CHAPTER 12

TIMSS 2023 Achievement Scaling Implementation

Bethany Fishbein
Liqun Yin
Ummugul Bezirhan

Introduction

Since 1995, TIMSS has been designed to provide international measures of students' mathematics and science achievement and to measure trends in achievement over time. TIMSS is based on broad definitions of mathematics and science achievement, recognizing different content areas within the subjects and covering a wide range of topics at each grade level assessed. The TIMSS assessments include items and tasks of varying contexts and difficulty levels accessible to students of wide-ranging abilities. Given this broad coverage, TIMSS uses a matrix-sampling booklet design, such that each student is administered only a subset of the entire TIMSS mathematics and science item pool.

TIMSS relies on item response theory (IRT) scaling to provide accurate measures of student proficiency distributions and trends. To provide unbiased estimates of student achievement and its relationship to contextual variables, the TIMSS psychometric analysis approach relies on latent regression population models with subsequent multiple imputations to obtain plausible values representing proficiency in mathematics and science.

This chapter describes the procedures for scaling the TIMSS 2023 achievement data. The TIMSS & PIRLS International Study Center implemented the psychometric analysis that includes IRT calibration and linking, population modeling, and imputation of plausible values of the TIMSS 2023 achievement data and conducted related analyses to ensure the quality and validity of the results. A detailed description of the TIMSS 2023 psychometric methodologies can be found in Chapter 11.

Consistent with previous assessments, the TIMSS 2023 psychometric analyses were based on a concurrent calibration of the TIMSS 2023 data with data from the previous TIMSS 2019 cycle. Plausible values (PVs) were imputed for all students in overall mathematics and science, in each of the content and cognitive subdomains, and in environmental knowledge.

TIMSS 2023 completed the transition of TIMSS from paper-and-pencil to digital format that began in TIMSS 2019 (Fishbein et al., 2018; Foy et al., 2020; von Davier et al., 2020) and

https://doi.org/10.6017/lse.tpisc.timss.rs4586

adopted a group-adaptive assessment design for the digital assessment to address the need for a broader range of assessment difficulty and more precise targeting of student ability. The 2023 cycle of TIMSS is the first fully digital assessment with new items developed only for computer-based administration. Six of the 14 item blocks in each subject and grade were developed and field tested for first-time use in TIMSS 2023 (see Chapter 1). Eight blocks were administered previously in digital format as part of eTIMSS 2019. These provide the basis for trend measurement.

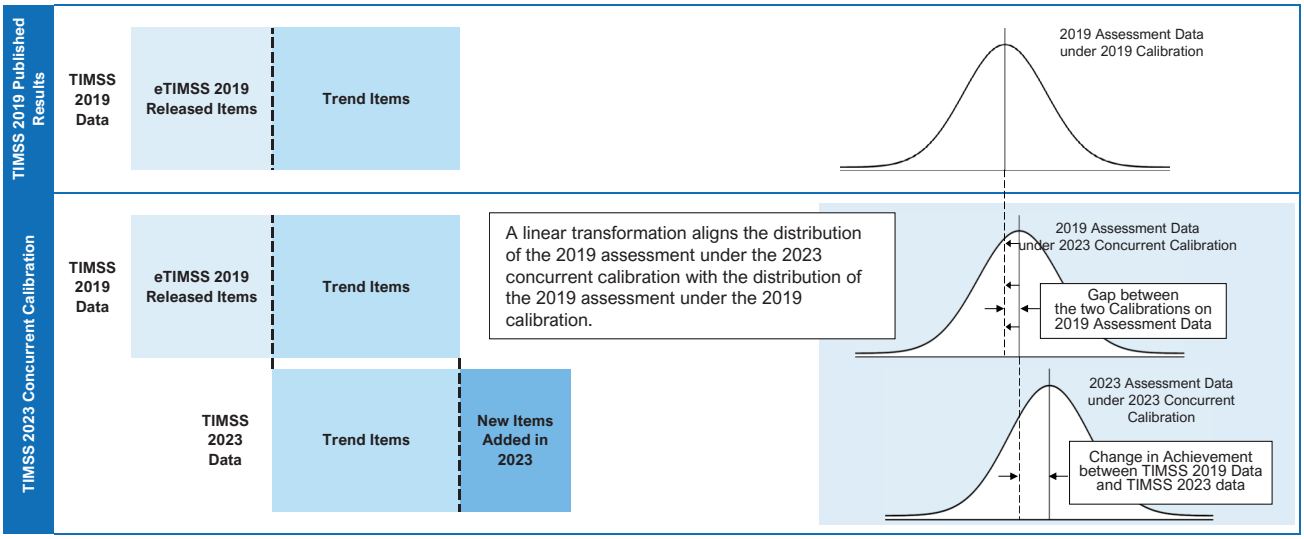
A few countries at each grade (three at the fourth grade and four at the eighth grade) administered the TIMSS Paper assessment version, which included only paper trend blocks from the 2019 assessment. Four of the fourth-grade countries administered the TIMSS Less Difficult Paper option, which contained some blocks composed of easier mathematics items. Achievement estimates for the TIMSS 2023 paper-based data were produced as additional, separate steps to the TIMSS 2023 psychometric analyses. The 2023 paper-based data is also reported on the same metric as the 2023 digital-based data.

Scaling the TIMSS 2023 Achievement Data

The TIMSS reporting metric was originally established for each grade and subject by setting the mean of the national average scores across all countries that participated in TIMSS 1995 to 500 and the standard deviation to 100. Successive framework updates and item release policies changed the composition of subsequent TIMSS rounds, yet the reporting scale was maintained utilizing an assessment design that provides a strong linkage over time. To enable measurement of trends over time, achievement data from successive TIMSS assessments were transformed to that same metric by linking each new data set to the immediate predecessor's scale. This was done by analyzing the data from each successive assessment, here 2023, with the data from the previous assessment, here 2019—a process known as concurrent calibration—and applying linear transformations to place the results from each successive assessment on the same scale as the results from the previous assessment. This means that TIMSS 2023 is linked to 1995 through a chain of linkages involving concurrent IRT calibrations that account for change in item selection and domain coverage over time due to assessment framework updates.

Exhibit 12.1 illustrates the general structure of the TIMSS 2023 concurrent calibration model. In linking the two successive assessments, concurrent calibration relies on retaining a large proportion of items from one assessment to the next ("trend items"). It is then possible to estimate the latent ability distributions of students in both assessments on a common scale. The difference between the two assessment distributions is the trend measure between them, although not yet on the TIMSS trend reporting metric until a set of transformations is applied.

Exhibit 12.1: TIMSS 2023 Concurrent Calibration Model



TIMSS 2023 item parameters were estimated for all digital items through the concurrent calibration of the data from the eTIMSS 2019 and TIMSS 2023 assessments. After item calibration, the TIMSS 2023 data underwent further analysis steps using latent regression models (e.g. Mislevy, 1984; Mislevy & Sheehan, 1987; von Davier et al., 2006; von Davier et al., 2009) to impute plausible values (PVs) for overall mathematics and science achievement, as well as for the content and the cognitive domains, plus an environmental knowledge subscale. The use of a latent regression IRT model to impute PVs is sometimes called *conditioning*, and contextual data from the student and parent questionnaires were used as predictors of achievement to increase the reliability of the imputations. Finally, the PVs were placed on the TIMSS trend reporting metric through a series of linear transformations.

The TIMSS 2023 achievement scaling implementation consisted of four major analysis phases conducted separately for each grade and subject:

1. **Item calibration**: In the first phase, the parameters of all TIMSS 2023 and eTIMSS 2019 digital items were estimated using multiple-group IRT models.
2. **Principal component analysis**: In the second phase, principal components were extracted from context data for each country and benchmark participants for use in population modeling.
3. **Latent regression population modeling**: In the third phase, latent regression models were estimated (conditioning) for each country's data to draw PVs of achievement.
4. **Scale transformation**: Finally, the imputed PVs were placed on the TIMSS reporting metric using linear transformations to report trends from previous assessments.

The TIMSS 2023 psychometric analysis procedures are described under four subsections according to these phases. Plausible values of achievement were imputed for all students in overall mathematics and overall science, as well as for several subdomain scales. At the fourth

grade, PVs were imputed for achievement in three mathematics content subdomains (Number, Measurement and Geometry, and Data) and three science content subdomains (Life Science, Physical Science, and Earth Science). At the eighth grade, estimates were produced for four mathematics content subdomains (Number, Algebra, Geometry and Measurement, and Data and Probability) and four science content subdomains (Biology, Chemistry, Physics, and Earth Science). At both grades, achievement estimates were produced for three cognitive domains (Knowing, Applying, and Reasoning). Finally, as a special environmental awareness initiative in TIMSS 2023 continued from TIMSS 2019 (Yin & Foy, 2021), PVs were imputed for fourth- and eighth-grade students in Environmental Knowledge.

Several quality checks and analyses were conducted iteratively throughout the analysis process. These analyses and their outcomes are described later in the "Validating the TIMSS 2023 Achievement Results" section of this chapter. IRT models, population models, and other theoretical foundations for the psychometric analysis procedures are described in Chapter 11.

Before the IRT calibration for the TIMSS 2023 achievement data, TIMSS conducted an extensive item-by-item review of classical item statistics for all countries to evaluate the quality of the assessment items and to identify any unexpected or problematic item behaviors. This review also included analyses of change with respect to percent correct and partial credit percentages, omit rates, item discrimination, and other classical item statistics for trend items relative to the 2019 assessment. These item review activities are described in Chapter 10.

Treatment of Item-Level Nonresponse (Omitted and Not-Reached)

Given the matrix-sampling design used by TIMSS, whereby a student is administered only a subset of the 14 item blocks in each subject, most item responses are missing by design for any given student. Students were assigned booklets randomly according to the design described in the TIMSS 2023 framework, so that the missing data introduced due to this process is ignorable (Little & Rubin, 1987; Rubin, 1976) in the analysis. However, missing data can also result from a student not answering an item, which can occur when the student receives the item but does not provide an answer, omits the item by mistake, needs more time to attempt the item, or other reasons. TIMSS considers an item to be "not reached" when—within the first or second part of a booklet—the item itself and the item immediately preceding it are not answered, and there are no other items completed in the remainder of that part of the booklet. All other skipped responses are considered "omitted."

The TIMSS & PIRLS International Study Center introduced a new mechanism for the treatment of item nonresponse for the psychometric analyses of the TIMSS 2023 achievement data. This approach is based on the strength of the evidence of missing data for estimating achievement. It assumes that item nonresponse occurs not at random but does not assume it occurs exclusively due to the low ability of the students. This strength-of-evidence approach avoids potential bias due to treating nonresponse deterministically as if all missing responses could be considered incorrect due to lack of knowledge. Research has shown that treating

omitted responses as incorrect leads to bias, in particular, to underestimation of achievement (Glas & Pimentel, 2006; Moustaki & Knott, 2000; Rose et al., 2010, 2017).

In the TIMSS 2023 analyses, both omitted and not-reached responses were ignored for estimating item parameters. To impute plausible values, nonresponse indicators were created to account for the non-randomly missing item responses, and their IRT parameters were estimated. The variables indicated whether each student answered all items (1) or had at least one missing response (0). A set of nonresponse indicators was created and used for each analyzed scale, including one for each of the subscales within each subject.

To mitigate any potential effects of the nonresponse indicators on item parameter estimation, a stepwise approach was adopted for estimating achievement item parameters and nonresponse indicators. First, only achievement items were included in the calibration. In the second step, a three-parameter (3PL) IRT model was applied to estimate the parameters for the nonresponse indicators, with all achievement item parameters fixed to the values estimated in the first step. This second step was carried out separately for the content domain indicators, the cognitive domain indicators, and the environmental knowledge indicators, respectively.

Parameters for the nonresponse indicator variables were estimated, treating them as common between TIMSS 2023 and eTIMSS 2019, except for the eighth-grade mathematics scale indicators. For the mathematics scales indicators, larger differences were observed in item nonresponse rates between the 2019 and 2023 cycles. Therefore, separate sets of nonresponse indicator parameters were estimated for the 2019 and 2023 eighth-grade data to mitigate any potential threats to the comparability of the results.

The nonresponse indicators were included as item response variables alongside achievement items for imputing plausible values. Specifically, the content domain nonresponse indicators were used for imputing PVs in overall mathematics and science as well as for the respective content subdomains, the cognitive domain indicators were used for imputing PVs for the cognitive subdomains, and the environmental indicators were used for imputing PVs for the environmental knowledge scale, along with the physical science indicator at the fourth grade and the chemistry and physics indicators at the eighth grade. This approach accounted for nonresponse in imputing PVs according to the dependency between missingness and achievement or to the extent that missingness and achievement are (negatively) correlated in the population.

For TIMSS 2023 paper options, the corresponding TIMSS 2019 paper item parameters were fixed for proficiency estimation, and the treatment of item-level nonresponse followed the same procedure as in TIMSS 2019 (Foy et al., 2020). When PVs were imputed, both not-reached and omitted item responses were treated as incorrect to be consistent with the 2019 data processing used to estimate the 2019 parameters, which is necessary to produce comparable trend results.

Phase 1: Item Calibration

Item calibration for TIMSS 2023 was conducted using the MIRT package (Chalmers, 2012) in the R statistical programming language (R Core Team, 2024). To meet the analytic goals for the TIMSS 2023 achievement data outlined in the overview, concurrent calibration IRT models were employed separately for each grade and subject to estimate the item parameters. Relying on the usual TIMSS concurrent calibration approach extended for multiple populations, data from TIMSS 2023 were scaled along with eTIMSS 2019 data to estimate item parameters for the items in both assessments.

Exhibits 12.2 through 12.5 show the number of items included in the TIMSS 2023 concurrent calibration by content and cognitive domain for both grades and subjects. For each grade and subject, one of the eight "trend" blocks containing PSIs developed in 2019 was not treated as common between cycles because they were an experimental extension of 2019 and not reported as part of the main TIMSS 2019 achievement scales (Fishbein & Foy, 2021). Instead, in each concurrent calibration model, the items in PSI block were treated as if they were new in 2023 and not treated as common between cycles.

Exhibit 12.2: Items for the TIMSS 2023 Concurrent Calibration – Grade 4 Mathematics

Domain	Items Unique in 2019		Items Common in 2019 and 2023		Items Unique in 2023		Total	
	Items	Points	Items	Points	Items	Points	Items	Points
Mathematics	87	93	84	90	99	102	270	285
Items by Content Domain								
Number	43	46	40	42	54	56	137	144
Measurement and Geometry	29	32	23	24	26	27	78	83
Data	15	15	21	24	19	19	55	58
Items by Cognitive Domain								
Knowing	31	31	28	28	30	30	89	89
Applying	36	38	38	42	47	49	121	129
Reasoning	20	24	18	20	22	23	60	67

Exhibit 12.3: Items for the TIMSS 2023 Concurrent Calibration – Grade 4 Science

Domain	Items Unique in 2019		Items Common in 2019 and 2023		Items Unique in 2023		Total	
	Items	Points	Items	Points	Items	Points	Items	Points
Science	86	88	82	85	91	100	259	273
Items by Content Domain								
Life Science	37	39	35	37	44	48	116	124
Physical Science	32	32	29	30	32	36	93	98
Earth Science	17	17	18	18	15	16	50	51
Items by Cognitive Domain								
Knowing	36	38	33	35	36	40	105	113
Applying	31	31	32	32	40	43	103	106
Reasoning	19	19	17	18	15	17	51	54
Environmental Knowledge Items								
Environmental Knowledge	19	19	19	19	25	26	63	64

Exhibit 12.4: Items for the TIMSS 2023 Concurrent Calibration – Grade 8 Mathematics

Domain	Items Unique in 2019		Items Common in 2019 and 2023		Items Unique in 2023		Total	
	Items	Points	Items	Points	Items	Points	Items	Points
Mathematics	103	108	103	109	97	98	303	315
Items by Content Domain								
Number	32	33	31	33	32	32	95	98
Algebra	28	29	33	33	25	26	86	88
Geometry and Measurement	21	23	22	26	20	20	63	69
Data and Probability	22	23	17	17	20	20	59	60
Items by Cognitive Domain								
Knowing	33	34	31	32	29	29	93	95
Applying	50	52	46	47	45	45	141	144
Reasoning	20	22	26	30	23	24	69	76

Exhibit 12.5: Items for the TIMSS 2023 Concurrent Calibration – Grade 8 Science

Domain	Items Unique in 2019		Items Common in 2019 and 2023		Items Unique in 2023		Total	
	Items	Points	Items	Points	Items	Points	Items	Points
Science	106	114	105	119	107	111	318	344
Items by Content Domain								
Biology	37	44	38	45	38	39	113	128
Chemistry	19	19	23	27	20	20	62	66
Physics	29	30	23	24	25	27	77	81
Earth Science	21	21	21	23	24	25	66	69
Items by Cognitive Domain								
Knowing	37	38	38	43	30	30	105	111
Applying	41	46	39	46	52	54	132	146
Reasoning	28	30	28	30	25	27	81	87
Environmental Knowledge Items								
Environmental Knowledge	22	23	22	24	36	37	80	84

Exhibits 12.6 and 12.7 show the sample sizes for the TIMSS 2023 concurrent calibration. The data from TIMSS 2023 trend calibration countries was combined with that of eTIMSS 2019, including only data from eTIMSS 2019 for those countries that also participated in TIMSS 2023. All student samples were weighted so that each country, within each assessment year, contributed equally to the item calibration. Cases were included in the calibration as long as they had at least one valid response in the corresponding subject. At the fourth grade, 37 TIMSS 2023 countries that also participated in TIMSS 2019 contributed data to the concurrent calibration, with 25 contributing data from both cycles. At the eighth grade, 31 countries that participated in both TIMSS 2023 and TIMSS 2019 contributed data to the concurrent calibration, with 20 countries providing comparable data from both cycles. Benchmarking participants did not contribute data to the calibration for either assessment cycle.

↓ Exhibit 12.6: Sample Sizes for the TIMSS 2023 Concurrent Calibration – Grade 4

↓ Exhibit 12.7: Sample Sizes for the TIMSS 2023 Concurrent Calibration – Grade 8

A multiple-group IRT model was utilized for item calibration, specifying country groups, resulting in 37 groups for each of the fourth-grade concurrent calibration models and 31 groups

for each of the eighth-grade models. Country groups were formed by combining data from across years, when available. While the item parameters were estimated to be equal across groups, the model allowed for estimating distinct ability distributions by country, to account for achievement differences between them properly.

Several types of IRT item functions were used simultaneously for the concurrent calibration. Multiple-choice items used in the TIMSS 2019 assessment were calibrated using the 3PL model. New multiple-choice items introduced in TIMSS 2023 were initially calibrated using the two-parameter logistic (2PL) model. Those that showed misfit were then recalibrated using the 3PL model. All other 1-point items were calibrated using the 2PL model, and polytomous items worth up to 2 points used the generalized partial credit model (GPCM).

The recent reduction in the use of the 3PL model for TIMSS is driven by many critical considerations, primarily identification issues inherent in the 3PL model due to its reliance on lower asymptote parameters, which often leads to non-unique solutions and potential biases in parameter estimation. The difficulty of estimating the lower asymptote parameter has been repeatedly highlighted in the literature (e.g., Kang & Cohen, 2007; Lord, 1968, 1975; Thissen & Wainer, 1985; von Davier, 2009; Whittaker et al., 2012) and was often observed in past TIMSS operational analysis. This behavior is especially prominent for items that are relatively easy or have low discriminatory power where there is not enough information to estimate the lower asymptote. Moreover, the 3PL model tends to produce monotone likelihoods because of the non-zero lower asymptote of each item. Consequently, this can lead to infinite ability estimates, even for students with non-extreme response patterns (von Davier, 2023).

The item parameters estimated from the concurrent calibration are presented in Appendices 12A to 12D. The estimated parameters for nonresponse indicators are presented in Appendix 12E.

The parameters resulting from this calibration were then used to estimate student proficiency for all countries and benchmarking entities participating in TIMSS 2023.

Phase 2: Principal Component Analysis

The second phase of the TIMSS 2023 psychometric analyses involved creating principal components for use in conditioning. Conditioning refers to applying a latent regression model that includes all available contextual information to improve the statistical properties of the estimated student proficiency plausible values. Ideally, all student-level contextual data would be included in the conditioning model, but because TIMSS has so many context variables that could be used in conditioning, TIMSS follows the practice established by NAEP and widely adopted in other large-scale studies of using principal component analysis (PCA) to reduce the number of variables necessary to represent variance in response data. Principal components for the TIMSS student context variables, including parent context variables at the fourth grade, were constructed as follows:

- Categorical variables with fewer than eight response options were dummy-coded to represent all response options, including responses coded as "not administered," "not applicable," and "omitted."
- Context variables with numerous response options (such as year of birth) were recoded using criterion scaling (Beaton, 1969; Beaton & Barone, 2017). This was done by replacing the response value with the mean interim achievement score of all students with that response value. Criterion scaling maximizes the correlation between the scaled variable and achievement. For TIMSS, the interim achievement score was the student-level average of the mathematics and science EAP scores produced by the item calibrations.
- Separately for each country, all the dummy-coded and criterion-scaled variables were included in a principal component analysis. The first principal components that accounted for 90% of the variance were initially retained as conditioning variables. Because the principal component analysis was performed separately for each country and benchmarking entity, different numbers of principal components were required to account for 90% of the common variance in each country's context variables. As an additional step, the number of principal components retained was trimmed not to exceed 5% of a country's unweighted student sample size.

In addition to the principal components, students' gender (dummy coded), the language of the test (dummy coded), an indicator of the classroom in the school to which a student belongs (criterion scaled), and an optional country-specific variable (dummy coded) were included as primary conditioning variables.

Exhibits 12.8 and 12.9 provide details on the conditioning variables used for proficiency estimation of the TIMSS 2023 data.

↓ Exhibit 12.8: Conditioning Variables Used for the TIMSS 2023 Data – Grade 4

↓ Exhibit 12.9: Conditioning Variables Used for the TIMSS 2023 Data – Grade 8

Phase 3: Latent Regression Population Modeling

Educational Testing Service's MGROUP programs (Rogers et al., 2006; Sheehan, 1985) were used to estimate the latent regression model and to impute plausible values for the TIMSS 2023 data. These programs take as input the students' responses to the items, the item parameters estimated at the calibration stage, and the conditioning variables. The program generates as output the estimated regression effects and the residual variance-covariance matrix, as well as imputed PVs that represent the posterior distribution of student proficiency given their achievement and contextual data (e.g., Mislevy, 1991; Thomas, 1993; von Davier et al., 2006; von Davier & Sinharay, 2013). More details on the latent regression model are available in Chapter 11.

Certain versions of the MGROUP set of programs allow multi-dimensional latent regression models using responses to all items across the proficiency scales and the correlations among the scales to improve the reliability of each individual scale. The multi-dimensional modeling feature implemented in MGROUP was used to impute PVs simultaneously for the overall mathematics and science scales using a two-dimensional model. A multi-dimensional model also was used to impute PVs separately for the content and cognitive subscales, and the environmental knowledge scale. The imputation of these PVs for the subscales relied on multi-dimensional IRT models using the item parameters estimated for the overall mathematics and science scales, as well as the same set of conditioning variables. In addition to dimensions for each of the subdomains, an additional dimension was included for the other overall subject.

Population models were estimated separately for each TIMSS 2023 country and benchmarking participant, and each eTIMSS 2019 country included in the calibration sample described in Phase 1. The latter ones were used to calculate the transformation constants to apply to the 2023 data to convert them from the logit metric to the reporting metric.

Phase 4: Scale Transformation

To provide results for the TIMSS 2023 assessments on the existing TIMSS achievement scales, the 2023 plausible values had to be transformed onto the existing TIMSS reporting metric. This process involved linear transformations of the PVs derived from the TIMSS data. The transformation constants for TIMSS 2023 were derived by finding the linear transformation equation that would convert the results obtained in Phase 3 for the eTIMSS 2019 countries onto their reported results in 2019. Only trend calibration countries were used for the calculation of the transformation constants. The equations used to derive transformation constants and to transform PVs are provided in Chapter 11. Separate transformation constants were calculated for each grade and subject and one for each of the five PVs.

Exhibits 12.10 and 12.11 show the TIMSS 2023 transformation constants for the TIMSS 2023 data. The same transformation constants were applied to all content and cognitive domain scales within a grade and subject, as well as the environmental knowledge scales.

Exhibit 12.10: Scale Transformation Constants for the TIMSS 2023 Data – Grade 4

Overall Mathematics	TIMSS 2019 Published Results		TIMSS 2019 Re-Scaled Results		A_{ik}	B_{ik}
	Mean	Standard Deviation	Mean	Standard Deviation		
PV1	527.90635	88.22644	-0.00622	1.04252	528.43303	84.62818
PV2	527.99506	88.52005	-0.00659	1.04232	528.55456	84.92619
PV3	528.44440	87.60418	-0.00576	1.04241	528.92879	84.03970
PV4	527.61923	88.35479	-0.00710	1.04254	528.22090	84.74995
PV5	527.29574	88.46631	-0.00284	1.04200	527.53716	84.90065

Overall Science	TIMSS 2019 Published Results		TIMSS 2019 Re-Scaled Results		A_{ik}	B_{ik}
	Mean	Standard Deviation	Mean	Standard Deviation		
PV1	522.57251	83.27920	-0.27084	0.81300	550.31537	102.43423
PV2	521.18142	83.64783	-0.26741	0.81296	548.69575	102.89329
PV3	521.30883	84.00746	-0.27144	0.81399	549.32287	103.20462
PV4	520.56156	84.18998	-0.26708	0.81371	548.19466	103.46427
PV5	522.41191	83.49539	-0.26686	0.81221	549.84555	102.80055

Exhibit 12.11: Scale Transformation Constants for the TIMSS 2023 Data – Grade 8

Overall Mathematics	TIMSS 2019 Published Results		TIMSS 2019 Re-Scaled Results		A_{ik}	B_{ik}
	Mean	Standard Deviation	Mean	Standard Deviation		
PV1	513.28507	99.69372	0.39222	0.81144	465.09685	122.86029
PV2	513.78805	100.64449	0.39248	0.81085	465.07185	124.12246
PV3	514.12696	101.27431	0.39316	0.81052	465.00145	124.95029
PV4	513.19556	101.64155	0.39319	0.81136	463.93904	125.27371
PV5	514.07390	101.13419	0.39156	0.81146	465.27297	124.63187

Overall Science	TIMSS 2019 Published Results		TIMSS 2019 Re-Scaled Results		A_{ik}	B_{ik}
	Mean	Standard Deviation	Mean	Standard Deviation		
PV1	512.25807	97.84370	0.13803	0.82464	495.88090	118.65006
PV2	511.99976	97.60301	0.14031	0.82498	495.40002	118.30916
PV3	512.99434	97.13450	0.13750	0.82381	496.78205	117.90938
PV4	511.64925	98.46462	0.13904	0.82334	495.02153	119.59157
PV5	512.60367	97.95987	0.14164	0.82391	495.76372	118.89645

The above transformation constants were also applied to the re-scaled 2019 plausible values to evaluate the variation of the linking adjustment between the 2023 and 2019 psychometric analysis models (see results in the "Validating the TIMSS 2023 Achievement Results" section).

For the TIMSS 2023 Paper assessment results, the scale transformation constants used for the TIMSS 2019 paper data were used to transform the corresponding plausible values onto the reporting metrics (Foy et al., 2020).

Validating the TIMSS 2023 Achievement Results

The psychometric analysis of the TIMSS 2023 achievement data included extensive steps throughout the process to ensure the quality of the results. In this section, two major aspects of the analysis are addressed:

- Evaluating item fit to the TIMSS 2023 assessment data
- Examining the variation in the trend linking error across countries

Evaluating Item Fit to the TIMSS 2023 Assessment Data

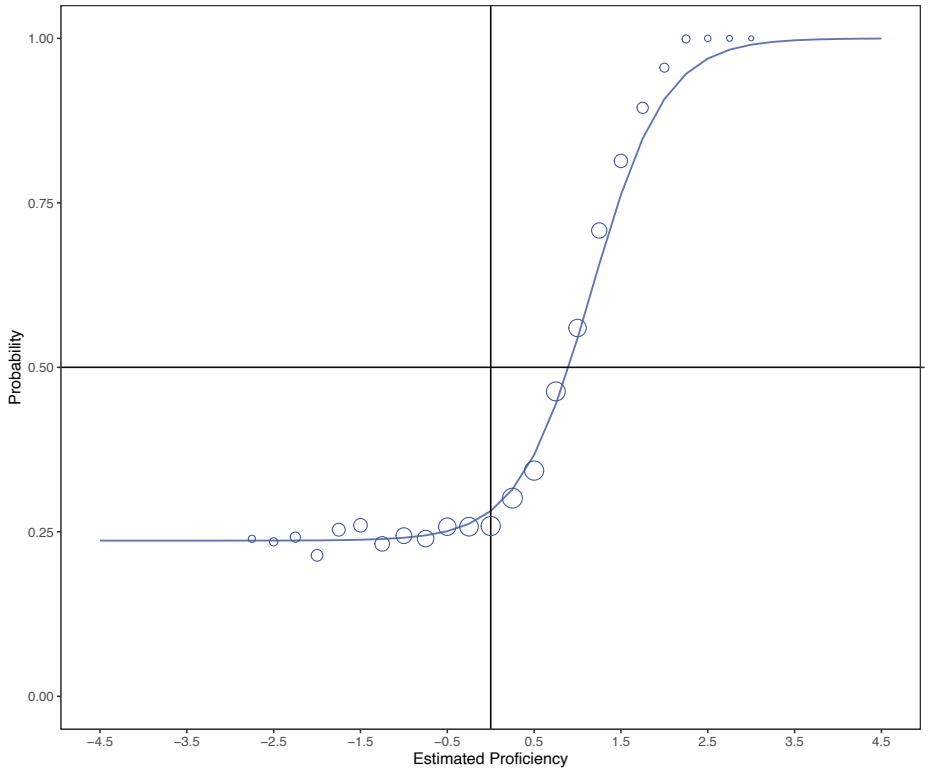
To evaluate the fit of the item parameters to the response data, a series of IRT-based checks were performed during the item calibration phase. These included examining graphical displays of item characteristic curves (ICCs) to check the empirical and fitted item response functions and to compare the empirical curves for trend items between the 2023 and 2019 cycles. In addition, quantitative inspections were conducted with the root mean square difference (RMSD) statistic.

Item Characteristic Curves

Item fit was assessed by visually comparing the item response function curves generated using the item parameters estimated from the data with the empirical item response curves calculated from the response data. The empirical functions are themselves based on an estimated latent ability distribution that uses the IRT model and, therefore, are also referred to as item functions based on pseudo counts. When the empirical results for an item fall near the fitted curves, the IRT model fits the data well and provides an accurate and reliable measurement of the underlying proficiency scale.

Plots of these response function curves are called item characteristic curves (ICC). ICC plots were examined for all TIMSS 2023 items. The plot in Exhibit 12.12 shows an example of the empirical and fitted item response functions for a dichotomously scored item worth one score point. The horizontal axis represents the proficiency scale on the logit metric, and the vertical axis represents the probability of a correct response. The fitted curve based on the estimated item parameters is shown as a solid line.

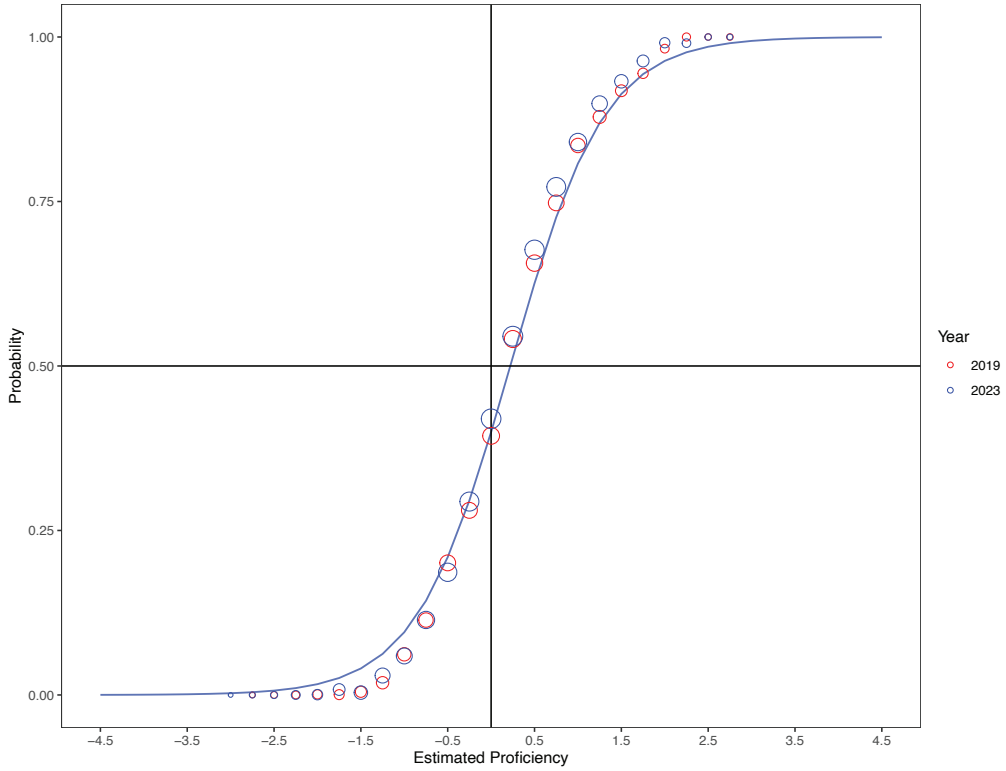
Exhibit 12.12: Example Item Response Function for a Dichotomous Item



Circles represent empirical results based on pseudo counts. The empirical results are obtained by first dividing the logit proficiency scale into intervals of equal size and then calculating the proportion of respondents within each of these segments that answer the item correctly. The center of each circle in the exhibits represents this empirical proportion of correct responses. The size of each circle is proportional to the number of students contributing to the empirical proportion correct in its corresponding interval.

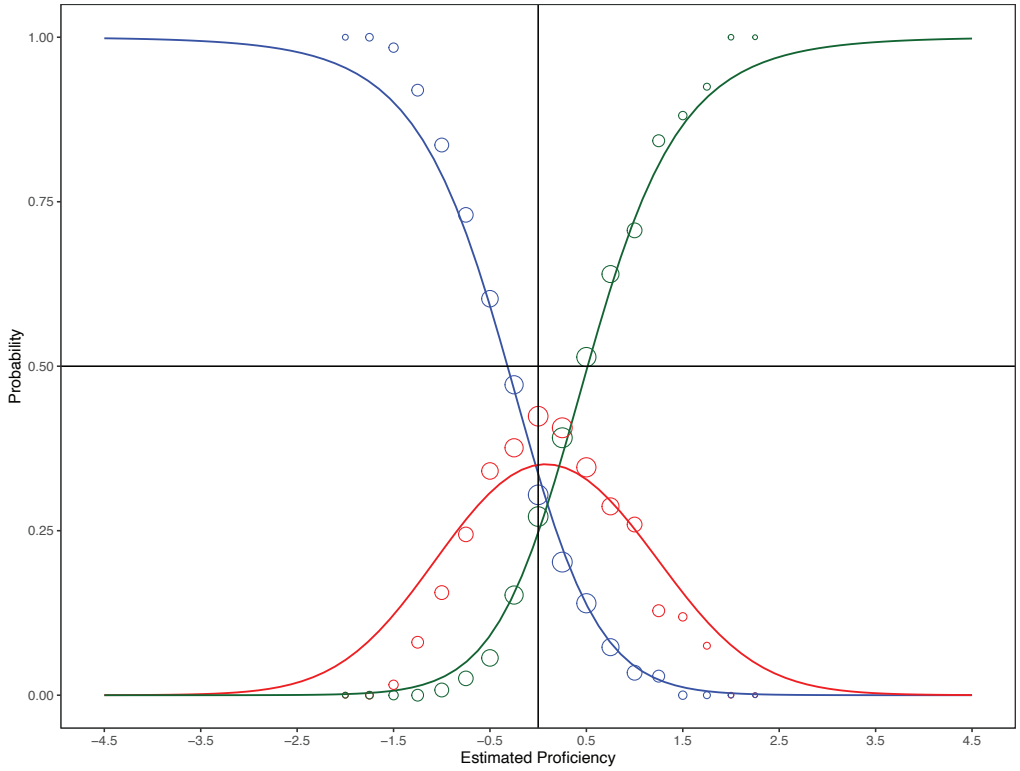
Although a single set of item parameters was estimated for each item in the concurrent calibration, two empirical curves were drawn for trend items, one for the data from each assessment cycle. Plotting both empirical curves from 2023 and 2019 allowed for a visual inspection of the invariance of the item parameters between cycles, a key aspect of the link to the trend scale. Exhibit 12.13 shows the ICC for a trend item, with its single fitted curve and two empirical curves: the red bubbles represent the empirical curve based on the TIMSS 2019 response data, while the blue bubbles represent the empirical curve based on the TIMSS 2023 response data.

Exhibit 12.13: Example Item Response Function for a Trend Dichotomous Item



The ICC plot in Exhibit 12.14 shows the empirical and fitted item response functions for a polytomous item worth two points. As in the dichotomous item plots above, the horizontal axis represents the proficiency scale in logits, but in this example, the vertical axis represents the probability of a response in the response categories. The fitted curves based on the estimated item parameters are shown as solid lines, and the circles represent the empirical results. The interpretation of the circles is the same as in Exhibit 12.12 and Exhibit 12.13. The curve starting at the top left of the chart shows the probability of a score of zero on the item. This probability should always decrease as proficiency increases. The bell-shaped curve shows the probability of a score of one point—partial credit, which should start low, approaching zero for low-ability students, reaching a maximum for medium-ability students, and decreasing for high-ability students. The curve ending at the top right corner of the chart shows the probability of a score of two points—full credit, starting low for low-ability students and increasing as proficiency increases.

Exhibit 12.14: Example Item Response Function for a Polytomous Item



Root Mean Square Deviation (RMSD)

In addition to the graphical model fit assessment, item fit was also checked using the root mean square deviation (RMSD) statistic. The RMSD is the square root of the average of squared differences between the empirical curve, shown as bubbles in the ICCs above, and the fitted curve, weighted by the number of students at each ability interval. Inspecting the RMSD values supplemented the inspection of the ICC.

RMSD values were computed for all TIMSS 2023 items and are reported in the item parameter tables in Appendices 12A to 12D. They are also presented graphically in Exhibits 12.15 to 12.18. In these exhibits, the items are sorted from smallest to largest RMSD values. For trend items with two RMSD values, the larger of the two determined the order. All items in the TIMSS 2023 IRT calibration had good RMSD fit statistics according to the following criteria.

Misfitting items were identified by using the median absolute deviation (MAD) outlier detection method on the RMSD statistic. MAD is a robust measure of dispersion used as a flagging rule instead of an arbitrary cut-off value (von Davier & Bezirhan, 2022). This method flags an item as a possible misfit if its distance from the median of the absolute distances of all other observations exceeds a predetermined threshold. For the TIMSS 2023 IRT calibration, a threshold of 2 was used to identify items that needed further evaluation or possible deletion.

Exhibit 12.15: RMSD Statistics for Items in the TIMSS 2023 Concurrent Calibration – Grade 4 Mathematics

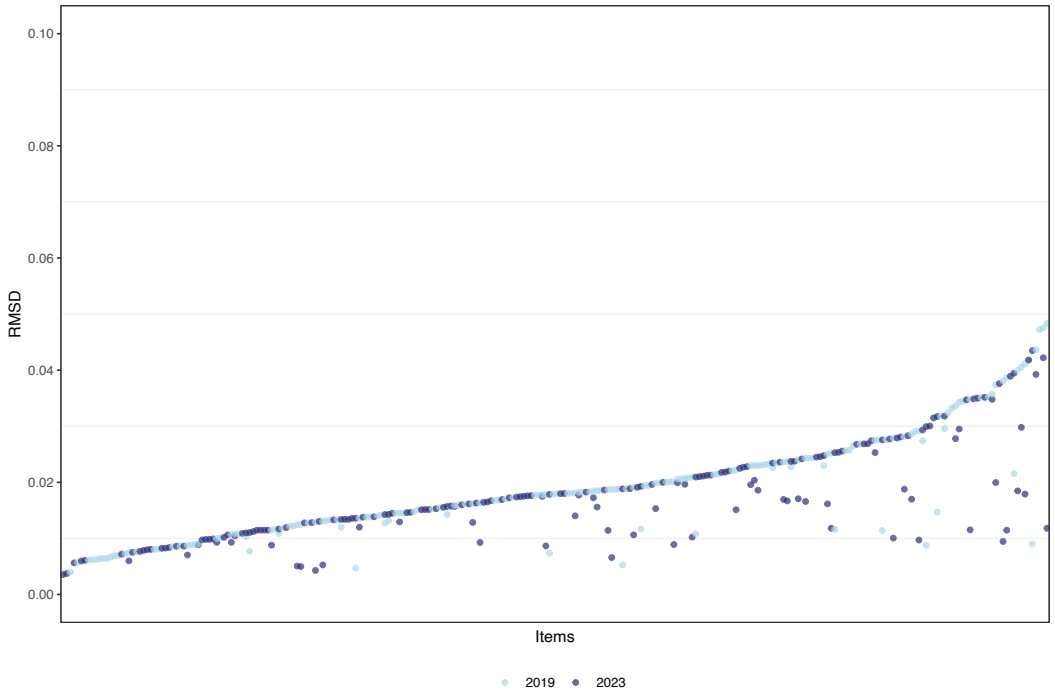


Exhibit 12.16: RMSD Statistics for Items in the TIMSS 2023 Concurrent Calibration – Grade 4 Science

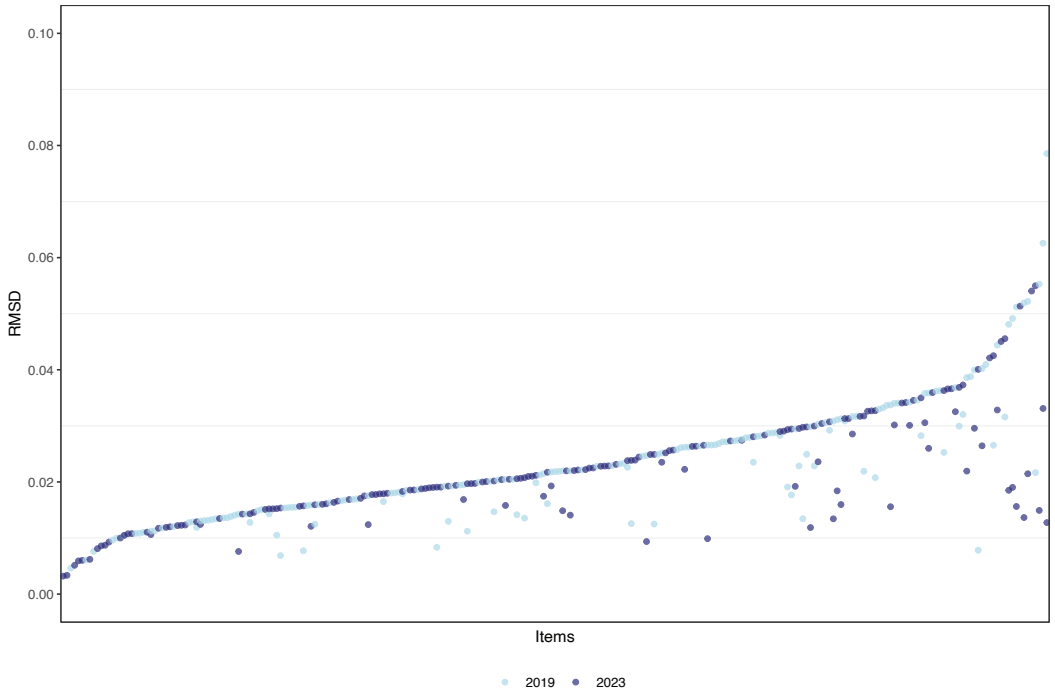


Exhibit 12.17: RMSD Statistics for Items in the TIMSS 2023 Concurrent Calibration – Grade 8 Mathematics

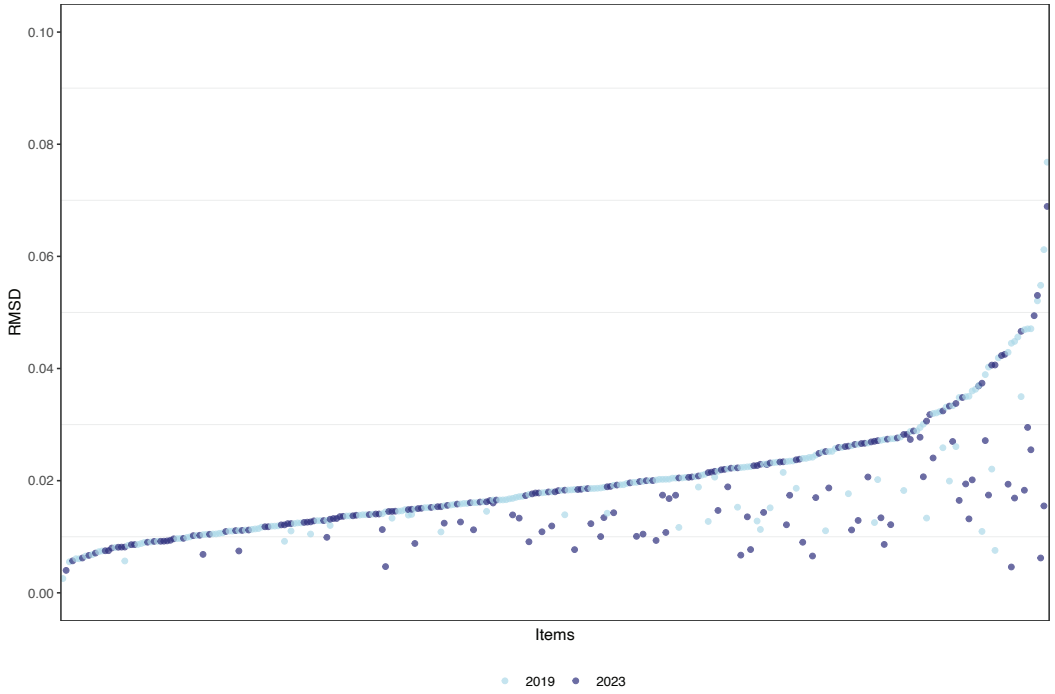
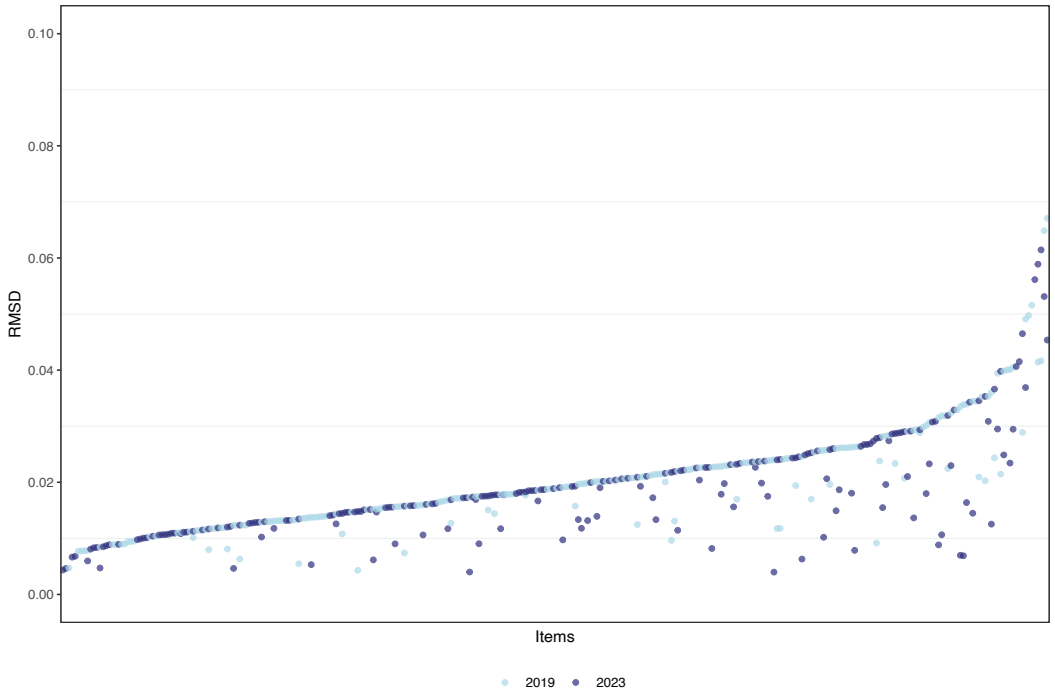


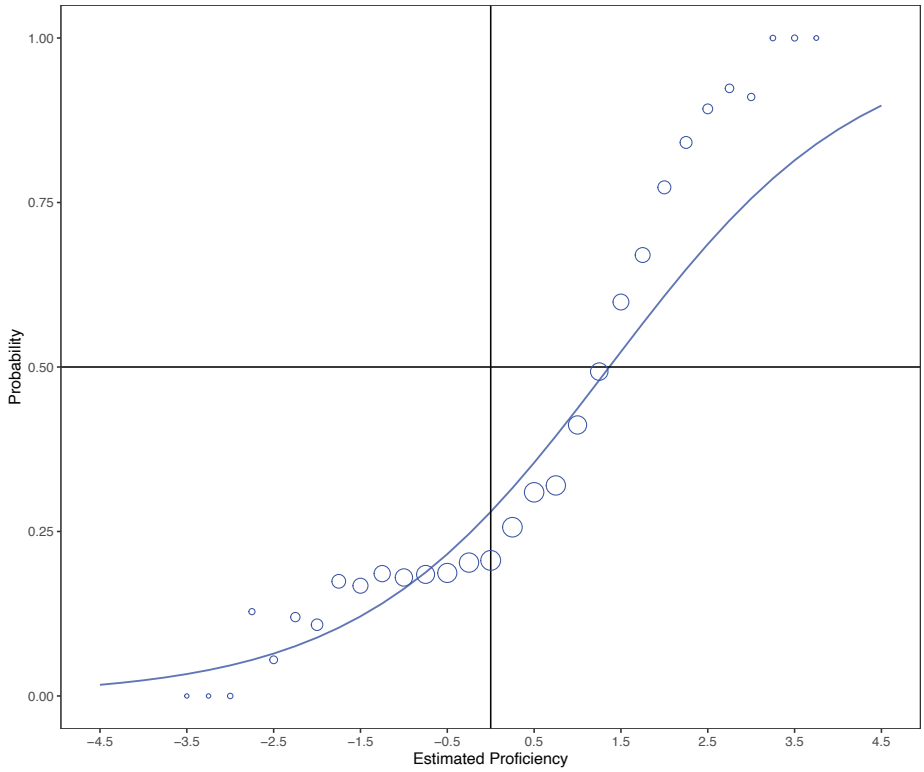
Exhibit 12.18: RMSD Statistics for Items in the TIMSS 2023 Concurrent Calibration – Grade 8 Science



The RMSD statistics of the new multiple-choice items were examined to identify potential misfitting items within this subset of items using the MAD outlier detection method. For the multiple-choice items flagged as outliers, the ICCs were then checked by comparing the empirical and fitted item response functions. If the graphical checks confirmed the misfit identified by RMSD and indicated that the 2PL model was inadequate, a 3PL model was then applied to that item. An example of such an item can be seen in Exhibit 12.19.

Through these rigorous checks on the new TIMSS 2023 items, three misfitting multiple-choice items were identified in each subject at the fourth grade, and four misfitting multiple-choice items were identified in science at the eighth grade. These items were subsequently estimated using a 3PL model in the final analysis.

Exhibit 12.19: Example Item Response Function for a Misfitting Multiple-Choice Item Estimated with 2PL



Examining the Variability in Trend Estimates

A key aspect of reporting the TIMSS 2023 results on the TIMSS trend scale is the ability to accurately re-estimate the TIMSS 2019 achievement results based on a concurrent calibration of the 2019 and 2023 data. This re-estimation using a concurrent calibration of trend and unique item parameters helps establish the linear transformation that places the TIMSS 2023 results on the TIMSS trend scale. Although this transformation was set globally to match the overall mean and standard deviation across the trend countries, it should also achieve an adequate alignment of the 2019 re-estimated results with the previous calibration for each trend country.

The difference between a trend country's TIMSS 2019 achievement mean published back in 2019 and re-estimated in the 2023–2019 concurrent calibration provides a measure of the quality of the link between the two assessments. However, TIMSS does not currently quantify this difference in its variance estimates.

Exhibits 12.20 to 12.23 provide results on the differences between the published 2019 results and the re-estimated 2019 results based on the concurrent calibration in 2023, referred to as trend linking error, associated with the TIMSS 2023 results. Overall, there was a good agreement between the countries' published and re-estimated 2019 results. Although there are small differences at the country level, most differences are within two points, and no standard errors exceed two points. These minor differences are expected, given that the eTIMSS 2019 data were calibrated under two different calibration models with different sets of countries and items contributing to the estimates of item parameters over time. The published 2019 results are based on a concurrent calibration with 2015 data to enable reporting on the trend scale, and the re-calibrated data are based on the joint calibration with 2023 data, but without the 2015 data. This is done to ensure the most recent framework updates, and the most recent new and trend blocks are the basis of reporting the 2023 results.

It is noteworthy that relatively larger linking differences, up to four points, occurred in a few countries, as shown in the exhibits. This is mainly due to the different treatments of omitted and not-reached responses in the two item calibrations. The omitted and not-reached responses were treated with the strength-of-evidence approach for the 2023–2019 analysis but were previously treated with the deterministic approach to equate item omissions with incorrect responses in the 2019–2015 analysis.

Exhibit 12.20: Trend Linking Error Variance between TIMSS 2019 and TIMSS 2023 Calibrations – Grade 4 Mathematics

Exhibit 12.21: Trend Linking Error Variance between TIMSS 2019 and TIMSS 2023 Calibrations – Grade 4 Science

Exhibit 12.22: Trend Linking Error Variance between TIMSS 2019 and TIMSS 2023 Calibrations – Grade 8 Mathematics

Exhibit 12.23: Trend Linking Error Variance between TIMSS 2019 and TIMSS 2023 Calibrations – Grade 8 Science

Summary

The psychometric analyses of the TIMSS 2023 achievement data were successful. They enabled the imputation of valid and reliable plausible values for reporting the results of the TIMSS 2023 assessments. The psychometric methods implemented and described in this chapter relied on past methods and experience for analyzing the TIMSS data over almost three decades. The use of multiple-group IRT models in concurrent calibrations enabled TIMSS to find international item

parameters that maximize fit across all countries. The successful conclusion of the analyses facilitated the successful linking of all TIMSS assessment data to the TIMSS trend scale such that results from the TIMSS 2023 assessments can be compared directly between countries without further adjustments. They also can be compared reliably with past TIMSS assessments.

References

Beaton, A. E. (1969). Criterion scaling of questionnaire items. *Socio-Economic Planning Sciences, 2*, 355–362.

Beaton, A. E., & Barone, J. L. (2017). Large-scale group-score assessment. In R. E. Bennett & M. von Davier (Eds.), *Advancing human assessment: The methodological, psychological and policy contributions of ETS* (pp. 233–284). Springer Science + Business Media. https://doi.org/10.1007/978-3-319-58689-2_8

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Fishbein, B., & Foy, P. (2021). Scaling the TIMSS 2019 problem solving and inquiry data. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and Procedures: TIMSS 2019 Technical Report* (pp. 17.1–17.51). Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/methods/chapter-17.html

Fishbein, B., Martin, M. O., Mullis, I. V. S., & Foy, P. (2018). The TIMSS 2019 item equivalence study: Examining mode effects for computer-based assessment and implications for measuring trends. *Large-scale Assessments in Education, 6*(1), 11. https://doi.org/10.1186/s40536-018-0064-z

Foy, P., Fishbein, B., von Davier, M., & Yin, L. (2020). Implementing the TIMSS 2019 scaling methodology. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and Procedures: TIMSS 2019 Technical Report* (pp. 12.1–12.146). Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/methods/chapter-12.html

Glas, C. A. W., & Pimentel, J. (2006). *Modeling nonignorable missing data processes in item calibration*. (LSAC Research Report Series; No. 04-07). Newton, PA: Law School Admission Council.

Kang, T. and Cohen, A.S. (2007) IRT Model Selection Methods for Dichotomous Items. *Applied Psychological Measurement, 31*, 331–358. https://doi.org/10.1177/0146621606292213

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: J. Wiley & Sons.

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three parameter logistic model. *Educational and Psychological Measurement, 28*, 989–1020.

Lord, F. M. (1975). *Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters*. (Research Bulletin RB-75-33). Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*, 359–381.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177–196.

Mislevy, R. J., & Sheehan, K. M. (1987). Marginal estimation procedures. In A. E. Beaton (Ed.), *Implementing the new design: The NAEP 1983–84 technical report* (No. 15-TR-20, pp. 293–360). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Moustaki, I., & Knott, M. (2000). Weighted for item non-response in attitude scales by using latent variable models with covariates. *Journal of the Royal Statistical Society, 163*(3), 445–459.

R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rogers, A., Tang, C., Lin, M.-J., & Kandathil, M. (2006). DGROUP [computer software]. Princeton, NJ: Educational Testing Service.

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)*. (ETS Research Report No. RR-10-11). Princeton, NJ: Educational Testing Service.

Rose, N., von Davier, M., & Nagengast, B. (2017). Modeling omitted and not-reached items in IRT models. *Psychometrika, 82*(3), 795–819.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592. https://www.jstor.org/stable/pdf/2335739.pdf

Sheehan, K. M. (1985). M-Group: Estimation of group effects in multivariate models [computer software, Version 3.2]. Princeton, NJ: Educational Testing Service.

Thissen, D., & Wainer, H. (1985). Some supporting evidence for Lord's guideline for estimating "c" theory (Research Report No. RR-85-15). Princeton, NJ: Educational Testing Service.

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics, 2*, 309 322.

von Davier, M. (2009). Is there Need for the 3PL Model? Guess what? *Measurement: Interdisciplinary Research and Perspectives, 7*(2), 110–114. https://doi.org/10.1080/15366360903117079

von Davier, M. (2023). Omitted response treatment using a modified Laplace smoothing for approximate Bayesian inference in item response theory. https://osf.io/preprints/psyarxiv/pc395/

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., Foy, P., Martin, M. O., & Mullis, I. V. S. (2020). Examining eTIMSS country differences between eTIMSS data and bridge data: A look at country-level mode of administration effects. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and Procedures: TIMSS 2019 Technical Report* (pp. 13.1–13.24). Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/methods/chapter-13.html

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? In M. von Davier & D. Hastedt (Eds.), *IERI Monograph Series: Issues and Methodologies in Large Scale Assessments* (Vol. 2, pp. 9–36). Retrieved from https://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02_Chapter_01.pdf

von Davier, M., & Sinharay, S. (2013). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 155–174). Boca Raton, FL: CRC Press.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). Statistical procedures used in the National Assessment of Educational Progress (NAEP): Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26: Psychometrics). Amsterdam, Netherlands: Elsevier.

Whittaker, T. A., Chang, W., & Dodd, B. G. (2012). The performance of IRT model selection methods with mixed-format tests. *Applied Psychological Measurement, 36*(3), 159–180. doi:10.1177/0146621612440305

Yin, L., & Foy, P. (2021). Constructing the TIMSS 2019 environmental awareness scales. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and Procedures: TIMSS 2019 Technical Report* (pp. 18.1–18.30). Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/methods/chapter-18.html

- Appendix 12A: Item Parameters from the TIMSS 2023 Concurrent Calibration – Grade 4 Mathematics
- Appendix 12B: Item Parameters from the TIMSS 2023 Concurrent Calibration – Grade 4 Science
- Appendix 12C: Item Parameters from the TIMSS 2023 Concurrent Calibration – Grade 8 Mathematics
- Appendix 12D: Item Parameters from the TIMSS 2023 Concurrent Calibration – Grade 8 Science
- Appendix 12E: Nonresponse Indicator Parameters from the TIMSS 2023 Concurrent Calibration