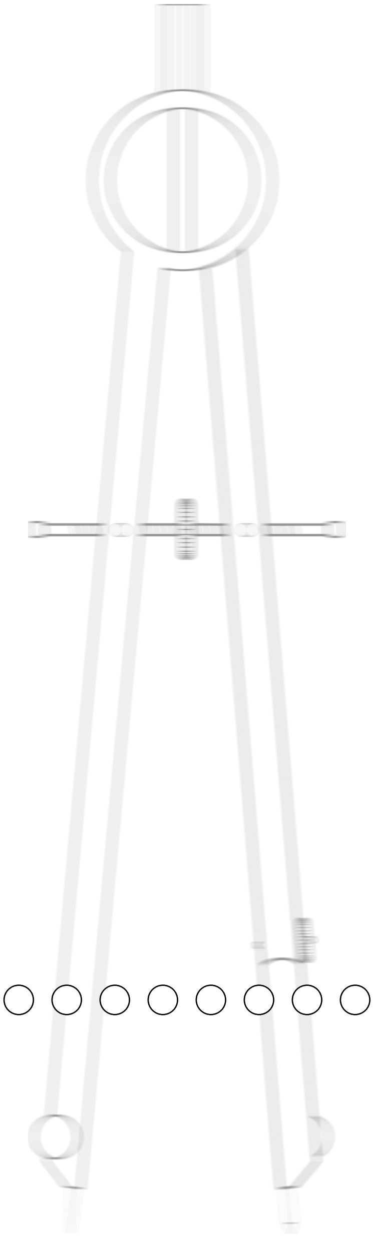
13

Scaling Methodology and Procedures for the TIMSS Mathematics and Science Scales

Kentaro Yamamoto
Edward Kulick



13 Scaling Methodology and Procedures for the TIMSS Mathematics and Science Scales

Kentaro Yamamoto
Edward Kulick

13.1 Overview

The TIMSS achievement test design made use of matrix sampling techniques to divide the assessment item pool so that each sampled student responded to just a portion of the items, thereby achieving wide coverage of the mathematics and science subject areas while keeping the response burden on individual students to a minimum.[1] TIMSS relied on a sophisticated form of psychometric scaling known as item response theory (IRT) scaling to combine the student responses in a way that provided accurate estimates of achievement. The TIMSS IRT scaling used multiple imputations or "plausible values" to obtain proficiency scores in mathematics and science and their content areas for all students, even though each student responded to only a part of the assessment item pool.

The TIMSS 1999 Benchmarking study used the same scaling and imputation methodology as the TIMSS 1999 International component. This chapter summarizes that methodology; further details can be found in the TIMSS 1999 Technical Report (see Yamamoto & Kulick, 2000).

13.2 TIMSS 1999 Benchmarking Scaling Methodology

Three distinct scaling models, depending on item type and scoring procedure, were used in analyzing the Benchmarking assessment data. Each is a *latent variable* model that describes the probability of a specific response 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 free-response items with just two response options, scored as correct or incorrect.

[1] The TIMSS 1999 achievement test design is described in chapter 2.

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 free-response items (i.e., those with more than two score points).

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

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

$$Pi1(x_i = 1|\theta_k, a_i, b_i, c_i) = c_i + \frac{(1 - c_i)}{1.0 + \exp(-1.7a_i(\theta_k - b_i))} \tag{1}$$

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;

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

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

The two-parameter logistic (2PL) model was used for the short free-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.

In scaling the Benchmarking data, the three- and two-parameter models were used in preference to the one-parameter Rasch model, primarily because they can more accurately account for the differences among items in their ability to discriminate between students of high and low ability. With the Rasch model, all items are assumed to have the same discriminating power, while the 2PL and 3PL models provide an extra item parameter to account for differences among items, and the 3PL model has a parameter that can be used to model guessing behavior among low-ability students.

Modeling item response functions as accurately as possible by using 2PL and 3PL models also reduces errors due to model mis-specification. The error is apparent when the model cannot exactly reproduce or predict the data using the estimated parameters. The difference between the observed data and those generated by the model is directly proportional to the degree of model mis-specification. Current psychometric convention does not allow model mis-specification errors to be represented in the proficiency scores. Instead, once item response parameters are estimated, they are treated as given and model mis-specification is ignored. For that reason it is generally preferable to use models that characterize the item response function as well as possible.

13.2.2 The IRT Model for Polytomous Items

Free-response items requiring an extended response were scored for partial credit, with zero, one, and two 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:

(3)

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

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 a person on scale k;

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

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

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

Indeterminacy of model parameters of the polytomous model are resolved by setting $d_{i,0}$=0 and setting

(4)

$$\sum_{l=1}^{m_i-1} d_{i,l} = 0$$

13.3 Item Parameter Estimation

For all of the IRT models, there is a linear indeterminacy between the values of item parameters and proficiency parameters; that is, 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 mean of 500 with standard deviation of 100.

IRT modeling relies on a number of assumptions, the most important being conditional independence. Under this assumption, item response probabilities depend only on θ_κ (a measure of 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. The joint probability of a particular response pattern x across a set of n items is then given by:

(5)

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

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 u_{il} is an indicator variable defined by

(6)

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

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 1999 Benchmarking analyses, estimates of both dichotomous and polytomous item parameters were obtained by the NAEP BILOG/ PARSCALE program, which combines Mislevy and Bock's (1982) BILOG and Muraki and Bock's (1991) PARSCALE computer programs. The item parameters in 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 θs for each respondent.

13.3.1 Evaluating Fit of IRT Models to the Data

The fit of the IRT models to the TIMSS 1999 data was examined within each scale by comparing the empirical item response functions with the theoretical item response function curves (see Exhibits 13.1 and 13.2). The theoretical curves are plots of the response functions generated by the model using values of the item parameters estimated from the data. The empirical results are calculated from the posterior distributions of the θs for each respondent who received the item. For dichotomous items the plotted values are the sums of these individual posteriors at each point on the proficiency scale for those students that responded correctly plus a fraction of the omitted responses, divided by the sum of the posteriors of all that were administered the item. 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.

Exhibit 13.1 **TIMSS 1999 Grade 8 Science Assessment Example Item Response Function—Dichotomous Item**

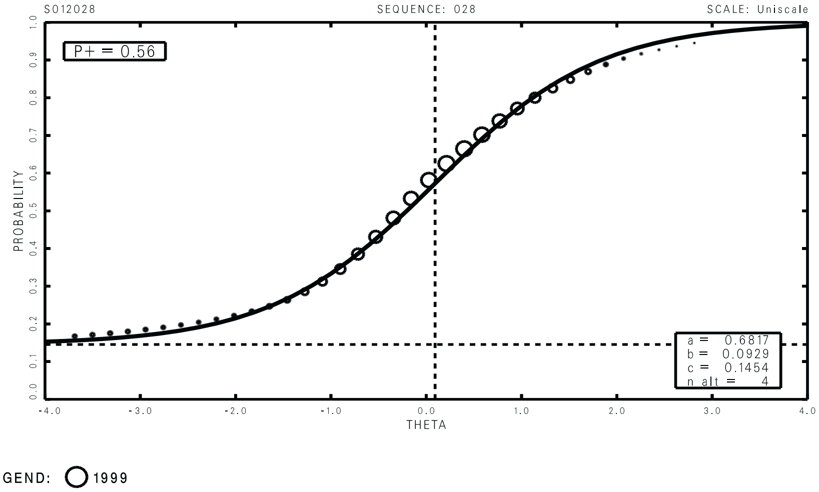


Exhibit 13.2 TIMSS 1999 Grade 8 Science Assessment Example Item Response Function—Polytomous Item

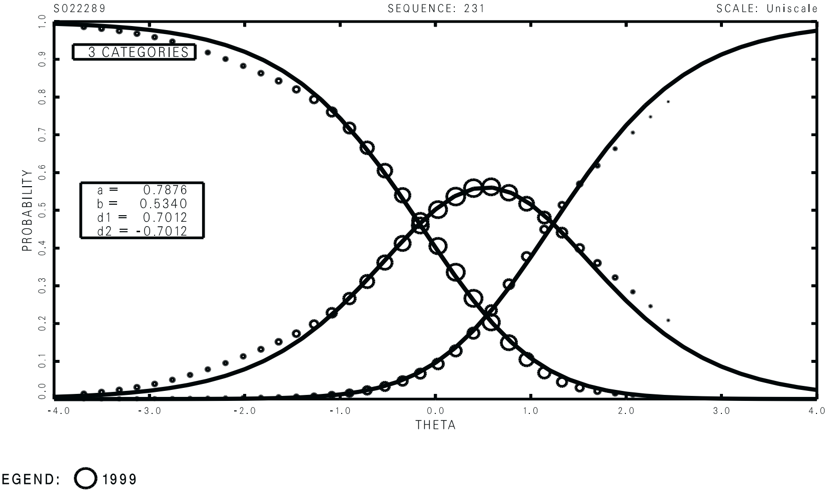


Exhibit 13.1 shows a plot of the empirical and theoretical item response functions for a dichotomous item. The horizontal axis represents the proficiency scale, and the vertical axis represents the probability of a correct response. The solid curve is the theoretical curve based on the estimated item parameters. The centers of the small circles represent the empirical proportions correct. 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 13.2 shows a plot of the empirical and theoretical item response functions for a polytomous item. Again, the horizontal axis represents the proficiency scale, but the vertical axis represents the probability of having a response fall in a given score category. The interpretation of the small circles is the same as in Exhibit 13.1. For items where the model fits the data well, the empirical and theoretical curves are close together.

13.4 Scaling Mathematics and Science Domains and Content Areas

In order to estimate student proficiency scores for the subject domains of mathematics and science, all items in each subject domain were calibrated together. This approach was chosen because it produced the best summary of student proficiency across the whole domain for each subject. Treating the entire mathematics or science item pool as a single domain maximizes the number of items per respondent, and the greatest amount of information possible is used to describe the proficiency distribution. This was found to be a more reliable way to compare proficiency across countries than to make a scale for each content area, such as algebra, geometry, etc., and then form a composite measure of mathematics by combining the content area scales.

A disadvantage of this approach is that differences in content scales may be underemphasized as they tend to regress toward the aggregated scale. Therefore, to enable comparisons of student proficiency on content scales, TIMSS provided separate scale scores of each content area in mathematics and science. If each content area is treated separately when estimating item parameters, differential profiles of content area proficiency can be examined, both across countries and across subpopulations within a country.

13.4.1 Omitted and Not-Reached Responses.

Apart from data that by design were not administered to a student, missing data could also occur when 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. In TIMSS 1999, 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 as not having been reached by students were treated as if they had not been administered. This approach was optimal for parameter estimation. However, since the time allotment for the TIMSS 1999 tests was generous, and enough for even marginally able respondents to complete the items, not reached items were considered to have incorrect responses when student proficiency scores were generated.

13.4.2 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, 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 θs.

For the distribution of proficiencies in large populations, more efficient estimates can be obtained from a matrix-sampling design like that used in TIMSS 1999. 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 imputed scores or 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).[2]

The following is a brief overview of the plausible-values approach. Let $\underline{y}$ represent the responses of all sampled students to background questions or background data of sampled students collected from other sources, and let $\underline{\theta}$ represent the proficiency of interest. If $\underline{\theta}$ were known for all sampled students, it would be possible to compute a statistic $t(\underline{\theta}, \underline{y})$ - 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, $\underline{\theta}$ values are not known even for sampled respondents. The solution to this problem is to follow Rubin (1987) by considering $\underline{\theta}$ as "missing data" and approximate $t(\underline{\theta}, \underline{y})$ by its expectation given $(\underline{x}, \underline{y})$, the data that actually were observed, as follows:

$$t^{*}(\underline{x}, \underline{y}) = E[t(\underline{\theta}, \underline{y}) | (\underline{x}, \underline{y})] = \int t(\underline{\theta}, \underline{y}) p(\underline{\theta} | (\underline{x}, \underline{y})) d\underline{\theta} \quad . \tag{7}$$

○○○

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

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 sampled student j. These values are referred to as imputations in the sampling literature, and as plausible values in large-scale surveys such as NAEP, NALS, and IALLS.[3] 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 uncertainty 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.

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 conditioning model is correctly specified, plausible values will provide consistent estimates of population proficiency, even though they are not generally unbiased estimates of the proficiencies of the individuals with whom they are associated.[4]

Plausible values for each respondent j are drawn from the conditional distribution $P(\underline{\theta}_j|(\underline{x}_j, \underline{y}_j, \Gamma, \Sigma))$, 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

$$P(\underline{\theta}_j|\underline{x}_j, \underline{y}_j, \Gamma, \Sigma) \propto P(\underline{x}_j|\underline{\theta}_j, \underline{y}_j, \Gamma, \Sigma)P(\underline{\theta}_j|\underline{y}_j, \Gamma, \Sigma) = P(\underline{x}_j|\underline{\theta}_j)P(\underline{\theta}_j|\underline{y}_j, \Gamma, \Sigma) \tag{8}$$

where $\underline{\theta}_j$ is a vector of scale values, $P(\underline{x}_j|\underline{\theta}_j)$ is the product over the scales of the independent likelihoods induced by responses to items within each scale, and $P(\underline{\theta}_j|(\underline{y}_j, \Gamma, \Sigma))$ is the multivariate joint

○○○

[3] U.S. National Assessment of Educational Progress (NAEP), U.S. National Adult Literacy Survey (NALS), the International Adult Literacy and Life Skills Survey (IALLS).

[4] For further discussion, see Mislevy, Beaton, Kaplan, & Sheehan (1992).

density of proficiencies of the scales, conditional on the observed value $\underline{y}_j$ 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.

13.4.3 Conditioning

A multivariate normal distribution was assumed for $P(\underline{\theta}_j, \underline{y}_j, \Gamma, \Sigma)$, 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% of the variance in the data are selected. These principal components are referred to as the conditioning variables and denoted as $\underline{y}^c$. The following model is then fit to the data:

(9) $$\theta = \Gamma'\underline{y}^c + \varepsilon$$

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

In order to be strictly correct for all functions Γ of $\underline{\theta}$, it is necessary that $p(\underline{\theta}|\underline{y})$ be correctly specified for all background variables in the survey. In Benchmarking, 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 $\underline{\theta}$ for these variables is nearly optimal. Estimates of functions Γ involving background variables not conditioned in this manner are subject to estimation error due to mis-specification. The nature of these errors is discussed in detail in Mislevy (1991).

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, $\underline{\theta}$, and variance Σ, of the posterior distribution in equation (7). For the multiple content area scales of TIMSS 1999, the computer program CGROUP (Thomas, 1993) was used. The program implemented a method to compute the moments using higher-order asymptotic corrections to a normal approximation. Case weights were employed in this step.

13.4.4 Generating Proficiency Scores

After completing the EM algorithm, the plausible values are drawn in a three-step process from the joint distribution of the values of Γ for all sampled. First, a value of Γ is drawn from a normal approximation to $P(\Gamma, \Sigma | (\underline{x}_j, \underline{y}_j))$ that fixes Σ at the value $\hat{\boldsymbol{\Sigma}}$ (Thomas, 1993). Second, conditional on the generated value of Γ (and the fixed value of $\Sigma = \hat{\boldsymbol{\Sigma}}$), the mean, $\underline{\theta}$, and variance, Σ_j^p, of the posterior distribution in equation (2) 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 $\underline{\theta}$ and variance Σ_j^p. These three steps are repeated five times, producing five imputations of $\underline{\theta}$ for each sampled respondent.

For respondents with an insufficient number of responses, the Γ and Σ described in the previous paragraph were fixed. Hence, all respondents — regardless of the number of items attempted — were assigned a set of plausible values for the various scales.

The plausible values could then be employed to evaluate equation (7) 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. As in step 1 above, evaluate the sampling variance of T, or $Var(T_1)$, with respect to respondents' first vectors of plausible values. Denote the result Var_1.
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, . . ., M, where M is the number of imputed values.
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:

(10)

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

5. An estimate of the variance of T is the sum of two components: an estimate of $Var(T_u)$ obtained as in step 4 and the variance among the T_us:

(11)
$$Var(T) = \frac{\sum_{u} Var_u}{M} + (1 + M^{-1})\frac{\sum_{u} (T_u - T)^2}{M - 1}$$

The first component in $Var(T)$ 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.

13.4.5 Working with Plausible Values

Plausible-values methodology was used in TIMSS 1999 to increase the accuracy of estimates of the proficiency distributions for various subpopulations and for the TIMSS 1999 population as a whole. This method correctly retains the uncertainty associated with proficiency estimates for individual respondents by using multiple imputed proficiency values rather than assuming that this type of uncertainty is zero — a more common practice. Yet, retaining this component of uncertainty requires that additional analytic procedures be used to estimate respondents' proficiencies, as follows.

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

(12)
$$v = \frac{1}{\frac{f_M^{\,2}}{M-1} + \frac{(1-f_M)^2}{d}}$$

where d is the degrees of freedom, and f is the proportion of total variance due to not observing $\underline{\theta}$ values:

(13)
$$f_M = \frac{(1 + M^{-1})B_M \cdot}{V_M}$$

Here B_M is the variance among M imputed values and V_M is the final estimate of the variance of T. When B is small relative to U^*, the reference distribution for incomplete-data statistics 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 U^* is a covariance matrix, and B is an average of squares and cross-products rather than simply an average of squares. In this case, the quantity $(T\text{-}t^*)V^1\ (T\text{-}t^*)'$ 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

(14)
$$f = \frac{(1 - M^{-1})\, Trace(BV^{-1})}{k} \ .$$

A chi-square distribution with k degrees of freedom can be used in place of f for the same reason that the normal distribution can approximate the t distribution.

Statistics t^*, the estimates of ability conditional on responses to cognitive items and student 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 and Johnson (1990), Mislevy (1991), and Mislevy and Sheehan (1987). To avoid such biases, the TIMSS 1999 analyses included nearly all student background variables.

13.5 Implementing the TIMSS Benchmarking Scaling Procedures

This section provides a synopsis of the IRT scaling and plausible-value methodology applied to the TIMSS 1999 data. Three major tasks were completed, as follows.

13.5.1 Rescaling of the TIMSS 1995 Data

TIMSS 1995 also made use of IRT scaling with plausible values (Adams, Wu, and Macaskill, 1997). The scaling model, however, relied on the one-parameter Rasch model rather than the more general two- and three-parameter models used in TIMSS 1999. Since a major goal of TIMSS 1999 was to measure trends by comparing results from both data collections, it was important that both sets of data be on the same scale. Accordingly it was decided as a first step to rescale the 1995 data using the scaling models from 1999.

The rescaling of the TIMSS 1995 data was conducted according to the method described in the TIMSS 1999 Technical Report (Yamamoto & Kulick, 2000). The scale was set so the distribution of eighth-grade proficiency scores in 1995 had a mean of 500 and a standard deviation of 100 for both the mathematics and science scales (Gonzalez, 1997). Setting the scale metric in this way produces slightly different means and standard deviations than in the original TIMSS 1995 results. Comparison of the original and rescaled 1995 proficiency scores is not appropriate because of this difference in the scale metric.

13.5.2 Scaling the 1999 Data and Linking to the 1995 Data

Since the achievement item pools used in 1995 and 1999 had about one-third of the items in common, the scaling of the 1999 data was designed to place both data sets on a common IRT scale. Although the common items administered in 1995 and 1999 formed the basis of the linkage, all of the items used in each data collection were included in the scaling since this increases the information for proficiency estimation and reduces measurement error.

The linking of the 1995 and 1999 scales was done at the mathematics and science domain levels only, since there were not enough common items to enable reliable linking within each content area.

13.5.3 Creating IRT Scales for Mathematics and Science Content Areas for 1995 and 1999 Data

IRT scales were also developed for each of the content areas in mathematics and science for both 1995 and 1999. Because there were few items common to the two assessments, and because of some differences in their composition, the two scales were not linked, but rather each was established independently.

For TIMSS 1999, the international mean for mathematics was 487 and the international mean for science was 488. The international mean for each content area was set to be equal to the subject area international mean.

13.5.4 Proficiency Scores for Benchmarking Students

Benchmarking plausible values for each student were generated using item statistics obtained from the international study. Consequently, the benchmarking plausible values are directly comparable to those obtained in the international study. For each student, five plausible values were produced for each of the five mathematics content areas (fractions and number sense; measurement; data representation, analysis, and probability; geometry; and algebra), as well as for mathematics overall. Similarly, plausible values were generated for each student in each of the six science content areas (earth science; life science; physics; chemistry; scientific inquiry; and the nature of science) and science overall.

13.6 Summary

IRT was used to model the TIMSS achievement data. TIMSS used two- and three-parameter IRT models, and plausible-value technology to reanalyze the 1995 achievement data and analyze the 1999 achievement data. Plausible-value methodology was used to generate proficiency estimates for each subject and each content area.

References

Adams, R.J., Wu, M.L., & Macaskill, G. (1997). Scaling methodology and procedures for the mathematics and science scales. In M.O. Martin & D. L. Kelly (Eds.), *TIMSS technical report volume II: Implementation and analysis* (pp. 111-145). Chestnut Hill, MA: Boston College.

Andersen, E.B. (1980). Comparing latent distributions. *Psychometrika, 45*, 121-134.

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* (pp. 397-479). Reading, MA: Addison-Wesley Publishing.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, *39*, 1-38.

Engelen, R.J.H. (1987). *Semiparametric estimation in the Rasch model.* (Department of Education Research Report No. 87-1). Twente, The Netherlands: University of Twente.

Gonzalez, E.J. (1997). Reporting student achievement in mathematics and science. In M. O. Martin & D. L. Kelly (Eds.), *TIMSS technical report volume II: Implementation and analysis* (pp. 147-174). Chestnut Hill, MA: Boston College.

Hoijtink, H. (1991). *Estimating the parameters of linear models with a latent dependent variable by nonparametric maximum likelihood.* (Research Bulletin No. HB-91-1040-EX). Groningen, The Netherlands: University of Groningen, Psychological Institute.

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

Laird, N.M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association, 73*, 805-811.

Lindsey, B., Clogg, C.C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association, 86*, 96-107.

Little, R.J.A., & Rubin, D.B. (1983). On jointly estimating parameters and missing data. *American Statistician, 37*, 218-220.

Little, R.J.A., & Rubin, D.B. (1987). *Statistical analysis with missing dat*a. New York, NY: John Wiley and Sons.

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

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

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., & Bock, R.D. (1982). *BILOG: Item analysis and test scoring with binary logistic models* (Computer program). Morresville, IN: Scientific Software.

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). Princeton, NJ: Educational Testing Service.

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.

Rubin, D.B. (1991). EM and beyond. *Psychometrika, 56*, 241-254.

Tanner, M., & Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association, 82*, 528-550.

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). Princeton, NJ: Educational Testing Service.

Van Der Linden, W.J., & Hambleton, R. (1996). *Handbook of modern item response theory.* 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* (pp. 237-263). Chestnut Hill, MA: Boston College.

Zwinderman, A.H. (1991). Logistic regression Rasch models. *Psychometrika*, *56*, 589-600.