CHAPTER 10

Reviewing the TIMSS Achievement Data

Lillian Tyack
Bethany Fishbein
Jessie Bristol
Tianzheng Mao
Eugenio Gonzalez

Introduction

The procedures described in this chapter aim to ensure that the data collected by the TIMSS are reliable, valid, and comparable across countries—three important quality criteria in international educational assessment. TIMSS employs state-of-the-art psychometric methods to review the statistical properties of the items used in the assessment, and takes appropriate actions if deviations are observed that may jeopardize reliability, validity, or comparability. These actions involve item treatments that are decided upon by psychometricians in close cooperation with content experts to ensure the utility of the TIMSS data is maximized for primary reporting and secondary analyses.

This chapter describes the general principles and the statistics used to evaluate the quality of the TIMSS achievement data, and, in the second half, how these were applied during the TIMSS 2023 data analysis. For each TIMSS assessment cycle, the TIMSS & PIRLS International Study Center conducts reviews of the achievement data using observed item statistics before applying model-based reviews of the item functioning as part of the psychometric analyses with item response theory (IRT) methods. During the review of observed item statistics, the data on each item is evaluated at the international level overall as well as within each participating country. Key diagnostic statistics are examined to detect items with unusual psychometric properties or reveal anomalous patterns in the data for a particular country. For trend items that are administered over multiple TIMSS cycles for trend measurement, analyses are conducted to identify and evaluate any potential differences in the measurement properties between the current and the previous cycles.

Aggregate-level analyses across items and countries also are conducted for additional quality assurance of the TIMSS data. For example, item position effects are evaluated to ensure student performance was not affected substantially by the position of the item blocks in the assessment booklets. Another example is the analysis by booklet difficulty—more difficult or less difficult—which allows detecting anomalous data patterns in particular countries and for evaluating the effectiveness of the group-adaptive assessment design. Extensive analyses of

https://doi.org/10.6017/lse.tpisc.timss.rs7695

each country's item statistics allow for detecting any irregular patterns relative to previous cycles or to the international distribution across countries.

TIMSS Item Review Procedures

Each TIMSS cycle, several item- and block-level statistics are estimated for each achievement item in the TIMSS fourth- and eighth-grade assessments. The review of item statistics is an ongoing process that takes place over several months, culminating in a comprehensive psychometric review meeting in the few months following the last data collection, prior to the final calibration of the items and calculation of the achievement scores.

During this psychometric review meeting, decisions are made about necessary adjustments to the treatment of the items and about areas requiring further review or analyses. Examples of these item treatments are item recodings of 2-point items into 1-point items if there is evidence that the scoring rule needs improvement, or item deletions in countries where items show substantial misfit due to suspected translation, display, or functionality issues. It is important to note that differential treatment of items is rare. In the vast majority of cases, the TIMSS data enters the psychometric analysis intact as captured during data collection, and as subsequently scored using the internationally agreed-upon scoring guidelines. The statistical and psychometric item review is conducted item-by-item simultaneously for all participating countries. Internal identification and country reports about translation errors or other technical problems are used as reference, when applicable. In addition, interactive tables and graphical displays of item statistics are reviewed to detect any incongruous and systematic patterns in a particular country's data that may warrant further investigation.

For example, in rare occasions, item statistics may indicate that an item is uncharacteristically difficult or almost impossible to solve in most or all countries, or it may show only weak statistical associations with other items in the TIMSS assessment (i.e., it shows a relatively low discriminating coefficient in a particular country). These types of deviations, while rare, can indicate a potential problem with translation, or some other technical issue that can be addressed in the item review by trained psychometricians. These deviations might be addressed by deleting the items for one, multiple, or all countries. Similarly, a human-scored constructed-response item with low scoring reliability can point to a problem in the implementation of the scoring guide in one or more countries. Scoring guide issues are typically resolved after the field test item review, so these types of problems tend to be uncommon at the international level in the main data collection item review. In rare instances where an item is found to be not functioning as expected for a particular country, the country's translation verification records and digital instrument archives are examined for potential flaws or inaccuracies that may have been missed during instrument verification. If errors are identified during this investigation, data for that item could be removed from the international database. For countries that administer the assessment in more than one language, errors may be detected in only one language version, and data would only be removed for students assessed in that language.

Following the comprehensive item review, National Research Coordinators (NRCs) from the participating countries and benchmarking entities are contacted with inquiries about concerns or anomalies detected in their data. Decisions about item deletions or recoding are then implemented in the international data files.

Item Review Statistics

Various item statistics are combined for internal and external review (e.g., by participating countries) in several *item almanacs* containing summary tables that include country-level statistics for each item that was administered.

Item statistics are displayed in various types of static tabular forms as well as in interactive and graphical displays. In addition to the country-by-country and international statistics for the data of the current cycle, trend item statistics are produced for the trend items administered in both the current and previous cycles of TIMSS.

Item statistics for all items include the number of students who were administered the item, an estimated item difficulty (the percentage of students that answered the item correctly), and an item discrimination index (point-biserial correlation between the response to the item and total score). Also included is an estimate of the item difficulty parameter under the Rasch model.

In the item almanacs, statistics for each item are displayed by country listed alphabetically, together with an international average calculated using all participating countries, and a reference average based on a pool of countries that have participated in TIMSS assessments since 1995 or 1999 and contributed to establishing the initial achievement scale metrics. The international and reference averages of the item difficulties and item discriminations guide the evaluation of the overall statistical properties of the items. The item almanacs also include item statistics for the benchmarking participants.

Statistics produced for multiple-choice items include 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 produced for constructed-response items, worth either 1 or 2 score points, include the percentage of students and point-biserial for each score level.

For constructed-response items scored by humans, a sample of responses are scored twice in each country to provide information about the scoring reliability in each country (see Chapter 7). Item statistics include the total number of responses that were scored twice and the raw agreement as a measure of scoring reliability, calculated as the percentage of score agreement between the two independent scorers on a subset of responses that were double-scored.

The definitions and detailed descriptions of the estimated item statistics are given below.

N (Cases): This is the number of students to whom the item was administered. Students with Omitted (OM) or Not-Reached (NR) codes on the item are included in this count. The number of students to whom the item was administered may differ

across items depending on the booklet rotation implemented within each country based on the group-adaptive design.

DIFF: The item difficulty is calculated as follows: For a 1-point item, including all multiple-choice items, it is the percentage of students who provided a fully correct response. For items worth a maximum of more than one score point, it is the average score on the item divided by the maximum score possible. Omitted and not-reached responses are not included in the calculation of this statistic.

DISC: The item discrimination is computed as the correlation between the item score and the total score (sum of score points across all items administered to a student for the particular subject, minus the item score). Items exhibiting good measurement properties should have a moderate to high positive correlation, indicating that those that do well on the test get the item right at a higher rate than those that do not do well on the test.

Percentages—P0, P1, P2, P3, P4, P5, P6, PM, PR: These statistics present the percentage of respondents choosing the different response options (for multiple-choice items), or the percentage of respondents by score points awarded (for constructed-response items), along with the percentage of students who omitted the response to the item or did not reach it. The percentages are computed based on N, the number of students to whom the item was administered, regardless of whether they responded to the item. These percentages sum to 100 percent.

Point-Biserials—PB0, PB1, PB2, PB3, PB4, PB5, PB6, PBM, PBR: These statistics present the point-biserial correlation between selecting each response option (for multiple-choice items) or score level (for constructed-response items), and the total score (the sum of score points across all items answered by the student in that subject).

RDIFF: This provides an estimate of the item difficulty based on the Rasch model applied to the achievement data of a given country. A positive value indicates a relatively difficult item, and a negative value indicates a relatively easy item. The average Rasch item difficulty across all items within each country is set to zero and therefore this serves as an indication of the relative difficulty of the item within the country.

Girls/Boys Percent Correct: These statistics provide the item difficulties (DIFF) separately for girls and for boys.

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

Reliability—ScorAgr: For human-scored constructed-response items, scoring agreement is calculated as the percent agreement between two scorers on the score point value assigned to the item response.

Reliability—CodeAgr: For human-scored constructed-response items, code agreement is calculated as the percent agreement between two scorers on the scoring code assigned to the item response.

Flags: a series of flags signaling the presence of one or more conditions that might indicate an item requires further review. The flags do not necessarily signify an actual problem, but rather serve to draw attention to potential sources of concern. The following conditions are flagged:

- Point-biserial not ordered: for multiple-choice items, indicates at least one wrong distracter has a positive point-biserial correlation, or for constructed-response items worth more than one point, indicates at least one lower score on the item was associated with a higher score on the overall percent correct score.
- Difficulty less than chance: for multiple-choice items, indicates the percent correct is less than the inverse of the number of possible response selections.
- Negative/low discrimination: indicates the discrimination coefficient (DISC) is less than 0.10.
- Easier than average or Harder than average: indicates the difficulty for the item within a country is significantly lower or higher than the international average difficulty for that item, using an alpha level of 0.01.
- Less than 10% of students in a score or response category: indicates less than 10% answered the item correctly or there are less than 10% of students in a response category.
- Difficulty greater than 95%: indicates the percent correct on an item was greater than 95%.
- Scoring reliability less than 85%: indicates the percent agreement on the scores assigned to an item by the two independent human scorers is less than 85%.

Scoring Reliability for Human-Scored Items

Constructed-response items contribute at least half of the score points in any TIMSS assessment. For some 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, detailed scoring guides are developed for each constructed-response item. The scoring guides provide descriptions and examples of acceptable responses for each score code.

Chapter 7 describes scoring procedures in TIMSS for both human- and automatically-scored items and describes how human-scoring reliability is assessed and documented within-country, over time (trend), and across countries.

Item-by-Country Interactions

Although countries exhibit some variation in performance across items, in general, countries with high average achievement on the assessment tend to perform relatively well across all items, and low-achieving countries tend to perform less well. When the opposite happens (e.g., a high-performing country has low performance on an item on which other countries did relatively 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 bias achievement estimates, lead to reduced measurement error, or both (e.g. Weeks et al., 2013; von Davier et al., 2019).

Two types of statistics with graphical displays are used to detect instances of country-level DIF. The first approach is based on comparing country-specific statistics using Rasch item difficulties calculated for each country. The second approach is conducted for each country's data but based on international item parameters estimated using two- and three-parameter generalized partial credit IRT models.

Rasch Method

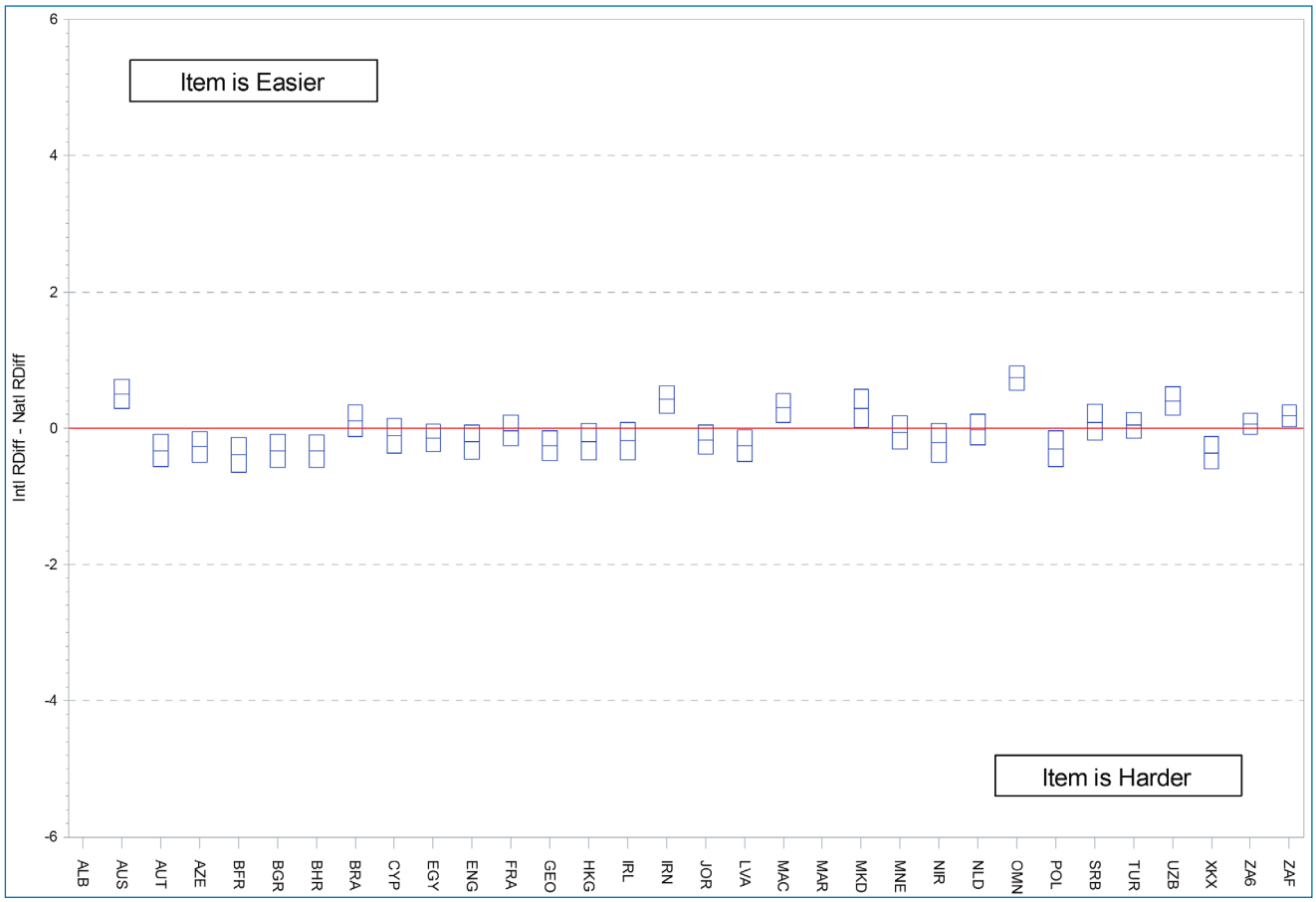
The first graphical display for a particular item, shown in Exhibit 10.1, displays the difference between each country's Rasch item difficulty and the international average Rasch item difficulty across all countries. When the absolute difference is greater than 2 logits, it is considered an *item-by-country interaction* and the item is flagged for additional review.

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

$$\begin{aligned} \text{Upper Limit} &= RDIFF_{i.} - RDIFF_{ik} + SE(RDIFF_{ik}) \cdot Z_b \\ \text{Lower Limit} &= RDIFF_{i.} - RDIFF_{ik} - SE(RDIFF_{ik}) \cdot Z_b \end{aligned} \tag{10.1}$$

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, using an alpha level of 0.05 adjusted for multiple comparisons.

Exhibit 10.1: Example Item-by-Country Rasch Plot



IRT Method

As an additional method to evaluate country DIF, international IRT parameters are used to generate model-based 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 are calculated for each item from the latent abilities estimated for each student that responds to the item. These country-level empirical ICCs are plotted alongside the international model-based ICCs (see example in Exhibit 10.2). These 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 10.2: Example Country-Level ICC Plot

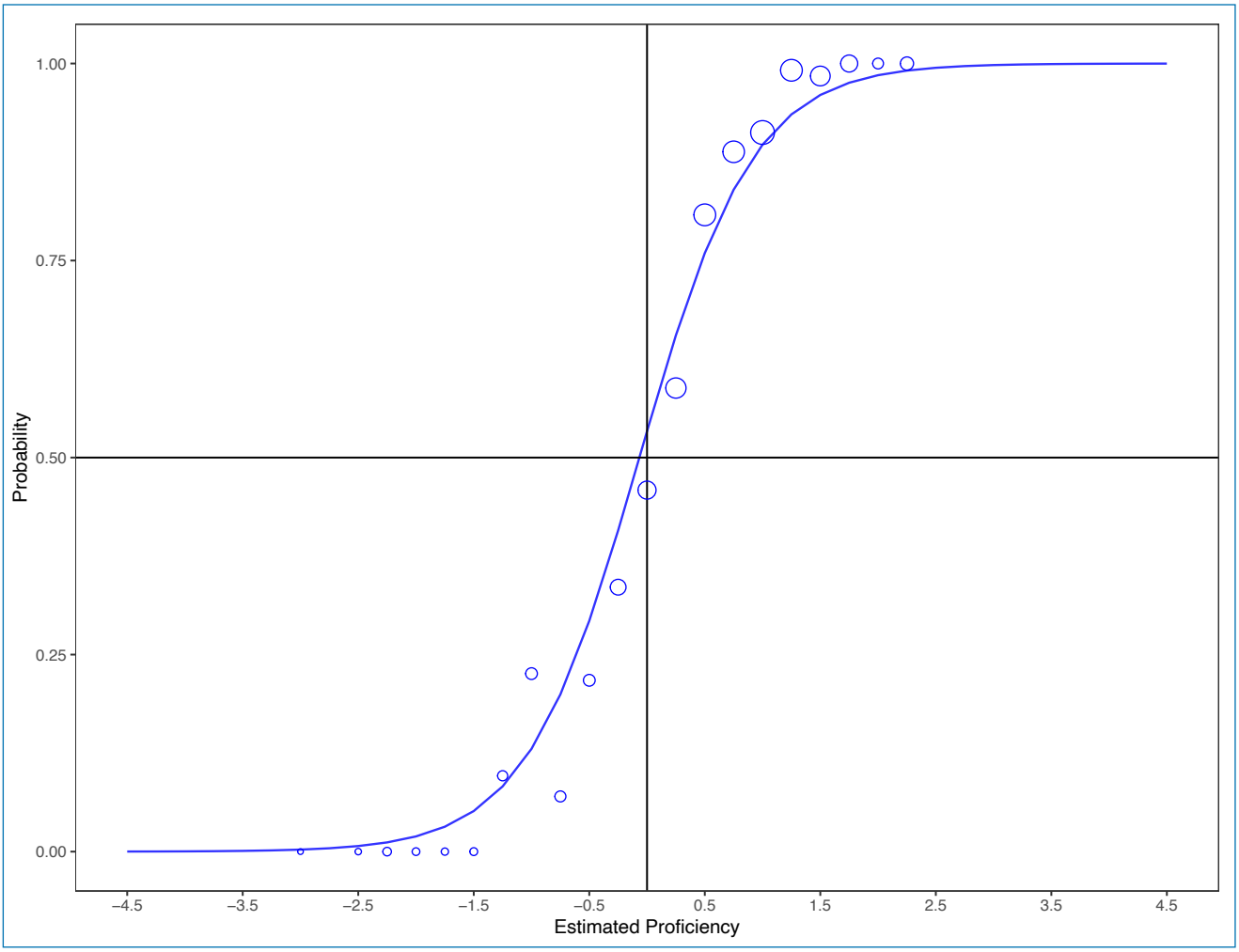


Exhibit 10.2 shows an example of a country's empirical ICC plotted with the fitted international curve. The horizontal axis (x-axis) represents the proficiency scale on the logit metric, and the vertical axis (y-axis) represents the probability of a correct response. The fitted curve based on the estimated 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 item 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, can be 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 between the empirical and the theoretical curve for the items, implying that the model with international item parameters is an accurate representation of the responses for that specific country.

The median absolute deviation (MAD) outlier detection method applied to the RMSD values is calculated for each country and item (von Davier & Bezirhan, 2022) and 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 a cut-off value. This method flags an item as a possible misfit for a country if the distance from the median of the absolute distances of all other observations exceeds a predetermined threshold (e.g., 4.5).

Review of Item Statistics for Measuring Trends

Each TIMSS assessment includes items from previous assessments in order to allow psychometric linking across assessment cycles, which enables reporting trends. Therefore, an important review step includes checking that these trend items have statistical properties in the current assessment cycle similar to those they had in the previous assessment cycle. The primary aim of reviewing these trend item statistics is to detect any unusual international-level changes in item statistics between administrations, which might indicate a problem in using the item to measure trends.

For TIMSS item review, trend item statistics are compared in tabular form for each item country-by-country. Plots are produced to visually compare changes between successive cycles in item statistics (item percent correct, Rasch difficulties, item discrimination, and nonresponse rates)—one set is produced for each item with points plotted for each country (Exhibit 10.3), and another set is produced for each country with points plotted for each item (Exhibit 10.4). Trend items are flagged for further review if systematic changes in item statistics are observed consistently across countries.

Exhibit 10.3 shows an example trend plot of countries' percent correct statistics for an item. Items with data points clustered close to the diagonal reference line indicate similar performance between assessment cycles across countries. Items where the statistics show consistent, systematic changes between cycles are flagged for further review.

Exhibit 10.3: Example Plot of Trends in Percent Correct Statistics by Item

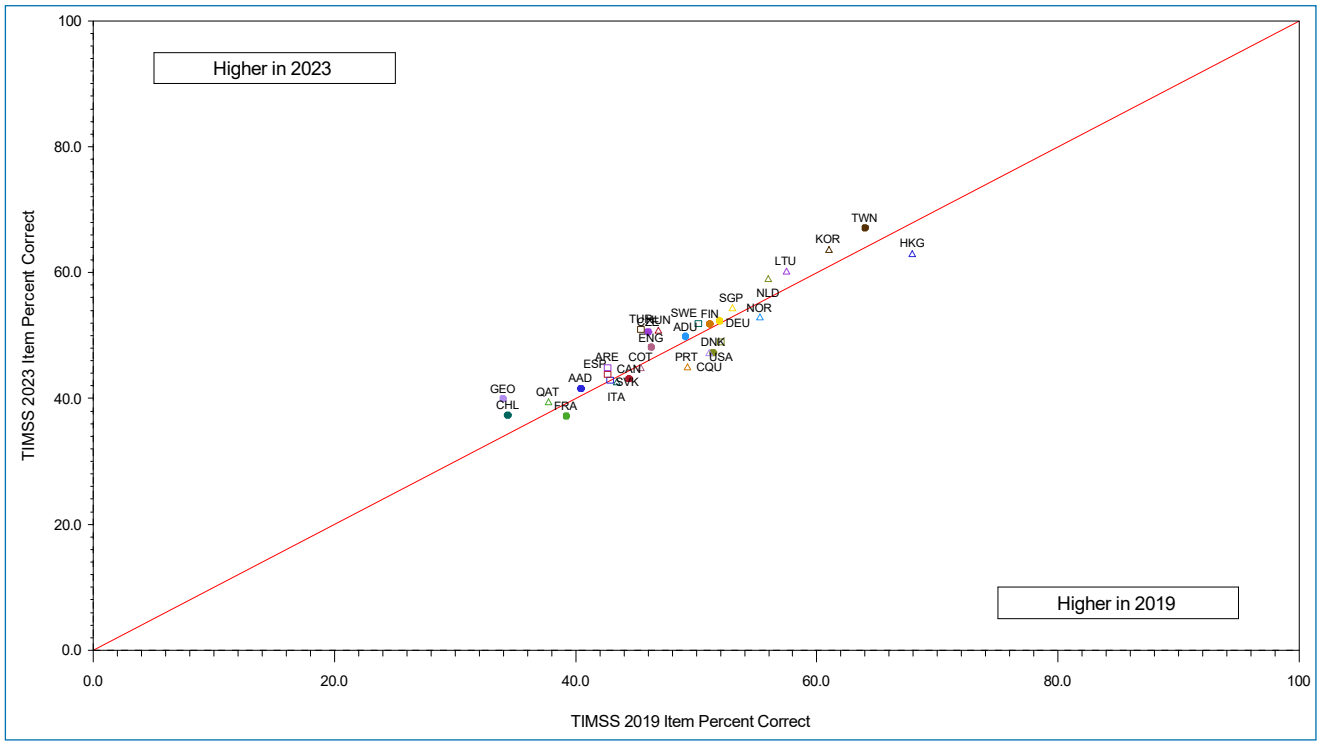


Exhibit 10.4: Example Plot of Trends in Percent Correct Statistics by Country

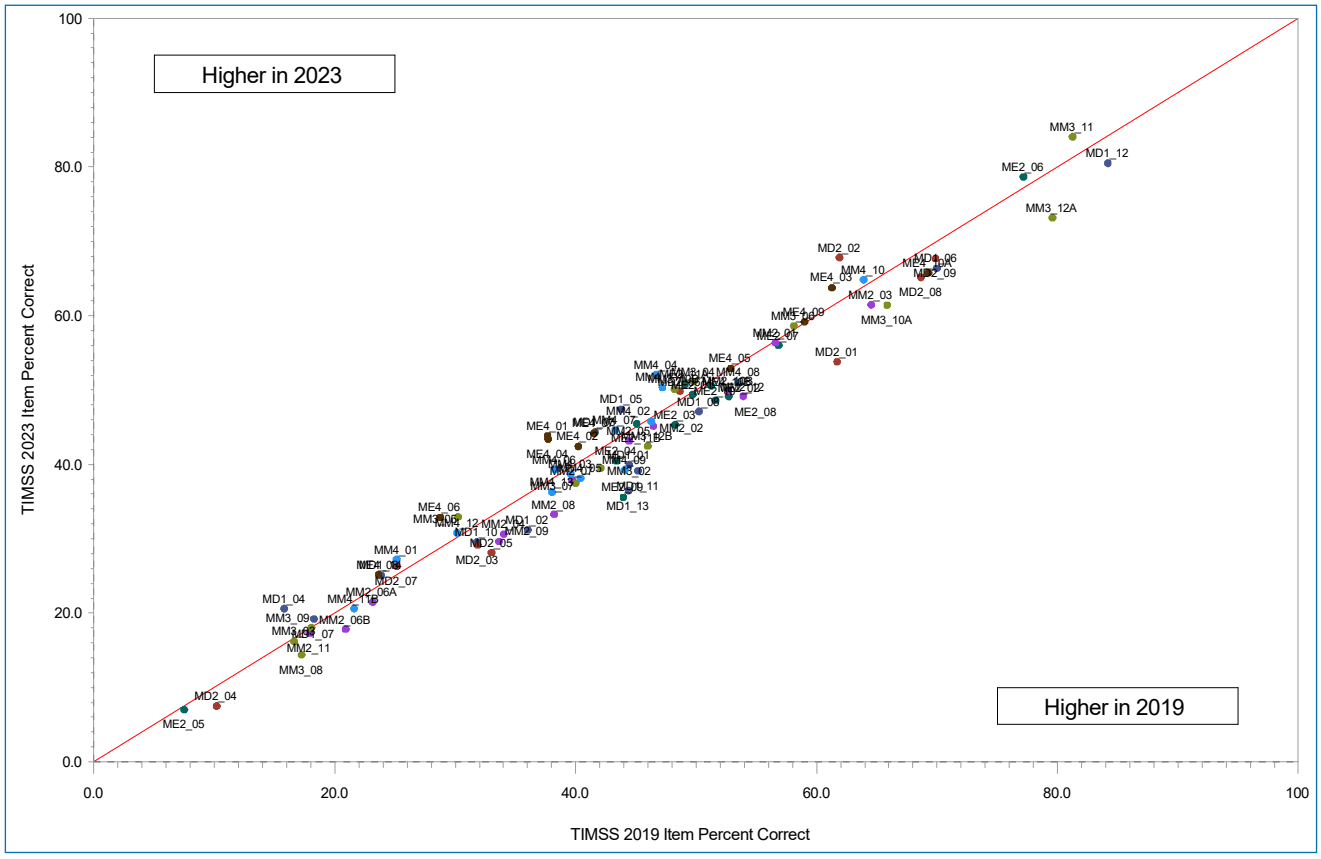


Exhibit 10.4 shows an example plot comparing trend item statistics for a particular country. At the country level, it is typical for sampling variance to cause small differences in statistics between assessments. Small, systematic changes in item difficulties for a country could be due to overall achievement that may have improved or declined. However, larger differences in item difficulty or systematic changes in item discrimination statistics could be indicative of a data problem or other issues warranting further review.

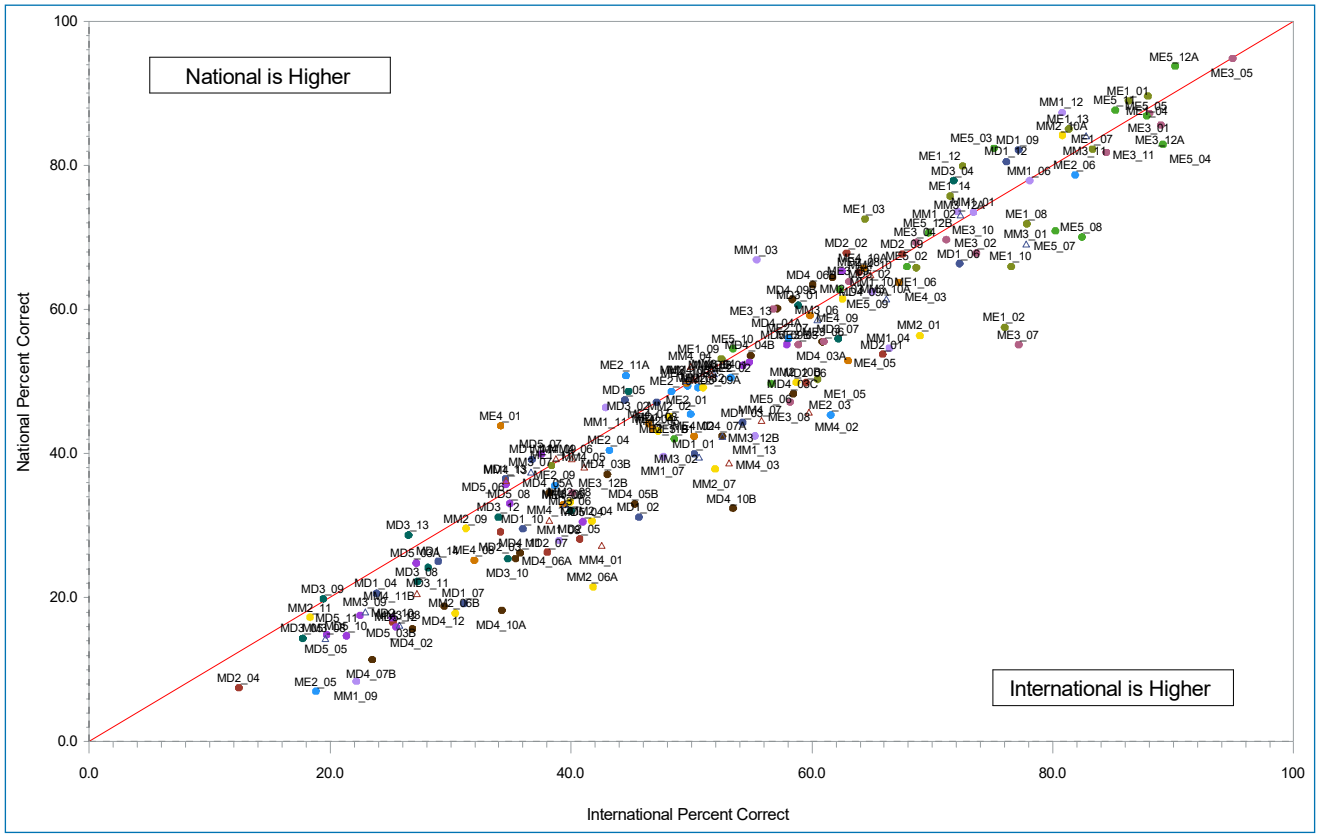
Detecting Anomalies in the TIMSS Achievement Data

To ensure that each participating country and benchmarking entity collected data adhering to TIMSS' quality standards, several analyses of item statistics at the country level are conducted. Several graphical displays are produced for each TIMSS participant: item percent correct, item point-biserial correlations, and item nonresponse rates (omitted or not reached). The graphs are examined to identify any anomalous patterns in any country's data relative to the international average or to their previous TIMSS performance, as described above. 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 is contacted to discuss the nature of the anomalies and resolve any issues.

The first set of graphical displays compares each country's performance to the international average for all items simultaneously where item performance is defined as item percent correct, item discrimination (point-biserial correlation), and item percent nonresponse. An example is shown in Exhibit 10.5 for item percent correct.

For each country, the graph plots the item percent correct of all items against their international averages for the current assessment cycle. 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 are noted for further investigation. Plots comparing national and international item discrimination and national and international nonresponse rates have similar patterns, but with data points more tightly clustered together since their range is smaller.

Exhibit 10.5: Example Plot of Item Percent Correct Across National and International by Country



These plots of national versus international item statistics are also compared against the same plots produced in the previous TIMSS cycle, as well as the trend comparison plots described earlier (see Exhibit 10.4). If the patterns for the assessments are unusually different, it might have indicated a problem in the current cycle's data. The plots are also examined separately for multiple-choice and constructed-response formatted items to ensure similar patterns. It is expected that the relationship between national and international statistics for both item types would also match that from the previous assessment cycle. In some cases, statistics may be compared for a selection of items administered in the field test one year before data collection. A large difference in item performance compared to the field test would be considered implausible and warrant further review.

TIMSS 2023 Achievement Item Review

Item statistics were produced for all countries and benchmarking participants in TIMSS 2023. The formal item review meeting took place at the TIMSS & PIRLS International Study Center at Boston College over one week in March 2024, with days devoted to each of the grade-subject combinations. Statistics for digital data were reviewed first, followed by data for the TIMSS 2023 paper options. The item statistics and graphical displays described above were produced and reviewed extensively prior to the in-person meetings to identify potential issues for group discussion. Statistics were summarized at the block level to evaluate how well targets were met for the group-adaptive assessment design.

To facilitate the TIMSS 2023 process for the digital version of the assessment, interactive Excel files were created for reviewing the item review statistics alongside IRT-based RMSD statistics. The interactive file flagged items based on various criteria, focused on identifying item-by-county interactions (DIF) using the MAD outlier detection method. The file allowed for filtering both by item and by country and with color-coded flags for easy identification of potential problems. In addition, item statistics were compared with those of TIMSS 2019 for countries that delivered both 2023 and 2019 assessments in the same mode of administration. When potential issues were flagged, multiple statistics were examined to determine if the potential issue should be investigated further.

Additional time in the meeting was dedicated to items with translation or technical issues reported by countries and to items identified as potentially problematic by study center staff during the extensive review prior to the adjudication meeting.

TIMSS 2023 Human-Scoring Scoring Reliability Outcomes

Scoring reliability for constructed-response items that required human judgment to score was considered during the item review. In addition to the within-country, trend, and cross-country reliability studies, scoring reliability was evaluated for a subset of TIMSS 2023 items using experimental artificial intelligence (AI) methods (see Chapter 7). The experimental process flagged a few potential issues prior to the item review meeting for investigation. During the item review meeting, the potential issues and findings from all data sources were discussed. NRCs were contacted to discuss and help resolve any potential issues.

Within-Country Scoring Reliability

The scoring agreement for each human-scored item in each country was examined as part of the item review process, singling out any countries where an item's scoring agreement was below 75% for further review. Appendix 10A presents the average and range of the within-country percentages of score-point agreement and diagnostic-code agreement across all TIMSS 2023 human-scored items. Score percent agreement across items was high on average across countries—94–97% on average internationally across the two subjects and two grades. There also was high agreement at the diagnostic-score level, where international average percent agreement ranged from 93% to 96%.

Trend Item Scoring Reliability Study

Steps were also taken to examine whether the TIMSS 2023 human-scored constructed-response items also used in eTIMSS 2019 were scored in the same way in both assessments. Each country with digital administration data from both cycles scored 200 responses for each of a set of items administered in both assessments. This included three mathematics items and seven science items at the fourth grade, and three mathematics items and six science items at the eighth grade.

There was a high degree of scoring consistency in TIMSS 2023. The international average exact agreement between the scores awarded in 2019 and those given by the 2023 scorers ranged from 88% to 96% across the two subjects and two grades. The average and range of scoring consistency over time can be found in Appendix 10B.

Cross-Country Scoring Reliability Study

Appendix 10C reports the results of the TIMSS 2023 cross-country scoring reliability study. Since participating TIMSS countries use many different languages in the assessment, it was not possible to establish the reliability of constructed-response scoring using responses from across all countries, and in every language. 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 each of three mathematics items and seven science items at the fourth grade, and for three mathematics items and six science items at the eighth grade.

In all, scorers from 52 fourth-grade countries and 40 eighth-grade countries that administered the digital assessment participated in the process. Making all possible comparisons among scorers gave a total of 1,326 comparisons at the fourth grade and 780 comparisons at the eighth grade for each student's response to each item. With 200 responses per item expected to be scored by each country, a maximum of 265,200 total comparisons at the fourth grade and 156,000 total comparisons at the eighth grade 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 TIMSS 2023 was high. The exact agreement between the scores awarded across countries ranged from 83% to 94% on average across the two subjects and two grades. There was similarly high agreement at the diagnostic code level.

Block Position and Booklet Effects

The TIMSS 2023 group-adaptive assessment design included the 14 mathematics and 14 science item blocks at each grade, where each item block is classified as difficult, medium, or easy depending on the average difficulty across items. Two blocks of each subject are combined into two levels of booklet difficulty. Each item block appears in two booklets, with each item block appearing in the first half of one booklet and the second half of another, paired with

another block of the same subject. In seven of the 14 booklets, two mathematics item blocks are given in Part 1 (positions 1 and 2) and two science blocks are given in Part 2 (positions 3 and 4). In the seven booklets, the science blocks appear in Part 1 and mathematics blocks appear in Part 2. When blocks of different difficulties are paired in the same booklet, the easier of the two always comes first. The TIMSS 2023 paper assessment options used a different design, with eight item blocks of each subject rotated in eight booklets.

To examine whether the particular position in which a block was administered affected student performance, item statistics were estimated for each of the four positions in which the blocks appeared in the booklet design. The results are reported in Appendix 10D for each assessment averaged across countries and for each country averaged across all items.

The results indicate minimal impact of block position on the TIMSS 2023 item statistics. On average, item blocks appearing in the second half of a booklet part (positions 2 and 4) were slightly more difficult and had slightly more omitted responses than item blocks appearing in the first half (positions 1 and 3). Similarly, items appearing in the second half of the parts had slightly higher not-reached rates than items appearing in the first half. Differences were larger for the TIMSS 2023 paper assessment options.

As an additional validation for the TIMSS 2023 group-adaptive design, nonresponse rates were produced for all countries by booklet type, along with distributions of ability estimates from the IRT model. Appendix 10E reports these results for each country and the proportion of more difficult and less difficult booklets administered in each country. For each subject and grade, average nonresponse rates were higher for more difficult booklets compared to 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.

Item Review Outcomes

Using all the information from the comprehensive collection of item analyses and reliability data that were computed and summarized for TIMSS 2023, 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. In particular, the following observations led to items being considered for possible deletion from the international database:

- A translation or localization error was detected for a particular country, but was not corrected before test administration
- A technical issue encountered by a country was reported that made the item challenging or impossible to answer during test administration
- The item had zero or negative discrimination, indicating it does not provide information to distinguish between low- and high-ability students
- For multiple-choice items, the item review revealed a faulty distracter paired with overall low discrimination

- 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 60%
- For trend items, an item performed substantially differently in 2023 compared to the eTIMSS 2019 administration

Multiple sources of evidence were required to warrant deleting the data for a particular item. When the item statistics indicated a problem with an item, the translation verification documentation was used to check the test booklets. However, if a question remained about potential translation or cultural issues, the NRC was consulted before deciding how the item should be treated.

Reviewing the TIMSS 2023 achievement data resulted in detecting a small number of items that were inappropriate for international comparisons. Among the few items singled out in the review process, most were removed due to differences attributable to translation problems, which were detected by an internal review of country adaptations or using the MAD outlier method. Very few items were identified as having severe differential item functioning after item review during IRT scaling. Score codes for some constructed-response items were recoded if the point-biserial correlations did not show the expected level of association with overall achievement.

Appendix 10F includes a list of deleted items and a list of recodes made to constructed-response items.

There also were some items in the assessments that were combined, or derived, for scoring purposes. See Appendix 10G for details about how score points were defined for each derived item.

References

von Davier, M., & Bezirhan, U. (2022). A robust method for detecting item misfit in large-scale assessments. *Educational and Psychological Measurement, 7*(2), 110–114. https://doi.org/10.1080/15366360903117079

von Davier, M., Yamamoto, K., Shin, H. J., Chen, H., Khorramdel, L., Weeks, J., & Kandathil, M. (2019). Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assessment in Education: Principles, Policy & Practice, 26*(4), 466–488. https://doi.org/10.1080/0969594X.2019.1586642

Weeks, J. P., von Davier, M., & Yamamoto, K. (2013). Design considerations for the Program for International Student Assessment. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 259–275). Boca Raton, FL: CRC Press.

- Appendix 10A: TIMSS 2023 Within-Country Scoring Reliability for Human-Scored Items
- Appendix 10B: TIMSS 2023 Trend Scoring Reliability for Human-Scored Items
- Appendix 10C: TIMSS 2023 Cross-Country Scoring Reliability for Human-Scored Items
- Appendix 10D: TIMSS 2023 Item Statistics by Booklet Position
- Appendix 10E: TIMSS 2023 Group Adaptive Design Outcomes
- Appendix 10F: Modifications to the TIMSS 2023 Achievement Data
- Appendix 10G: Derived Items in TIMSS 2023