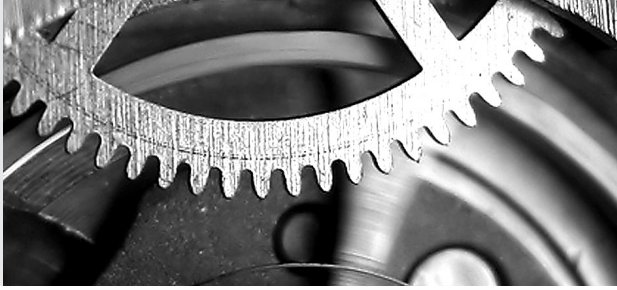
Chapter 11



Scaling the Data from the TIMSS 2007 Mathematics and Science Assessments

Pierre Foy, Joseph Galia, and Isaac Li

11.1 Overview

The TIMSS 2007 goals of broad coverage of the mathematics and science curriculum and of measuring trends across assessments necessitated a complex matrix-sampling booklet design,[1] with individual students responding to just a subset of the mathematics and science items in the assessment, and not the entire assessment item pool. Given the complexities of the data collection and the need to have student scores on the entire assessment for analysis and reporting purposes, TIMSS 2007 relied on Item Response Theory (IRT) scaling to describe student achievement on the assessment and to provide accurate measures of trends from previous assessments. The TIMSS IRT scaling approach used multiple imputation—or "plausible values"—methodology to obtain proficiency scores in mathematics and science for all students, even though each student responded to only a part of the assessment item pool. To enhance the reliability of the student scores, the TIMSS scaling combined student responses to the items they were administered with information about students' backgrounds, a process known as "conditioning."

This chapter first reviews the psychometric models and the conditioning and plausible values methodology used in scaling the TIMSS 2007 data, and then describes how this approach was applied to the TIMSS 2007 data and to the data from the previous TIMSS 2003 study, in order to measure trends in achievement. It also describes how "bridging" data, specifically collected in TIMSS 2007 to examine for any possible differences between the booklet designs from 2003 and 2007, were used in the scaling to preserve the TIMSS trend measures. The TIMSS scaling was conducted jointly by the TIMSS & PIRLS International Study Center

1 The TIMSS 2007 assessment design is described in Chapter 2.

at Boston College and Educational Testing Service, using software from Educational Testing Service.[2]

11.2 TIMSS 2007 Scaling Methodology[3]

The IRT scaling approach used by TIMSS was developed originally by Educational Testing Service for use in the U.S. National Assessment of Educational Progress. It is based on psychometric models that were first used in the field of educational measurement in the 1950s and have become popular since the 1970s for use in large-scale surveys, test construction, and computer adaptive testing.[4] This approach also has been used to scale IEA's PIRLS data to measure progress in reading literacy.

Three distinct IRT models, depending on item type and scoring procedure, were used in the analysis of the TIMSS 2007 assessment data. Each is a "latent variable" model that describes the probability that a student will respond in a specific way to an item in terms of the student's proficiency, which is an unobserved, or "latent", trait, and various characteristics (or "parameters") of the item. A three-parameter model was used with multiple-choice items, which were scored as correct or incorrect, and a two-parameter model for constructed-response items with just two response options, which also were scored as correct or incorrect. Since each of these item types has just two response categories, they are known as dichotomous items. A partial credit model was used with polytomous constructed-response items, i.e., those with more than two response options.

11.2.1 Two- and Three-Parameter IRT Models for Dichotomous Items

The fundamental equation of the three-parameter (3PL) model gives the probability that a student whose proficiency on a scale k is characterized by the unobservable variable θ_k will respond correctly to item i as:

$$(1) \qquad P\left(x_i = 1 \mid \theta_k, a_i, b_i, c_i\right) = c_i + \frac{1 - c_i}{1 + \exp\left(-1.7 \cdot a_i \cdot (\theta_k - b_i)\right)} \equiv P_{i,1}\left(\theta_k\right)$$

2 TIMSS is indebted to Matthias Von Davier, Ed Kulick, Scott Davis, and John Barone of Educational Testing Service for their advice and support.

3 This section describing the TIMSS scaling methodology has been adapted with permission from Chapter 14 of the *TIMSS 1999 Technical Report* (Yamamoto and Kulick, 2000).

4 For a description of IRT scaling see Birnbaum (1968); Lord and Novick (1968); Lord (1980); Van Der Linden and Hambleton (1996). The theoretical underpinning of the multiple imputation methodology was developed by Rubin (1987), applied to large-scale assessment by Mislevy (1991), and studied further by Mislevy, Johnson and Muraki (1992), and Beaton and Johnson (1992). The procedures used in TIMSS have been used in several other large-scale surveys, including Progress in International Reading Literacy Study (PIRLS), the U.S. National Assessment of Educational Progress (NAEP), the U.S. National Adult Literacy Survey (NALS), the International Adult Literacy Survey (IALS), and the International Adult Literacy and Life Skills Survey (IALLS).

where

x_i is the response to item i, 1 if correct and 0 if incorrect;

θ_k is the proficiency of a student on a scale k (note that a student with higher proficiency has a greater probability of responding correctly);

a_i is the slope parameter of item i, characterizing its discriminating power;

b_i is the location parameter of item i, characterizing its difficulty;

c_i is the lower asymptote parameter of item i, reflecting the chances of students with very low proficiency selecting the correct answer.

The probability of an incorrect response to the item is defined as:

$$(2) \qquad P_{i,0} = P\left(x_i = 0 \mid \theta_k, a_i, b_i, c_i\right) = 1 - P_{i,1}\left(\theta_k\right)$$

The two-parameter (2PL) model was used for the constructed-response items that were scored as either correct or incorrect. The form of the 2PL model is the same as Equations (1) and (2) with the c_i parameter fixed at zero.

11.2.2 IRT Model for Polytomous Items

In TIMSS 2007, as in previous study cycles, constructed-response items requiring an extended response were scored for partial credit, with 0, 1, and 2 as the possible score levels. These polytomous items were scaled using a generalized partial credit model (Muraki, 1992). The fundamental equation of this model gives the probability that a student with proficiency θ_k on scale k will have, for the i^{th} item, a response x_i that is scored in the l^{th} of m_i ordered score categories as:

$$(3) \qquad P\left(x_i = l \mid \theta_k, a_i, b_i, d_{i,1}, \cdots, d_{i,m_i-1}\right) = \frac{\exp\left(\sum_{v=0}^{l} 1.7 \cdot a_i \cdot \left(\theta_k - b_i + d_{i,v}\right)\right)}{\sum_{g=0}^{m_i-1} \exp\left(\sum_{v=0}^{g} 1.7 \cdot a_i \cdot \left(\theta_k - b_i + d_{i,v}\right)\right)} \equiv P_{i,l}\left(\theta_k\right)$$

where

m_i is the number of response categories for item i, usually 3;
x_i is the response to item i, ranging between 0 and $m_i - 1$;
θ_k is the proficiency of a student on a scale k;
a_i is the slope parameter of item i;
b_i is its location parameter, characterizing its difficulty;
$d_{i,l}$ is the category l threshold parameter ($l = 0, ..., m_i - 1$).

The indeterminacy of model parameters in the polytomous model is resolved by setting $d_{i,0} = 0$ and $\sum_{j=1}^{m_i - 1} d_{i,j} = 0$.

For all of the IRT models there is a linear indeterminacy between the values of item parameters and proficiency parameters, i.e., mathematically equivalent but different values of item parameters can be estimated on an arbitrarily linearly transformed proficiency scale. This linear indeterminacy can be resolved by setting the origin and unit size of the proficiency scale to arbitrary constants, such as a mean of 500 and a standard deviation of 100, as was done for TIMSS back in 1995. The indeterminacy is most apparent when the scale is set for the first time.

IRT modeling relies on a number of assumptions, the most important being conditional independence. Under this assumption, item response probabilities depend only on θ_k (a measure of a student's proficiency) and the specified parameters of the item, and are unaffected by the demographic characteristics or unique experiences of the students, the data collection conditions, or the other items presented in the test. Under this assumption, the joint probability of a particular response pattern x across a set of n items is given by:

$$(4) \qquad P\left(x \mid \theta_k, \textit{item parameters}\right) = \prod_{i=1}^{n} \prod_{l=0}^{m_i - 1} P_{i,l}\left(\theta_k\right)^{u_{i,l}}$$

where $P_{i,l}\left(\theta_k\right)$ is of the form appropriate to the type of item (dichotomous or polytomous), m_i is equal to 2 for dichotomously scored items, and $u_{i,l}$ is an indicator variable defined as:

$$(5) \qquad u_{i,l} = \begin{cases} 1 & \textit{if response } x_i \textit{ is in category } l; \\ 0 & \textit{otherwise}. \end{cases}$$

Replacing the hypothetical response pattern with the real scored data, the above function can be viewed as a likelihood function to be maximized by a given set of item parameters. In TIMSS 2007, the item parameters for each scale were estimated independently of the parameters of other scales. Once items were calibrated in this manner, a likelihood function for the proficiency θ_k was induced from student responses to the calibrated items. This likelihood function for the proficiency θ_k is called the posterior distribution of the θ's for each student.

11.2.3 Proficiency Estimation Using Plausible Values

Most cognitive skills testing is concerned with accurately assessing the performance of individual students for the purposes of diagnosis, selection, or placement. Regardless of the measurement model used, whether classical test theory or item response theory, the accuracy of these measurements can be improved—that is, the amount of measurement error can be reduced—by increasing the number of items given to the individual. Thus, it is common to see achievement tests designed to provide information on individual students that contain more than 70 items. Since the uncertainty associated with each θ in such tests is negligible, the distribution of θ, or the joint distribution of θ with other variables, can be approximated using each individual's estimated θ.

For the distribution of proficiencies in large populations, however, more efficient estimates can be obtained from a matrix-sampling design like that used in TIMSS. This design solicits relatively few responses from each sampled student while maintaining a wide range of content representation when responses are aggregated across all students. With this approach, however, the advantage of estimating population characteristics more efficiently is offset by the inability to make precise statements about individuals. The uncertainty associated with individual θ estimates becomes too large to be ignored. In this situation, aggregations of individual student scores can lead to seriously biased estimates of population characteristics (Wingersky, Kaplan, & Beaton, 1987).

Plausible values methodology was developed as a way to address this issue. Instead of first computing estimates of individual θ's and then aggregating these to estimate population parameters, the plausible values approach uses all available data, students' responses to the items they were administered together with all background data, to estimate directly the characteristics of student populations and subpopulations. Although these

directly estimated population characteristics could be used for reporting purposes, instead the usual plausible values approach is to generate multiple imputed scores, called plausible values, from the estimated ability distributions and to use these in analyses and reporting, making use of standard statistical software. By including all available background data in the model, a process known as "conditioning", relationships between these background variables and the estimated proficiencies will be appropriately accounted for in the plausible values. Because of this, analyses conducted using plausible values will provide an accurate representation of these underlying relationships. A detailed review of the plausible values methodology is given in Mislevy (1991).[5]

The following is a brief overview of the plausible values approach. Let y represent the responses of all sampled students to background questions or background data of sampled students collected from other sources, and let θ represent the proficiency of interest. If θ were known for all sampled students, it would be possible to compute a statistic $t(\theta,y)$, such as a sample mean or sample percentile point, to estimate a corresponding population quantity T.

Because of the latent nature of the proficiency, however, θ values are not known even for sampled students. The solution to this problem is to follow Rubin (1987) by considering θ as "missing data" and approximate $t(\theta,y)$ by its expectation given (x,y), the data that actually were observed, as follows:

$$
\begin{aligned}
t^*(x,y) &= E\left[\, t(\underline{\theta},\underline{y}) \mid \underline{x},\underline{y} \,\right] \\
&= \int t(\underline{\theta},\underline{y})\, p(\underline{\theta} \mid \underline{x},\underline{y})\; d\underline{\theta}
\end{aligned}
\tag{6}
$$

It is possible to approximate t^* using random draws from the conditional distribution of the scale proficiencies given the student's item responses x_j, the student's background variables y_j, and model parameters for the items. These values are referred to as imputations in the sampling literature, and as plausible values in large-scale surveys such as PIRLS, TIMSS, NAEP, NALS, and IALLS. The value of θ for any student that would enter into the computation of t is thus replaced by a randomly selected value from his or her conditional distribution. Rubin (1987) proposed repeating this process several times so that the uncertainly associated with imputation can be quantified. For example, the average of multiple estimates of t, each computed

5 Along with theoretical justifications, Mislevy presents comparisons with standard procedures; discusses biases that arise in some secondary analyses; and offers numerical examples.

from a different set of plausible values, is a numerical approximation of t^* of the above equation; the variance among them reflects the uncertainty due to not observing $\underline{\theta}$. It should be noted that this variance does not include the variability of sampling from the population. That variability is estimated separately by a jackknife variance estimation procedure, which is presented later in this chapter.

Plausible values are not intended to be estimates of individual student scores, but rather are imputed scores for like students—students with similar response patterns and background characteristics in the sampled population—that may be used to estimate population characteristics correctly. When the underlying model is correctly specified, plausible values will provide consistent estimates of population characteristics, even though they are not generally unbiased estimates of the proficiencies of the individuals with whom they are associated. Taking the average of the plausible values still will not yield suitable estimates of individual student scores.[6]

Plausible values for each student j are drawn from the conditional distribution $P\left(\theta_j \middle| x_j, y_j, \Gamma, \Sigma\right)$, where Γ is a matrix of regression coefficients for the background variables, and Σ is a common variance matrix of residuals. Using standard rules of probability, the conditional probability of proficiency can be represented as:

$$(7)\quad P\left(\theta_j \middle| x_j, y_j, \Gamma, \Sigma\right) \propto P\left(x_j \middle| \theta_j, y_j, \Gamma, \Sigma\right) P\left(\theta_j \middle| y_j, \Gamma, \Sigma\right) = P\left(x_j \middle| \theta_j\right) P\left(\theta_j \middle| y_j, \Gamma, \Sigma\right)$$

where θ_j is a vector of scale values, $P\left(x_j \middle| \theta_j\right)$ is the product over the scales of the independent likelihoods induced by responses to items within each scale, and $P\left(\theta_j \middle| y_j, \Gamma, \Sigma\right)$ is the multivariate joint density of proficiencies for the scales, conditional on the observed values y_j of background responses and parameters Γ and Σ. Item parameter estimates are fixed and regarded as population values in the computations described in this section.

11.2.4 Conditioning

A multivariate normal distribution was assumed for $P\left(\theta_j \middle| y_j, \Gamma, \Sigma\right)$, with a common variance Σ, and with a mean given by a linear model with regression parameters Γ. Since in large-scale studies like TIMSS there are many hundreds of background variables, it is customary to conduct a principal components analysis to reduce the number of variables to be

6 For further discussion, see Mislevy, Beaton, Kaplan, and Sheehan (1992).

used in Γ. Typically, components accounting for 90 percent of the variance in the data are selected. These principal components are referred to as the conditioning variables and denoted as y^c. The following model is then fit to the data:

$$\theta = \Gamma' y^c + \varepsilon \tag{8}$$

where ε is normally distributed with mean zero and variance Σ. As in a regression analysis, Γ is a matrix each of whose columns is the effects for each scale and Σ is the matrix of residual variance between scales.

Note that in order to be strictly correct for all functions Γ of θ, it is necessary that $P\left(\theta \middle| y\right)$ be correctly specified for all background variables in the survey. Estimates of functions Γ involving background variables not conditioned in this manner are subject to estimation error due to misspecification. The nature of these errors is discussed in detail in Mislevy (1991). In TIMSS 2007, however, principal component scores based on nearly all background variables were used. Those selected variables were chosen to reflect high relevance to policy and to education practices. The computation of marginal means and percentile points of θ for these variables is nearly optimal.

The basic method for estimating Γ and Σ with the Expectation and Maximization (EM) procedure is described in Mislevy (1985) for a single scale case. The EM algorithm requires the computation of the mean θ, and variance Σ, of the posterior distribution in equation (7).

11.2.5 Generating Proficiency Scores

After completing the EM algorithm, plausible values for all sampled students are drawn from the joint distribution of the values of Γ in a three-step process. First, a value of Γ is drawn from a normal approximation $P\left(\Gamma, \Sigma \middle| x_j, y_j\right)$ to that fixes Σ at the value $\hat{\Sigma}$ (Thomas, 1993). Second, conditional on the generated value of Γ (and the fixed value of $\Sigma = \hat{\Sigma}$), the mean θ_j and variance Σ_j^p of the posterior distribution in equation (7), where p is the number of scales, are computed using the methods applied in the EM algorithm. In the third step, the proficiency values are drawn independently from a multivariate normal distribution with mean θ_j and variance Σ_j^p

These three steps are repeated five times, producing five imputations of θ_j for each sampled student.

For students with an insufficient number of responses, the Γ's and Σ's described in the previous paragraph are fixed. Hence, all students—regardless of the number of items attempted—are assigned a set of plausible values.

The plausible values can then be employed to evaluate equation (6) for an arbitrary function T as follows:

- Using the first vector of plausible values for each student, evaluate T as if the plausible values were the true values of θ. Denote the result as T_1.
- Evaluate the sampling variance of T_1, or Var_1, with respect to students' first vector of plausible values.
- Carry out steps 1 and 2 for the second through fifth vectors of plausible values, thus obtaining T_u and Var_u for $u = 2, \ldots, 5$.
- The best estimate of T obtainable from the plausible values is the average of the five values obtained from the different sets of plausible values:

$$\hat{T} = \frac{\sum_u T_u}{5}$$

- An estimate of the variance of $\hat{T}$ is the sum of two components: an estimate of Var_u obtained by averaging as in the previous step, and the variance among the T_u's.

  Let $\bar{U} = \frac{\sum_u Var_u}{M}$, and let $B_M = \frac{\sum_u \left(T_u - \hat{T}\right)^2}{M-1}$ be the variance among the M plausible values. Then the estimate of the total variance of $\hat{T}$ is:

(9) $$Var\left(\hat{T}\right) = \bar{U} + \left(1 + M^{-1}\right) B_M$$

The first component in $Var\left(\hat{T}\right)$ reflects the uncertainty due to sampling students from the population; the second reflects the uncertainty due to the

fact that sampled students' θ's are not known precisely, but only indirectly through *x* and *y*.

11.2.6 Working with Plausible Values

The plausible values methodology was used in TIMSS 2007 to ensure the accuracy of estimates of the proficiency distributions for the TIMSS population as a whole and particularly for comparisons between subpopulations. A further advantage of this method is that the variation between the five plausible values generated for each student reflects the uncertainty associated with proficiency estimates for individual students. However, retaining this component of uncertainty requires that additional analytical procedures be used to estimate students' proficiencies.

If the θ values were observed for all sampled students, the statistic $\left(t-T\right)\Big/U^{1/2}$ would follow a *t*-distribution with *d* degrees of freedom. Then the incomplete-data statistic $\left(T-\hat{T}\right)\Big/\left[Var\left(\hat{T}\right)\right]^{1/2}$ is approximately *t*-distributed, with degrees of freedom (Johnson & Rust, 1992) given by:

$$\nu = \frac{1}{\frac{f_M^{\,2}}{M-1}+\frac{\left(1-f_M\right)^2}{d}} \tag{10}$$

where *d* is the degrees of freedom for the complete-data statistic, and f_M is the proportion of total variance due to not observing the θ values:

$$f_M = \frac{\left(1+M^{-1}\right)B_M}{Var\left(\hat{T}\right)} \tag{11}$$

When B_M is small relative to $\bar{U}$, the reference distribution for the incomplete-data statistic differs little from the reference distribution for the corresponding complete-data statistic. If, in addition, *d* is large, the normal approximation can be used instead of the *t*-distribution.

For a *k*-dimensional function *T*, such as the *k* coefficients in a multiple regression analysis, each *U* and $\bar{U}$ is a covariance matrix, and B_M is an average of squares and cross-products rather than simply an average of squares. In this case, the quantity $\left(\underline{T}-\underline{\hat{T}}\right)Var^{-1}\left(\underline{\hat{T}}\right)\left(\underline{T}-\underline{\hat{T}}\right)'$ is approximately

F-distributed with degrees of freedom equal to k and $\dot{\nu}$, with $\dot{\nu}$ defined as above but with a matrix generalization of f_M:

$$f_M = \left(1 + M^{-1}\right) Trace\left[B_M Var^{-1}\left(\hat{T}\right)\right] \Big/ k \qquad (12)$$

For the same reason that the normal distribution can approximate the *t*-distribution, a chi-square distribution with k degrees of freedom can be used in place of the *F*-distribution for evaluating the significance of the above quantity $\left(\underline{T} - \hat{\underline{T}}\right) Var^{-1}\left(\hat{\underline{T}}\right)\left(\underline{T} - \hat{\underline{T}}\right)$.

Statistics $\hat{T}$, the estimates of proficiency conditional on responses to cognitive items and background variables, are consistent estimates of the corresponding population values T, as long as background variables are included in the conditioning variables. The consequences of violating this restriction are described by Beaton & Johnson (1990), Mislevy (1991), and Mislevy & Sheehan (1987). To avoid such biases, the TIMSS 2007 analyses included all student background variables, as well as the class means to preserve between-class differences—the between- and within-classroom variance structure essential for hierarchical modeling.

11.3 Implementing the Scaling Procedures for the TIMSS 2007 Assessment Data

The application of IRT scaling and plausible values methodology to the TIMSS 2007 assessment data involved four major tasks: calibrating the achievement test items (estimating model parameters for each item), creating principal components from the student questionnaire data for use in conditioning; generating IRT scale scores (proficiency scores) for overall mathematics and science and for each of the mathematics and science content and cognitive domains; and placing the proficiency scores on the metric used to report the results from previous assessments.

The TIMSS eighth-grade reporting metric was established in 1995 by setting the average of the mean scores of the countries that participated in TIMSS 1995 at the eighth grade to 500 and the standard deviation to 100. To enable comparisons between 2007, 2003, 1999 and 1995, the TIMSS 2007, TIMSS 2003, and TIMSS 1999 eighth-grade data also were placed on this metric. This was done by concurrently scaling the assessment data from each successive TIMSS cycle with the assessment data from the previous cycle

and applying linear transformations to set the scores from each successive cycle on the same metric as the scores from the previous cycle. Placing the TIMSS 2007 eighth-grade results on this common metric permitted trend results from four points in time: 1995, 1999, 2003, and 2007.

The TIMSS fourth-grade reporting metric was set in much the same way as was done for the eighth grade, with the notable exception that TIMSS 1999 did not have a fourth-grade assessment. The TIMSS 2003 fourth-grade data were placed directly on the 1995 fourth-grade scale, which also had a mean of 500 and standard deviation of 100 based on the countries that participated in TIMSS 1995 at the fourth grade. This enabled comparisons between results from 1995 and 2003. Subsequently, the TIMSS 2007 fourth-grade data were put on the 1995 metric to produce trend results from all three survey cycles: 1995, 2003, and 2007. In 2007, as in previous TIMSS cycles, scale metrics were aligned for trend reporting only for overall mathematics and overall science; there were insufficient trend items from previous survey cycles to reliably measure trends in the content and cognitive domains.

11.3.1 The Bridging Study

In 2003, TIMSS introduced a new assessment design, consisting of a series of interlinked student booklets, each containing six blocks of assessment items.[7] From examination of the TIMSS 2003 data, it was apparent that not all students had sufficient time to complete their 2003 assessment booklets. This led to a "position effect",[8] whereby items positioned later in a booklet appeared to be more difficult than the same items positioned earlier in the booklet. The position effect was detectable because of the counterbalanced design of the 2003 assessment booklets. A new booklet design was introduced in TIMSS 2007, providing more time for students to respond to the items. Unlike the TIMSS 2003 booklets, which each contained six blocks of items, the TIMSS 2007 booklets each comprised just four of these blocks, to be completed in the same amount of time (i.e., 72 minutes at the fourth grade and 90 minutes at the eighth grade). Concerned that the 2007 assessment booklets might appear easier because students had more time, TIMSS implemented a "bridging study" to see if this was indeed the case. The bridging study involved the administration of a subset of the TIMSS 2003 assessment booklets at both grades in 2007 to establish a bridge between the 2003 and 2007 assessments. The data from the bridging study would

7 The TIMSS 2003 assessment design is described in the *TIMSS Assessment Frameworks and Specifications 2003 – 2nd Edition* (Mullis, et al., 2003).

8 The TIMSS 2003 position effect is described in the *TIMSS 2003 Technical Report* (Martin, et al., 2004, p. 264).

reveal if the change in booklet design from 2003 to 2007 had any effect on the difficulty of the achievement items, and if so, would provide a basis for maintaining the measurement of trends by adjusting for this effect.

It was important to establish that a subset of 2003 booklets could be a suitable representation of the TIMSS 2003 assessment as a whole. This evaluation was done by re-scaling the 2003 data using items only from four selected 2003 booklets: booklets 5, 6, 11, and 12. These were selected to maximize the number of common item blocks between the 2003 and 2007 assessments. A comparison of the resulting national average scale scores to the ones published in the 2003 international reports, showed that virtually all differences were well within sampling error. As well, an examination of Cronbach's alpha reliability coefficients across the set of items in these four booklets revealed that they remained as high, or nearly so, when compared to the reliability coefficients across all TIMSS 2003 items.

By inserting them into the rotation of the fourteen 2007 assessment booklets, the four bridge booklets were administered alongside the TIMSS 2007 assessment booklets to randomly equivalent samples of students in all trend countries (countries that participated in both TIMSS 2003 and TIMSS 2007).[9] All item blocks in the bridge booklets also were part of the TIMSS 2003 assessment, and four mathematics and four science blocks in the bridge booklets (at each grade level) also were included in the TIMSS 2007 assessment booklets. Presenting the same items using the 2007 bridge booklets and the 2007 assessment booklets allowed TIMSS to isolate the effect of changing the booklet design, and to provide enough data to adjust for this effect, as necessary.

A comparison of the average percent correct statistics of the common items in the 2007 bridge booklets and 2007 assessment booklets confirmed that the items were easier, on average, in the TIMSS 2007 assessment booklets, particularly at the eighth grade, as shown in Exhibit 11.1. The percent correct averaged across all fourth-grade mathematics items were 0.3% higher in the 2007 assessment booklets; the fourth-grade science items were 0.9% higher. The percent correct averaged across the eighth-grade mathematics items were 1.2% higher; the eighth-grade science items were 1.1% higher. Thus, because of the change in booklet design, the trend items in the TIMSS 2007 assessment booklets could not be assumed to have behaved as they had in the TIMSS 2003 booklets. The bridging data

9 The assignment of TIMSS 2007 bridge booklets and TIMSS 2007 assessment booklets was done automatically by the WinW3S software, as described in Chapter 6.

show what could have been expected if the booklet design had not been changed. Consequently, it was necessary to incorporate this effect into the trend scaling. The trend scaling of overall mathematics and overall science was performed by combining the assessment data from the TIMSS 2003 assessment booklets, the TIMSS 2007 bridge booklets, and the TIMSS 2007 assessment booklets using all items from the bridge booklets as trend items from the 2003 assessment and freeing all items in the 2007 assessment booklets to have their own IRT model parameters.

Exhibit 11.1 Overall Percent Correct and Percent Not Reached for Common Items in TIMSS 2007 Bridge Booklets and Assessment Booklets

Grade and Subject		Number of Common Items	TIMSS 2007 Bridge Booklets		TIMSS 2007 Assessment Booklets	
			Overall Percent Correct	Overall Percent Not Reached	Overall Percent Correct	Overall Percent Not Reached
Fourth Grade (19 Countries)	Mathematics	47	53.4	1.2	53.7	2.1
	Science	47	58.1	0.4	59.0	1.9
Eighth Grade (32 Countries)	Mathematics	52	44.6	0.2	45.8	1.3
	Science	57	43.6	0.1	44.7	1.2

11.3.2 Calibrating the TIMSS 2007 Assessment Data

As described in the TIMSS 2007 Assessment Frameworks (Mullis, Martin, Ruddock, O'Sullivan, Arora, & Erberber, 2005), the TIMSS 2007 achievement test design consisted of a total of 14 mathematics blocks and 14 science blocks at each grade, distributed across 14 assessment booklets. Each block contained either mathematics or science items, drawn from a range of content and cognitive domains. The 14 mathematics blocks were designated M01 through M14, and the 14 science blocks S01 through S14. All odd-numbered item blocks were previously used in the 2003 assessment and all even-numbered blocks consisted of newly-developed items for the 2007 assessment. Each assessment booklet contained four blocks—two mathematics and two science blocks. Two of the blocks (one mathematics and one science) were new in 2007 and two had previously been used in 2003.

The TIMSS 2007 test administration also included the four bridge booklets for trend countries, i.e., countries that also had participated in the 2003 assessment. Thus each sampled student in a trend country completed either one of the fourteen 2007 assessment booklets, or one of the four 2007 bridge booklets. Students in "non-trend" countries completed one of the fourteen 2007 assessment booklets. The booklets were distributed among the students in each sampled class according to a scheme that ensured

comparable random samples of students responded to each booklet, including the bridge booklets in trend countries.

In line with the TIMSS assessment framework, IRT scales were constructed for reporting overall student achievement in mathematics and science, as well as for reporting separately for each of the mathematics and science content and cognitive domains. Item calibration for the content and cognitive domains was conducted by the TIMSS & PIRLS International Study Center using the commercially-available Parscale software (Muraki & Bock, 1991). Item calibration for the overall mathematics and science scales was performed by ETS using their in-house version of Parscale and included data from the TIMSS 2003 assessment, the TIMSS 2007 assessment and the 2007 bridging study. The calibration was conducted using all available data from each country's TIMSS student samples and from all three assessments. All student samples were weighted so that each country contributed equally to the item calibration.

The first step in constructing the scales for TIMSS 2007 was to estimate the IRT model item parameters for each item on each of the scales. The trend scales for overall mathematics and science typically are based on a concurrent item calibration approach. The general concurrent calibration approach consists of three steps that look to build a linkage between the item calibration that was done in the previous assessment—called the previous calibration—and the current assessment. The first step consists of establishing a common set of item parameters for the two assessments through a concurrent calibration of both sets of assessment data, and setting common items to have the same item parameter estimates. It is then possible to obtain the mean and standard deviation of the latent ability distribution of students in both assessments under the concurrent calibration. The difference between these two distributions is the change in achievement from the previous to the current assessment. However, this difference is in the logit metric, and not the metric of the previous assessment, which would be necessary to measure growth.

The second step is to find the linear transformation that transforms the distribution of the previous assessment data under the concurrent calibration to match the distribution of these data under the previous calibration. The third step is to apply this same transformation to the current assessment data scaled using the concurrent calibration. This places the current assessment data on the metric of the previous assessment.

Exhibit 11.2 illustrates how the concurrent calibration approach customarily has been applied in the context of TIMSS trend scaling. The observed gap between both calibrations on the previous assessment data is generally small and arises from slight differences in the item parameter estimations, which in turn are due mostly to the previous assessment data being calibrated with other assessment data in the two calibrations. The linear transformation removes this gap by shifting the two distributions from the concurrent calibration, such that the distribution of the previous assessment from the concurrent calibration aligns with the distribution of the previous assessment from the previous calibration, while preserving the gap between the previous and current assessment data under the concurrent calibration. This latter gap is the change in achievement between the previous and current assessments that TIMSS seeks to measure as its trend.

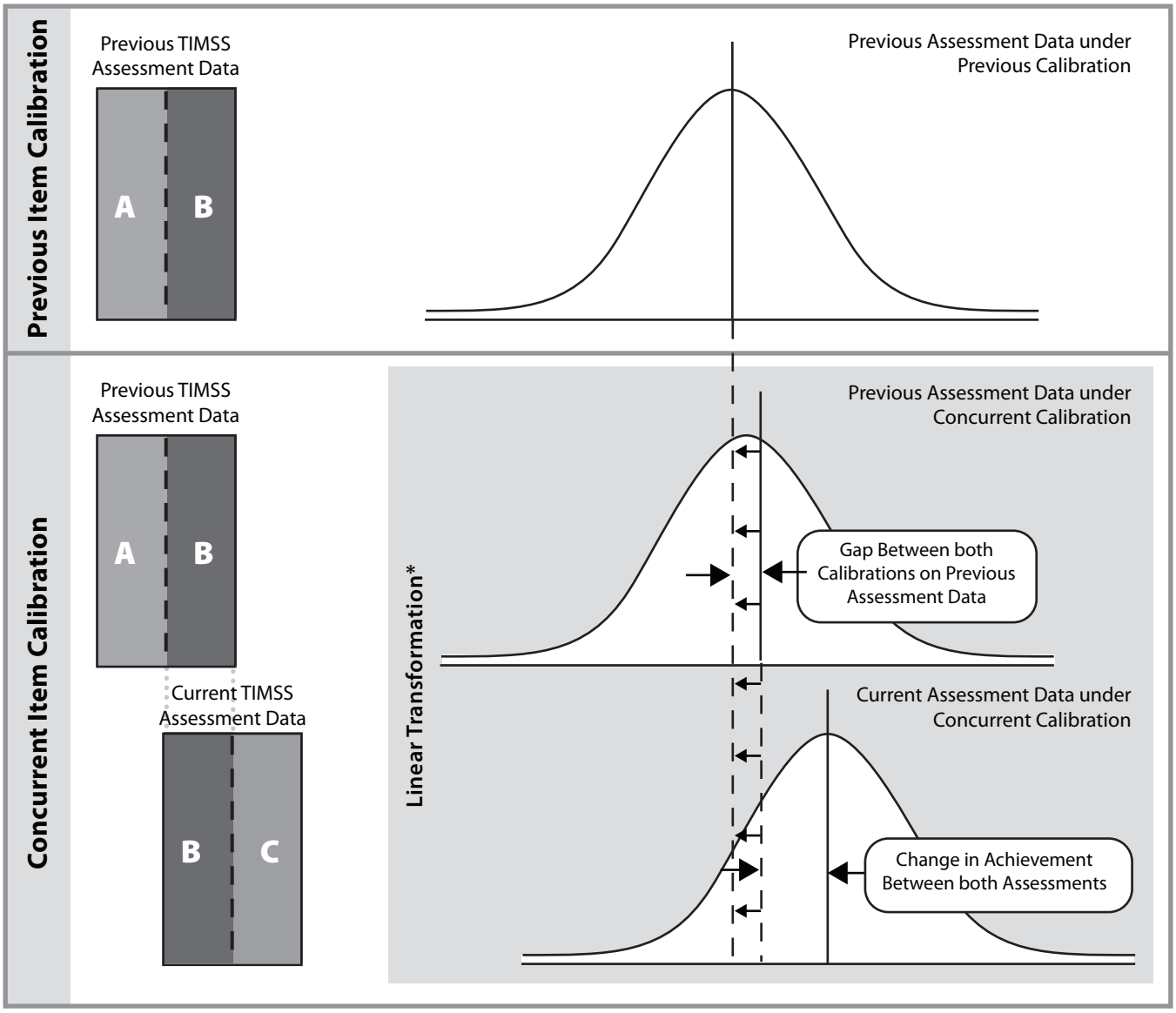
Because the bridging study demonstrated that the common items did not behave similarly across the 2003 and 2007 assessment booklets, it was necessary to adapt the concurrent calibration approach to include the 2007 bridging data. Accordingly, the 2007 concurrent calibration included the original 2003 data, the 2007 bridging data, and the 2007 data. Only countries that participated in both 2003 and 2007 were included in this concurrent calibration. All of the items contained in the 2007 bridge booklets also were contained in the 2003 booklets, so that these received the same item parameters in the concurrent scaling. This constituted the link between the 2003 assessment and the 2007 bridging data. The 2007 bridge booklets and the 2007 assessment booklets were administered to randomly equivalent samples of the 2007 assessment populations, which constituted the link between the 2007 bridging data and the 2007 assessment data.

Having estimated the item parameters from the concurrent calibration, new achievement distributions were generated by applying these item parameters to the 2003 assessment data, the 2007 bridging data, and the 2007 assessment data. Following the procedure outlined above, the next step was to identify the linear transformation that transformed the 2003 assessment distribution generated by the concurrent calibration item parameters to match the 2003 assessment distribution generated by the item parameters from the original 2003 calibration, and to apply this same transformation to the 2007 bridging data distribution (also generated by the concurrent calibration item parameters). An additional step, however, was required to establish a second linear transformation to make the distribution of the

2007 assessment data match the now-transformed distribution of the 2007 bridging data. This was done on the basis that both the 2007 assessment data and the 2007 bridging data came from randomly equivalent samples of the same 2007 assessment population.

Exhibit 11.3 demonstrates how this modified concurrent calibration approach was implemented in TIMSS 2007. As was explained in Exhibit 11.2, the gap between both calibrations on the 2003 assessment data was due largely to minor differences in the estimated item parameters arising from the fact that the 2003 assessment data were combined with the 1999 assessment data (the 1995 assessment data at the fourth grade) in the 2003 calibration and combined with the 2007 bridging data and 2007 assessment data in the 2007 calibration. The first linear transformation served to remove this gap while preserving the gap between the 2003 assessment data and the 2007 bridging data under the 2007 concurrent calibration, which was the change in achievement used to determine the TIMSS measure of trend. Finally, the gap between the 2007 bridging data and 2007 assessment data was primarily the result of minor sampling differences across the national samples of students between the two sets of data and was removed by the second linear transformation, which aligned the distribution of the 2007 assessment data with the distribution of the 2007 bridging data.

Exhibit 11.2 Concurrent Calibration Model Used Traditionally for TIMSS



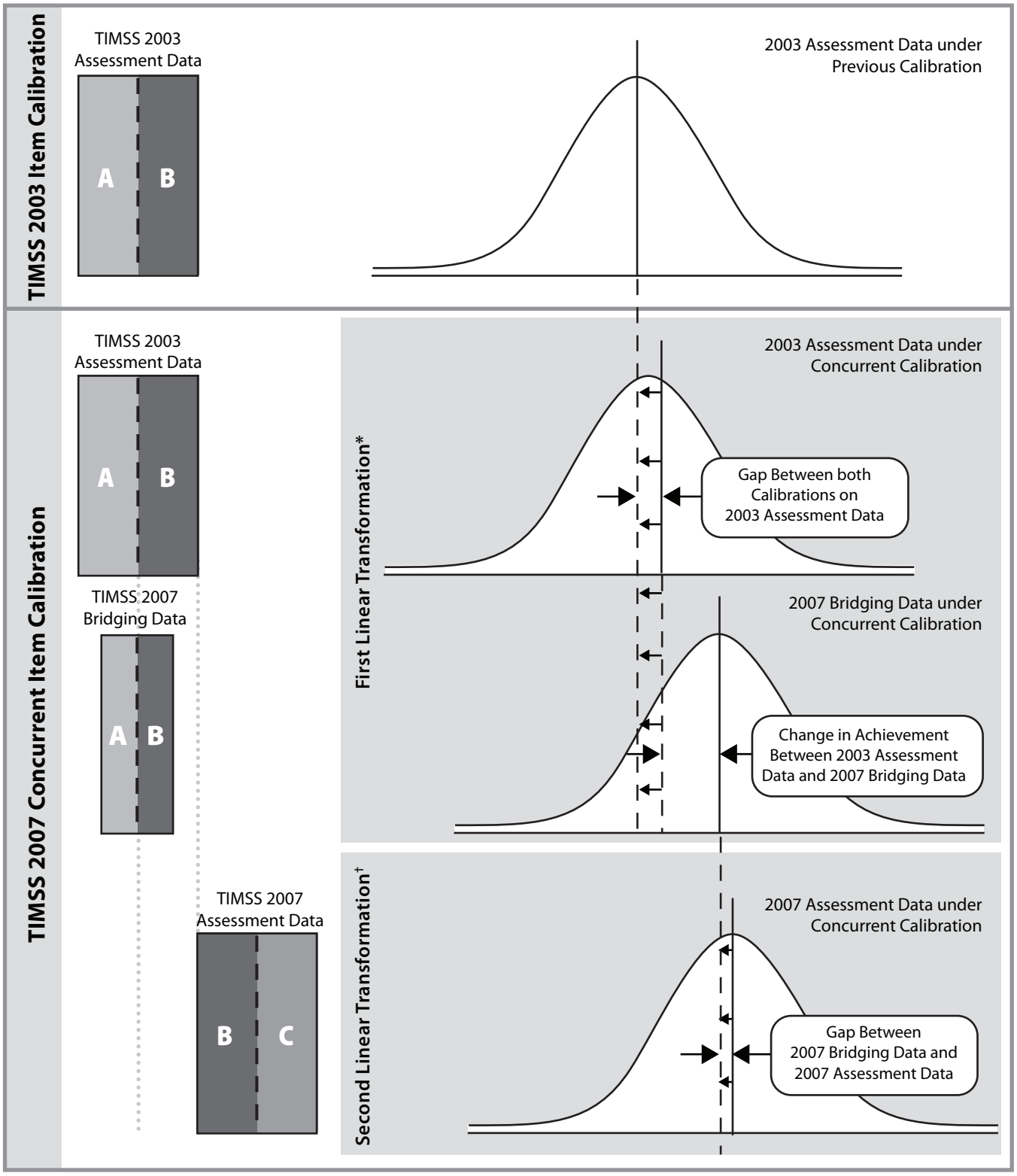
A *Item Blocks Released after Previous Assessment*

B *Item Blocks Secured for Future Assessments*

C *Item Blocks Developed in Current Assessment*

* *The two distributions under the concurrent calibration are transformed through a linear transformation such that the distribution of the previous assessment under concurrent calibration aligns with the distribution of the previous assessment under the previous calibration*

Exhibit 11.3 Concurrent Calibration Model Used for TIMSS 2007



A *Item Blocks Released after 2003 Assessment*

B *Item Blocks Secured for Future Assessments*

C *Item Blocks Developed in 2007 Assessment*

* *The distributions of the 2003 assessment and 2007 bridging under the concurent calibration are transformed through a linear transformation such that the distribution of the 2003 assessment under concurrent calibration aligns with the distribution of the 2003 assessment under the previous calibration*

† *The distribution of the 2007 assessment is aligned with the distribution of the 2007 bridging through a second linear transformation*

Exhibit 11.4 shows the distribution of items included in the TIMSS 2007 concurrent calibrations for reporting trends in overall mathematics and science at both grades. All data were included from the 2003 and 2007 assessments, as well as the data from the 2007 bridge booklets to account for the modified TIMSS 2007 assessment design. Items were categorized as items unique to the TIMSS 2003 assessment, items in the TIMSS 2007 bridge booklets—which by design also were included in the TIMSS 2003 assessment and constituted the set of common items—and items in the TIMSS 2007 assessment booklets. Taking eighth-grade mathematics as an example, the TIMSS 2007 assessment booklets contributed 214 items worth 236 points, the TIMSS 2007 bridge booklets contributed 151 items worth 165 points (these same items were also in the TIMSS 2003 assessment booklets), and there were 216 items worth 237 points unique to the TIMSS 2003 assessment booklets.

Exhibit 11.4 Items Included in the TIMSS 2007 Concurrent Item Calibrations of Overall Mathematics and Science

TIMSS 2007 Trend Scales		Items in TIMSS 2007 Assessment Booklets		Items in TIMSS 2007 Bridge Booklets		Items Unique to TIMSS 2003 Assessment Booklets		TOTAL	
		Number	Points	Number	Points	Number	Points	Number	Points
Fourth Grade	Mathematics	177	188	125	130	171	179	473	497
	Science	170	189	119	130	159	175	448	494
Eighth Grade	Mathematics	214	236	151	165	216	237	581	638
	Science	210	231	151	163	202	220	563	614

At the fourth grade, to construct separate overall mathematics and science scales for reporting trends, as well as performance generally in 2007, concurrent item calibrations were conducted using data from the 21 countries that participated in both 2003 and 2007 assessments. These calibrations included 93,863 student records from the 2003 assessment, 25,952 records from the 2007 bridging study, and 91,204 records from the 2007 assessment, for a total of 211,019 student records. The item parameters established in these calibrations were used subsequently for estimating student scores for all 37 countries and 7 benchmarking entities that participated in 2007.

At the eighth grade, concurrent item calibrations for the overall mathematics and science scales were conducted using data from the 33 countries that participated in both 2003 and 2007 assessments. They included 158,477 student records from the 2003 assessment, 41,377 records

from the 2007 bridging study, and 145,349 records from the 2007 assessment, for a total of 345,203 student records. The item parameters established in these calibrations were used subsequently for estimating student scores for all 50 countries and 7 benchmarking entities that participated in 2007. All countries and their samples included in these calibrations for reporting trends are presented in Exhibit 11.5.

Because there were insufficient items to construct reliable scales for measuring trends in each of the content and cognitive domains, scales for these domains were constructed using 2007 data only. At the fourth grade, separate calibrations were conducted for each of the three mathematics and three science content domains and the three mathematics and three science cognitive domains. These calibrations were based on 160,922 student records from the 36 countries that participated in the 2007 assessment.[10] Similarly at the eighth grade, separate calibrations were conducted for each of the four mathematics and four science content domains and the three mathematics and three science cognitive domains. These calibrations were based on 220,788 student records from the 49 countries that participated in the 2007 assessment at the eighth grade.[10] All countries and their samples included in the item calibrations for the content and cognitive domains are presented in Exhibit 11.6.

Item calibrations for the content and cognitive domains included only the items from the TIMSS 2007 assessment booklets. Exhibit 11.7 and Exhibit 11.8 show the number of items and score points included in each content and cognitive domain at the fourth and eighth grades, respectively.

Exhibits D.1 through D.30 in Appendix D present the item parameters generated from all item calibrations. In Exhibits D.1 through D.4, items where the parameters were freed in 2003, to address the position effect in 2003, have an "F" in the second character position of the item label. All items from the TIMSS 2007 assessment booklets have the letter "Z" in the second character position of the item label. As a by-product of the calibrations, interim scores in mathematics, science, and all content and cognitive domains were produced for use in constructing conditioning variables.

10 Data from Mongolia and the seven benchmarking participants were not included in these item calibrations.

Exhibit 11.5 Sample Sizes for Item Calibrations of Overall Mathematics and Science for Countries Participating in both TIMSS 2003 and TIMSS 2007

Country	Fourth Grade			Eighth Grade		
	TIMSS 2003 Assessment Booklets	TIMSS 2007 Bridge Booklets	TIMSS 2007 Assessment Booklets	TIMSS 2003 Assessment Booklets	TIMSS 2007 Bridge Booklets	TIMSS 2007 Assessment Booklets
Armenia	5,674	1,139	4,079	5,726	1,307	4,689
Australia	4,321	1,186	4,108	4,791	1,164	4,069
Bahrain	—	—	—	4,199	1,210	4,230
Botswana	—	—	—	5,150	1,197	4,208
Bulgaria	—	—	—	4,117	1,141	4,019
Chinese Taipei	4,661	1,192	4,131	5,379	1,155	4,046
Cyprus	—	—	—	4,002	1,255	4,399
Egypt	—	—	—	7,095	1,871	6,582
England	3,585	1,208	4,316	2,830	1,159	4,025
Ghana	—	—	—	5,100	1,498	5,294
Hong Kong SAR	4,608	1,072	3,791	4,972	986	3,470
Hungary	3,319	1,155	4,048	3,302	1,183	4,111
Indonesia	—	—	—	5,762	967	3,374
Iran, Islamic Rep. of	4,352	1,087	3,833	4,942	1,115	3,981
Israel	—	—	—	4,318	926	3,294
Italy	4,282	1,277	4,470	4,278	1,242	4,408
Japan	4,535	1,274	4,487	4,856	1,221	4,312
Jordan	—	—	—	4,489	1,492	5,251
Korea, Rep. of	—	—	—	5,309	1,208	4,240
Latvia	2,451	1,101	3,908	—	—	—
Lebanon	—	—	—	3,814	1,073	3,786
Lithuania	4,422	1,134	3,980	4,964	1,141	3,991
Malaysia	—	—	—	5,314	1,285	4,466
Morocco	4,264	1,090	3,894	—	—	—
Netherlands	2,937	962	3,349	—	—	—
New Zealand	4,254	1,405	4,940	—	—	—
Norway	4,342	1,165	4,108	4,133	1,317	4,627
Palestinian Nat'l Auth.	—	—	—	5,357	1,253	4,378
Romania	—	—	—	4,104	1,201	4,198
Russian Federation	3,963	1,277	4,464	4,667	1,277	4,472
Scotland	3,936	1,123	3,929	3,516	1,156	4,070
Serbia	—	—	—	4,296	1,153	4,045
Singapore	6,668	1,440	5,041	6,018	1,329	4,599
Slovenia	3,126	1,244	4,351	3,578	1,150	4,043
Sweden	—	—	—	4,256	1,473	5,215
Tunisia	4,334	1,160	4,081	4,931	1,175	4,080
United States	9,829	2,261	7,896	8,912	2,097	7,377
Total	**93,863**	**25,952**	**91,204**	**158,477**	**41,377**	**145,349**

Exhibit 11.6 Sample Sizes for Scaling the Content and Cognitive Domains for All Countries Participating in TIMSS 2007

Country	Fourth Grade		Eighth Grade	
	Item Calibration	Proficiency Estimation	Item Calibration	Proficiency Estimation
Algeria	4,223	4,223	5,447	5,447
Armenia	4,079	4,079	4,689	4,689
Australia	4,108	4,108	4,069	4,069
Austria	4,859	4,859	—	—
Bahrain	—	—	4,230	4,230
Bosnia and Herzegovina	—	—	4,220	4,220
Botswana	—	—	4,208	4,208
Bulgaria	—	—	4,019	4,019
Chinese Taipei	4,131	4,131	4,046	4,046
Colombia	4,801	4,801	4,873	4,873
Cyprus	—	—	4,399	4,399
Czech Republic	4,235	4,235	4,845	4,845
Denmark	3,519	3,519	—	—
Egypt	—	—	6,582	6,582
El Salvador	4,166	4,166	4,063	4,063
England	4,316	4,316	4,025	4,025
Georgia	4,108	4,108	4,178	4,178
Germany	5,200	5,200	—	—
Ghana	—	—	5,294	5,294
Hong Kong SAR	3,791	3,791	3,470	3,470
Hungary	4,048	4,048	4,111	4,111
Indonesia	—	—	4,203	4,203
Iran, Islamic Rep. of	3,833	3,833	3,981	3,981
Israel	—	—	3,294	3,294
Italy	4,470	4,470	4,408	4,408
Japan	4,487	4,487	4,312	4,312
Jordan	—	—	5,251	5,251
Korea, Rep. of	—	—	4,240	4,240
Kazakhstan	3,990	3,990	—	—
Kuwait	3,803	3,803	4,091	4,091
Latvia	3,908	3,908	—	—
Lebanon	—	—	3,786	3,786
Lithuania	3,980	3,980	3,991	3,991
Malaysia	—	—	4,466	4,466
Malta	—	—	4,670	4,670
Mongolia	—	4,523	—	4,499
Morocco	3,894	3,894	3,060	3,060
Netherlands	3,349	3,349	—	—
New Zealand	4,940	4,940	—	—
Norway	4,108	4,108	4,627	4,627
Oman	—	—	4,752	4,752
Palestinian Nat'l Auth.	—	—	4,378	4,378
Qatar	7,019	7,019	7,184	7,184
Romania	—	—	4,198	4,198
Russian Federation	4,464	4,464	4,472	4,472
Saudi Arabia	—	—	4,243	4,243
Scotland	3,929	3,929	4,070	4,070
Serbia	—	—	4,045	4,045
Singapore	5,041	5,041	4,599	4,599
Slovak Republic	4,963	4,963	—	—
Slovenia	4,351	4,351	4,043	4,043
Sweden	4,676	4,676	5,215	5,215
Syrian Arab Republic	—	—	4,650	4,650
Thailand	—	—	5,412	5,412
Tunisia	4,134	4,134	4,080	4,080
Turkey	—	—	4,498	4,498
Ukraine	4,292	4,292	4,424	4,424
United States	7,896	7,896	7,377	7,377
Yemen	5,811	5,811	—	—
Benchmarking Participants				
Alberta, Canada	—	4,037	—	—
Basque Country, Spain	—	—	—	2,296
British Columbia, Canada	—	4,153	—	4,256
Dubai, UAE	—	3,064	—	3,195
Massachusetts, US	—	1,747	—	1,897
Minnesota, US	—	1,846	—	1,777
Ontario, Canada	—	3,496	—	3,448
Quebec, Canada	—	3,885	—	3,956
Total	**160,922**	**187,673**	**220,788**	**246,112**

Exhibit 11.7 TIMSS 2007 Items by Content and Cognitive Domains at the Fourth Grade

TIMSS 2007 Scales in the Content and Cognitive Domains at the Fourth Grade			Items in TIMSS 2007 Assessment Booklets	
			Number	Points
Mathematics	Overall		177	188
	Content Domains	Number	91	96
		Geometric Shapes and Measures	60	64
		Data Display	26	28
	Cognitive Domains	Knowing	68	71
		Applying	70	74
		Reasoning	39	43
Science	Overall		170	189
	Content Domains	Life Science	71	81
		Physical Science	64	66
		Earth Science	35	42
	Cognitive Domains	Knowing	74	84
		Applying	63	68
		Reasoning	33	37

Exhibit 11.8 TIMSS 2007 Items by Content and Cognitive Domains at the Eighth Grade

TIMSS 2007 Scales in the Content and Cognitive Domains at the Eighth Grade			Items in TIMSS 2007 Assessment Booklets	
			Number	Points
Mathematics	Overall		214	236
	Content Domains	Number	63	72
		Algebra	64	69
		Geometry	47	49
		Data and chance	40	46
	Cognitive Domains	Knowing	81	83
		Applying	88	97
		Reasoning	45	56
Science	Overall		210	231
	Content Domains	Biology	75	86
		Chemistry	41	45
		Physics	54	57
		Earth Science	40	43
	Cognitive Domains	Knowing	83	87
		Applying	84	95
		Reasoning	43	49

11.3.3 Omitted and Not-Reached Responses

Apart from missing data on items that by design were not administered to a student, missing data could also occur because a student did not answer an item—whether because the student did not know the answer, omitted it by mistake, or did not have time to attempt the item. An item was considered not reached when—within part 1 or part 2 of the booklet—the item itself and the item immediately preceding it were not answered, and there were no other items completed in the remainder of that part of the booklet.

In TIMSS 2007, as in previous TIMSS assessments, not-reached items were treated differently in estimating item parameters and in generating student proficiency scores. In estimating the values of the item parameters, items in the TIMSS 2007 assessment booklets that were considered not to have been reached by students were treated as if they had not been administered. This approach was considered optimal for parameter estimation. Because of the position effect described earlier, items located in positions 3 and 6 of the test booklets in the TIMSS 2003 assessment data and TIMSS 2007 bridging data that were considered not to have been reached by the students were treated as incorrect. However, not-reached items were always considered as incorrect responses when student proficiency scores were generated.

11.3.4 Evaluating Fit of IRT Models to the TIMSS 2007 Data

After the item calibrations were completed, checks were performed to verify that the item parameters obtained from Parscale adequately reproduced the observed distribution of student responses across the proficiency continuum. The fit of the IRT models to the TIMSS 2007 data was examined by comparing the item response function curves generated using the item parameters estimated from the data with the empirical item response functions calculated from the posterior distributions of the θ's for each student that responded to the item. When the empirical results fall near the fitted curves for any given item, the IRT model fits the data well and leads to more accurate and reliable measurement of the underlying proficiency scale. Graphical plots of these response function curves are called item characteristic curves (ICC).

Exhibit 11.9 shows an ICC plot of the empirical and fitted item response functions for a dichotomous item. In the plot, the horizontal axis represents the proficiency scale, 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. Empirical results are represented by triangles. The empirical results were obtained by first dividing the proficiency scale into intervals of equal size and then counting the number of students responding to the item whose EAP scores from Parscale fell in each interval. Then the proportion of students in each interval that responded correctly to the item was calculated. In the exhibit, the center of each triangle represents this empirical proportion of correct responses. The size of each triangle is proportional to the number of students contributing to the estimation of its empirical proportion correct.

Exhibit 11.9 TIMSS 2007 Mathematics Assessment Example Item Response Function for a Dichotomous Item

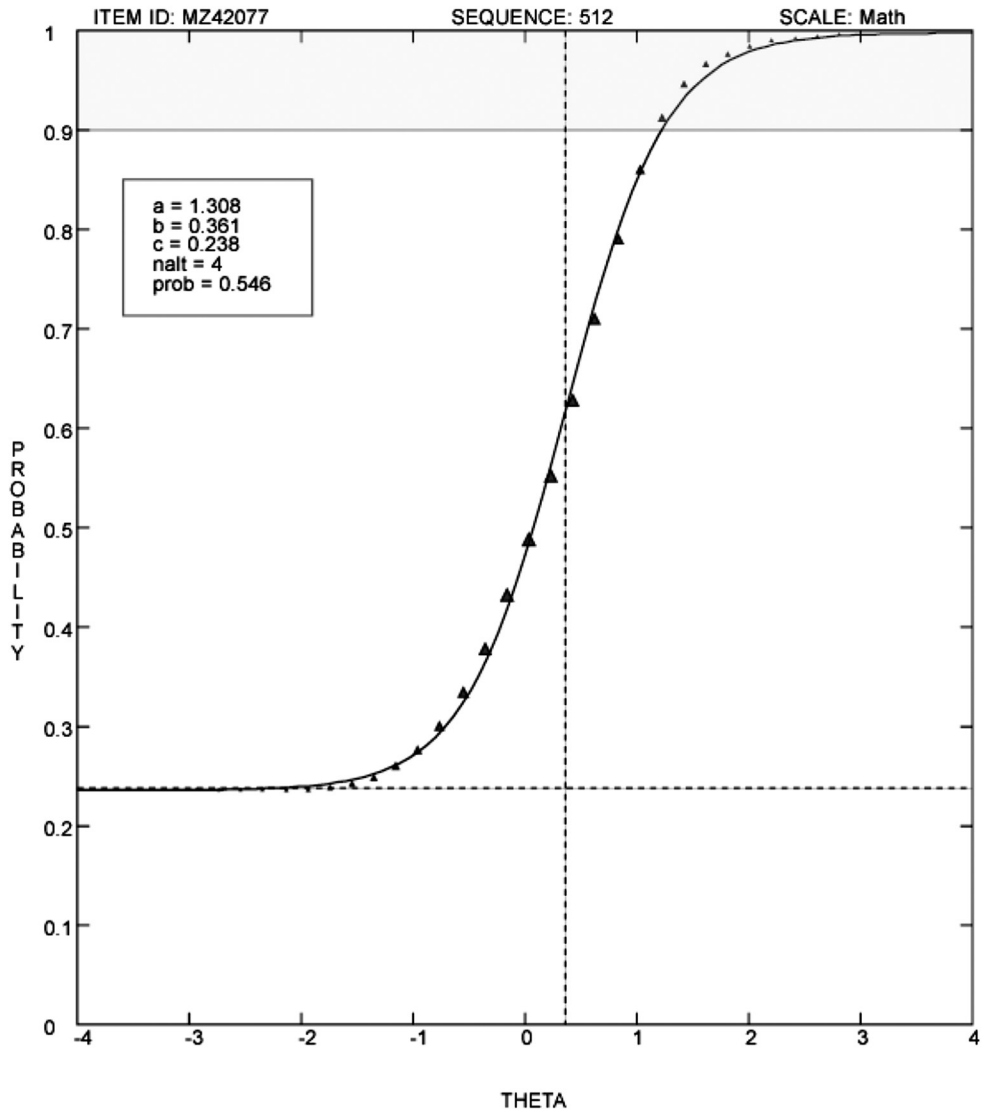


Exhibit 11.10 TIMSS 2007 Mathematics Assessment Example Item Response Function for a Polytomous Item

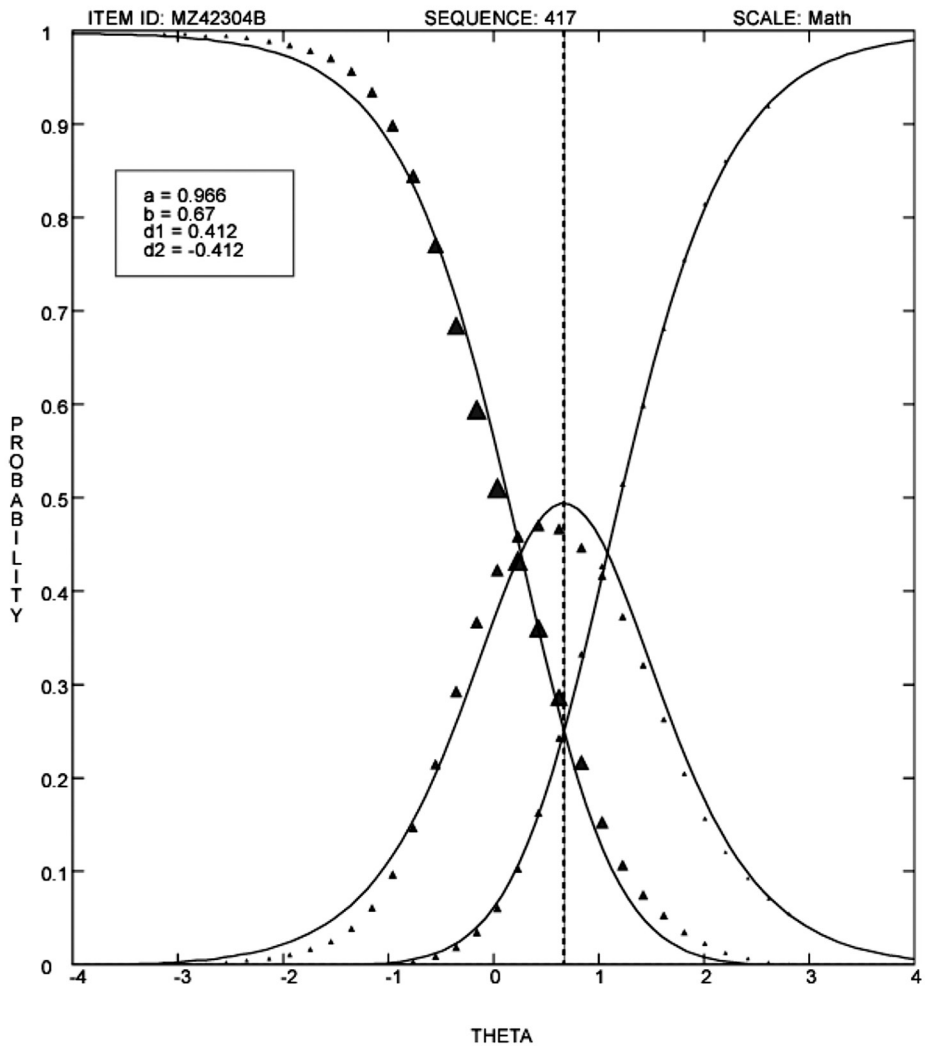


Exhibit 11.10 contains an ICC plot of the empirical and fitted item response functions for a polytomous item. As for the dichotomous item plot, the horizontal axis represents the proficiency scale, but the vertical axis represents the probability of having a response in a given response category. The fitted curves based on the estimated item parameters are shown as solid lines. Empirical results are represented by triangles. The interpretation of the triangles is the same as in Exhibit 11.9. The curve starting at the top left of the chart plots the probability of a score of zero on the item, which decreases as θ increases. The bell-shaped curve shows the probability of a score of one point—starting low 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 θ increases.

11.3.5 Variables for Conditioning the TIMSS 2007 Data

Because there were so many background variables that could be used in conditioning, TIMSS followed the practice established by NAEP and followed by other large-scale studies of using principal components analysis to reduce the number of variables while explaining most of their common variance. Principal components for the TIMSS 2007 background data were constructed as follows:

- For categorical variables (questions with a small number of fixed response options), a "dummy coded" variable was created for each response option, with a value of one if the option was chosen and zero otherwise. If a student omitted or was not administered a particular question, all dummy coded variables associated with that question were assigned the value zero.
- Background variables with numerous response options (such as year of birth or number of people who live in the home) were recoded using criterion scaling.[11] This was done by replacing each response option with an interim achievement score. For the overall mathematics and science scales, the interim achievement scores were the average across the interim mathematics and science scores produced from the item calibration. For the content domain scales, the interim achievement scores from the calibration in each subject were averaged to form a composite mathematics and a composite science score, and the average of these composite scores was used as the interim achievement score.
- Separately for each TIMSS country, all the dummy-coded and criterion-scaled variables were included in a principal components analysis. Those principal components accounting for 90 percent of the variance of the background variables were retained for use as conditioning variables. Because the principal components analysis was performed separately for each country, different numbers of principal components were required to account for 90% of the common variance in each country's background variables.[12]

11 The process of generating criterion-scaled variables is described in Beaton (1969).

12 The criterion was reduced to 80% when applied to the TIMSS 2007 bridging data because of the smaller student sample sizes.

In addition to the principal components, student gender (dummy coded), the language of the test (dummy coded), an indicator of the classroom in the school to which the student belonged (criterion scaled), and an optional country-specific variable (dummy coded) were included as primary conditioning variables, thereby accounting for most of the variance between students and preserving the between- and within-classrooms variance structure in the scaling model. Exhibit 11.11 and Exhibit 11.12 show the total number of variables that were used in the principal component analysis and the number of principal components selected within each country. Conditioning variables were needed for the TIMSS 2007 assessment data of all participants, as well as for the TIMSS 2007 bridging data and the TIMSS 2003 assessment data of all trend countries.

Exhibit 11.11 Number of Variables and Principal Components for Conditioning in TIMSS 2007 at the Fourth Grade

Country	TIMSS 2003 Assessment Booklets			TIMSS 2007 Bridge Booklets			TIMSS 2007 Assessment Booklets		
	Number of Primary Conditioning Variables	Total Number of Principal Components	Number of Principal Components Retained	Number of Primary Conditioning Variables	Total Number of Principal Components	Number of Principal Components Retained	Number of Primary Conditioning Variables	Total Number of Principal Components	Number of Principal Components Retained
Algeria	—	—	—	—	—	—	2	285	172
Armenia	2	291	178	2	287	114	2	287	172
Australia	2	301	166	2	293	110	2	293	163
Austria	—	—	—	—	—	—	2	293	168
Chinese Taipei	2	313	172	2	293	116	2	293	165
Colombia	—	—	—	—	—	—	2	285	168
Czech Republic	—	—	—	—	—	—	2	293	168
Denmark	—	—	—	—	—	—	2	285	159
El Salvador	—	—	—	—	—	—	2	293	173
England	2	295	165	2	291	115	2	291	165
Georgia	—	—	—	—	—	—	2	289	171
Germany	—	—	—	—	—	—	2	293	163
Hong Kong SAR	2	313	171	3	291	110	3	293	160
Hungary	2	307	172	2	291	115	2	291	166
Iran, Islamic Rep. of	2	305	172	2	293	115	2	293	170
Italy	2	311	173	2	236	110	2	237	152
Japan	2	313	175	2	293	116	2	293	165
Kazakhstan	—	—	—	—	—	—	3	291	158
Kuwait	—	—	—	—	—	—	2	285	171
Latvia	3	313	173	2	293	110	2	293	164
Lithuania	2	290	163	2	293	114	2	293	166
Moldova, Rep. of	—	—	—	—	—	—	3	291	145
Mongolia	—	—	—	—	—	—	3	277	165
Morocco	2	297	177	2	291	118	2	291	174
Netherlands	2	289	164	2	285	108	2	285	160
New Zealand	8	311	174	7	293	120	7	293	168
Norway	2	313	177	2	293	114	2	293	165
Qatar	—	—	—	—	—	—	3	291	176
Russian Federation	2	241	134	2	293	114	2	293	167
Scotland	2	295	168	2	291	115	2	291	166
Singapore	2	301	170	2	293	118	2	293	164
Slovak Republic	—	—	—	—	—	—	3	293	169
Slovenia	2	313	172	2	293	119	2	293	168
Sweden	—	—	—	—	—	—	2	293	166
Tunisia	2	311	184	2	293	123	2	293	176
Ukraine	—	—	—	—	—	—	3	291	169
United States	8	287	168	7	283	125	7	283	166
Yemen	—	—	—	—	—	—	2	285	180
Benchmarking Participants									
Alberta, Canada	—	—	—	—	—	—	3	287	162
British Columbia, Canada	—	—	—	—	—	—	3	287	162
Dubai, UAE	—	—	—	—	—	—	3	291	163
Massachusetts, US	—	—	—	—	—	—	2	281	155
Minnesota, US	—	—	—	—	—	—	2	283	156
Ontario, Canada	3	291	160	3	287	103	3	287	159
Quebec, Canada	3	291	165	3	287	108	3	287	162

Exhibit 11.12 Number of Variables and Principal Components for Conditioning in TIMSS 2007 at the Eighth Grade

Country	TIMSS 2003 Assessment Booklets			TIMSS 2007 Bridge Booklets			TIMSS 2007 Assessment Booklets		
	Number of Primary Conditioning Variables	Total Number of Principal Components	Number of Principal Components Retained	Number of Primary Conditioning Variables	Total Number of Principal Components	Number of Principal Components Retained	Number of Primary Conditioning Variables	Total Number of Principal Components	Number of Principal Components Retained
Algeria	—	—	—	—	—	—	3	811	391
Armenia	2	891	430	3	892	233	3	892	445
Australia	2	417	225	3	399	139	3	399	217
Bahrain	3	429	242	4	396	152	4	396	226
Bosnia and Herzegovina	—	—	—	—	—	—	5	895	453
Botswana	2	424	248	3	399	162	3	399	237
Bulgaria	2	913	409	3	899	179	3	899	375
Chinese Taipei	2	432	231	3	396	139	3	396	208
Colombia	—	—	—	—	—	—	3	388	225
Cyprus	2	897	420	3	897	218	3	898	407
Czech Republic	—	—	—	—	—	—	3	900	460
Egypt	4	418	249	4	396	167	4	396	237
El Salvador	—	—	—	—	—	—	3	399	230
England	2	410	216	3	375	135	3	381	207
Georgia	—	—	—	—	—	—	3	895	416
Ghana	2	410	245	3	399	163	3	399	236
Hong Kong SAR	2	432	233	3	399	135	3	399	211
Hungary	2	907	437	3	898	241	3	899	445
Indonesia	2	633	336	3	899	231	3	901	421
Iran, Islamic Rep. of	2	424	243	3	399	151	3	399	228
Israel	3	432	241	4	396	145	4	396	222
Italy	2	430	234	3	325	137	3	326	198
Japan	2	425	231	3	394	139	3	395	212
Jordan	2	432	247	3	396	154	3	396	229
Korea, Rep. of	2	432	234	3	377	141	3	396	214
Kuwait	—	—	—	—	—	—	3	386	221
Lebanon	2	745	376	4	734	194	4	734	361
Lithuania	3	811	392	3	900	233	3	900	442
Malaysia	2	412	231	3	396	150	3	397	220
Malta	—	—	—	—	—	—	3	897	409
Moldova, Rep. of	—	—	—	—	—	—	4	867	319
Mongolia	—	—	—	—	—	—	4	897	425
Morocco	—	—	—	—	—	—	3	891	403
Norway	2	429	236	3	396	146	3	396	217
Oman	—	—	—	—	—	—	4	396	231
Palestinian Nat'l Auth.	3	432	252	3	392	157	3	392	231
Qatar	—	—	—	—	—	—	4	394	227
Romania	3	919	453	4	899	231	4	901	438
Russian Federation	2	915	446	3	898	225	3	897	431
Saudi Arabia	—	—	—	—	—	—	3	387	226
Scotland	2	410	224	3	381	141	3	381	210
Serbia	2	919	444	3	837	226	3	894	435
Singapore	2	420	233	3	398	145	3	398	214
Slovenia	2	766	372	3	786	223	3	786	395
Sweden	2	916	398	3	901	218	3	901	396
Syrian Arab Republic	—	—	—	—	—	—	3	901	464
Thailand	—	—	—	—	—	—	3	399	224
Tunisia	2	410	242	3	399	159	3	399	234
Turkey	—	—	—	—	—	—	3	396	227
Ukraine	—	—	—	—	—	—	4	901	439
United States	8	404	229	8	389	160	8	389	222
Benchmarking Participants									
Basque Country, Spain	3	429	230	4	377	122	4	377	202
British Columbia, Canada	—	—	—	—	—	—	4	388	215
Dubai, UAE	—	—	—	—	—	—	3	397	217
Massachusetts, US	—	—	—	—	—	—	3	389	209
Minnesota, US	—	—	—	—	—	—	3	389	204
Ontario, Canada	3	410	219	4	388	128	4	388	209
Quebec, Canada	3	410	223	4	388	136	4	388	212

11.3.6 Generating IRT Proficiency Scores for the TIMSS 2007 Data

Educational Testing Service's MGROUP program (Sheehan, 1985)[13] was used to generate the IRT proficiency scores. This program takes as input the students' responses to the items they were given, the item parameters estimated at the calibration stage, and the conditioning variables, and generates as output the plausible values that represent student proficiency. A useful feature of MGROUP is its ability to perform multi-dimensional scaling using the responses to all items across the scales and the correlations among the scales to improve the reliability of each individual scale. Because the redesigned TIMSS 2007 assessment booklets were balanced in terms of their mathematics and science content, TIMSS was able to capitalize on this feature for the first time in 2007. In this way, the overall mathematics and science scales were established simultaneously using a two-dimensional MGROUP run. This feature of MGROUP also was used to generate multi-dimensional scales across the mathematics content domains, the mathematics cognitive domains, the science content domains, and the science cognitive domains.

In addition to generating plausible values for the TIMSS 2007 assessment data, the parameters estimated at the calibration stage also were used to generate plausible values on the overall mathematics and science scales using the fourth-grade 2003 assessment data and 2007 bridging data for the 21 trend countries that also participated in the TIMSS 2003 fourth-grade assessment, and the eighth-grade 2003 assessment data and 2007 bridging data for the 33 countries that also participated in the 2003 eighth-grade assessment. These additional plausible values were then used to establish the two successive linear transformations necessary to place the TIMSS 2007 assessment on the TIMSS trend scale.

In all, a total of 209 (86 at the fourth grade and 123 at the eighth grade) two-dimensional MGROUP runs were required for the overall mathematics and science scales, and 404 (176 at the fourth grade and 228 at the eighth grade) multidimensional MGROUP runs for the content and cognitive scales. Exhibit 11.13 shows the sizes of the student samples—2003 assessment data, 2007 bridging data, and 2007 assessment data—for which proficiency scores using the 2007 item parameters were generated on the overall mathematics and science scales. At the fourth grade, scores on the 2003 assessment data were generated for 103,865 students, scores on the 2007 bridging data were generated for 28,098 students, and scores on the

13 The MGROUP program was provided by ETS under contract to the TIMSS & PIRLS International Study Center at Boston College. It is now commercially available as DESI.

2007 assessment data for 187,673 students. At the eighth grade, scores on the 2003 assessment data were generated for 169,619 students, scores on the 2007 bridging data for 44,350 students, and scores on the 2007 assessment data for 246,112 students. Exhibit 11.6, presented previously, shows that a total of 187,673 students received proficiency scores on the 2007 assessment data in the content and cognitive domains at the fourth grade and 246,112 students at the eighth grade.

Exhibit 11.13 Sample Sizes for TIMSS 2007 Proficiency Estimation of Overall Mathematics and Science

Country	Fourth Grade			Eighth Grade		
	TIMSS 2003 Assessment Booklets	TIMSS 2007 Bridge Booklets	TIMSS 2007 Assessment Booklets	TIMSS 2003 Assessment Booklets	TIMSS 2007 Bridge Booklets	TIMSS 2007 Assessment Booklets
Algeria	—	—	4,223	—	—	5,447
Armenia	5,674	1,139	4,079	5,726	1,307	4,689
Australia	4,321	1,186	4,108	4,791	1,164	4,069
Austria	—	—	4,859	—	—	—
Bahrain	—	—	—	4,199	1,210	4,230
Bosnia and Herzegovina	—	—	—	—	—	4,220
Botswana	—	—	—	5,150	1,197	4,208
Bulgaria	—	—	—	4,117	1,141	4,019
Chinese Taipei	4,661	1,192	4,131	5,379	1,155	4,046
Colombia	—	—	4,801	—	—	4,873
Cyprus	—	—	—	4,002	1,255	4,399
Czech Republic	—	—	4,235	—	—	4,845
Denmark	—	—	3,519	—	—	—
Egypt	—	—	—	7,095	1,871	6,582
El Salvador	—	—	4,166	—	—	4,063
England	3,585	1,208	4,316	2,830	1,159	4,025
Georgia	—	—	4,108	—	—	4,178
Germany	—	—	5,200	—	—	—
Ghana	—	—	—	5,100	1,498	5,294
Hong Kong SAR	4,608	1,072	3,791	4,972	986	3,470
Hungary	3,319	1,155	4,048	3,302	1,183	4,111
Indonesia	—	—	—	5,762	1,202	4,203
Iran, Islamic Rep. of	4,352	1,087	3,833	4,942	1,115	3,981
Israel	—	—	—	4,318	926	3,294
Italy	4,282	1,277	4,470	4,278	1,242	4,408
Japan	4,535	1,274	4,487	4,856	1,221	4,312
Jordan	—	—	—	4,489	1,492	5,251
Kazakhstan	—	—	3,990	—	—	—
Korea, Rep. of	—	—	—	5,309	1,208	4,240
Kuwait	—	—	3,803	—	—	4,091
Latvia	3,687	1,101	3,908	—	—	—
Lebanon	—	—	—	3,814	1,073	3,786
Lithuania	4,422	1,134	3,980	4,964	1,141	3,991
Malaysia	—	—	—	5,314	1,285	4,466
Malta	—	—	—	—	—	4,670
Mongolia	—	—	4,523	—	—	4,499
Morocco	4,264	1,090	3,894	—	—	3,060
Netherlands	2,937	962	3,349	—	—	—
New Zealand	4,308	1,405	4,940	—	—	—
Norway	4,342	1,165	4,108	4,133	1,317	4,627
Oman	—	—	—	—	—	4,752
Palestinian Nat'l Auth.	—	—	—	5,357	1,253	4,378
Qatar	—	—	7,019	—	—	7,184
Romania	—	—	—	4,104	1,201	4,198
Russian Federation	3,963	1,277	4,464	4,667	1,277	4,472
Saudi Arabia	—	—	—	—	—	4,243
Scotland	3,936	1,123	3,929	3,516	1,156	4,070
Serbia	—	—	—	4,296	1,153	4,045
Singapore	6,668	1,440	5,041	6,018	1,329	4,599
Slovak Republic	—	—	4,963	—	—	—
Slovenia	3,126	1,244	4,351	3,578	1,150	4,043
Sweden	—	—	4,676	4,256	1,473	5,215
Syrian Arab Republic	—	—	—	—	—	4,650
Thailand	—	—	—	—	—	5,412
Tunisia	4,334	1,174	4,134	4,931	1,175	4,080
Turkey	—	—	—	—	—	4,498
Ukraine	—	—	4,292	—	—	4,424
United States	9,829	2,261	7,896	8,912	2,097	7,377
Yemen	—	—	5,811	—	—	—
Benchmarking Participants						
Alberta, Canada	—	—	4,037	—	—	—
Basque Country, Spain	—	—	—	2,514	645	2,296
British Columbia, Canada	—	—	4,153	—	—	4,256
Dubai, UAE	—	—	3,064	—	—	3,195
Massachusetts, US	—	—	1,747	—	—	1,897
Minnesota, US	—	—	1,846	—	—	1,777
Ontario, Canada	4,362	1,021	3,496	4,217	989	3,448
Quebec, Canada	4,350	1,111	3,885	4,411	1,104	3,956
Total	**103,865**	**28,098**	**187,673**	**169,619**	**44,350**	**246,112**

11.3.7 Transforming the Mathematics and Science Scores to Measure Trends

To provide results for TIMSS 2007 that would be comparable to results from previous TIMSS assessments, the 2007 proficiency scores (plausible values) for overall mathematics and science had to be transformed to the metric used in 1995, 1999, and 2003. This was accomplished through two successive linear transformations as part of the concurrent calibration approach.

First, the means and standard deviations of the mathematics and science 2003 scores produced in 2007—the plausible values from the TIMSS 2003 assessment data based on the 2007 concurrent item calibrations—were made to match the means and standard deviations of the scores reported in the TIMSS 2003 assessment—the plausible values produced in 2003 using the 2003 item calibrations—by applying the appropriate linear transformations. These linear transformations were given by:

$$PV_{k,i}^{*} = A_{k,i} + B_{k,i} \cdot PV_{k,i} \quad (13)$$

where

$PV_{k,i}$ was the plausible value i of scale k prior to transformation;

$PV_{k,i}^{*}$ was the plausible value i of scale k after transformation;

and $A_{k,i}$ and $B_{k,i}$ were the linear transformation constants.

The linear transformation constants were obtained by first computing the international means and standard deviations of the proficiency scores for the overall mathematics and science scales using the plausible values produced in 2003 based on the 2003 item calibrations for the trend countries. Next, the same calculations were done using the plausible values from the TIMSS 2003 assessment data based on the 2007 item calibrations for the same set of countries. The linear transformation constants were defined as:

$$\begin{aligned} B_{k,i} &= \sigma_{k,i} / \sigma_{k,i}^{*} \\ A_{k,i} &= \mu_{k,i} - B_{k,i}\ \mu_{k,i}^{*} \end{aligned} \quad (14)$$

where

$\mu_{k,i}$ was the international mean of scale k based on plausible value i released in 2003;

$\mu_{k,i}^{*}$ was the international mean of scale *k* based on plausible value *i* from the TIMSS 2003 assessment data based on the 2007 concurrent item calibrations;

$\sigma_{k,i}$ was the international standard deviation of scale *k* based on plausible value *i* released in 2003;

$\sigma_{k,i}^{*}$ was the international standard deviation of scale *k* based on plausible value *i* from the TIMSS 2003 assessment data based on the 2007 concurrent item calibrations.

Exhibit 11.14 shows the linear transformation constants that were computed in this first step. Once the linear transformation constants had been established, all of the mathematics and science plausible values generated on the TIMSS 2007 bridging data were transformed by applying the linear transformations.

Exhibit 11.14 Linear Transformation Constants Applied to the TIMSS 2007 Bridge Scores

Scale		Plausible Value	TIMSS 2003 Scores Using 2003 Item Calibrations		TIMSS 2003 Scores Using 2007 Item Calibrations		$A_{k,i}$	$B_{k,i}$
			Mean	Standard Deviation	Mean	Standard Deviation		
Fourth Grade	Mathematics	PV1	498.12622	104.81269	–0.06579	0.99477	505.05840	105.36413
		PV2	498.31619	103.90056	–0.06546	0.99426	505.15723	104.50041
		PV3	498.14926	104.01856	–0.06582	0.99533	505.02747	104.50692
		PV4	498.51640	104.36297	–0.06712	0.99476	505.55795	104.91235
		PV5	498.33038	103.88447	–0.06510	0.99498	505.12714	104.40824
	Science	PV1	495.05010	109.62454	–0.05554	0.98941	501.20328	110.79794
		PV2	494.22197	109.40731	–0.05360	0.98730	500.16177	110.81421
		PV3	494.23251	110.17620	–0.05360	0.98717	500.21478	111.60831
		PV4	494.34316	109.52188	–0.05348	0.98990	500.26064	110.63879
		PV5	495.13090	109.68009	–0.05185	0.98629	500.89740	111.20455
Eighth Grade	Mathematics	PV1	476.14829	105.92163	0.00510	0.98871	475.60194	107.13090
		PV2	476.39770	107.36384	0.00539	0.99167	475.81398	108.26543
		PV3	476.33494	107.48064	0.00480	0.99012	475.81336	108.55323
		PV4	475.96981	107.31753	0.00481	0.98907	475.44768	108.50350
		PV5	476.42089	107.00376	0.00551	0.99005	475.82554	108.07918
	Science	PV1	481.84829	105.24281	0.00707	0.98023	481.08890	107.36518
		PV2	481.99746	105.50264	0.00785	0.98128	481.15317	107.51570
		PV3	482.40244	104.91097	0.00804	0.97856	481.54006	107.20928
		PV4	482.08413	105.81120	0.00856	0.97901	481.15912	108.08008
		PV5	482.51302	104.94370	0.00939	0.97924	481.50676	107.16884

Next, the means and standard deviations of the mathematics and science proficiency scores on the TIMSS 2007 assessment data were made to match the means and standard deviations of the now-transformed scores on the TIMSS 2007 bridging data by applying appropriate linear transformations. These linear transformations were derived using the same equations given above, with the linear transformation constants obtained by first computing the international means and standard deviations of the now-transformed scores on the TIMSS 2007 bridging data for the overall mathematics and science scales across the trend countries, and then the same calculations using the plausible values generated on the TIMSS 2007 assessment data across the trend countries.

Exhibit 11.15 shows the linear transformation constants that were computed in this second step. Once these linear transformation constants had been established, all of the 2007 mathematics and science proficiency scores—the plausible values generated on the TIMSS 2007 assessment data—for all participating countries and benchmarking participants were transformed by applying the linear transformations. This provided mathematics and science student achievement scores for the TIMSS 2007 assessment that were directly comparable to the scores from the 1995, 1999 (only at the eighth grade), and 2003 assessments.

Exhibit 11.15 Linear Transformation Constants Applied to the TIMSS 2007 Proficiency Scores

Scale		Plausible Value	Transformed TIMSS 2007 Bridge Scores		TIMSS 2007 Proficiency Scores		$A_{k,i}$	$B_{k,i}$
			Mean	Standard Deviation	Mean	Standard Deviation		
Fourth Grade	Mathematics	PV1	506.17533	108.02573	–0.01243	1.04972	507.45462	102.90944
		PV2	506.23088	107.63611	–0.01115	1.04540	507.37904	102.96198
		PV3	506.62376	107.29968	–0.01037	1.04678	507.68705	102.50484
		PV4	506.15659	108.10783	–0.01021	1.04853	507.20928	103.10455
		PV5	506.19823	107.37574	–0.01337	1.04727	507.56872	102.52942
	Science	PV1	504.92173	112.88966	0.01118	1.01466	503.67776	111.25838
		PV2	503.55827	112.77187	0.01470	1.00669	501.91179	112.02242
		PV3	503.42470	113.64933	0.01197	1.00968	502.07753	112.55966
		PV4	503.36473	112.95516	0.01129	1.01015	502.10236	111.82060
		PV5	504.79464	112.70603	0.01263	1.01355	503.38990	111.19905
Eighth Grade	Mathematics	PV1	474.29429	109.44201	–0.01422	1.01544	475.82719	107.77822
		PV2	474.61572	110.62798	–0.01264	1.01579	475.99222	108.90822
		PV3	474.52757	111.06244	–0.01359	1.01350	476.01716	109.58307
		PV4	474.22239	110.91719	–0.01266	1.01656	475.60358	109.11081
		PV5	475.17257	110.29007	–0.01343	1.01490	476.63216	108.67084
	Science	PV1	481.92084	105.72417	0.00330	0.97876	481.56437	108.01886
		PV2	482.06417	105.48861	0.00376	0.97833	481.65864	107.82554
		PV3	482.56974	104.81989	0.00504	0.97830	482.03002	107.14473
		PV4	481.56147	106.10752	0.00105	0.98092	481.44803	108.17180
		PV5	482.65436	105.06218	0.00228	0.97759	482.40927	107.47102

11.3.8 Setting the Metric for the Mathematics and Science Content and Cognitive Domain Scales

As described earlier, the IRT scales for the mathematics and science content and cognitive domains had no provision for measuring trends, so there was no need to establish links to previous assessment metrics. Instead, the plausible values for each content and cognitive domain scale were transformed to the same metric as its respective overall subject scale in 2007. For example, in eighth-grade mathematics, the mean and standard deviation for the number, algebra, geometry, and data and chance scales were set to have the same mean and standard deviation as the 2007 eighth-grade mathematics scale. Setting linear transformation constants was done in the same manner as described in the previous section, with the exception that the means and standard deviations of the overall subject scales were averaged across the five plausible values. Exhibits 11.16 through 11.19 show the transformations that were applied to all the content and cognitive domains. Taking fourth-grade mathematics as an example, the plausible values of all fourth-grade

mathematics content and cognitive domains were transformed to have a mean of 472.9372 and a standard deviation of 123.6880, the international mean and standard deviation for overall mathematics across the 36 fourth-grade countries.

Exhibit 11.16 Linear Transformation Constants for the TIMSS 2007 Fourth-Grade Mathematics Content and Cognitive Domains

Scale		Plausible Values	Mean	Standard Deviation		
Mathematics		PV1	472.7558	123.9167		
		PV2	472.8534	123.9992		
		PV3	473.3264	123.3602		
		PV4	472.7947	123.8875		
		PV5	472.9556	123.2766		
		Overall	**472.9372**	**123.6880**	$A_{k,i}$	$B_{k,i}$
Content Domains	Number	PV1	–0.1044	1.1129	484.5396	111.1409
		PV2	–0.1036	1.1094	484.4879	111.4913
		PV3	–0.1052	1.1138	484.6169	111.0519
		PV4	–0.1049	1.1126	484.6034	111.1682
		PV5	–0.1059	1.1145	484.6843	110.9775
	Geometric Shapes and Mesures	PV1	–0.1654	1.1350	490.9571	108.9716
		PV2	–0.1661	1.1340	491.0578	109.0680
		PV3	–0.1654	1.1366	490.9363	108.8250
		PV4	–0.1635	1.1351	490.7560	108.9620
		PV5	–0.1663	1.1366	491.0385	108.8265
	Data Display	PV1	–0.2348	1.2274	496.5946	100.7747
		PV2	–0.2298	1.2283	496.0757	100.7013
		PV3	–0.2376	1.2257	496.9106	100.9138
		PV4	–0.2318	1.2256	496.3298	100.9167
		PV5	–0.2263	1.2204	495.8715	101.3540
Cognitive Domains	Knowing	PV1	–0.1233	1.0819	487.0345	114.3231
		PV2	–0.1207	1.0824	486.7352	114.2708
		PV3	–0.1198	1.0801	486.6588	114.5102
		PV4	–0.1212	1.0777	486.8516	114.7689
		PV5	–0.1217	1.0788	486.8850	114.6544
	Applying	PV1	–0.1649	1.1292	491.0036	109.5366
		PV2	–0.1661	1.1313	491.0923	109.3315
		PV3	–0.1648	1.1281	491.0032	109.6410
		PV4	–0.1635	1.1301	490.8277	109.4483
		PV5	–0.1656	1.1278	491.0955	109.6730
	Reasoning	PV1	–0.1643	1.1908	490.0010	103.8737
		PV2	–0.1640	1.1930	489.9417	103.6768
		PV3	–0.1653	1.1929	490.0784	103.6869
		PV4	–0.1627	1.1931	489.8072	103.6689
		PV5	–0.1591	1.1895	489.4796	103.9819

Exhibit 11.17 Linear Transformation Constants for the TIMSS 2007 Fourth-Grade Science Content and Cognitive Domains

Scale		Plausible Values	Mean	Standard Deviation		
Science		PV1	476.8554	127.8734		
		PV2	475.2254	128.4317		
		PV3	475.0733	128.9199		
		PV4	475.1666	128.1879		
		PV5	476.6620	128.0548		
		Overall	**475.7965**	**128.2935**	$A_{k,i}$	$B_{k,i}$
Content Domains	Life Science	PV1	–0.0991	1.0267	488.1824	124.9529
		PV2	–0.0989	1.0222	488.2120	125.5117
		PV3	–0.1012	1.0243	488.4686	125.2512
		PV4	–0.0995	1.0261	488.2427	125.0362
		PV5	–0.1015	1.0258	488.4969	125.0695
	Physical Science	PV1	–0.1244	1.0591	490.8606	121.1338
		PV2	–0.1270	1.0588	491.1865	121.1670
		PV3	–0.1236	1.0580	490.7812	121.2581
		PV4	–0.1268	1.0616	491.1226	120.8508
		PV5	–0.1250	1.0602	490.9197	121.0130
	Earth Science	PV1	–0.1738	1.1588	495.0349	110.7096
		PV2	–0.1759	1.1559	495.3152	110.9871
		PV3	–0.1729	1.1604	494.9164	110.5598
		PV4	–0.1759	1.1589	495.2658	110.7030
		PV5	–0.1727	1.1595	494.9014	110.6414
Cognitive Domains	Knowing	PV1	–0.0979	1.0077	488.2655	127.3159
		PV2	–0.1015	1.0130	488.6458	126.6496
		PV3	–0.1000	1.0098	488.4998	127.0429
		PV4	–0.0996	1.0124	488.4181	126.7196
		PV5	–0.1000	1.0101	488.4992	127.0106
	Applying	PV1	–0.1053	1.0206	489.0330	125.7006
		PV2	–0.1064	1.0213	489.1652	125.6169
		PV3	–0.1074	1.0243	489.2459	125.2444
		PV4	–0.1070	1.0193	489.2690	125.8602
		PV5	–0.1051	1.0216	488.9907	125.5752
	Reasoning	PV1	–0.1044	1.1160	487.7931	114.9562
		PV2	–0.1061	1.1128	488.0338	115.2891
		PV3	–0.1028	1.1156	487.6219	114.9956
		PV4	–0.1075	1.1136	488.1796	115.2033
		PV5	–0.1054	1.1165	487.9106	114.9097

Exhibit 11.18 Linear Transformation Constants for the TIMSS 2007 Eighth-Grade Mathematics Content and Cognitive Domains

Scale		Plausible Values	Mean	Standard Deviation		
Mathematics		PV1	450.7160	111.5804		
		PV2	450.8086	112.6485		
		PV3	450.5763	113.2416		
		PV4	450.2712	113.1116		
		PV5	451.3883	112.4931		
		Overall	**450.7521**	**112.6151**	$A_{k,i}$	$B_{k,i}$
Content Domains	Number	PV1	–0.0300	1.0349	454.0165	108.8143
		PV2	–0.0335	1.0350	454.4005	108.8096
		PV3	–0.0323	1.0346	454.2695	108.8481
		PV4	–0.0309	1.0338	454.1148	108.9336
		PV5	–0.0344	1.0346	454.4955	108.8470
	Algebra	PV1	–0.0044	1.0900	451.2025	103.3148
		PV2	–0.0044	1.0906	451.2070	103.2605
		PV3	–0.0038	1.0905	451.1481	103.2675
		PV4	–0.0056	1.0910	451.3284	103.2175
		PV5	–0.0098	1.0935	451.7566	102.9858
	Geometry	PV1	–0.0828	1.0827	459.3668	104.0144
		PV2	–0.0802	1.0803	459.1119	104.2434
		PV3	–0.0824	1.0820	459.3264	104.0787
		PV4	–0.0814	1.0795	459.2466	104.3209
		PV5	–0.0798	1.0808	459.0657	104.1960
	Data and Chance	PV1	–0.0674	1.0645	457.8778	105.7897
		PV2	–0.0717	1.0606	458.3616	106.1821
		PV3	–0.0697	1.0633	458.1391	105.9131
		PV4	–0.0716	1.0597	458.3578	106.2665
		PV5	–0.0706	1.0603	458.2455	106.2111
Cognitive Domains	Knowing	PV1	–0.0671	1.0395	458.0263	108.3317
		PV2	–0.0717	1.0415	458.5033	108.1275
		PV3	–0.0670	1.0393	458.0155	108.3590
		PV4	–0.0656	1.0409	457.8484	108.1890
		PV5	–0.0672	1.0402	458.0238	108.2634
	Applying	PV1	–0.0495	1.0458	456.0790	107.6794
		PV2	–0.0516	1.0464	456.3011	107.6224
		PV3	–0.0517	1.0472	456.3152	107.5379
		PV4	–0.0508	1.0483	456.2053	107.4299
		PV5	–0.0519	1.0449	456.3465	107.7801
	Reasoning	PV1	–0.0441	1.0749	455.3733	104.7632
		PV2	–0.0414	1.0752	455.0850	104.7371
		PV3	–0.0474	1.0745	455.7219	104.8029
		PV4	–0.0463	1.0752	455.6066	104.7384
		PV5	–0.0469	1.0723	455.6766	105.0259

Exhibit 11.19 Linear Transformation Constants for the TIMSS 2007 Eighth-Grade Science Content and Cognitive Domains

Scale		Plausible Values	Mean	Standard Deviation		
Science		PV1	465.4845	106.0061		
		PV2	465.6370	105.7173		
		PV3	466.0839	105.1221		
		PV4	465.1039	106.3709		
		PV5	466.2519	105.4609		
		Overall	**465.7122**	**105.7354**	$A_{k,i}$	$B_{k,i}$
Content Domains	Biology	PV1	–0.0398	0.8496	470.6701	124.4517
		PV2	–0.0401	0.8504	470.7007	124.3389
		PV3	–0.0415	0.8476	470.8869	124.7403
		PV4	–0.0413	0.8480	470.8605	124.6949
		PV5	–0.0422	0.8514	470.9481	124.1906
	Chemistry	PV1	–0.0654	1.0273	472.4460	102.9242
		PV2	–0.0656	1.0270	472.4647	102.9571
		PV3	–0.0652	1.0240	472.4397	103.2609
		PV4	–0.0649	1.0288	472.3856	102.7784
		PV5	–0.0650	1.0311	472.3790	102.5499
	Physics	PV1	–0.0827	0.9906	474.5414	106.7348
		PV2	–0.0842	0.9905	474.7044	106.7466
		PV3	–0.0805	0.9882	474.3278	107.0004
		PV4	–0.0774	0.9886	473.9906	106.9579
		PV5	–0.0821	0.9865	474.5151	107.1846
	Earth Science	PV1	–0.0951	1.0407	475.3735	101.6041
		PV2	–0.0920	1.0419	475.0517	101.4861
		PV3	–0.0922	1.0393	475.0911	101.7377
		PV4	–0.0962	1.0372	475.5150	101.9418
		PV5	–0.0939	1.0436	475.2263	101.3174
Cognitive Domains	Knowing	PV1	–0.0454	0.8542	471.3322	123.7832
		PV2	–0.0443	0.8545	471.1986	123.7342
		PV3	–0.0428	0.8545	471.0059	123.7376
		PV4	–0.0435	0.8535	471.1056	123.8873
		PV5	–0.0448	0.8553	471.2555	123.6291
	Applying	PV1	–0.0596	0.8704	472.9576	121.4767
		PV2	–0.0606	0.8681	473.0985	121.7949
		PV3	–0.0600	0.8684	473.0205	121.7655
		PV4	–0.0594	0.8702	472.9248	121.5055
		PV5	–0.0585	0.8696	472.8282	121.5880
	Reasoning	PV1	–0.0815	1.0554	473.8798	100.1850
		PV2	–0.0838	1.0618	474.0618	99.5821
		PV3	–0.0822	1.0580	473.9259	99.9417
		PV4	–0.0794	1.0586	473.6424	99.8829
		PV5	–0.0801	1.0576	473.7238	99.9776

11.4 Capturing the Uncertainty in the TIMSS Student Achievement Scores

To obtain estimates of students' proficiency in mathematics and science that were both accurate and cost-effective, TIMSS 2007 made extensive use of probability sampling techniques to sample students from national eighth- and fourth-grade student populations, and applied matrix sampling methods to target individual students with a subset of the entire set of assessment materials. Statistics computed from these student samples were used to estimate population parameters. This approach made an efficient use of resources, in particular keeping student response burden to a minimum, but at a cost of some variance or uncertainty in the statistics. To quantify this uncertainty, each statistic in the TIMSS 2007 international reports (Martin et al., 2008; Mullis et al., 2008) is accompanied by an estimate of its standard error. These standard errors incorporate components reflecting the uncertainty due to generalizing from student samples to the entire eighth- or fourth-grade student populations (sampling variance), and to inferring students' performance on the entire assessment from their performance on the subset of items that they took (imputation variance).

11.4.1 Estimating Sampling Variance

The TIMSS 2007 sampling design applied a stratified multistage cluster-sampling technique to the problem of selecting efficient and accurate samples of students while working with schools and classes. This design capitalized on the structure of the student population (i.e., students grouped in classrooms within schools) to derive student samples that permitted efficient and economical data collection. Unfortunately, however, such a complex sampling design complicates the task of computing standard errors to quantify sampling variability.

When, as in TIMSS, the sampling design involves multistage cluster sampling, there are several options for estimating sampling errors that avoid the assumption of simple random sampling (Wolter, 1985). The jackknife repeated replication technique (JRR) was chosen by TIMSS because it is computationally straightforward and provides approximately unbiased estimates of the sampling errors of means, totals, and percentages.

The variation on the JRR technique used in TIMSS 2007 is described in Johnson and Rust (1992). It assumes that the primary sampling units (PSUs) can be paired in a manner consistent with the sampling design, with

each pair regarded as members of a pseudo-stratum for variance estimation purposes. When used in this way, the JRR technique appropriately accounts for the combined effect of the between- and within-PSU contributions to the sampling variance. The general use of JRR entails systematically assigning pairs of schools to sampling zones, and randomly selecting one of these schools to have its contribution doubled and the other to have its contribution zeroed, so as to construct a number of "pseudo-replicates" of the original sample. The statistic of interest is computed once for the entire original sample, and once again for each jackknife pseudo-replicate sample. The variation between the estimates for each of the jackknife replicate samples and the original sample estimate is the jackknife estimate of the sampling error of the statistic.

11.4.2 Constructing Sampling Zones for Sampling Variance Estimation

To apply the JRR technique used in TIMSS 2007, the sampled schools were paired and assigned to a series of groups known as sampling zones. This was done at Statistics Canada by working through the list of sampled schools in the order in which they were selected and assigning the first and second participating schools to the first sampling zone, the third and fourth participating schools to the second zone, and so on. In total, 75 zones were used, allowing for 150 schools per country. When more than 75 zones were constructed, they were collapsed to keep the total number to 75.

Sampling zones were constructed within design domains, or explicit strata. When there was an odd number of schools in an explicit stratum, either by design or because of school non-response, the students in the remaining school were randomly divided to make up two "quasi" schools for the purposes of calculating the jackknife standard error.[14] Each sampling zone then consisted of a pair of schools or "quasi" schools. Exhibit 11.20 shows the number of sampling zones in each country.

Within each sampling zone, both schools were assigned an indicator (u_j), coded randomly to 0 or 1, such that one school had a value of 0, and the other a value of 1. This indicator determined whether the weights for the sampled students in the school in this zone were to be doubled ($u_j = 1$) or zeroed ($u_j = 0$) for the purposes of creating the pseudo-replicate samples.

14 If the remaining school consisted of 2 sampled classrooms, each classroom became a "quasi" school.

Exhibit 11.20 Number of Sampling Zones Used in All TIMSS 2007 Countries

Country	TIMSS 2007 Sampling Zones	
	Fourth Grade	Eighth Grade
Algeria	75	75
Armenia	74	74
Australia	75	75
Austria	75	—
Bahrain	—	75
Bosnia and Herzegovina	—	75
Botswana	—	75
Bulgaria	—	75
Chinese Taipei	75	75
Colombia	72	75
Cyprus	—	75
Czech Republic	72	74
Denmark	69	—
Egypt	—	75
El Salvador	75	73
England	72	69
Georgia	75	71
Germany	75	—
Ghana	—	75
Hong Kong SAR	64	61
Hungary	73	72
Indonesia	—	75
Iran, Islamic Rep. of	75	75
Israel	—	74
Italy	75	75
Japan	75	74
Jordan	—	75
Kazakhstan	71	—
Korea, Rep. of	—	75
Kuwait	75	75
Latvia	74	—
Lebanon	—	68
Lithuania	75	72
Malaysia	—	75
Malta	—	75
Mongolia	75	75
Morocco	75	68
Netherlands	71	—
New Zealand	75	—
Norway	73	70
Oman	—	75
Palestinian Nat'l Auth.	—	75
Qatar	75	75
Romania	—	75
Russian Federation	61	63
Saudi Arabia	—	75
Scotland	70	65
Serbia	—	74
Singapore	75	75
Slovak Republic	75	—
Slovenia	74	74
Sweden	75	75
Syrian Arab Republic	—	75
Thailand	—	75
Tunisia	75	75
Turkey	—	74
Ukraine	73	74
United States	75	75
Yemen	73	—
Benchmark Participants		
Alberta, Canada	73	—
Basque Country, Spain	—	65
British Columbia, Canada	75	75
Dubai, UAE	75	75
Massachusetts, US	24	24
Minnesota, US	25	25
Ontario, Canada	75	75
Quebec, Canada	75	75

11.4.3 Computing Sampling Variance Using the JRR Method

To compute a statistic *t* from the sample of a country, the formula for the sampling variance estimate of the statistic *t*, based on the JRR algorithm used in TIMSS 2007, is given by the following equation:

$$Var_{jrr}(t) = \sum_{h=1}^{H} \left[t(J_h) - t(S) \right]^2$$

where H is the total number of sampling zones in the sample of the country under consideration. The term $t(S)$ corresponds to the statistic of interest for the whole sample computed with the overall sampling weights (as described in Chapter 9). The term $t(J_h)$ denotes the same statistic using the h^{th} jackknife replicate sample J_h and its set of replicate sampling weights, which are identical to the overall sampling weights, except for the students in the h^{th} sampling zone. For the students in the h^{th} zone, all students belonging to one of the randomly selected schools of the pair were removed, and the students belonging to the other school in the zone were included twice. In practice, this was accomplished by recoding to zero the weights for the students in the school to be excluded from the replication, and multiplying by two the weights of the remaining students within the h^{th} pair. Each sampled student was assigned a vector of 75 replicate sampling weights W_{hi}, where h took values from 1 to 75. If W_{0i} was the overall sampling weight of student *i*, the h replicate weights for that student were computed as

$$W_{hi} = W_{0i} \cdot k_{hi}$$

where

$$k_{hi} = \begin{cases} 2 \cdot u_j & \text{if student } i \text{ is in school } j \text{ of sampling zone } h \\ 1 & \text{otherwise} \end{cases}$$

The school-level indicators u_j determined which students in a sampling zone would get zero weights and which ones would get double weights, on the basis of the school within the pair from which the students were sampled. The process of setting the k_{hi} values for all sampled students and across

all sampling zones is illustrated in Exhibit 11.21. Thus, the computation of the JRR variance estimate for any statistic in TIMSS 2007 required the computation of the statistic up to 76 times for any given country: once to obtain the statistic for the full sample based on the overall weights W_{0i}, and up to 75 times to obtain the statistics for each of the jackknife replicate samples J_h using a set of replicate weights W_{hi}.

Exhibit 11.21 Construction of Replicate Weights Across Sampling Zones in TIMSS 2007

Sampling Zone	School Replicate Indicator (u_i)	Replicate Factors for Computing JRR Replicate Sampling Weights (k_{hi})						
		1	2	3	⋯	h	⋯	75
1	0	0	1	1	⋯	1	⋯	1
	1	2						
2	0	1	0	1	⋯	1	⋯	1
	1		2					
3	0	1	1	0	⋯	1	⋯	1
	1			2				
⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋯	⋮
h	0	1	1	1	⋯	0	⋯	1
	1					2		
⋮	⋮	⋮	⋮	⋮	⋯	⋮	⋱	⋮
75	0	1	1	1	⋯	1	⋯	0
	1							2

In the TIMSS 2007 analyses, 75 replicate weights were computed for each country regardless of the number of actual zones within the country. If a country had fewer than 75 zones, then the additional replicate weights where h was greater than the number of zones within the country were all made equal to the overall sampling weight. Although this involved some redundant computations, having 75 replicate weights for each country had no effect on the magnitude of the error variance computed using the jackknife formula and it simplified the computation of standard errors for numerous countries at a time. All standard errors presented in the TIMSS 2007 international reports were computed using SAS programs developed at the TIMSS & PIRLS International Study Center.

11.4.4 Estimating Imputation Variance

The TIMSS 2007 item pool was far too extensive to be administered in its entirety to any one student, and so a matrix-sampling test design was developed whereby each student was given a single test booklet containing only a part of the entire assessment.[15] The results for all of the booklets were then aggregated using item response theory to provide results for the entire assessment. Since each student responded to just a subset of the assessment items, multiple imputation (the generation of plausible values) was used to derive reliable estimates of student performance on the assessment as a whole. Since every student proficiency estimate incorporates some uncertainty arising from the use of IRT models, TIMSS followed the customary procedure of generating five estimates for each student and using the variability among them as a measure of this imputation uncertainty, or error. In the TIMSS 2007 international reports, the imputation error for each variable has been combined with the sampling error for that variable to provide a standard error that incorporates both.

The general procedure for estimating the imputation variance using plausible values is described in Mislevy, Beaton, Kaplan, and Sheenan (1992). First, compute the statistic t for each set of M plausible values. The statistics t_m, where $m = 1, 2, \ldots, 5$, can be anything estimable from the data, such as a mean, the difference between means, percentiles, and so forth.

Once the statistics t_m are computed, the imputation variance is then calculated as:

$$Var_{imp} = \left(1 + \frac{1}{M}\right) Var\left(t_1, \cdots, t_m\right)$$

where M is the number of plausible values used in the calculation, and $Var\left(t_1, ..., t_M\right)$ is the usual variance of the M estimates computed using each plausible value.

11.4.5 Combining Sampling and Imputation Variance

The standard errors of all proficiency statistics reported by TIMSS include both sampling and imputation variance components. These standard errors were computed using the following formula:

15 The TIMSS 2007 assessment design is described in Chapter 2.

$$Var\left(t_{pv}\right) = Var_{jrr}\left(t_1\right) + Var_{imp}$$

where $Var_{jrr}(t_1)$ is the sampling variance computed for the first plausible value[16] and Var_{imp} is the imputation variance. The *TIMSS 2007 User Guide for the International Database* (Foy & Olson, 2009) contains programs in SAS and SPSS that compute each of these variance components for the TIMSS 2007 data. Furthermore, the IDB Analyzer—software provided with the international database—automatically computes standard errors as described in this section.

Exhibits 11.22 through 11.25 show basic summary statistics for overall mathematics and science achievement in the TIMSS 2007 assessment for the fourth and eighth grades. Each exhibit presents the student sample size, the mean and standard deviation averaged across the five plausible values, the jackknife sampling error for the mean, and the overall standard error for the mean, which includes the imputation error. Appendix E contains tables showing the same summary statistics for the mathematics and science content and cognitive domains at the fourth and eighth grades.

16 Under ideal circumstances and with unlimited computing resources, the JRR sampling variance would be computed for each of the plausible values and the imputation variance as described here. This would require computing the same statistic up to 380 times (once overall for each of the five plausible values using the overall sampling weights, and then 75 times more for each plausible value using the complete set of replicate weights). An acceptable shortcut, however, is to compute the JRR sampling variance component using only one plausible value (the first one), and then the imputation variance using the five plausible values. Using this approach, a statistic needs to be computed only 80 times.

Exhibit 11.22 Summary Statistics and Standard Errors for Proficiency in Mathematics at the Fourth Grade

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Algeria	4,223	377.645	89.560	5.058	5.176
Armenia	4,079	499.513	89.523	4.245	4.286
Australia	4,108	516.062	83.306	3.468	3.509
Austria	4,859	505.389	67.937	1.905	2.005
Chinese Taipei	4,131	575.819	69.225	1.633	1.733
Colombia	4,801	355.450	90.178	4.794	4.974
Czech Republic	4,235	486.399	71.458	2.665	2.781
Denmark	3,519	523.106	70.835	2.335	2.403
El Salvador	4,166	329.906	90.819	3.463	4.104
England	4,316	541.465	86.044	2.856	2.882
Georgia	4,108	438.458	88.430	4.180	4.207
Germany	5,200	525.155	68.149	2.224	2.254
Hong Kong SAR	3,791	606.802	67.126	3.429	3.584
Hungary	4,048	509.720	91.160	3.505	3.547
Iran, Islamic Rep. of	3,833	402.422	83.522	3.617	4.054
Italy	4,470	506.750	77.025	3.132	3.135
Japan	4,487	568.157	76.075	2.093	2.121
Kazakhstan	3,990	549.348	83.807	7.117	7.146
Kuwait	3,803	315.535	99.299	3.412	3.646
Latvia	3,908	537.200	71.904	2.089	2.306
Lithuania	3,980	529.799	75.761	2.288	2.372
Morocco	3,894	341.305	95.265	4.509	4.668
Netherlands	3,349	534.952	61.346	2.130	2.145
New Zealand	4,940	492.475	86.135	2.216	2.313
Norway	4,108	473.216	76.222	2.430	2.543
Qatar	7,019	296.268	90.067	0.974	1.043
Russian Federation	4,464	544.045	83.370	4.909	4.911
Scotland	3,929	494.449	78.926	2.182	2.214
Singapore	5,041	599.406	84.146	3.716	3.744
Slovak Republic	4,963	495.975	84.937	4.428	4.468
Slovenia	4,351	501.843	71.399	1.628	1.811
Sweden	4,676	502.574	66.482	2.385	2.527
Tunisia	4,134	327.435	110.809	4.406	4.469
Ukraine	4,292	469.003	84.479	2.893	2.912
United States	7,896	529.009	75.329	2.395	2.448
Yemen	5,811	223.683	110.136	5.637	5.968
Benchmarking Participants					
Alberta, Canada	4,037	505.320	66.059	2.938	2.952
British Columbia, Canada	4,153	505.219	71.314	2.543	2.749
Dubai, UAE	3,064	444.334	89.598	1.896	2.141
Massachusetts, US	1,747	572.484	69.772	3.468	3.513
Minnesota, US	1,846	554.117	77.714	5.823	5.863
Ontario, Canada	3,496	511.614	68.001	3.008	3.100
Quebec, Canada	3,885	519.103	67.347	2.999	3.028

Exhibit 11.23 Summary Statistics and Standard Errors for Proficiency in Science at the Fourth Grade

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Algeria	4,223	353.819	101.883	5.810	6.024
Armenia	4,079	484.387	118.784	5.529	5.684
Australia	4,108	527.397	80.497	3.149	3.341
Austria	4,859	525.627	77.410	2.182	2.520
Chinese Taipei	4,131	556.696	77.353	1.911	2.002
Colombia	4,801	400.305	97.459	5.397	5.446
Czech Republic	4,235	515.052	75.607	2.895	3.124
Denmark	3,519	516.917	76.937	2.709	2.854
El Salvador	4,166	389.583	93.202	3.191	3.368
England	4,316	541.527	80.219	2.790	2.852
Georgia	4,108	417.637	84.662	4.094	4.556
Germany	5,200	527.554	79.119	2.283	2.403
Hong Kong SAR	3,791	554.181	67.885	3.460	3.498
Hungary	4,048	536.226	84.807	3.113	3.346
Iran, Islamic Rep. of	3,833	435.639	97.424	4.071	4.275
Italy	4,470	535.217	81.368	3.090	3.172
Japan	4,487	547.780	69.631	1.672	2.066
Kazakhstan	3,990	532.830	74.326	5.481	5.631
Kuwait	3,803	348.151	123.080	4.096	4.367
Latvia	3,908	541.895	66.857	2.142	2.288
Lithuania	3,980	514.205	65.196	1.807	2.366
Morocco	3,894	297.447	123.744	5.580	5.864
Netherlands	3,349	523.176	59.870	2.209	2.610
New Zealand	4,940	504.066	90.091	2.369	2.626
Norway	4,108	476.551	76.659	2.488	3.484
Qatar	7,019	294.396	129.491	1.240	2.559
Russian Federation	4,464	546.231	80.524	4.636	4.781
Scotland	3,929	500.409	76.241	2.002	2.275
Singapore	5,041	586.654	93.044	3.905	4.091
Slovak Republic	4,963	525.691	87.247	4.634	4.765
Slovenia	4,351	518.393	76.172	1.887	1.936
Sweden	4,676	524.810	73.575	2.763	2.876
Tunisia	4,134	318.474	141.383	5.524	5.907
Ukraine	4,292	473.814	82.912	2.605	3.085
United States	7,896	538.574	83.990	2.579	2.714
Yemen	5,811	197.365	130.062	6.650	7.188
Benchmarking Participants					
Alberta, Canada	4,037	542.588	73.632	3.655	3.828
British Columbia, Canada	4,153	536.690	72.661	2.476	2.691
Dubai, UAE	3,064	459.648	107.310	2.601	2.752
Massachusetts, US	1,747	570.894	69.230	3.845	4.253
Minnesota, US	1,846	551.478	79.542	6.056	6.089
Ontario, Canada	3,496	535.869	78.245	3.289	3.722
Quebec, Canada	3,885	517.122	66.651	2.415	2.664

Exhibit 11.24 Summary Statistics and Standard Errors for Proficiency in Mathematics at the Eighth Grade

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Algeria	5,447	386.752	59.250	1.493	2.142
Armenia	4,689	498.680	84.735	3.438	3.505
Australia	4,069	496.232	79.426	3.874	3.934
Bahrain	4,230	398.071	83.601	1.320	1.567
Bosnia and Herzegovina	4,220	455.863	77.801	2.678	2.697
Botswana	4,208	363.539	76.579	1.981	2.268
Bulgaria	4,019	463.630	101.605	4.857	4.965
Chinese Taipei	4,046	598.301	105.505	4.337	4.533
Colombia	4,873	379.636	78.935	3.600	3.632
Cyprus	4,399	465.477	89.319	1.569	1.648
Czech Republic	4,845	503.807	73.686	2.313	2.392
Egypt	6,582	390.557	100.247	3.409	3.571
El Salvador	4,063	340.441	72.822	2.664	2.756
England	4,025	513.404	83.579	4.790	4.816
Georgia	4,178	409.617	96.464	5.889	5.950
Ghana	5,294	309.370	91.597	4.150	4.364
Hong Kong SAR	3,470	572.487	93.734	5.675	5.793
Hungary	4,111	516.895	84.678	3.417	3.474
Indonesia	4,203	397.110	87.341	3.692	3.808
Iran, Islamic Rep. of	3,981	403.380	86.095	3.968	4.116
Israel	3,294	463.251	98.873	3.866	3.949
Italy	4,408	479.626	76.231	2.925	3.037
Japan	4,312	569.810	85.416	2.063	2.407
Jordan	5,251	426.893	102.208	4.037	4.117
Korea, Rep. of	4,240	597.266	92.069	2.471	2.707
Kuwait	4,091	353.670	78.636	2.196	2.316
Lebanon	3,786	449.061	74.637	3.827	3.984
Lithuania	3,991	505.818	79.744	2.218	2.324
Malaysia	4,466	473.886	79.248	5.005	5.029
Malta	4,670	487.752	91.772	0.868	1.210
Morocco	3,060	380.784	80.326	2.753	2.970
Norway	4,627	469.216	65.665	1.918	1.976
Oman	4,752	372.434	94.944	2.848	3.370
Palestinian Nat'l Auth.	4,378	367.155	102.436	3.399	3.549
Qatar	7,184	306.791	93.360	0.727	1.374
Romania	4,198	461.318	99.748	4.038	4.099
Russian Federation	4,472	511.734	83.079	4.045	4.101
Saudi Arabia	4,243	329.337	76.433	2.174	2.852
Scotland	4,070	487.406	79.727	3.606	3.705
Serbia	4,045	485.767	89.451	3.077	3.316
Singapore	4,599	592.785	92.958	3.732	3.814
Slovenia	4,043	501.476	71.618	1.996	2.110
Sweden	5,215	491.300	70.052	2.093	2.260
Syrian Arab Republic	4,650	394.838	82.402	3.407	3.765
Thailand	5,412	441.390	91.617	4.897	4.951
Tunisia	4,080	420.413	66.519	2.343	2.433
Turkey	4,498	431.810	108.742	4.680	4.753
Ukraine	4,424	462.162	89.231	3.600	3.621
United States	7,377	508.454	76.736	2.773	2.830
Benchmarking Participants					
Basque Country, Spain	2,296	498.559	68.590	2.723	2.990
British Columbia, Canada	4,256	509.449	72.443	3.016	3.032
Dubai, UAE	3,195	460.616	96.176	2.257	2.370
Massachusetts, US	1,897	547.130	79.234	4.510	4.559
Minnesota, US	1,777	532.450	67.764	4.299	4.411
Ontario, Canada	3,448	517.232	70.214	3.485	3.518
Quebec, Canada	3,956	528.110	68.410	3.221	3.512

Exhibit 11.25 Summary Statistics and Standard Errors for Proficiency in Science at the Eighth Grade

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Algeria	5,447	408.060	62.603	1.488	1.738
Armenia	4,689	487.960	101.142	5.511	5.755
Australia	4,069	514.788	80.324	3.610	3.648
Bahrain	4,230	467.448	86.027	1.411	1.718
Bosnia and Herzegovina	4,220	465.745	79.444	2.772	2.815
Botswana	4,208	354.534	99.425	2.537	3.054
Bulgaria	3,079	470.284	102.622	5.676	5.892
Chinese Taipei	4,046	561.003	89.274	3.603	3.686
Colombia	4,873	417.182	76.652	3.466	3.515
Cyprus	4,399	451.624	85.319	1.655	2.044
Czech Republic	4,845	538.878	71.394	1.892	1.919
Egypt	6,582	408.242	99.381	3.356	3.563
El Salvador	4,063	387.274	69.770	2.745	2.926
England	4,025	541.505	85.398	4.458	4.479
Georgia	4,178	420.902	83.326	4.603	4.768
Ghana	5,294	303.272	108.360	5.006	5.356
Hong Kong SAR	3,470	530.209	80.969	4.847	4.919
Hungary	4,111	539.034	76.583	2.840	2.919
Indonesia	4,203	426.990	74.181	3.168	3.366
Iran, Islamic Rep. of	3,981	458.929	81.340	3.484	3.594
Israel	3,294	467.922	100.906	4.304	4.338
Italy	4,408	495.147	77.517	2.773	2.818
Japan	4,312	553.815	77.108	1.852	1.897
Jordan	5,251	481.721	97.720	3.945	3.962
Korea, Rep. of	4,240	553.139	75.862	1.939	2.034
Kuwait	4,091	417.956	89.241	2.552	2.818
Lebanon	3,786	413.611	96.812	5.808	5.932
Lithuania	3,991	518.559	78.205	2.266	2.550
Malaysia	4,466	470.801	88.199	5.981	6.027
Malta	4,670	457.167	113.859	1.238	1.365
Morocco	3,060	401.831	78.550	2.597	2.898
Norway	4,627	486.758	73.272	2.059	2.187
Oman	4,752	422.502	95.744	2.911	2.964
Palestinian Nat'l Auth.	4,378	404.126	110.930	3.456	3.504
Qatar	7,184	318.854	125.866	0.927	1.734
Romania	4,198	461.900	87.893	3.672	3.850
Russian Federation	4,472	529.570	77.651	3.819	3.883
Saudi Arabia	4,243	403.245	77.978	2.213	2.448
Scotland	4,070	495.732	81.116	3.319	3.397
Serbia	4,045	470.307	84.720	3.007	3.151
Singapore	4,599	567.250	103.889	4.373	4.448
Slovenia	4,043	537.544	72.017	2.133	2.213
Sweden	5,215	510.690	78.033	2.477	2.557
Syrian Arab Republic	4,650	451.976	74.713	2.678	2.885
Thailand	5,412	470.614	82.735	4.268	4.297
Tunisia	4,080	444.898	60.475	1.921	2.117
Turkey	4,498	454.159	91.892	3.648	3.711
Ukraine	4,424	485.063	83.992	3.418	3.459
United States	7,377	519.989	82.274	2.832	2.857
Benchmarking Participants					
Basque Country, Spain	2,296	497.706	72.028	2.746	2.956
British Columbia, Canada	4,256	525.717	70.793	2.660	2.685
Dubai, UAE	3,195	488.865	94.001	2.601	2.762
Massachusetts, US	1,897	556.041	79.367	4.354	4.554
Minnesota, US	1,777	538.510	71.850	4.716	4.762
Ontario, Canada	3,448	526.128	69.455	3.574	3.648
Quebec, Canada	3,956	506.589	68.973	2.897	3.054

References

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

Beaton, A.E., & Johnson, E.G. (1990). The average response method of scaling. *Journal of Educational Statistics*, 15, 9–38.

Beaton, A.E., & Johnson, E.G. (1992). Overview of the scaling methodology used in the national assessment. *Journal of Educational Measurement*, 26(2), 163–175.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley Publishing.

Foy, P. & Olson, J.F. (2009). *TIMSS 2007 user guide for the international database*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Johnson, E.G., & Rust, K.F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics*, 17(2), 175–190.

Lord, F.M. (1980). *Applications of items response theory to practical testing problems*. Hillsdales, NJ: Lawrence Erlbaum Associates.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Martin, M.O., Mullis, I.V.S., & Chrostowski, S.J. (2004). *TIMSS 2003 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M.O., Mullis, I.V.S., & Foy, P. (with Olson, J.F., Erberber, E., Preuschoff, C., & Galia, J.). (2008). *TIMSS 2007 international science report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mislevy, R.J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, 80, 993–97.

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

Mislevy, R.J., Beaton, A., Kaplan, B.A., & Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29(2), 133–161.

Mislevy, R.J., Johnson, E.G. & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, 17(2), 131–154.

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

Mullis, I.V.S., Martin, M.O., & Foy, P. (with Olson, J.F., Preuschoff, C., Erberber, E., Arora, A., & Galia, J.). (2008). *TIMSS 2007 international mathematics report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I.V.S., Martin, M.O., Ruddock, G.J., O'Sullivan, C.Y., Arora, A., & Erberber, E. (2005). *TIMSS 2007 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I.V.S., Martin, M.O., Smith, T.A., Garden, R.A., Gregory, K.D., Gonzalez, E.J., Chrostowki, S.J., & O'Connor, K.M. (2003). *TIMSS assessment frameworks and specifications 2003* (2nd ed.). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

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

Muraki, E., & Bock, R.D. (1991). *PARSCALE: Parameter scaling of rating data* [Software Version 4.1]. Chicago, IL: Scientific Software, Inc.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Sheehan, K. M. (1985). *M-GROUP: Estimation of group effects in multivariate models* [Software Version 3.2]. 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, 309–22.

Van Der Linden, W.J. & Hambleton, R. (1996). *Handbook of modern item response theory*. New York. Springer-Verlag.

Wingersky, M., Kaplan, B.A., & Beaton, A.E. (1987). Joint estimation procedures. In A. E. Beaton (Ed.), *Implementing the new design: The NAEP 1983–84 technical report* (pp.285–92) (No. 15-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Wolter, K. M. (1985). *Introduction to variance estimation*. New York: Springer-Verlag.

Yamamoto, K., & Kulick, E. (2000). Scaling methodology and procedures for the TIMSS mathematics and science scales. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS 1999 technical report*. Chestnut Hill, MA: Boston College.