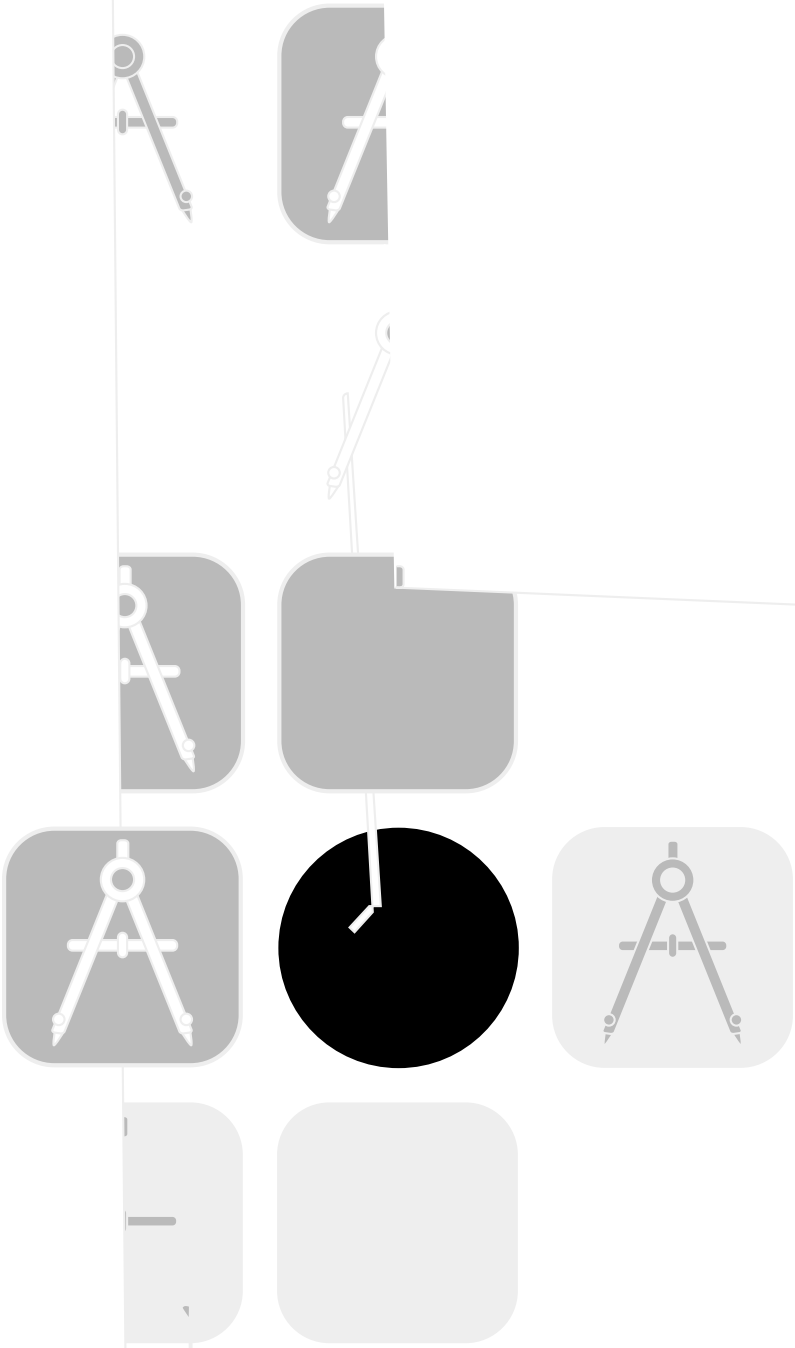
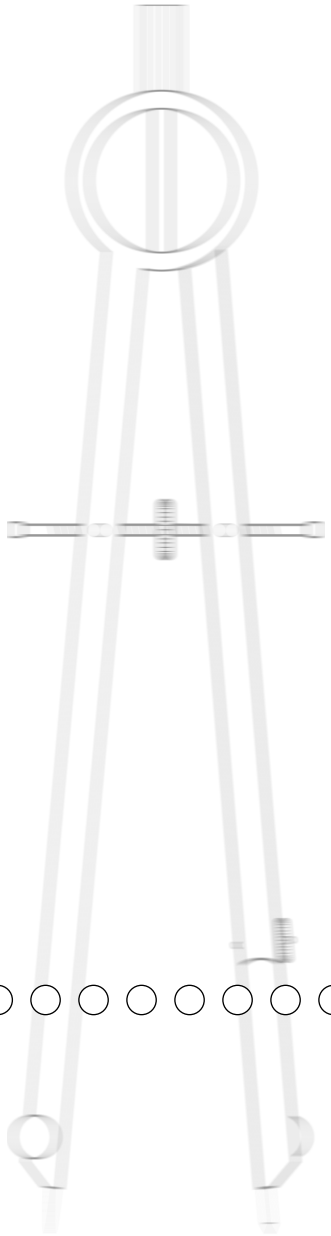


Scaling Methodology and Procedures for the TIMSS Mathematics and Science Scales

Kentaro Yamamoto
Edward Kulick



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

Kentaro Yamamoto
Edward Kulick

14.1 Overview

The TIMSS achievement test design makes use of matrix-sampling techniques to divide the assessment item pool so that each sampled student responds 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 relies on a sophisticated form of psychometric scaling known as IRT (Item Response Theory) scaling to combine the student responses in a way that provides accurate estimates of achievement. The TIMSS IRT scaling uses the multiple imputation or "plausible values" method 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.

This chapter first reviews the psychometric models used in scaling the TIMSS 1999 data, and the multiple imputation or "plausible values" methodology that allows such models to be used with sparse item-sampling in order to produce proficiency scale values of respondents. Next, the procedures followed in applying these models to the TIMSS 1999 data are described.

14.2 TIMSS 1999 Scaling Methodology

The psychometric models used in the TIMSS analysis are not new. A similar model has been used in the field of educational measurements since the 1950s and the approach has been even more popular since the 1970s in large-scale surveys, test construction, and computer adaptive testing.[2] (Birnbaum, 1968; Lord and Novick, 1968; Lord, 1980; Van Der Linden and Hambleton,1996).

Three distinct scaling models, depending on item type and scoring procedure, were used in the analysis of the 1999 TIMSS assessment data. Each is a "latent variable" model that describes the probability that a student will respond in a specific way to an

○○○

1. The TIMSS 1999 achievement test design is described in Chapter 2.

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

14.2.1 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:

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

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);
- 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.

2. 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). Other researchers have published widely on related aspects of the methodology; see, for example, Dempster, Laird, and Rubin (1977); Little and Rubin (1983, 1987); Andersen (1980); Engelen (1987); Hoijtink (1991); Laird (1978); Lindsey, Clogg, and Grego (1991); Zwinderman (1991); Tanner and Wong (1987); and Rubin (1987, 1991). The procedures used in TIMSS have also been used in several other large-scale surveys, including 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).

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

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

The two-parameter (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.

The two- and three-parameter models were used in scaling the TIMSS 1999 data in preference to the one-parameter Rasch model used in TIMSS 1995, 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 in discriminating power. However, the accuracy of representing item response functions by 2PL and 3PL models does not come without cost. Since more item parameters must be estimated, larger amounts of data — and consequently larger sample sizes — are required to obtain the same degree of confidence in the estimated item parameters. However, the TIMSS 1999 database is more than large enough to provide the required level of confidence.

Modeling item response functions as accurately as possible by using 2PL and 3PL models also reduces errors due to model mis-specification. Any mathematical modeling of data without saturated parameters contains errors not accounted for by the model. 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 preferable to use models that characterize the item response function as well as possible.

14.2.2 The IRT Model for Polytomous Items

In TIMSS 1999, free-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:

(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 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

(4)

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

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 mean of 500 with standard deviation of 100. 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_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. Under this assumption, the joint probability of a particular response pattern x across a set of n items is given by:

(5)
$$P(x|\theta_k, item\ parameters) = \prod_{i=1}^{n} \prod_{l=0}^{m_i - 1} P_{il}(\theta_k)^{u_{il}}$$

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 u_{il} is an indicator variable defined by

(6)
$$U_{il} = \begin{cases} 1\ if\ response\ x_i\ is\ in\ category\ l \\ 0\ otherwise. \end{cases}$$

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 1999 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.

14.2.3 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 14.1 and 14.2). 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 14.1 TIMSS 1999 Grade 8 Science Assessment Example Item Response Function Dichotomous Item

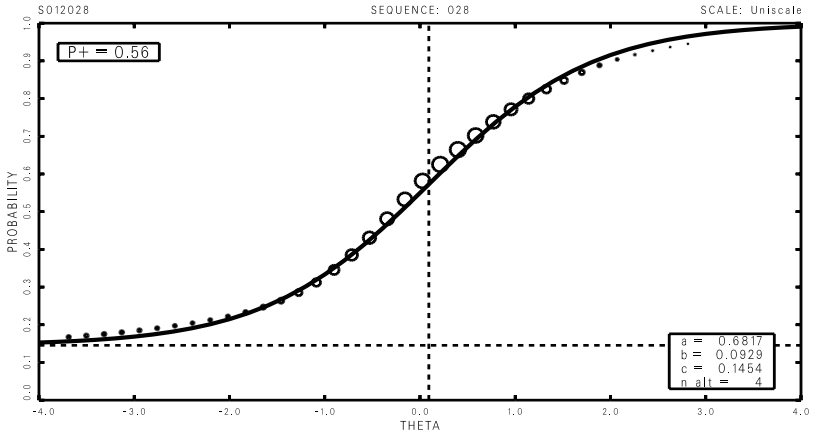


Exhibit 14.2 TIMSS 1999 Grade 8 Science Assessment Example Item Response Function Polytomous Item

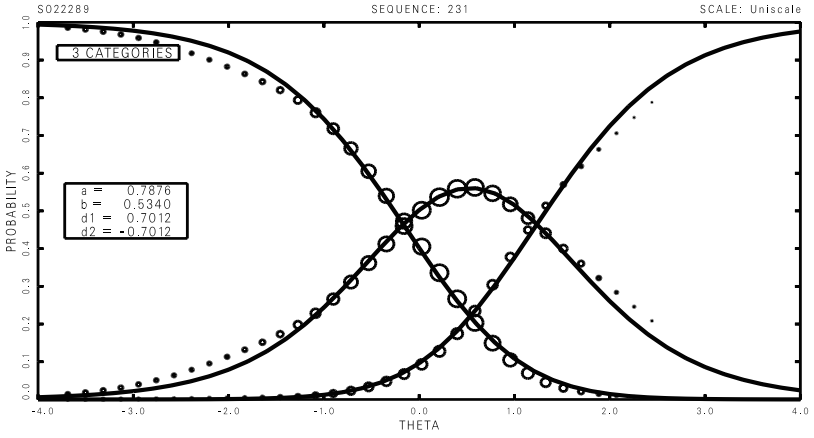


Exhibit 14.1 contains a plot of the empirical and theoretical item response functions for a dichotomous item. In the plot, the horizontal axis represents the proficiency scale, and the vertical axis represents the probability of a correct response. The solid curve is the theoretical curve based on the 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 14.2 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. The interpretation of the small circles is the same as in Exhibit 14.1. For items where the model fits the data well, the empirical and theoretical curves are close together.

14.2.4 Scaling Mathematics and Science Domains and Content Areas

In order to estimate student proficiency scores in TIMSS 1999 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 of the content areas such as algebra, geometry, etc., and then form a composite measure of mathematics by combining the content area scales. The domain-scaling approach was also found to be more reliable for assessing change from TIMSS 1995 to TIMSS 1999.

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.

14.2.5 Omitted and Not-Reached Responses.

Apart from missing data on 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. 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 not to have 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 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.

14.2.6 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 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 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)[3].

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:

(7)

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

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 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 pro-

○○○

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

cess 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.

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[4].

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

(8)
$$P(x_{ij} = 1|\theta_k,a_i,b_i,c_i) = c_i + \frac{(1 - c_i)}{1 + \exp(-1.7\,a_i(\theta_k - b_i))}$$

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 of 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.

14.2.7 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

○○○

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

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 y^c. The following model is then fit to the data.

$$\theta = \Gamma' y^c + \varepsilon, \quad (9)$$

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 $p(\theta|\boldsymbol{y})$ be correctly specified for all background variables in the survey. In TIMSS 1999, 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. Estimates of functions Γ involving background variables not conditioned on in this manner are subject to estimation error due to mis-specification. The nature of these errors was 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, θ, and variance, Σ, of the posterior distribution in (4). 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.

14.2.8 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 $P(\Gamma,\Sigma|x_j,y_j)$ 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, 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 pro-

ficiency values are drawn independently from a multivariate normal distribution with mean θ and variance Σ_j^p. These three steps are repeated five times, producing five imputations of θ for each sampled respondent.

For respondents with an insufficient number of responses, the Γ and Σs 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 (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. As in step 1 above, evaluate the sampling variance of T, or $\mathrm{Var}(T_1)$, 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 $\mathrm{Var}(T_u)$ 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.

14.2.9 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 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. However, retaining this component of uncertainty requires that additional analytic procedures be used to estimate respondents' proficiencies, as follows.

If θ values were observed for sampled respondents, the statistic $(t-T)/U^{1/2}$ would follow a *t*-distribution with d degrees of freedom. Then the incomplete-data statistic $(t^*-T)/(\mathrm{Var}(t^*))^{1/2}$ 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 θ values:

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

where 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 $(T-t^*)V^{-1}(T-t^*)'$ 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 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 background variables.

14.3 Implementing the TIMSS 1999 Scaling Procedures

This section describes how the IRT scaling and plausible value methodology was applied to the TIMSS 1999 data. This consisted of three major tasks, as follows.

Re-scaling of the 1995 TIMSS data. TIMSS in 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 1999. Since a major goal of TIMSS 1999 was to measure trends since 1995 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.

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. Item-level linking of two or more scales in this way is one of the most powerful methods of scale linking and is well suited to IRT methods. This is one of the benefits of using the IRT scaling procedures.

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 1995 and 1999 scales were not linked, but rather each was established independently.

14.3.1 Re-scaling of the 1995 TIMSS Data

The re-scaling of 1995 TIMSS followed, as much as possible, the procedures used in the original 1995 analyses, while using two- and three-parameter scaling models in place of the more restrictive one-parameter Rasch model. Item parameter estimates were obtained using an "international calibration sample" that consisted of random samples of 600 eighth-grade students from each of the 37 countries that participated in TIMSS in 1995 (plus 300 from Israel). The calibration samples were drawn with probability proportional to size of sampling weight in each country, so that the sample accurately reflected the distribution of students in the population. The 1995 estimated item parameters for mathematics may be found in Exhibit E.1 in Appendix E and for science in Exhibit E.2.

Using the re-estimated item parameters from the two- and three-parameter and polytomous IRT models, the conditioning analyses were completed, with a conditioning model similar to the one used in 1995. Following that approach, and separately within each country, responses to background variables were summarized through a principal components analysis. Enough principal components were created to account for at least 90% of the variability in the original set of background variables. In addition to the principal components, several background variables were explicitly included in the conditioning model. These included student gender and the school mean on a simple Rasch based measure of student achievement in the subject (mathematics or science) being scaled. Additionally, the conditioning for mathematics included the Rasch score for science, and the conditioning for science, the score for mathematics. Exhibit 14.3 shows the total number of conditioning variables used in the re-scaling for each country.

Exhibit 14.3 Number of Conditioning Variables for TIMSS 1995 Re-scaling

Country	Sample size	Number of Principal Components	Number of conditioning variables
Australia	12852	317	648
Austria	5786	361	735
Belgium (Flemish)	5662	473	952
Belgium (French)	4883	425	858
Bulgaria	3771	2	12
Canada	16581	348	711
Colombia	5304	357	723
Czech Republic	6672	540	1089
Cyprus	5852	358	731
Germany	5763	484	976
Denmark	4370	434	876
Spain	7596	349	707
France	6014	367	745
England	3579	261	527
Greece	7921	556	1119
Hong Kong	6752	319	646
Hungary	5978	524	1057
Ireland	6203	360	726
Iran, Islamic Rep.	7429	328	664
Iceland	3730	492	997
Israel	1415	308	621
Japan	10271	257	520
Korea, Rep. of	5827	346	697
Kuwait	1655	303	610
Lithuania	5056	540	1088
Latvia (LSS)	4976	477	962
Netherlands	4084	451	915
Norway	5736	340	691
New Zealand	6867	352	712
Philippines	11847	379	766
Portugal	6753	413	838
Romania	7471	572	1153
Russian Federation	8160	563	1131
Scotland	5776	220	448
Singapore	8285	334	675
Slovak Republic	7101	521	1050
Slovenia	5606	475	964
Sweden	8855	595	1201
Switzerland	11722	358	727
Thailand	11643	351	710
United States	10973	350	712
South Africa	9792	387	793

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. Instead, a reporting metric that has desirable properties is usually selected. In the original 1995 scaling, a metric was chosen for reporting TIMSS results such that the combined proficiency distribution for seventh and eighth grade students had a mean of 500 and a standard deviation of 100 (Gonzalez, 1997).

In the re-scaling of the 1995 data, the transformation procedures to establish the reporting metric were slightly different. Since the 1999 assessment consisted of eighth-grade students only (not both seventh- and eighth-grade students as in 1995), and since a major goal of the re-scaling was to establish a trend line to 1999, a metric was chosen for the re-scaled 1995 data that had desirable properties for the proficiency distribution of eighth-grade students. Accordingly, the scale was set so that the distribution of eighth-grade students in 1995 had a mean of 500 and a standard deviation of 100. The same metric transformation was applied to the re-scaled seventh- grade data from 1995. This procedure was followed for both the mathematics and science scales. Extreme scale values were truncated, i.e., plausible values below 5 were set to 5 and plausible values above 995 were set to 995.

Setting the scale metric as described above produces slightly lower means and slightly higher standard deviations than the original 1995 eighth-grade results. This is solely the result of the decision to base the metric on the eighth-grade distribution only rather than on the combined seventh- and eighth-grade distributions. Comparisons between the original and re-scaled 1995 proficiency scores are not appropriate because of this difference in the scale metric.

14.3.2 Scaling the 1999 Data and Linking to the 1995 Data

The linking of the 1995 and 1999 scales was conducted at the mathematics and science domain levels only, since there were not enough common items to enable reliable linking within each mathematics or science content area. As may be seen from Exhibit 14.4, about one-third of the items were common to both assessments (48 items in mathematics and 48 in science), which was enough to provide a reliable link between the 1995 and 1999 assessments.

Exhibit 14.4 Numbers of items Common and Unique to TIMSS 1995 and TIMSS 1999

Subject	Items	TIMSS 1995	TIMSS 1999
Mathematics	Unique to TIMSS 1995	111	
	Unique to TIMSS 1999		115
	Common to both TIMSS 1995 and TIMSS 1999	48	
	Total	159	163
	Grand Total for Mathematics	274	
Science	Unique to TIMSS 1995	94	
	Unique to TIMSS 1999		106
	Common to both TIMSS 1995 and TIMSS 1999	48	
	Total	142	154
	Grand Total for Science	248	

Calibration samples of 1,000 students per country per assessment were selected from each of the 25 countries that participated in both assessments, using the same method as in 1995. All 274 of the mathematics items (common items and items unique to one or the other assessment) were scaled together to provide new item parameter estimates that fit both calibration samples (1995 and 1999). The same procedure was followed for all 248 of the science items. Estimated item parameters from this joint 1995-1999 scaling may be found in Exhibit E.3 in Appendix E for mathematics and in Exhibit E.4 for science.

These item parameters estimates were used to generate plausible values for all of the 38 TIMSS 1999 countries, including those that participated only in 1999.[5] A new set of principal components was calculated for the TIMSS 1999 data in each country for use in conditioning. Exhibit 14.5 shows the total number of conditioning variables used for the TIMSS 1999 for each country. Plausible values were generated for all countries for both assessments using the new, jointly estimated item parameters.

○○○

5. In addition to its eighth-grade sample, Chile also surveyed a seventh-grade sample that was scaled with the 1999 item parameters.

Exhibit 14.5 Number of Variables and Principal Components for Conditioning TIMSS 1999

Country	Sample size	Total number of conditioning variables	Total number of principal components only
Australia	4032	374	348
Belgium (Flemish)	5259	485	479
Bulgaria	3272	582	575
Canada	8770	385	364
Chile	5907	410	405
Chinese Taipei	5772	379	374
Cyprus	3116	394	385
Czech Republic	3453	557	551
England	2960	298	292
Finland	2920	548	538
Hong Kong, SAR	5179	392	384
Hungary	3183	584	578
Indonesia	5848	403	397
Iran, Islamic Rep.	5301	406	400
Israel	4195	405	398
Italy	3328	380	374
Japan	4745	362	354
Jordan	5052	415	409
Korea, Rep. of	6114	408	388
Latvia (LSS)	2873	522	514
Lithuania	2361	342	336
Morocco	5402	681	675
Moldova	3711	599	593
Macedonia, Rep. of	4023	576	569
Malaysia	5577	386	381
Netherlands	2962	437	430
New Zealand	3613	338	332
Philippines	6601	422	415
Romania	3425	597	589
Russian Federation	4332	657	608
Singapore	4966	370	365
Slovak Rep.	3497	408	401
South Africa	8146	441	426
Slovenia	3109	584	578
Thailand	5732	398	390
Tunisia	5051	418	413
Turkey	7841	449	405
United States	9072	409	392

The final step in the scaling of the data was to locate both the 1995 and the 1999 data on the same scale. This was done by calculating transformation constants that matched the means and standard deviations of the re-scaled 1995 plausible values, which were on the required scale, with the means and standard deviations of the jointly-scaled 1995-1999 plausible values for the same set of countries, which were on an independent scale. This procedure was used for the countries that participated in both assessments.[6] The transformation constants, which were applied as $A*\theta+B$, are shown in Exhibit 14.6 for mathematics and science.

Exhibit 14.6 Transformation Constants for TIMSS 1999 Mathematics and Science Domain Scales

TIMSS 1995 and TIMSS 1999	A	B
Mathematics	99.593	510.169
Science	102.188	508.961

These linear transformations were then applied to the plausible values of the TIMSS 1999 students to place their results on the same scale as the 1995 data. If the transformation is accurate it should produce practically identical means in each country for both the re-scaled 1995 plausible values and the plausible values based on the joint 1995-1999 scaling. Exhibit 14.7 presents a comparison of the results for mathematics from both sets of data, and Exhibit 14.8 shows the same results for science. Both exhibits indicate that the differences between the proficiency means of the re-scaled 1995 data and the jointly-scaled 1995-1999 data are very small for every country. They are on average less than 20% of the standard error of measurement, implying that no systematic errors exist and that the differences can be considered ignorable.

○○○

6. Because they did not satisfy all sampling guidelines in 1995, Israel, South Africa, and Thailand were omitted from the calculation of transformation constants.

Exhibit 14.7 Comparison of 1995 TIMSS Reanalysis and Univariate Linking Mathematics Scale

Country	Mean from 1995 Re-Scaling	Mean for 1995 from Joint 1995-1999 Scaling	Difference
Singapore	609	609	0.0
Korea	581	581	0.5
Hong Kong, SAR	569	569	0.4
Japan	581	581	0.2
Belgium (Flemish)	550	549	-1.1
Netherlands	529	529	0.0
Hungary	527	526	-0.4
Canada	521	520	-1.0
Slovenia	531	530	-0.9
Russian Federation	524	523	-0.4
Australia	519	518	-0.5
Czech Republic	546	544	-1.1
Bulgaria	527	527	0.0
Latvia (LSS)	488	489	0.5
United States	492	492	-0.4
England	498	496	-1.3
New Zealand	501	501	-0.3
Lithuania	472	472	0.2
Italy	491	492	0.7
Cyprus	468	468	0.2
Romania	474	474	0.6
Iran, Islamic Rep.	418	423	4.3
Mean:	518.750	518.750	
Standard Deviation:	92.363	92.363	

Exhibit 14.8 Comparison of 1995 TIMSS Reanalysis and Univariate Linking Science Scale

Country	Mean from 1995 Re-Scaling	Mean for 1995 from Joint 1995-1999 Scaling	Difference
Australia	527	526	0.0
Belgium (Flemish)	533	533	0.0
Bulgaria	545	544	-1.0
Canada	514	516	1.9
Cyprus	452	452	0.2
Czech Republic	555	553	-1.5
England	533	533	-0.8
Hong Kong, SAR	510	512	1.9
Hungary	537	536	-0.5
Iran, Islamic Rep.	463	464	0.7
Italy	497	500	3.2
Japan	554	552	-2.5
Korea, Rep. of	546	545	-0.8
Latvia (LSS)	476	475	-0.8
Lithuania	464	465	1.3
Netherlands	541	542	1.0
New Zealand	511	512	1.2
Romania	471	471	0.3
Russian Federation	523	522	-0.9
Singapore	580	578	-2.4
Slovenia	541	538	-2.9
United States	513	515	2.7
Mean:	517.511	517.511	
Standard Deviation:	91.587	91.587	

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

The primary function of the IRT scales in the mathematics and science content areas is to portray student achievement in each country in terms of a profile of relative performance in each area. Such profiles should, for example, show countries where performance in algebra was relatively better than in geometry, or in life science than in chemistry. Although it would have been desirable to establish a link from 1995 and 1999 in each content area, there were not enough common items in the two assessments to do this reliably. However, the numbers of items in each

content area were considered sufficient to develop content area scales for each assessment separately. The five content areas in mathematics and six areas in science for which scales were developed are presented in Exhibit 14.9.

Exhibit 14.9 Number of Items in Mathematics and Science Content Areas (1995 and 1999 Combined)

Mathematics Content Areas	No. of Items	Science Content Areas	No. of Items
Fractions/Number	104	Earth Science	34
Measurement	39	Life Science	70
Data Representation	33	Physics	64
Geometry	41	Chemistry	39
Algebra	57	Environmental and Resource Issues	17
		Scientific Inquiry and the Nature of Science	12
Total	274	Total	236

The calibration samples used for the joint 1995-1999 scaling were also used to estimate the item parameters for each of the content area scales (shown in Exhibit E.5 in Appendix E for mathematics and in Exhibit E.6 for science). The principal components produced for the conditioning of the joint 1995-1999 mathematics and science domain scales were used for the 1999 content area plausible value analyses as well. Plausible values were generated for all countries for both assessments using the new, jointly estimated item parameters under multivariate conditions.

The indeterminacy of the content area scales in mathematics was resolved by setting the mean of each mathematics content area scale over all of the 38 TIMSS 1999 countries to be the same as the mean of the domain scale for mathematics. The same approach was taken for science. The transformation constants used to do this are presented in Exhibit 14.10.

It should be noted that since there were far fewer items in each content area scale than in the domain scales (for example, 57 algebra items compared with 274 mathematics items), a relatively greater proportion of the variance in the content area scales was due to measurement error. In the scaling, the total variance for content area scales and domain scales was set to be equal, and

therefore the measurement error plays a relatively greater role in the variance of the content area scales. This implies that the content area scale means of each country tend to be regressed toward the grand mean, and that the regression is more noticeable for very high- or very low-achieving countries.

Exhibit 14.10 Transformation Constants for TIMSS 1999 Content Area Scales

TIMSS 1999	A	B
Mathematics Scales		
Algebra	82.454	511.536
Data Representation	71.222	506.175
Fractions and Number	85.000	511.931
Geometry	65.933	506.741
Measurement	74.404	511.959
Science Scales		
Chemistry	64.553	505.616
Earth Science	69.688	508.543
Life Science	78.839	507.671
Physics	78.111	507.558
Environmental and Resource Issues	64.201	503.668
Scientific Inquiry and the Nature of Science	48.504	516.944

14.4 Summary

Item Response Theory was used to model the TIMSS achievement data. In order to better monitor trends in mathematics and science achievement, TIMSS used 2- and 3-parameter IRT, and plausible-value technology to re-analyze the 1995 achievement data, and analyze the 1999 achievement data. The procedures used to link the 1995 and 1999 achievement data were described.

References

Adams, R.J., Wu, M.L., & Macaskill, G. (1997). "Scaling Methodology and Procedures for the Mathematics and Science Scales" in M.O. Martin and D. L. Kelly (Eds.), *TIMSS Technical Report Volume II: Implementation and Analysis*. 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*. 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.* Research Report 87-1. Twente, the Netherlands: Department of Education, 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.* 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 HB-91-1040-EX. Groningen, The Netherlands: Psychological Institute, University of Groningen.

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

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 data*. 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.* Redding, 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]. Moresville, 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). (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.

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) (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.

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