
CHAPTER 15

# Creating and Interpreting the TIMSS Context Questionnaire Scales

Liqun Yin
Katherine A. Reynolds

## Introduction

In addition to collecting the student response data required to estimate students' mathematics and science achievement, TIMSS administers extensive context questionnaires. These questionnaires enable reporting on key subgroups of the population and help in understanding factors associated with acquiring and developing mathematics and science knowledge and skills. To this end, TIMSS gathers data about students, their homes, their schools, and their teachers to better understand the contexts in which learning occurs.

Many sets of items in the TIMSS Context Questionnaires are reported as scales that measure a common underlying latent construct. These contextual variables facilitate the exploration of factors that might be related to mathematics and science achievement across and within countries.

The use of item response theory (IRT) methods for reporting context data, specifically the Rasch partial credit model (Masters, 1982; Masters & Wright, 1997), was introduced in TIMSS 2011 and has continued to be used through TIMSS 2023. Many constructs have been of interest over several assessment cycles. Many scales are updated from cycle to cycle based on input from National Research Coordinators (NRCs) and the context experts represented in the TIMSS Questionnaire Item Review Committee (QIRC). New context scales are introduced to address recent research questions and gather valuable information in critical areas (e.g., measuring environmental attitudes). Some scales that retain many of the same items across TIMSS assessments are linked over time to enable reporting trends on a common metric.

To facilitate interpretation of the context scale results, questionnaire respondents (students, or their parents, teachers, or principals) are classified based on their responses into regions corresponding to high, middle, and low levels of the construct. For each TIMSS cycle, the context scales are included in the TIMSS International Database as continuous scale scores and categorical scale region variables.

This chapter describes the procedures for constructing, interpreting, and validating the TIMSS Context Questionnaire Scales and details the process for transforming and reporting


https://doi.org/10.6017/lse.tpisc.timss.rs2550

scales. The chapter concludes with details about the implementation of procedures in TIMSS 2023.

## Procedure for Creating TIMSS Context Questionnaire Scales

The TIMSS Context Questionnaire Scales are created using the Rasch partial credit model (PCM; Masters, 1982). Partial credit IRT analysis is based on a statistical model that relates the probability that a person will choose a particular response for an item to that person's location on an underlying construct. The PCM was the basis for Muraki's (1992) generalized partial credit model (GPCM), described in Chapter 11 for polytomous achievement items, but with a uniform discrimination across all scale items. As such, the PCM is given by equation (11.3), where the slope parameters $a_i$ for all items are equal to 1.

There are several steps followed when analyzing the TIMSS context questionnaire responses, including item calibration, evaluation of item fit, estimation of scale scores, scale transformation onto a reporting metric, and creation of scale regions.

### Item Calibration

The estimation of the item parameters, a procedure also known as item calibration, is conducted using the combined data from all countries that participate at each grade. Each country initially contributes equally to the calibration through the use of "senate weights," which sum to 500 for each country's entire student data. However, only cases with at least two valid item responses on a scale are included for the calibration, without any weight adjustment, after removing cases that do not meet this criterion. Therefore, countries with missing responses may contribute less data to the calibration for a scale compared to countries without missing responses.

### Evaluating Item Fit

Item fit statistics are used to evaluate how well the model fits the data across all countries contributing to the calibration, including the Rasch infit statistic (Wright & Masters, 1982) and the Q-index (Rost & von Davier, 1994). The Rasch infit item statistic is a residual-based measure of how well the estimated difficulty (location) parameter fits the data, with values ranging from 0 to infinity. A value of 1.0 corresponds to optimal fit to the Rasch model. The TIMSS & PIRLS International Study Center uses a value of 1.4 as an upper bound to indicate potential misfit, based on Adams and Khoo (1996) and Preuschoff (2010).

The Q item fit index (Rost & von Davier, 1994) for the ordinal Rasch model is used to evaluate the fit of an item with regard to the conditional probability of its observed response vector and does not depend on the item parameters. The Q-index is standardized and ranges from 0 to 1 with a midpoint of 0.5. A value of 0 indicates a perfect fit to a Guttman pattern (a more extreme, deterministic prediction than the Rasch model). In contrast, a value of 1 indicates the least likely response, the anti-Guttman pattern. The midpoint of 0.5 indicates random response behavior, such that the item is independent of the latent trait.

In case of item misfit, the content of the item is evaluated, and item response patterns are examined to check for problems in the data. In rare cases, item response categories may be collapsed to create the scale, or items may be removed.

## Estimating Scale Scores

After estimating international item parameters, context scale scores for each respondent are estimated using weighted maximum likelihood estimation (Warm, 1989). Cases with valid responses to at least two items on a scale are assigned scale scores.

## Scale Transformation

Each scale requires its own set of transformation constants. Scale scores in the logit metric, as obtained from the PCM calibration, are converted into a reporting metric with a mean of 10 and a standard deviation of 2 based on the calibration data. The metric of the scale is set during the TIMSS cycle when the scale is first introduced. In subsequent cycles, the current cycle results may be linked to the previously established metric if the source items remain unchanged or the scale is subjected to only minor modifications. In this case, the current cycle item parameters are first transformed to the metric used during the previous cycle, and this transformation is applied to the current cycle logit metric scale scores. Then, the same transformation used in the previous cycle is applied. This process involves quality checks to ensure the same underlying construct is being measured over time. The metric for a previously existing scale may be reset for the current cycle if the scale items appear to function differently over time, or if the source items were revised, for example, by adding, removing, or changing items or by revising response options. The later section “Transforming Scale Scores onto Existing Reporting Metrics” provides more information about this process.

## Creating Scale Regions

Scale regions are created for each context scale that relate to both raw score points as well as reporting scale score cut points. Two cut points on the reporting scale divide the scale into high, medium, and low regions, each with a content-referenced interpretation based on the corresponding most likely response categories that can be found based on the cut points describing the region. A respondent at the cut point between low and medium will, in expectation, produce responses in lower response categories than a respondent located at the cut point between medium and high score regions.

Interpretation of the regions is content-referenced to the extent that the boundaries of the regions can be described in terms of identifiable combinations of item responses. The region cut point boundaries are defined based on score points that correspond to identified item response combinations. Then, respondents are classified into the regions based on the corresponding scale scores. A property of a Rasch scale is that each raw score is associated with a unique scale score. Raw scores are quantified by assigning a numeric value to each item response category such that the lowest response category is set to 0, the next set to 1, and so

on. Summing the values across the items results in a raw score that has an equivalent scale score on the reporting metric.

For each scale, the particular response combinations that define the regions' boundaries, or cut points, are identified using one of two methods. For most context scales, the raw cut points are established using a judgment-based method, where item response combinations for the boundaries are determined by TIMSS & PIRLS International Study Center content experts, considering what constitutes a high or low region on each individual scale based on the possible item response patterns. For example, on a Likert scale, the cut point for the "high" region may correspond to the raw score of "agreeing a lot" to at least half of the items, and "agreeing a little" with the other half. Similarly, the cut point for the "low" region may correspond to "disagreeing a lot" with half of the items and "disagreeing a little" with the other half. The sums of the numeric values across the two response patterns result in two raw score cut points, and their equivalent scale scores are used to classify respondents into the regions for reporting.

For scales where the content-referenced cut-score definitions produce score regions that contain very few or no students, the method based on expert judgement, but not on the observed range of responses, is replaced by a statistical approach that helps identify three optimal score regions. Identifying homogeneous groups of respondents based on categorical data is often done by means of latent class analysis (LCA; Lazarsfeld & Henry, 1968; von Davier & Lee, 2019), which can be used to form cut scores and for standard setting (e.g., Brown, 2007; Jiao et al., 2011; Yin et al., 2024). The TIMSS LCA-based cut score method (LCA-CS method) is an alternative data-driven method to identify raw cut points for some of the TIMSS Context Questionnaire Scales. This method builds on the classical latent class analysis (Lazarsfeld, 1955), a latent variable modeling technique for categorical data that identifies groups based on a statistical optimality criterion. Once the cut points are determined using this method, the subsequent procedures of assigning respondents to categories mirror those of the judgment-based cut point specification method.

Given the identified raw cut scores by the LCA-CS method, the minimum responses needed to meet or exceed the cut scores can be determined by calculating the expected responses for each item based on the Rasch model and estimated item parameters. This involves selecting the most likely response for each item given the associated scale cut score, starting with the response category with the highest probability across all items, then moving to the next highest probability on another item until the total raw scores of expected responses are achieved to have the same values as the identified raw cut scores. Note that any response pattern that matches the raw score associated with the scale cut score is compatible with this approach, just as in the judgement-based approach. However, defining the most likely responses given a cut score can facilitate meaningful interpretation of the results.

## Transforming Scale Scores onto Existing Reporting Metrics

Scales with no changes or with minor modifications from prior TIMSS cycles are transformed to a common reporting metric established in the cycle when the scale was first used.

A context questionnaire scale is transformed to the previously established reporting metric if it has a sufficient number of common items—at least two-thirds—shared with the corresponding scale in the previous TIMSS cycle. In addition, it is required that the scale stem and the number and verbal anchoring of response options do not change across cycles. Although similar, the current cycle logit metric will not be identical to the previous one, even with all common items, due to different participating countries and observed response distribution differences. Therefore, the current cycle scores need to be transformed to place them on the previously established metric if it can be shown that item functioning does not change substantially across cycles. Changes to item functioning might occur between cycles, for example, due to shifts in the way students interpret and respond to items over time.

To validate the transformation of the current cycle results onto existing scales, the TIMSS & PIRLS International Study Center conducts extensive analyses to examine item behavior across the current and previous cycles. Item parameter estimates for the common items are compared across the two cycles by examining the differences between the previous cycle item parameter estimates and the current cycle item parameter estimates after transformation to the previous cycle logit metric so that they are directly comparable. Item functioning is considered stable across cycles if item parameter differences are less than 0.1.

The transformation of scale scores onto a previously reported scale is a two-step process. The first transformation places the current cycle logit scale scores on the previous cycle logit metric, where the transformations are obtained from the item parameter estimates for the common items between the two cycles. The second transformation places the resulting logit scale scores on the previous cycle logit metric to the (10,2) TIMSS scale reporting metric.

While the resulting trend scale scores for the context scales are comparable to those from previous cycles, these are often based on relatively small numbers of items and therefore, comparisons across cycles are not always very reliable. In the case of short context scales, if the functioning of one or two items on the scale changes over cycles, a relatively large percentage of the scale is affected. In addition, because new item parameters are estimated for each cycle, although transformed to the existing metric, the corresponding scale score cut points used to categorize respondents into reporting categories may change slightly. However, such a change would be unusual without adding or dropping items, and may signal during quality control that a new metric should be established for the current cycle.

The scale cut points derived for the scale in the current cycle are not always comparable to the previous cycles when raw cut points have been changed, for example when items are added or dropped from a scale. Raw cut points are identified in each cycle for such scales, which correspond to different scale scores used for the categorization. Raw cut points for a scale most often change due to the number of items changing between cycles. However, there

may be other exceptional cases where raw cut points are changed, which are documented in the later "TIMSS 2023 Context Scaling Implementation" section.

## Evaluating Reliability and Validity of the TIMSS Context Questionnaire Scales

As one part of the evidence that the TIMSS Context Questionnaire Scales provide comparable measurement quality across countries, reliability coefficients are estimated for each scale for all countries and benchmarking participants. In addition, a principal component analysis (Hotelling, 1933) of the responses to the scale is conducted within each country, and the results are examined alongside the estimates of Cronbach's Alpha measure of internal consistency (Peterson, 1994; Taber, 2017).

The relationship of context scale scores with mathematics and science achievement is also an important aspect of validity for the TIMSS context questionnaire scales. This is examined with country-level estimates of the Pearson correlation of context scales and achievement as well as the proportion of variance in achievement accounted for by the scales. In addition, $\eta^2$ are calculated to quantify the proportion of variance in achievement accounted for by the differences between the scale regions.

## TIMSS 2023 Context Scaling Implementation

Psychometric analyses of the TIMSS 2023 context questionnaire data were conducted using the ConQuest 2.0 software (Wu et al., 2007). To remove scale indeterminacy in calibration, the "items constraint" in ConQuest was used to set the mean of the item difficulty (location) parameters to zero. Item calibration was conducted on the combined data from all countries participating in the TIMSS 2023 computer-based assessment. As a result, 52 countries contributed data to the item calibration for the context scales at the fourth grade, while 40 countries contributed to the item calibration at the eighth grade. Item fit was examined for all scale items, and items were removed from the scale in cases of poor fit.

Scale regions were created by defining raw score cut points for each scale based on combinations of item responses, as described in the earlier "Creating Scale Regions" subsection. Most TIMSS 2023 scales used the judgment-based method as used in previous TIMSS cycles. Raw cut points for creating scale regions were identified using the LCA-CS method for a small number of context scales with highly skewed distributions across countries. This was done in cases where the expert-provided cut points led to regions that contained very few or no students and made reporting for those regions impossible. This approach applied to context scales with highly skewed distributions in TIMSS 2023 improved classification accuracy and reporting by ensuring that respondents are categorized in all three scale regions.

Some scales included items that required reverse coding before assigning a "raw score" to each response category for scaling. Such items included those stated in a negative way, such that agreeing with the item would imply that the respondent has a lower level of the underlying latent construct. For example, the Likert item "Mathematics is harder for me than for many of my classmates" in the *Students Confident in Mathematics* scale at the eighth grade was reverse-coded before estimating the model parameters, producing scale scores, and determining cut points for the scale regions.

Exhibits 15.1 and 15.2 list all the TIMSS 2023 Context Questionnaire Scales included in the TIMSS 2023 International Database. Columns indicate the assessment year the scale was first established and whether the scale results were included in *TIMSS 2023 International Results in Mathematics and Science*.

**Exhibit 15.1: List of TIMSS 2023 Context Questionnaire Scales – Grade 4**

| Scale Name | Respondent | Scale Score Variable Name | Scale Index Variable Name | Year Scale Metric Established | Included in TIMSS 2023 International Results |
|---|---|---|---|---|---|
| Digital Self-Efficacy | Students | ASBGSEC | ASDGSEC | 2023 | ✓ |
| Students Value Environmental Preservation | Students | ASBGVEP | ASDGVEP | 2023 | ✓ |
| Sense of School Belonging | Students | ASBGSSB | ASDGSSB | 2023 | ✓ |
| Student Bullying | Students | ASBGSB | ASDGSB | 2023 | ✓ |
| Students Like Learning Mathematics | Students | ASBGSLM | ASDGSLM | 2023 | ✓ |
| Instructional Clarity in Mathematics Lessons | Students | ASBGICM | ASDGICM | 2023 | ✓ |
| Disorderly Behavior during Mathematics Lessons | Students | ASBGDML | ASDGDML | 2023 | ✓ |
| Students Confident in Mathematics | Students | ASBGSCM | ASDGSCM | 2023 | ✓ |
| Students Like Learning Science | Students | ASBGSLS | ASDGSLS | 2023 | ✓ |
| Instructional Clarity in Science Lessons | Students | ASBGICS | ASDGICS | 2023 | ✓ |
| Disorderly Behavior during Science Lessons | Students | ASBGDSL | ASDGDSL | 2023 | ✓ |
| Students Confident in Science | Students | ASBGSCS | ASDGSCS | 2023 | ✓ |
| Home Early Literacy Activities Before Primary School* | Parents | ASBHELA | ASDHELA | 2011 | |

**Exhibit 15.1: List of TIMSS 2023 Context Questionnaire Scales – Grade 4 (Continued)**

| Scale Name | Respondent | Scale Score Variable Name | Scale Index Variable Name | Year Scale Metric Established | Included in TIMSS 2023 International Results |
|---|---|---|---|---|---|
| Home Early Numeracy Activities Before Primary School* | Parents | ASBHENA | ASDHENA | 2011 | |
| Home Early Literacy and Numeracy Activities Before Primary School* | Parents | ASBHELN | ASDHELN | 2011 | ✓ |
| Could Do Early Literacy Tasks When Beginning Primary School | Parents | ASBHELT | ASDHELT | 2015 | |
| Could Do Early Numeracy Tasks When Beginning Primary School | Parents | ASBHENT | ASDHENT | 2015 | |
| Could Do Literacy and Numeracy Tasks When Beginning Primary School | Parents | ASBHLNT | ASDHLNT | 2015 | ✓ |
| Parents' Perceptions of Their Child's School | Parents | ASBHPSP | ASDHPSP | 2015 | |
| Home Socioeconomic Status | Parents | ASBHSES | ASDHSES | 2019 | ✓ |
| Home Resources for Learning | Parents/ Students | ASBGHRL | ASDGHRL | 2011 | |
| Instruction Affected by Mathematics Resource Shortages | Principals | ACBGMRS | ACDGMRS | 2011 | |
| Instruction Affected by Science Resource Shortages | Principals | ACBGSRS | ACDGSRS | 2011 | |
| School Emphasis on Academic Success - Principals' Reports | Principals | ACBGEAS | ACDGEAS | 2015 | ✓ |
| School Discipline | Principals | ACBGDAS | ACDGDAS | 2011 | ✓ |
| Schools Where Students Begin Primary Grades with Literacy and Numeracy Skills | Principals | ACBGLNS | ACDGLNS | 2015 | ✓ |
| School Emphasis on Academic Success - Teachers' Reports | Teachers | ATBGEAS | ATDGEAS | 2015 | |
| Safe and Orderly School | Teachers | ATBGSOS | ATDGSOS | 2011 | ✓ |
| Teachers' Job Satisfaction | Teachers | ATBGTJS | ATDGTJS | 2015 | |
| Teaching Limited by Students Not Ready for Instruction | Teachers | ATBGLSN | ATDGLSN | 2015 | ✓ |

* Indicates the LCA-CS method was used to specify the raw cut points for the scale regions.

## Exhibit 15.2: List of TIMSS 2023 Context Questionnaire Scales – Grade 8

| Scale Name | Respondent | Scale Score Variable Name | Scale Index Variable Name | Year Scale Metric Established | Included in TIMSS 2023 International Results |
|---|---|---|---|---|---|
| Home Educational Resources* | Students | BSBGHER | BSDGHER | 2011 | ✓ |
| Digital Self-Efficacy | Students | BSBGSEC | BSDGSEC | 2023 | ✓ |
| Students Value Environmental Preservation | Students | BSBGVEP | BSDGVEP | 2023 | ✓ |
| Sense of School Belonging | Students | BSBGSSB | BSDGSSB | 2023 | ✓ |
| Student Bullying | Students | BSBGSB | BSDGSB | 2023 | ✓ |
| Students Like Learning Mathematics | Students | BSBGSLM | BSDGSLM | 2023 | ✓ |
| Insructional Clarity in Mathematics Lessons | Students | BSBGICM | BSDGICM | 2023 | ✓ |
| Disorderly Behavior during Mathematics Lessons | Students | BSBGDML | BSDGDML | 2023 | ✓ |
| Students Confident in Mathematics | Students | BSBGSCM | BSDGSCM | 2023 | ✓ |
| Students Value Mathematics | Students | BSBGSVM | BSDGSVM | 2011 | ✓ |
| Students Like Learning Science | Students | BSBGSLS | BSDGSLS | 2023 | ✓ |
| Instructional Clarity in Science Lessons | Students | BSBGICS | BSDGICS | 2023 | ✓ |
| Disorderly Behavior during Science Lessons | Students | BSBGDSL | BSDGDSL | 2023 | ✓ |
| Students Confident in Science | Students | BSBGSCS | BSDGSCS | 2023 | ✓ |
| Students Value Science | Students | BSBGSVS | BSDGSVS | 2011 | ✓ |
| Students Like Learning Biology | Students | BSBGSLB | BSDGSLB | 2023 | ✓ |
| Instructional Clarity in Biology Lessons | Students | BSBGICB | BSDGICB | 2023 | ✓ |
| Disorderly Behavior during Biology Lessons | Students | BSBGDBL | BSDGDBL | 2023 | ✓ |
| Students Confident in Biology | Students | BSBGSCB | BSDGSCB | 2023 | ✓ |
| Students Like Learning Chemistry | Students | BSBGSLC | BSDGSLC | 2023 | ✓ |
| Instructional Clarity in Chemistry Lessons | Students | BSBGICC | BSDGICC | 2023 | ✓ |

**Exhibit 15.2: List of TIMSS 2023 Context Questionnaire Scales – Grade 8 (Continued)**

| Scale Name | Respondent | Scale Score Variable Name | Scale Index Variable Name | Year Scale Metric Established | Included in TIMSS 2023 International Results |
|---|---|---|---|---|---|
| Disorderly Behavior during Chemistry Lessons | Students | BSBGDCL | BSDGDCL | 2023 | ✓ |
| Students Confident in Chemistry | Students | BSBGSCC | BSDGSCC | 2023 | ✓ |
| Students Like Learning Physics | Students | BSBGSLP | BSDGSLP | 2023 | ✓ |
| Instructional Clarity in Physics Lessons | Students | BSBGICP | BSDGICP | 2023 | ✓ |
| Disorderly Behavior during Physics Lessons | Students | BSBGDPL | BSDGDPL | 2023 | ✓ |
| Students Confident in Physics | Students | BSBGSCP | BSDGSCP | 2023 | ✓ |
| Students Like Learning Earth Science | Students | BSBGSLE | BSDGSLE | 2023 | ✓ |
| Instructional Clarity in Earth Science Lessons | Students | BSBGICE | BSDGICE | 2023 | ✓ |
| Disorderly Behavior during Earth Science Lessons | Students | BSBGDEL | BSDGDEL | 2023 | ✓ |
| Students Confident in Earth Science | Students | BSBGSCE | BSDGSCE | 2023 | ✓ |
| Instruction Affected by Mathematics Resource Shortages | Principals | BCBGMRS | BCDGMRS | 2011 | |
| Instruction Affected by Science Resource Shortages | Principals | BCBGSRS | BCDGSRS | 2011 | |
| School Emphasis on Academic Success - Principals' Reports | Principals | BCBGEAS | BCDGEAS | 2015 | ✓ |
| School Discipline | Principals | BCBGDAS | BCDGDAS | 2011 | ✓ |
| School Emphasis on Academic Success - Teachers' Reports | Teachers | BTBGEAS | BTDGEAS | 2015 | |
| Safe and Orderly School | Teachers | BTBGSOS | BTDGSOS | 2011 | ✓ |
| Teachers' Job Satisfaction | Teachers | BTBGTJS | BTDGTJS | 2015 | |
| Teaching Limited by Students Not Ready for Instruction | Teachers | BTBGLSN | BTDGLSN | 2015 | ✓ |

* Indicates the LCA-CS method was used to specify the raw cut points for the scale regions.

Existing scales brought forward to TIMSS 2023 with no changes or with minor modifications were transformed to place the context scale results from multiple cycles on a common metric, established either in 2011, 2015, or 2019 (Martin et al., 2012; Martin et al., 2016; Yin & Fishbein, 2020). Out of 30 context scales at the fourth grade, 18 were transformed to a metric established in a prior cycle. At the eighth grade, 11 out of 39 scales were transformed to existing metrics. Among the remaining scales, some were newly developed for this cycle, such as the *Students Value Environmental Preservation* scale. Other scales existed in the previous cycles but underwent significant updates, including modifications to item text or the addition or removal of multiple items. Consequently, there were insufficient numbers of trend items to support reliable transformations for these updated scales, leading to their classification as new scales in TIMSS 2023.

As described earlier, the cut points used for TIMSS 2023 to categorize the students into three reporting categories based on their scale scores may be different from those used in prior cycles. Setting cut scores based on the current cycle's metric, even though linked to the previous cycle, accounted for possible effects resulting from any changes in items which may result in a change of the observed range of responses in the new cycle and the number of component variables when scales are modified across cycles. As such, the procedure primarily depended on cycle-specific similarities in response patterns.

Four context scales (*Early Literacy Activities, Early Numeracy Activities,* and *Early Literacy and Numeracy Activities* at the fourth grade, and *Home Educational Resources* at the eighth grade) with highly skewed distributions in TIMSS 2023 were subjected to the LCA-CS method for identifying cut points. Historically, categories for these scales based on human judgment led to groups containing very few or no respondents due to cut scores being too extreme relative to how respondents answered the questions. The cut points identified based on LCA-CS as applied to the TIMSS 2023 data reflected the additional information available in the current cycle and improved reporting by ensuring that categories are informative. The new categorizations in TIMSS 2023 were not applied to data from previous cycles, and new categorizations are not fully comparable to prior categorizations. Users interested in comparing percentages of students classified into scale regions between cycles should apply the TIMSS 2023 cut scores to the data from previous TIMSS cycles, if the scales remained unchanged, to create comparable regions.

Detailed information on the TIMSS 2023 Context Questionnaire scales, including item parameter estimates and scale statistics described above, can be found in the following downloadable exhibits:

- Exhibits 15.3 and 15.4 provide a list of items comprising each scale and their response categories.

**↓ Exhibit 15.3: TIMSS 2023 Context Scale Descriptions – Grade 4**

**↓ Exhibit 15.4: TIMSS 2023 Context Scale Descriptions – Grade 8**

- Exhibits 15.5 and 15.6 provide international item parameters and item fit statistics. For each item, the delta parameter $\delta_i$ shows the estimated overall location of the item on the scale, and the tau parameters $\tau_{ij}$ show the location of the steps, expressed as deviations from delta ($b_{ij} = \delta_i - \tau_i$).

  **⭳ Exhibit 15.5: TIMSS 2023 Context Scale Item Parameters and Item Fit Statistics – Grade 4**

  **⭳ Exhibit 15.6: TIMSS 2023 Context Scale Item Parameters and Item Fit Statistics – Grade 8**

- Exhibits 15.7 and 15.8 report the scale transformation constants applied to the international distribution of logit scores to put the TIMSS 2023 estimates on the TIMSS (10,2) reporting metric.

  **⭳ Exhibit 15.7: TIMSS 2023 Context Scale Transformation Constants – Grade 4**

  **⭳ Exhibit 15.8: TIMSS 2023 Context Scale Transformation Constants – Grade 8**

- Exhibits 15.9 and 15.10 provide the equivalence tables of raw and transformed scale scores with the cut points used to create the scale regions.

  **⭳ Exhibit 15.9: TIMSS 2023 Context Scale Equivalence Tables of Raw and Transformed Scale Scores – Grade 4**

  **⭳ Exhibit 15.10: TIMSS 2023 Context Scale Equivalence Tables of Raw and Transformed Scale Scores – Grade 8**

- Exhibits 15.11 and 15.12 report country-level Cronbach's Alpha reliability coefficients and principal components analysis results.

  **⭳ Exhibit 15.11: TIMSS 2023 Context Scale Reliability and Principal Component Analysis – Grade 4**

  **⭳ Exhibit 15.12: TIMSS 2023 Context Scale Reliability and Principal Component Analysis – Grade 8**

- Exhibits 15.13 and 15.14 report country-level estimates of the relationship between the results of each scale and mathematics and science achievement.

  **⭳ Exhibit 15.13: TIMSS 2023 Context Scale Relationships with Achievement – Grade 4**

## ⭳ Exhibit 15.14: TIMSS 2023 Context Scale Relationships with Achievement – Grade 8

# References


Adams, R. J., & Khoo, S. T, (1996). Quest: The interactive test analysis system. Camberwell, Australia: Australian Council for Educational Research.

Brown, R. S. (2007). Using latent class analysis to set academic performance standards. *Educational Assessment*, *12*(3–4), 283–301.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*, 417–441, 498–520.

Jiao, H., Lissitz, R. W., Macready, G., Wang, S., & Liang, S. (2011). Exploring levels of performance using the mixture Rasch model for standard setting. *Psychological Test and Assessment Modeling*, *53*(4), 499–522.

Lazarsfeld, P. F. (1955). Recent developments in latent structure analysis. *Sociometry*, *18*(4), 391–403. https://doi.org/10.2307/2785875

Lazarsfeld, P. F., & Henry, N.W. (1968). *Latent structure analysis*. Houghton Mifflin, New York.

Martin, M. O., Mullis, I. V., Foy, P., & Arora, A. (2012). Creating and interpreting the TIMSS and PIRLS 2011 context questionnaire scales. In M. O. Martin & I. V. S. Mullis (Eds.), *Methods and procedures in TIMSS and PIRLS*. Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/methods/pdf/TP11_Context_Q_Scales.pdf

Martin, M. O., Mullis, I. V. S., Hooper, M., Yin, L., Foy, P., & Palazzo, L. (2016). Creating and Interpreting the TIMSS 2015 Context Questionnaire Scales. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015 Technical Report* (pp. 15.1–15.312). Boston College, TIMSS & PIRLS International Study Center. http://timss.bc.edu/publications/timss/2015-methods/chapter-15.html

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In: W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. Berlin: Springer.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176.

Peterson, A. R. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research, 21*(2), 381–391.

Preuschoff, A. C. (2010). *Using TIMSS and PIRLS to construct global indicators of effective environments for learning* [Doctoral dissertation, Boston College]. https://eric.ed.gov/?id=ED523015

Rost, J. & von Davier, M. (1994). A conditional item-fit index for Rasch models. *Applied Psychological Measurement, 18*(2), 171–182. https://doi.org/10.1177/014662169401800206

Taber, K. (2017). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48,* 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

von Davier, M., & Lee, Y. S. (2019). Introduction: From latent classes to cognitive diagnostic models. In M. von Davier & Y. S. Lee (Eds.), *Handbook of diagnostic classification models*. Springer, Cham. https://doi.org/10.1007/978-3-030-05584-4_1

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis: Rasch measurement. Chicago: MESA Press.

Wu, M. L., Adams, R. J, Wilson, M. R., & Haldane, S. (2007). Conquest 2.0 [computer software]. Camberwell, Australia: Australian Council for Educational Research.

Yin, L., Bezirhan, U., & von Davier, M. (2024). *Improving Context Scale Interpretation Using LCA for Cut Scores* [Paper presentation]. International Meeting of the Psychometric Society 2024, Prague, Czech Republic. https://www.psychometricsociety.org/imps-2024

Yin, L., & Fishbein, B. (2020). Creating and interpreting the TIMSS 2019 context questionnaire scales. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and Procedures: TIMSS 2019 Technical Report* (pp. 16.1–16.331). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/methods/chapter-16.html