100**	**91**	**78**	**98**

[1] For advanced mathematics, the Russian Federation participated in 2015 with an expanded population that included the more specialized students assessed in 1995 and 2008.

TIMSS Advanced 2015 Trend Scoring Reliability for the Physics Constructed Response Items

Country	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Minimum	Maximum		Minimum	Maximum
Italy	93	88	97	89	79	92
Norway	91	74	99	87	73	96
Russian Federation	92	83	99	88	72	96
Slovenia	90	81	98	82	69	92
Sweden	93	76	98	87	75	93
International Avg.	**92**	**80**	**98**	**87**	**73**	**94**

Appendix 11C: Country Adaptations to Items and Item Scoring

TIMSS Advanced Advanced Mathematics

Deleted Items

ALL COUNTRIES

MA13014, M1_04 (attractive distracter)

ITALY

MA23185, M5_02 (negative discrimination)

PORTUGAL

MA23185, M5_02 (negative discrimination)

SLOVENIA

MA33140, M2_07 (translation error)

Constructed Response Items with Category Recodes

ALL COUNTRIES

MA33039, M2_11 (recode 20 to 10, 21 to 11, 10 to 79)

TIMSS Advanced Physics

Deleted Items

ALL COUNTRIES

PA13020, P1_10 (low discrimination)

PA23131, P5_06 (outside scope of TIMSS Advanced 2015 Assessment Framework)

PA33056, P9_01 (poor discrimination)

FRANCE

PA33120, P4_07 (translation error)

PA33119, P6_07 (translation error)

PA33111B, P9_05B (translation error)

Constructed Response Items with Category Recodes

ALL COUNTRIES

PA23088, P3_10 (recode 11 to 71)

PA33073, P6_02 (recode 70 to 12)

Appendix 11D: Score Points for Derived Items in TIMSS Advanced 2015

TIMSS Advanced Advanced Mathematics
MA33225, M2_02 – Item parts A, B, C, D, and E are combined to create a 2-point item, where 2 score points are awarded if all parts are correct and 1 score point is awarded if 4 parts are correct
MA33118, M8_06 – Item parts A, B, and C are combined to create a 2-point item, where 2 score points are awarded if all parts are correct and 1 score point is awarded if 2 parts are correct
MA33236, M9_08 – Item parts A, B, C, and D are combined to create a 2-point item, where 2 score points are awarded if all parts are correct and 1 score point is awarded if 3 parts are correct

TIMSS Advanced Physics
PA33102A, P2_05 – Item parts A, B, and C are combined to create a 1-piont item, where 1 score point is awarded if all parts are correct
PA33008, P8_10 – Item parts A, B, and C are combined to create a 2-point item, where 2 score points are awarded if all parts are correct and 1 score point is awarded if 2 parts are correct