Chapter 11

Scaling Methods and Procedures for the TIMSS 2003 Mathematics and Science Scales

Eugenio J. Gonzalez, Joseph Galia, and Isaac Li

11.1 Overview

As described in Chapter 1, the TIMSS 2003 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 2003 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 multiple imputation or "plausible values" methodology used in scaling the TIMSS 2003 data, and then describes how this approach was applied to the TIMSS 2003 data and to the data from the previous TIMSS 1999 and TIMSS 1995 studies, in order to measure trends in achievement. The TIMSS

1 The TIMSS 2003 achievement test design is described in Chapter 2.

scaling was conducted at the TIMSS & PIRLS International Study Center at Boston College, using software from Educational Testing Service.[2]

11.2 TIMSS 2003 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 scaling models, depending on item type and scoring procedure, were used in the analysis of the TIMSS 2003 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 respondent'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 score points.

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 person whose proficiency on a scale k is characterized by the unobservable variable θ will respond correctly to item i:

$$P(x_i = 1 \mid \theta_k, a_i, b_i, c_i) = c_i + \frac{1 - c_i}{1 + \exp(-1.7 a_i (\theta_k - b_i))} \equiv P_{i1}(\theta_k) \qquad (1)$$

where

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

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

2 TIMSS is indebted to Matthias Von Davier, Ed Kulick, 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 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 imputed value 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 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).

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

b_i is its location parameter, characterizing its difficulty;

c_i is its lower asymptote parameter, reflecting the chances of respondents of very low proficiency selecting the correct answer.

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

$$P_{i0} \equiv P(x_i = 0 \mid \theta_k, a_i, b_i, c_i) = 1 - P_{i1}(\theta_k) \tag{2}$$

The two-parameter (2PL) model was used for the short constructed-response items that were scored as 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 The IRT Model for Polytomous Items

In TIMSS 2003, as in TIMSS 1995 and TIMSS 1999, 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 person 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:

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

where

m_i is the number of response categories for item *i*;

x_i is the response to item *i*, possibilities ranging between 0 and m_i-1;

θ_k is the proficiency of person on a scale *k*;

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

b_i is its location parameter, characterizing its difficulty;

$d_{i,l}$ is category *l* threshold parameter.

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

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 with a standard deviation of 100, as was done for TIMSS 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 person proficiency) and the specified parameters of the item, and are unaffected by the demographic characteristics or unique experiences of the respondents, 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:

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

where $P_{il}(\theta_k)$ is of the form appropriate to the type of item (dichotomous or polytomous), m_i is equal to 2 for the dichotomously scored items and is equal to 3 for the polytomous items, and u_{il} is an indicator variable defined by

$$u_{il} = \begin{cases} 1 & if\ response\ x_i\ is\ in\ category\ l \\ 0 & otherwise \end{cases}$$

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 2003 analyses, estimates of both dichotomous and polytomous item parameters were obtained using the commercially available Parscale software (Muracki & Bock, 1991; version 4.1). 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 respondent.

11.2.3 Proficiency Estimation Using Plausible Values

Most cognitive skills testing is concerned with accurately assessing the performance of individual respondents 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 individual θ's.

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 respondent while maintaining a wide range of content representation when responses are aggregated across all respondents. 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 by using all available data to estimate directly the characteristics of student populations and subpopulations, and then generating multiple imputed scores, called plausible values, from these distributions that can be used in analyses with standard statistical software. A detailed review of plausible values methodology is given in Mislevy (1991).

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 respondents. 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:

$$t^*(x,y) = E\left[t(\underline{\theta}, \underline{y}) \mid \underline{x}, \underline{y}\right]$$

$$= \int t(\underline{\theta}, \underline{y})\, p(\underline{\theta} \mid \underline{x}, \underline{y})\, d\underline{\theta}$$

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 student. These values are referred to as imputations in the sampling literature, and as plausible values in large-scale surveys such as TIMSS, NAEP, NALS, and IALLS. The value of θ for any respondent 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 by "multiple imputation". For example, the average of multiple estimates of t, each computed from a different set of plausible values, is a numerical approximation of t^* of the above equation; the variance among them reflects 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 jackknife variance estimation procedures, which are discussed in Chapter 12.

Note that plausible values are not test scores for individuals in the usual sense, but rather are imputed values 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.[5]

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

$$P(\theta_j|x_j,y_j,\Gamma,\Sigma) \propto P(x_j|\theta_j,y_j,\Gamma,\Sigma)P(\theta_j|y_j,\Gamma,\Sigma) = P(x_j|\theta_j)P(\theta_j|y_j,\Gamma,\Sigma) \quad (3)$$

where θ_j is a vector of scale values, $P(x_j,\theta_j)$ is the product over the scales of the independent likelihoods induced by responses to items within each scale, and $P(\theta_j|y_j,\Gamma,\Sigma)$ is the multivariate joint density of proficiencies for the scales, conditional on the observed value 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.

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

11.2.4 Conditioning

A multivariate normal distribution was assumed for $P(\theta_j | y_j, \Gamma, \Sigma)$, 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 to be used in Γ. Typically, components representing 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.

$$\theta = \Gamma' y^c + \varepsilon,$$

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 $P(\theta|y)$ be correctly specified for all background variables in the survey. Estimates of functions Γ involving background variables not conditioned on in this manner are subject to estimation error due to misspecification. The nature of these errors was discussed in detail in Mislevy (1991). In TIMSS 2003, 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 (3).

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 to $P(\Gamma, \Sigma | x_j, y_j)$ 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 (3), 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 respondent.

For respondents with an insufficient number of responses, the Γ and Σs described in the previous paragraph are fixed. Hence, all respondents - regardless of the number of items attempted - are assigned a set of plausible values.

The plausible values could then be employed to evaluate equation (1) for an arbitrary function T as follows:

1. Using the first vector of plausible values for each respondent, evaluate T as if the plausible values were the true values of θ. Denote the result T_1.
2. Evaluate the sampling variance of T, or Var(T_1), with respect to respondents' first vectors of plausible values.
3. 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, . . ., 5.
4. 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:
$$\hat{T} = \frac{\sum_u T_u}{5}$$
5. An estimate of the variance of $\hat{T}$ is the sum of two components: an estimate of Var(T_u) obtained by averaging as in step 4, and the variance among the T_us. Let $\overline{U} = \frac{\sum_u Var_u}{M}$, 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 final estimate of the variance of $\hat{T}$ is:
$$Var\left(\hat{T}\right) = \overline{U} + \left(1 + M^{-1}\right)B_M$$

The first component in $Var\left(\hat{T}\right)$ reflects uncertainty due to sampling respondents from the population; the second reflects uncertainty due to the fact that sampled respondents' θs are not known precisely, but only indirectly through x and y.

11.2.6 Working with Plausible Values

Plausible values methodology was used in TIMSS 2003 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 respondent reflects the uncertainty associated with proficiency estimates for individual respondents. However, retaining this component of uncertainty requires that additional analytical procedures be used to estimate respondents' proficiencies, as follows.

If θ values were observed for all sampled respondents, the statistic $(t-T)/U^{1/2}$ would follow a *t*-distribution with d degrees of freedom. Then the incomplete-data statistic $(T-\hat{T})/[Var(\hat{T})]^{1/2}$ is approximately *t*-distributed, with degrees of freedom (Johnson & Rust, 1993) given by

$$\nu = \frac{1}{\frac{f_M{}^2}{M-1} + \frac{(1-f_M)^2}{d}}$$

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 θ values:

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

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

For *k*-dimensional *t*, such as the *k* coefficients in a multiple regression analysis, each U and $\overline{U}$ 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 $(\underline{T}-\underline{\hat{T}})Var^{-1}(\underline{\hat{T}})(\underline{T}-\underline{\hat{T}})'$ is approximately *F*-distributed with degrees of freedom equal to k and ν, with ν defined as above but with a matrix generalization of f_M:

$$f_M = (1+M^{-1})Trace[B_M Var^{-1}(\hat{T})]/k .$$

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 $(\underline{T}-\underline{\hat{T}})Var^{-1}(\underline{\hat{T}})(\underline{T}-\underline{\hat{T}})'$.

Statistics $\hat{T}$, the estimates of ability 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 2003 analyses included nearly all background variables.

11.3 Implementing the Scaling Procedures for the TIMSS 2003 Assessment Data

The application of IRT scaling and plausible value methodology to the TIMSS 2003 assessment data involved four major tasks: calibrating the achievement test items (estimating model parameters for each item), creating principal components from the questionnaire data for use in conditioning; generating IRT scale scores (proficiency scores) for mathematics and science and for each of the mathematics and science content domains; and placing the proficiency scale scores on the metric used to report the results from previous assessments. The TIMSS eighth-grade reporting metric was established 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 1999 and 1995, the TIMSS 1999 eighth-grade data also were placed on this metric. Placing the 2003 eighth-grade results on this metric permitted trend results from three points in time: 1995, 1999, and 2003. Since TIMSS did not collect data at the fourth grade in 1999, the TIMSS 2003 fourth-grade data were placed directly on the 1995 fourth-grade scale, providing comparisons between results from 1995 and 2003. Scale metrics were aligned for trend reporting only for mathematics and science overall; there were insufficient trend items from 1995 and 1999 to measure trends in content areas reliably.

11.3.1 Calibrating the TIMSS 2003 Test Items

As described in Chapter 2, the TIMSS 2003 achievement test design consisted of a total of 14 mathematics blocks and 14 science blocks at each grade, distributed across 12 student 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. Each student booklet contained six blocks, which were chosen according to a matrix-sampling scheme that kept the number of booklets as few as possible while maximizing the number of times blocks were paired together in a booklet. Half of the booklets contained four mathematics blocks and two science blocks, and half four science blocks and two mathematics blocks. Each sampled student completed one of the twelve student booklets. During the testing sessions, each student responded to three blocks of items, took a short break, and then responded to the other three blocks. 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.

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 domains.

The first step in constructing these scales was to estimate the IRT model item parameters for each item on each of the scales. This item calibration was conducted using the commercially-available Parscale software (Muraki & Bock, 1991; version 4.1). Item calibration for the overall mathematics and science scales, which were used to measure trends from 1995 and 1999, included data from 1995 for fourth grade and from 1999 for eighth grade. The calibration was conducted using a self-weighting random sample of 1000 students from each country's TIMSS student sample from each assessment year. This ensured that the data from each country and each assessment year contributed equally to the item calibration, while keeping the amount of data to be analyzed to a reasonable size.

Several calibrations were conducted. At the eighth grade, to construct separate overall mathematics and science scales for reporting trends, as well as performance generally in 2003, item calibrations were conducted using data from the 29 countries that participated in both 1999 and 2003 assessments. These calibrations each included 29,000 student records from the 1999 assessment and 29,000 records from the 2003 assessment, for a total of 58,000 student records. The item parameters established in these calibrations were used subsequently for estimating student scores for all 49 countries and 4 benchmarking entities that participated in 2003.

At the fourth grade, item calibrations for the overall mathematics and science scales for reporting trends, as well as performance generally in 2003, were conducted using data from the 15 countries that participated in both 1995 and 2003 assessments. These calibrations each included 15,000 student records from the 1999 assessment and 15,000 records from the 2003 assessment, for a total of 30,000 student records. As for the eighth grade, the item parameters established in these calibrations were used subsequently for estimating student scores for all 26 countries and 3 benchmarking entities that participated in 2003.

Because there were insufficient items to construct reliable scales for measuring trends in each of the content domains, scales for these domains were constructed using 2003 data only. At the eighth grade, separate calibrations were conducted for each of the five mathematics and five science content domains. These calibrations were based on 46,000 student records, 1,000 from each of the 46 countries that participated in the 2003 assess-

ment.[6] Similarly at the fourth grade, separate calibrations were conducted for each of the five mathematics and three science content domains. These calibrations were based on 26,000 student records, 1,000 from each of the 26 countries that participated in the 2003 assessment at the fourth grade. Although, because of the matrix-sampling design, not all students responded to every item, there were at least 2,000 student responses to each item in all calibrations.

All items in the TIMSS 2003 assessment were included in the item calibrations. However, a non-trivial position effect was detected during routine quality control checks on the data. As described in Chapter 2, TIMSS has a complicated booklet design, with blocks of items appearing in different positions in different booklets. For example, the items in block M1 appear as the first block in Booklet 1, as the second block in Booklet 6, and as the third block in Booklet 12. This allows the booklets to be linked together efficiently, but also to monitor and counterbalance any position effect. The counterbalanced booklet design made it possible to detect an unexpectedly strong position effect in the data as the item statistics for each country were reviewed. More specifically, this position effect occurred because some students in all countries did not reach all the items in the third block position, which was the end of the first half of each booklet before the break. The same effect was evident for the sixth block position, which was the last block in the booklets. The IRT scaling addressed this problem by treating items in the third and sixth block positions as if they were unique, even though they also appeared in other positions. For example, the mathematics items in block M1 from Booklet 1 (the first position) and from Booklet 6 (second position) were considered to be the same items for scaling and reporting purposes, but those in Booklet 12 (the third position) were scaled as items that were different and unique. This technique is also known as “splitting” the items, or “freeing” the item parameters.

Exhibits D.1 through D.22 in Appendix D present the item parameters generated from the calibrations. Items where the parameters have been freed have an "F" in the second character position of the item label. As a by-product of the calibrations, interim scores in mathematics, science, and the content domains for use in constructing conditioning variables were produced.

11.3.2 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

6 Data from the four Benchmarking participants were not included in the item calibration.

not reached when (within part 1 or part 2 of the booklet) the item itself and the item immediately preceding were not answered, and there were no other items completed in the remainder of the booklet.

In TIMSS 2003, 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 that were considered not to have been reached by students, and that were located in positions 1, 2, 4, and 5 of the test booklet, were treated as if they had not been administered. Items that were considered not to have been reached by the students, and that were located in positions 3 and 6 of the test booklet were treated as incorrect. This approach was considered optimal for parameter estimation. However, not-reached items were always considered as incorrect responses when student proficiency scores were generated.

11.3.3 Evaluating Fit of IRT Models to the TIMSS 2003 Data

After the calibrations were completed, checks were performed to verify that the item parameters obtained from Parscale adequately reproduced the observed distribution of responses across the proficiency continuum. The fit of the IRT models to the TIMSS 2003 data was examined by comparing the theoretical 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 θs for each respondent that received the item.

Exhibit 11.1 shows a plot of the empirical and theoretical 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. Values from the theoretical curve based on the estimated item parameters are shown as crosses. Empirical results are represented by circles. The centers of the circles represent the empirical proportions correct. The plotted values are the sums of these individual posteriors at each point on the proficiency scale for those students that responded correctly to the item, plus a fraction of the omitted responses, divided by the sum of the posteriors of all that were administered the item. The size of the circles is proportional to the sum of the posteriors at each point on the proficiency scale for all of those who received the item; this is related to the number of respondents contributing to the estimation of that empirical proportion correct.

Exhibit 11.2 contains a plot of the empirical and theoretical item response functions for a polytomous item. As for the dichotomous item plot

Exhibit 11.1 **TIMSS 2003 Mathematics Assessment Example Response Function for a Dichotomous Item**

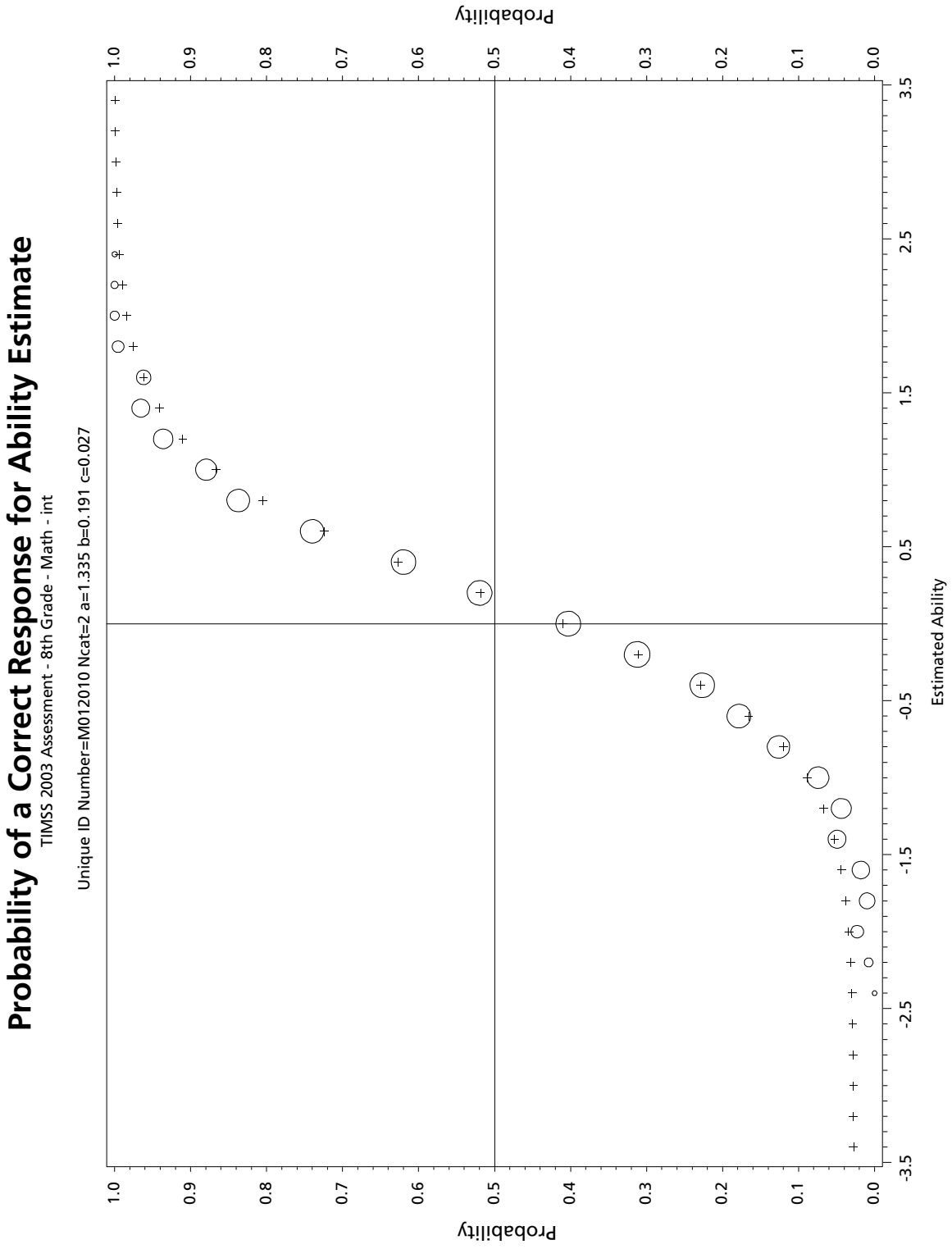
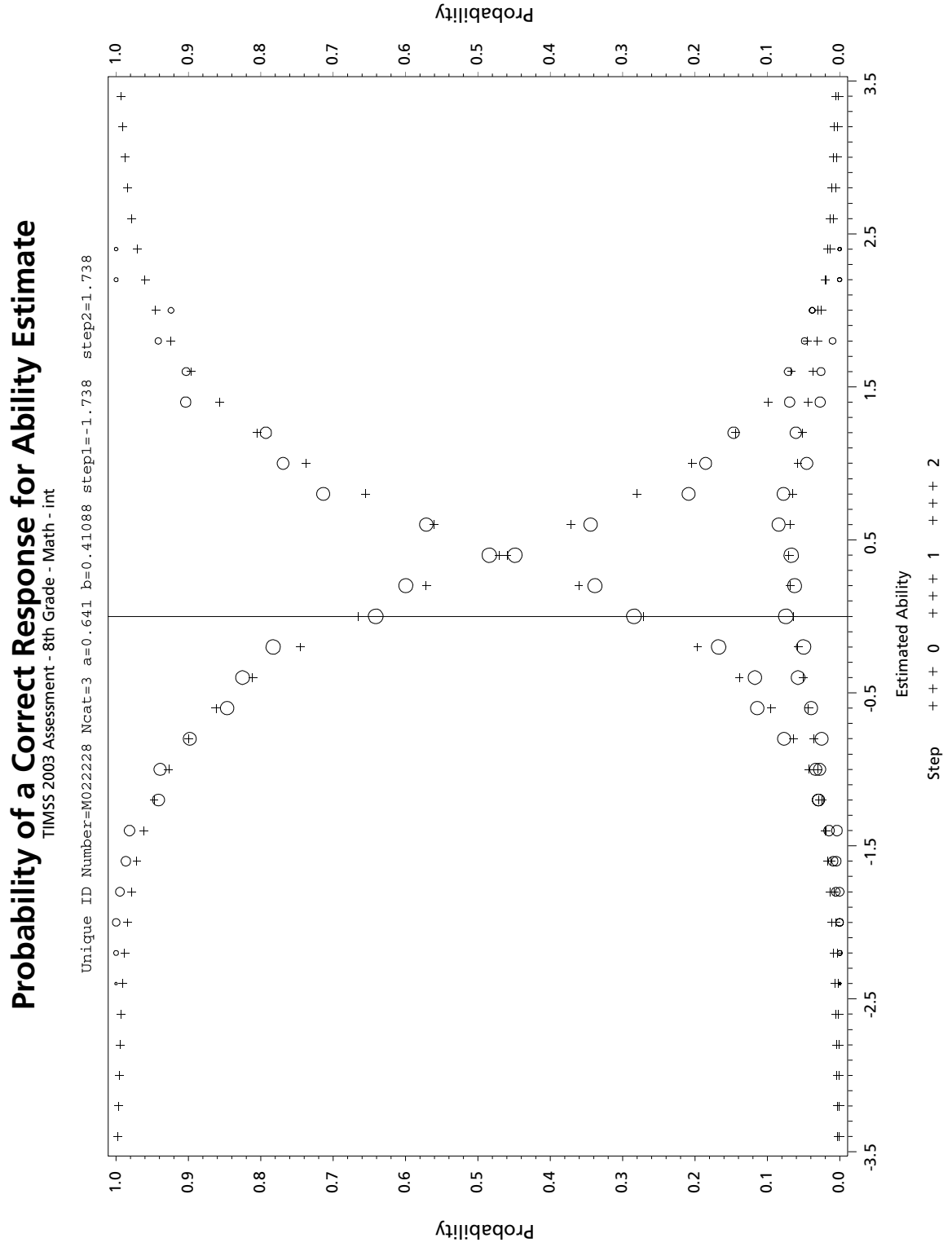


Exhibit 11.2 **TIMSS 2003 Mathematics Assessment Example Response Function for a Polytomous Item**



above, the horizontal axis represents the proficiency scale, but the vertical axis represents the probability of having a response fall in a given score category. For polytomous items, the sums for those who scored in the category of interest is divided by the sum for all those that were administered the item. The interpretation of the circles is the same as in Exhibit 11.2.

11.3.4 Variables for Conditioning the TIMSS 2003 Data

Because there were so many background variables that could be used in conditioning, TIMSS followed the practice established in 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 2003 background data were constructed as follows:

1. 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.
2. 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.[7] 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.
3. 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. Exhibit 11.3 and Exhibit 11.4 show the total number of variables that were used in the principal component analysis and the number of principal components selected to account for 90% of the background variance within each country.

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

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 conditioning variables.

Exhibit 11.3 **Number of Variables and Principal Components for Conditioning TIMSS 2003 Fourth Grade Data**

Country	Sample Size	Total number of conditioning variables	Total number of principal components only
ARM	5674	291	283
AUS	4321	301	216
BFL	4712	305	235
COT	4362	291	218
CQU	4350	291	217
CYP	4328	291	216
ENG	3585	295	179
HKG	4608	313	230
HUN	3319	307	165
IRN	4352	305	217
ITA	4282	311	214
JPN	4535	313	226
LTU	4422	290	221
LVA	3687	313	184
MAR	4263	297	213
MDA	3981	307	199
NLD	2937	289	146
NOR	4342	313	217
NZL	4308	311	215
PHL	4572	303	228
RUS	3963	305	198
SCO	3936	295	196
SGP	6668	301	333
SVN	3126	313	156
TUN	4334	311	216
TWN	4661	313	233
USA	9829	287	491
YEM	4205	313	210

Exhibit 11.4 **Number of Variables and Principal Components for Conditioning TIMSS 2003 Eighth Grade Data**

Country	Sample Size	Total Number of Conditioning Variables	Total Number of Principal Components Only
ARM	5726	893	286
AUS	4791	417	225
BFL	4970	762	248
BGR	4117	913	205
BHR	4199	432	209
BSQ	2514	431	125
BWA	5150	424	248
CHL	6377	416	240
COT	4217	410	210
CQU	4411	410	220
CYP	4002	897	200
EGY	7095	418	249
ENG	2830	410	141
EST	4040	903	202
GHA	5100	410	245
HKG	4972	432	233
HUN	3302	907	165
IDN	5762	897	288
IRN	4942	424	244
ISR	4318	432	215
ITA	4278	430	213
JOR	4489	432	224
JPN	4856	426	231
KOR	5309	432	234
LBN	3814	745	190
LTU	4964	811	248
LVA	3630	679	181
MAR	3160	408	158
MDA	4033	913	201
MKD	3893	919	194
MYS	5314	412	231
NLD	3065	735	153
NOR	4133	429	206
NZL	3801	430	190
PHL	6917	422	243
PSE	5357	432	251
ROM	4104	919	205
RUS	4667	912	233
SAU	4295	426	214
SCG	4296	919	214
SCO	3516	410	175

Exhibit 11.4 **Number of Variables and Principal Components for Conditioning TIMSS 2003 Eighth Grade Data** *(...Continued)*

Country	Sample Size	Total Number of Conditioning Variables	Total Number of Principal Components Only
SGP	6018	420	233
SVK	4215	912	210
SVN	3578	766	178
SWE	4256	916	212
SYR	4895	418	240
TUN	4931	410	242
TWN	5379	432	231
USA	8912	404	229
ZAF	8952	432	255

11.3.5 Generating IRT Proficiency Scores for the TIMSS 2003 Data

Educational Testing Service's MGROUP program (ETS, 1998; version 3.1)[8] 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. Four MGROUP runs were conducted at each grade level using the 2003 assessment data: one unidimensional run for the overall mathematics scale, one unidimensional run for the overall science scale, one multidimensional run for the mathematics content domain scales, and one multidimensional run for the science content domain scales.

In addition to generating plausible values for the TIMSS 2003 data, the parameters estimated at the calibration stage also were used to generate plausible values on the overall mathematics and science scales using the 1999 eighth-grade data for the 29 trend countries that participated in the TIMSS 1999 eighth-grade assessment and the 1995 fourth-grade data for the 15 countries that participated in the 1995 fourth-grade assessment. These plausible values for the trend countries were called "bridge scores."

Plausible values generated by the conditioning program are initially on the same scale as the item parameters used to estimate them. This scale metric is generally not useful for reporting purposes since it is somewhat arbitrary, ranges between approximately –3 and +3, and has a mean of zero across all countries.

8 The MGROUP program was provided by ETS under contract to the TIMSS and PIRLS International Study Center at Boston College.

11.3.6 Transforming the Mathematics and Science Scores to Measure Trends from 1995 and 1999

To provide results for TIMSS 2003 that would be comparable to results from previous TIMSS' assessments, the 2003 proficiency scores (plausible values) for overall mathematics and science had to be transformed to the metric used in 1995 and 1999. To accomplish this, the means and standard deviations of the mathematics and science "bridge scores" were made to match the means and standard deviations of the scores reported in the earlier assessments by applying the appropriate linear transformations. Once the linear transformation constants had been established, all of the mathematics and science scores from the 2003 assessment were transformed by applying the same linear transformations. This provided mathematics and science student achievement scores for the TIMS 2003 assessment that were directly comparable to the scores from the 1995 and 1999 assessments.

11.3.7 Setting the Metric for the Mathematics and Science Content Domain Scales

As described earlier, the IRT scales for the mathematics and science content domains had no provision for measuring trends, and so there was no need to establish links to previous assessment metrics. Instead, the plausible values for each content domain scale were transformed to the same metric as the overall subject scale in 2003. For example, in eighth-grade mathematics, the mean and standard deviation for the number, algebra, measurement, geometry, and data scales were set to have the same mean and standard deviation as the 2003 eighth-grade mathematics scale.

References

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.

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

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

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

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.

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*. Chicago, IL: Scientific Software, Inc.

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

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics*, 2, 309-22.

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.

Van Der Linden, W.J. & Hambleton, R. (1996). *Handbook of Modern Item Response Theory*. New York. Springer-Verlag.