Chapter 11

Scaling Methods and Procedures for

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 *h* is characterized by the unobservable variable θ will respond correctly to item *f*:

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

where

u_f is the response to item *f*, 1 if correct and 0 if incorrect;

θ_h is the proficiency of a person on a scale *h* (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).

λ_f is the slope parameter of item *f*, characterizing its discriminating power;

f

TI = TI = O ч G ' _ T Į g I

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

It is possible to approximate θ^* using random draws from the conditional distribution of the scale proficiencies given the student's item responses u_g, the student's background variables v_g, and model parameters for the student.

If θ values were observed for all sampled respondents, the statistic

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

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

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

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

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 C** ƙ ˙ ῶ ƺ א µ ʊ ɵe ƙ ſ

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	49		

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	6 ø 6		

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, !