7 Scaling Methodology and Procedures for the Mathematics and Science Scales

Raymond J. Adams
Margaret L. Wu
Greg Macaskill
Australian Council for Educational Research

The principal method by which student achievement is reported in TIMSS is through scale scores derived using Item Response Theory (IRT) scaling. With this approach, the performance of a sample of students in a subject area can be summarized on a common scale or series of scales even when different students have been administered different items. The common scale makes it possible to report on relationships between students' characteristics (based on their responses to the background questionnaires) and their overall performance in mathematics and science.

Because of the need to achieve broad coverage of both mathematics and science within a limited amount of student testing time, each student was administered relatively few items within each content area of each subject. In order to achieve reliable indices of student proficiency in this situation, it was necessary to make use of multiple imputation or "plausible values" methodology. Further information on plausible value methods may be found in Mislevy (1991), and in Mislevy, Johnson, and Muraki (1992). The proficiency scale scores or plausible values assigned to each student are actually random draws from the estimated ability distribution of students with similar item response patterns and background characteristics. The plausible values are intermediate values that may be used in statistical analyses to provide good estimates of parameters of student populations. Although intended for use in place of student scores in analyses, plausible values are designed primarily to estimate population parameters, and are not optimal estimates of individual student proficiency.

This chapter provides details of the IRT model used in TIMSS to scale the achievement data. For those interested in the technical background of the scaling, the chapter describes the model itself and the method of estimating the parameters of the model.

7.1 THE TIMSS SCALING MODEL

The scaling model used in TIMSS was the multidimensional random coefficients logit model as described by Adams, Wilson, and Wang (1997), with the addition of a multivariate linear model imposed on the population distribution. The scaling was done with the *ConQuest* software (Wu, Adams, and Wilson, 1997) that was developed in part to meet the needs of the TIMSS study.

The multidimensional random coefficients model is a generalization of the more basic unidimensional model.

7.1.1 The Unidimensional Random Coefficients Model

Assume that **I** items are indexed i=1,...,I with each item admitting $K_i + 1$ response alternatives $k = 0,1,...,K_i$. Use the vector valued random variable, $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{iK_i})$,

$$\text{where } X_{ij} = \begin{cases} 1 & \text{if response to item } i \text{ is in category } j \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

to indicate the $K_i + 1$ possible responses to item i.

A response in category zero is denoted by a vector of zeroes. This effectively makes the zero category a reference category and is necessary for model identification. The choice of this as the reference category is arbitrary and does not affect the generality of the model. We can also collect the $\mathbf{X}_i$ together into the single vector $\mathbf{X}' = (\mathbf{X}_1', \mathbf{X}_2',...,\mathbf{X}_\mathbf{i}')$, which we call the response vector (or pattern). Particular instances of each of these random variables are indicated by their lower-case equivalents: $\mathbf{x}$, $\mathbf{x}_i$ and x_{ik}.

The items are described through a vector $\xi^T = (\xi_1, \xi_2, \ldots, \xi_p)$ of p parameters. Linear combinations of these are used in the response probability model to describe the empirical characteristics of the response categories of each item. These linear combinations are defined by design vectors $\mathbf{a}_{jk}$, ($j = 1,\ldots,I$; $k = 1,\ldots K_i$) each of length p, which can be collected to form a design matrix $\mathbf{A}' = (\mathbf{a}_{11}, \mathbf{a}_{12}, \ldots, \mathbf{a}_{1K_1}, \mathbf{a}_{21}, \ldots, \mathbf{a}_{2K_2}, \ldots, \mathbf{a}_{iK_i})$. Adopting a very general approach to the definition of items, in conjunction with the imposition of a linear model on the item parameters, allows us to write a general model that includes the wide class of existing Rasch models, for example, the item bundles models of Wilson and Adams (1995).

An additional feature of the model is the introduction of a scoring function, which allows the specification of the score or "performance level" that is assigned to each possible response to each item. To do this we introduce the notion of a response score b_{ij}, which gives the performance level of an observed response in category j of item i. The b_{ij} can be collected in a vector as $\mathbf{b}^T = (b_{11}, b_{12}, \ldots, b_{1K_1}, b_{21}, b_{22}, \ldots, b_{2K_2}, \ldots, b_{iK_i})$. (By definition, the score for a response in the zero category is zero, but other responses may also be scored zero.)

In the majority of Rasch model formulations there has been a one-to-one match between the category to which a response belongs and the score that is allocated to the response. In the simple logistic model, for example, it has been standard practice to use the labels 0 and 1 to indicate both the categories of performance and the scores. A similar practice has been followed with the rating scale and partial credit models, where each different possible response is seen as indicating a different level of performance, so that the category indicators 0, 1, 2, etc. that are used serve as both scores and labels. The use of **b** as a scoring function allows a more flexible relationship between the qualitative aspects of a response and the level of performance that it reflects. Examples of where this is applicable are given in Kelderman and Rijkes (1994) and Wilson (1992). A primary reason for implementing this feature in the model was to facilitate the analysis of the two-digit coding scheme that was used in the TIMSS short-answer and extended-response items. In the final analyses, however, only the first digit of the coding was used in the scaling, so this facility in the model and scaling software was not used in TIMSS.

Letting θ be the latent variable, the item response probability model is written as:

$$\Pr(\mathbf{X}_{ij} = 1; \mathbf{A}, \mathbf{b}, \xi \mid \theta) = \frac{\exp(b_{ij}\theta + \mathbf{a}_{ij}^T \xi)}{\sum_{k=1}^{K_i} \exp(b_{ik}\theta + \mathbf{a}_{ij}^T \xi)} \tag{2}$$

and a response vector probability model as

$$f(\mathbf{x}; \xi \mid \theta) = \Psi(\theta, \xi) \exp[\mathbf{x}^T(\mathbf{b}\theta + \mathbf{A}\xi)] \tag{3}$$

with

$$\Psi(\theta, \xi) = \left\{ \sum_{\mathbf{z} \in \Omega} \exp[\mathbf{z}^T(\mathbf{b}\theta + \mathbf{A}\xi)] \right\}^{-1} \tag{4}$$

where Ω is the set of all possible response vectors.

7.1.2 The Multidimensional Random Coefficients Multinomial Logit Model

The multidimensional form of the model is a straightforward extension of the model that assumes that a set of D traits underlie the individuals' responses. The D latent traits define a D-dimensional latent space, and the individuals' positions in the D-dimensional latent space are represented by the vector $\theta = (\theta_1, \theta_2, \ldots, \theta_D)$. The scoring function of response category k in item i now corresponds to a D by 1 column vector rather than a scalar as in the unidimensional model. A response in category k in dimension d of item i is scored b_{ikd}. The scores across D dimensions can be collected into a column vector $\mathbf{b}_{ik} = (b_{ik1}, b_{ik2}, \ldots, b_{ikD})^T$, again be collected into the scoring sub-matrix for item i, $\mathbf{B}_i = (\mathbf{b}_{i1}, \mathbf{b}_{i2}, \ldots, \mathbf{b}_{ik_i})^T$, and then be collected into a scoring matrix $\mathbf{B} = (\mathbf{B}_1^T, \mathbf{B}_2^T, \ldots \mathbf{B}_I^T)^T$ for the whole test. If the item parameter vector, ξ, and the design matrix, $\mathbf{A}$, are defined as they were in the unidimensional model, the probability of a response in category k of item i is modeled as

$$\Pr(\mathbf{X}_{ij} = 1; \mathbf{A}, \mathbf{B}, \xi \mid \theta) = \frac{\exp(\mathbf{b}_{ij}\theta + \mathbf{a}_{ij}^T \xi)}{\sum_{k=1}^{K_i} \exp(\mathbf{b}_{ik}\theta + \mathbf{a}_{ik}^T \xi)} \tag{5}$$

And for a response vector we have:

$$f(\mathbf{x}; \xi \mid \theta) = \Psi(\theta, \xi) \exp[\mathbf{x}'(\mathbf{B}\theta + \mathbf{A}\xi)] \tag{6}$$

with

$$\Psi(\theta,\xi) = \left\{\sum_{\mathbf{z}\in\Omega} \exp[\mathbf{z}^T(\mathbf{B}\theta + \mathbf{A}\xi)]\right\}^{-1} \tag{7}$$

The difference between the unidimensional model and the multidimensional model is that the ability parameter is a scalar, θ, in the former, and a D by 1 column vector, θ, in the latter. Likewise, the scoring function of response k to item i is a scalar, b_{ik}, in the former, whereas it is a D by 1 column vector, $\mathbf{b}_{ik}$, in the latter.

7.2 THE POPULATION MODEL

The item response model is a conditional model in the sense that it describes the process of generating item responses conditional on the latent variable, θ. The complete definition of the TIMSS model, therefore, requires the specification of a density, $f_\theta(\theta;\alpha)$, for the latent variable, θ. We use α to symbolize a set of parameters that characterize the distribution of θ. The most common practice when specifying unidimensional marginal item response models is to assume that the students have been sampled from a normal population with mean μ and variance σ^2. That is:

$$f_\theta(\theta;\alpha) \equiv f_\theta(\theta;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2}\right] \tag{8}$$

or equivalently

$$\theta = \mu + E \tag{9}$$

where $E \sim N(0,\sigma2)$.

Adams, Wilson, and Wu (1997) discuss how a natural extension of (8) is to replace the mean, μ, with the regression model, $\mathbf{Y}_n^T\beta$, where $\mathbf{Y}_n$ is a vector of u fixed and known values for student n, and β is the corresponding vector of regression coefficients. For example, $\mathbf{Y}_n$ could be constituted of student variables such as gender, socio-economic status, or major. Then the population model for student n becomes

$$\theta_n = \mathbf{Y}_n^T\beta + E_n \tag{10}$$

where we assume that the E_n are independently and identically normally distributed with mean zero and variance σ^2 so that (10) is equivalent to

$$f_\theta(\theta_n;\mathbf{Y}_n,\mathrm{b},\sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}}\exp\left[-\frac{1}{2\sigma^2}(\theta_n - \mathbf{Y}_n^T\beta)^T(\theta_n - \mathbf{Y}_n^T\beta)\right] \tag{11}$$

a normal distribution with mean $\mathbf{Y}_n^T\beta$ and variance σ^2. If (11) is used as the population model then the parameters to be estimated are β, σ^2, and ξ.

The TIMSS scaling model takes the generalization one step further by applying it to the vector valued θ rather than the scalar valued θ, resulting in the multivariate population model

$$f_\theta(\theta_n;\mathbf{W}_n,\gamma,\Sigma) = (2\pi)^{-\frac{D}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\theta_n-\gamma\mathbf{W}_n)^T\Sigma^{-1}(\theta_n-\gamma\mathbf{W}_n)\right] \quad (12)$$

where γ is a $u \times D$ matrix of regression coefficients, Σ is a $D \times D$ variance-covariance matrix and $\mathbf{W}_n$ is a $u \times 1$ vector of fixed variables. If (12) is used as the population model then the parameters to be estimated are γ, Σ, and ξ. In TIMSS we refer to the W_n variables as conditioning variables.

7.3 ESTIMATION

The ConQuest software uses maximum likelihood methods to provide estimates of γ, Σ, and ξ. Combining the conditional item response model (6) and the population model (12) we obtain the unconditional or marginal response model

$$f_x(\mathbf{x};\xi,\gamma,\Sigma) = \int_\theta f_x(\mathbf{x};\xi|\theta)f_\theta(\theta;\gamma,\Sigma)d\theta \quad (13)$$

and it follows that the likelihood is

$$\Lambda = \prod_{n=1}^{N} f_x(\mathbf{x}_n;\xi,\gamma,\Sigma) \quad (14)$$

where N is the total number of sampled students.

Differentiating with respect to each of the parameters and defining the marginal posterior as

$$h_\theta(\theta_n;\mathbf{W}_n\xi,\gamma,\Sigma|\mathbf{x}_n) = \frac{f_x(\mathbf{x}_n;\xi|\theta_n)f_\theta(\theta_n;\mathbf{W}_n,\gamma,\Sigma)}{f_x(\mathbf{x}_n;\mathbf{W}_n\xi,\gamma,\Sigma)} \quad (15)$$

provides the following system of likelihood equations:

$$\mathbf{A}'\sum_{n=1}^{N}\left[\mathbf{x}_n - \int_{\theta_n} E_{\mathbf{z}}(\mathbf{z}|\theta_n)h_\theta(\theta_n;\mathbf{Y}_n,\xi,\gamma,\Sigma|\mathbf{x}_n)d\theta_n\right] = \mathbf{0} \quad (16)$$

$$\hat{\gamma} = \left(\sum_{n=1}^{N}\bar{\theta}_n\mathbf{W}_n^T\right)\left(\sum_{n=1}^{N}\mathbf{W}_n\mathbf{W}_n^T\right)^{-1} \quad (17)$$

and

$$\hat{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}\int_{\theta_n}(\theta_n-\gamma\mathbf{W}_n)(\theta_n-\gamma\mathbf{W}_n)^T h_\theta(\theta_n;\mathbf{Y}_n,\xi,\gamma,\Sigma|\mathbf{x}_n)d\theta_n \quad (18)$$

where

$$E_z(\mathbf{z}|\theta_n) = \Psi(\theta_n, \xi) \sum_{\mathbf{z} \in \Omega} \mathbf{z} \exp[\mathbf{z}'(\mathbf{b}\theta_n + \mathbf{A}\xi)] \tag{19}$$

and

$$\bar{\theta}_n = \int_{\theta_n} \theta_n h_\theta(\theta_n; \mathbf{Y}_n, \xi, \gamma, \Sigma|\mathbf{x}_n) d\theta_n \tag{20}$$

The system of equations defined by (16), (17), and (18) is solved using an EM algorithm (Dempster, Laird, and Rubin, 1977) following the approach of Bock and Aitken (1981).

7.3.1 Quadrature and Monte Carlo Approximations

The integrals in equations (16), (17) and (18) are approximated numerically using either quadrature or Monte Carlo methods. In each case we define, Θ_p, p=1,...,P a set of P D-dimensional vectors (which we call nodes), and for each node we define a corresponding weight $W_p(\gamma,\Sigma)$. The marginal item response probability (13) is then approximated using

$$f_{\mathbf{x}}(\mathbf{x}; \xi, \gamma, \Sigma) = \sum_{p=1}^{P} f_{\mathbf{x}}(\mathbf{x}; \xi|\Theta_p) W_p(\gamma, \Sigma) \tag{21}$$

and the marginal posterior (15) is approximated using

$$h_\Theta(\Theta_q; \mathbf{W}_n, \xi, \gamma, \Sigma|\mathbf{x}_n) = \frac{f_{\mathbf{x}}(\mathbf{x}_n; \xi|\Theta_q) W_q(\gamma, \Sigma)}{\sum_{p=1}^{P} f_{\mathbf{x}}(\mathbf{x}_n; \xi|\Theta_p) W_p(\gamma, \Sigma)} \tag{22}$$

for q=1,...,P.

The EM algorithm then proceeds as follows:

Step 1. Prepare a set of nodes and weights depending upon $\gamma^{(t)}$ and $\Sigma^{(t)}$ the estimates of γ and Σ at iteration t.

Step 2. Calculate the discrete approximation of the marginal posterior density of θ_n given $\mathbf{x}_n$ at iteration t using

$$h_\Theta(\Theta_p; \mathbf{W}_n, \xi^{(t)} \gamma^{(t)}, \Sigma^{(t)}|\mathbf{x}_n) = \frac{f_{\mathbf{x}}(\mathbf{x}_n; \xi^{(t)}|\Theta_p) W_p(\gamma^{(t)}, \Sigma^{(t)})}{\sum_{p=1}^{P} f_{\mathbf{x}}(\mathbf{x}_n; \xi^{(t)}|\Theta_p) W_p(\gamma^{(t)}, \Sigma^{(t)})} \tag{23}$$

where $\xi^{(t)}$, $\gamma^{(t)}$, $\Sigma^{(t)}$ and are estimates of $\xi^{(t)}$, $\gamma^{(t)}$, and $\Sigma^{(t)}$ at iteration t.

Step 3. Use a Newton-Raphson method to solve the following to produce estimates of $\hat{\xi}^{(t+1)}$.

$$A'\sum_{n=1}^{N}\left[\mathbf{x}_n - \sum_{r=1}^{P} E_{\mathbf{z}}(\mathbf{z}|\Theta_r)h_{\Theta}(\Theta_r;\mathbf{W}_n, \xi^{(t)}, \gamma^{(t)}, \Sigma^{(t)}|\mathbf{x}_n)\right] = \mathbf{0} \tag{24}$$

Step 4. Estimate $\gamma^{(t+1)}$ and $\Sigma^{(t+1)}$ using

$$\hat{\gamma}^{(t+1)} = \left(\sum_{n=1}^{N}\bar{\Theta}^n\mathbf{W}_n^T\right)\left(\sum_{n=1}^{N}\mathbf{W}_n\mathbf{W}_n^T\right)^{-1} \tag{25}$$

and

$$\hat{\Sigma}^{(t+1)} = \frac{1}{N}\sum_{n=1}^{N}\sum_{r=1}^{P}(\Theta_r - \gamma^{(t+1)}\mathbf{W}_n)(\Theta_r - \gamma^{(t+1)}\mathbf{W}_n)^T h_{\Theta}(\Theta_r;\mathbf{Y}_n, \xi^{(t)}, \gamma^{(t)}, \Sigma^{(t)}|\mathbf{x}_n) \tag{26}$$

where

$$\bar{\Theta}^n = \sum_{r=1}^{P}\Theta_r h_{\Theta}(\Theta_r;\mathbf{W}_n, \xi^{(t)}, \gamma^{(t)}, \Sigma^{(t)}|\mathbf{x}_n) \tag{27}$$

Step 5. Return to Step 1.

The difference between the quadrature and Monte Carlo methods lies in the way the nodes and weights are prepared. For the quadrature case we begin by choosing a fixed set of Q points, $(\Theta_{d1}, \Theta_{d2},\ldots,\Theta_{dQ})$ for each latent dimension and then define a set of Q^D nodes that are indexed r = $1,\ldots,Q^D$, and are given by the Cartesian coordinates

$$\Theta_r = (\Theta_{1j_1}, \Theta_{2j_2}, \ldots, \Theta_{dj_d}) \text{ with } j_1 = 1,\ldots Q; j_2 = 1,\ldots,Q; \ldots; j_d = 1,\ldots,Q \quad .$$

The weights are then chosen to approximate the continuous latent population density (12). That is,

$$W_p = K(2\pi)^{-\frac{d}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\Theta_p - \gamma\mathbf{W}_n)^T\Sigma^{-1}(\Theta_p - \gamma\mathbf{W}_n)\right] \tag{28}$$

where K is a scaling factor to ensure that the sum of the weights is 1.

In the Monte Carlo case the nodes are drawn at random from the standard multivariate normal distribution and at each iteration the nodes are rotated using standard methods so that they become random draws from a multivariate normal distribution with mean $\gamma\mathbf{W}_n$ and variance Σ. In the Monte Carlo case the weight for all nodes is $1/P$.

For further information on the quadrature approach to estimating the model see Adams, Wilson, and Wang (1997), and for further information on the Monte Carlo method see Volodin and Adams (1997). In the TIMSS scaling the Bock-Aitken quadrature approach was used for unidimensional models and the Volodin Monte Carlo methods was used when scaling in high dimensions.

7.3.2 Latent Estimation and Prediction

The marginal item response (13) does not include parameters for the latent values θ_n and hence the estimation algorithm does not result in estimates of the latent values. For TIMSS, expected a posteriori estimates (EAP) of each student's latent achievement was produced. The EAP prediction of the latent achievement for case n is

$$\theta_n^{EAP} = \sum_{r=1}^{P} \Theta_r h_\Theta(\Theta_r; \mathbf{W}_n, \hat{\xi}, \hat{\gamma}, \hat{\Sigma} | \mathbf{x}_n) \tag{29}$$

Variance estimates for these predictions were estimated using

$$\mathrm{var}(\theta_n^{EAP}) = \sum_{r=1}^{P} (\Theta_r - \theta_n^{EAP})(\Theta_r - \theta_n^{EAP})^T h_\Theta(\Theta_r; \mathbf{W}_n \hat{\xi}, \hat{\gamma}, \hat{\Sigma} | \mathbf{x}_n) \tag{30}$$

7.3.3 Drawing Plausible Values

Plausible values are random draws from the marginal posterior of the latent distribution, (15), for each student. For details on the use of plausible values the reader is referred to Mislevy (1991) and Mislevy et al. (1992).

Unlike previously described methods for drawing plausible values (Beaton, 1987; Mislevy et al., 1992) ConQuest does not assume normality of the marginal posterior distributions. Recall from (15) that the marginal posterior is given by

$$h_\theta(\theta_n; \mathbf{W}_n, \xi, \gamma, \Sigma | \mathbf{x}_n) = \frac{f_{\mathbf{x}}(\mathbf{x}_n; \xi | \theta_n) f_\theta(\theta_n; \mathbf{W}_n, \gamma, \Sigma)}{\int_\theta f_{\mathbf{x}}(\mathbf{x}; \xi | \theta) f_\theta(\theta, \gamma, \Sigma) d\theta} \tag{31}$$

The ConQuest procedure begins drawing M vector valued random deviates, $\{\varphi_{nm}\}_{m=1}^{M}$ from the multivariate normal distribution $f\theta(\theta_n, \mathbf{W}_n\gamma, \Sigma)$ for each case n. These vectors are used to approximate the integral in the denominator of (31) using the Monte Carlo integration

$$\int_\theta f_{\mathbf{x}}(\mathbf{x}; \xi | \theta) f_\theta(\theta, \gamma, \Sigma) d\theta \approx \frac{1}{M} \sum_{m=1}^{M} f_{\mathbf{x}}(\mathbf{x}; \xi | \varphi_{mn}) \equiv \Im \tag{32}$$

At the same time the values

$$p_{mn} = f_{\mathbf{x}}(\mathbf{x}_n;\xi|\varphi_{mn})f_{\theta}(\varphi_{mn};\mathbf{W}_n,\gamma,\Sigma) \tag{33}$$

are calculated, so that we obtain the set of pairs $\left(\varphi_{nm}, \frac{p_{mn}}{\mathfrak{J}}\right)_{m=1}^{M}$ which can be used as an approximation to the posterior density (31). The probability that φ_{nj} could be drawn from this density is given by

$$q_{nj} = \frac{p_{mn}}{\sum_{m=1}^{M} p_{mn}} \tag{34}$$

At this point, L uniformly distributed random numbers, $\{\eta_i\}_{i=1}^{L}$, are generated and for each random draw the vector φ_{ni_0} that satisfies the condition

$$\sum_{s=1}^{i_0-1} q_{sn} < \eta_i \le \sum_{s=1}^{i_0} q_{sn} \tag{35}$$

is selected as a plausible vector.

7.4 SCALING STEPS

The item response model described above was fit to the data in two steps. In the first step the items were calibrated using a subsample of students drawn from the samples of the participating countries. These samples were called the *international calibration samples.* In a second step the model was fit separately for each country with the item parameters fixed at values estimated in the first step.

There were three principal reasons for using an international calibration sample for estimating international item parameters. First, it seemed unnecessary to estimate parameters using the complete data set; second, drawing equal-sized subsamples from each country for inclusion in the international calibration sample ensured that each country was given equal weight in the estimation of the international parameters; and third, the drawing of appropriately weighted samples meant that weighting would not be necessary in the international scaling runs.

7.4.1 Drawing the International Calibration Sample

At the time when the international scaling of the data commenced the TIMSS database of item response data contained information from 25 Population 1 countries and 39 Population 2 countries. Those countries are listed in Table 7.1.

For each target population, samples of 600 tested students were selected from the database for each participating country. This generally lead to roughly equal samples from each target grade. For Israel, where only the upper grade was tested, the sample size was reduced to 300 tested students. The sampled students were selected using a probability-proportional-to-size systematic selection method. The overall sampling weights were used as measures of size for this purpose. This resulted in equal selection probabilities, within national samples, for the students in the calibration samples. The Population 1 and 2 international calibration samples contained 14,700 and 23,100 students, respectively.

Table 7.1 Countries Included in the International Item Calibration

Population 1[1]	Population 2[2]	
Australia	Australia	Netherlands
Austria	Austria	New Zealand
Canada	Belgium (Flemish)	Norway
Cyprus	Belgium (French)	Portugal
Czech Republic	Bulgaria	Romania
England	Canada	Russian Federation
Greece	Colombia	Scotland
Hong Kong	Cyprus	Singapore
Hungary	Czech Republic	Slovak Republic
Iceland	Denmark	Slovenia
Iran	England	Spain
Ireland	France	Sweden
Israel*	Germany	Switzerland
Japan	Greece	United States
Korea	Hong Kong	
Latvia	Hungary	
Mexico	Iceland	
Netherlands	Iran	
New Zealand	Ireland	
Norway	Israel*	
Portugal	Japan	
Scotland	Korea	
Singapore	Latvia	
Slovenia	Lithuania	
United States	Mexico	

*A sample of 600 students was drawn from each country, excepting for Israel where only 300 students where drawn because Israel sampled students from only the higher of the two grade levels.

Note: Mexico's data was used to estimate the international item parameters, although Mexico subsequently withdrew its results from the international reports. Although results for Kuwait, the Philippines, and South Africa were reported in the international reports, their data were not used to estimate the international parameters.

[1] Third and fourth grades in most countries.

[2] Seventh and eighth grades in most countries.

7.4.2 International Scaling Results

Tables 7.2, 7.3, 7.4, and 7.5 show the basic statistics that resulted from international scaling for mathematics and science at Populations 1 and 2. The number of respondents shown for each item is the number of cases that were considered valid for calibration purposes. There are two reasons why this value is not equal to the total number of students in the calibration samples. First, the test rotation design was such that only items in cluster A were administered to all students (see Adams and Gonzalez, 1996), and second, items to which students did not respond because there were deemed to be "not reached" were treated as missing data in the calibration phase of the analysis. The percent correct figures that are reported were computed by summing the total of the scores achieved by all students who provided valid responses and dividing that by the number of students multiplied by the maximum score that could be achieved for that item; for most but not all items, the maximum possible scores was one. The difficulty estimate and asymptotic errors are in the logit metric, which is the natural metric for the ConQuest scaling software. The mean square fit statistic is an index of the fit of the data to the assumed scaling model; the statistic used here was derived by Wu (1997). Under the null hypothesis that the data and model are consistent, the expected value of these statistics is one. Values that are less than one are usually associated with items that have greater than average discrimination, while values that are greater than one may result from lower than average discrimination, guessing, or some other deviation from the model.

7.4.3 Fit of the Scaling Model

Tables 7.1 and 7.2 show the international item statistics and parameter estimates for Population 1 mathematics and science, respectively. Table 7.3 and 7.4 show the corresponding information for Population 2. The mean square fit statistics reported in Tables 7.2, 7.3, 7.4, and 7.5 show that the vast majority of items fit the Rasch model very well. Items with mean squares greater than one in Population 1 mathematics were B06, H05, I08, K07, M07, U01, and V04a. The reasons for the misfit of these items vary. For item B06, misfit is caused by the fact that the item does not discriminate as well as the other items. This may be seen in Figure 7.1 showing the modeled and empirical item characteristic curves for this item. For item H05, the modeled and empirical item characteristics curves are shown in Figure 7.2. There appear to be two reasons for this misfit of this item: first, it is slightly less discriminating than was assumed by the model; but second, interestingly, some students in the middle of the latent ability distribution did not perform as well as was expected, and these students would receive considerable weight in the estimation of the weighted mean square. Item K07 (Figure 7.3) was amongst the most difficult items, and it was multiple choice, so it is not surprising that some students are likely to have attempted to guess the correct response. A closer review of the item shows that one of the distracters proved to be attractive to some higher-achieving students – in fact the point-biserial for the distracter is positive for quite a few countries. This item survived the review process because of a policy decision to retain as many items as possible. Items M07, U01, and V04a all misfit because they had a slightly lower than modeled discrimination.

The items that had mean square statistics less than one were all found to be more discriminating than was modeled. Misfit of this form is not usually deemed to be of concern. Interestingly, however, the majority of the most discriminating items are short-answer or extended-response type. This may well be due to the fact that it is unlikely that students would have guessed the answers to these questions.

Table 7.2 Population 1 Mathematics: Item Statistics and Parameter Estimates for the International Calibration Sample

Item Label	Number of Respondents in International Calibration Sample	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
ASMMA01	14637	79.78	-1.345	0.022	1.03
ASMMA02	14627	51.47	0.218	0.018	1.08
ASMMA03	14618	58.61	-0.131	0.018	1.01
ASMMA04	14603	78.33	-1.242	0.022	0.95
ASMMA05	14571	79.29	-1.307	0.022	1.06
ASMMB05	7283	60.66	-0.233	0.026	0.99
ASMMB06	7268	48.98	0.343	0.026	1.13
ASMMB07	7259	40.52	0.761	0.026	1.06
ASMMB08	7247	82.03	-1.511	0.033	0.94
ASMMB09	7240	56.66	-0.029	0.026	0.99
ASMMC01	5460	80.44	-1.401	0.037	0.89
ASMMC02	5447	64.79	-0.448	0.031	0.96
ASMMC03	5441	48.87	0.350	0.030	0.99
ASMMC04	5437	75.19	-1.040	0.034	0.90
ASMMD05	5463	74.81	-1.007	0.034	0.93
ASMMD06	5451	51.73	0.213	0.030	1.09
ASMMD07	5438	83.41	-1.609	0.039	0.94
ASMMD08	5428	55.56	0.031	0.030	0.99
ASMMD09	5411	47.92	0.405	0.030	1.01
ASMME01	5439	63.28	-0.373	0.031	0.98
ASMME02	5420	71.25	-0.807	0.033	0.90
ASMME03	5425	59.54	-0.177	0.031	0.96
ASMME04	5413	31.98	1.216	0.032	1.10
ASMMF05	5471	66.02	-0.515	0.031	0.97
ASMMF06	5459	59.11	-0.161	0.030	1.09
ASMMF07	5451	59.70	-0.189	0.030	0.99
ASMMF08	5445	60.39	-0.223	0.030	1.04
ASMMF09	5437	68.81	-0.662	0.032	1.03
ASMMG01	5181	36.33	0.987	0.032	1.07
ASMMG02	5170	69.83	-0.707	0.033	1.06
ASMMG03	5152	87.36	-1.990	0.044	0.93
ASMMG04	5365	47.83	0.414	0.030	1.01
ASMMH05	5434	45.99	0.464	0.030	1.15
ASMMH06	5422	48.51	0.345	0.030	1.06
ASMMH07	5407	66.08	-0.518	0.031	1.01
ASMMH08	5394	64.29	-0.422	0.031	0.99
ASMMH09	5383	65.11	-0.464	0.031	1.00
ASMMI01	1864	49.79	0.306	0.051	1.01
ASMMI02	1859	31.47	1.239	0.055	1.07
ASMMI03	1859	53.68	0.118	0.051	1.03
ASMMI04	1859	81.33	-1.461	0.064	0.90
ASMMI05	1859	46.26	0.480	0.051	0.98
ASMMI06	1858	69.48	-0.697	0.055	0.97
ASMMI07	1858	54.36	0.086	0.051	0.89
ASMMI08	1854	49.30	0.334	0.051	1.17
ASMMI09	1853	61.09	-0.247	0.052	1.00
ASMMJ01	1814	84.45	-1.768	0.069	0.94

Table 7.2 Population 1 Mathematics: Item Statistics and Parameter Estimates for the International Calibration Sample (Continued 1)

Item Label	Number of Respondents in International Calibration Sample	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
ASMMJ02	1811	60.41	-0.237	0.053	1.07
ASMMJ03	1728	68.34	-0.698	0.057	0.85
ASMMJ04	1804	39.36	0.831	0.054	0.97
ASMMJ05	1724	36.31	0.966	0.056	1.05
ASMMJ06	1799	66.54	-0.555	0.055	1.08
ASMMJ07	1796	53.73	0.112	0.053	0.97
ASMMJ08	1793	44.28	0.588	0.053	0.99
ASMMJ09	1783	71.34	-0.817	0.058	0.99
ASMMK01	1803	62.17	-0.287	0.054	1.07
ASMMK02	1797	77.13	-1.147	0.061	0.98
ASMMK03	1796	45.38	0.560	0.053	0.98
ASMMK04	1718	44.88	0.628	0.054	0.93
ASMMK05	1787	73.81	-0.928	0.059	1.07
ASMMK06	1784	58.18	-0.069	0.053	0.99
ASMMK07	1779	22.60	1.848	0.062	1.18
ASMMK08	1771	68.83	-0.629	0.056	0.91
ASMMK09	1764	35.43	1.088	0.055	1.04
ASSML01	1791	42.16	0.699	0.053	0.85
ASMML02	1789	45.95	0.515	0.052	1.01
ASMML03	1714	83.84	-1.638	0.070	0.97
ASMML04	1784	66.70	-0.512	0.055	0.94
ASMML05	1782	38.61	0.886	0.053	1.07
ASMML06	1705	47.39	0.423	0.053	1.06
ASMML07	1772	41.25	0.755	0.053	0.89
ASMML08	1767	28.13	1.459	0.058	0.96
ASMML09	1761	59.68	-0.144	0.053	1.02
ASMMM01	1829	73.32	-0.905	0.057	1.04
ASSMM02	1826	47.21	0.415	0.051	0.91
ASMMM03	1819	58.71	-0.133	0.052	1.03
ASSMM04	1811	36.33	0.953	0.053	1.01
ASMMM05	1805	36.95	0.923	0.053	1.13
ASMMM06	1798	65.46	-0.463	0.054	0.95
ASMMM07	1788	36.86	0.934	0.053	1.15
ASMMM08	1779	84.54	-1.664	0.069	0.92
ASMMM09	1778	65.75	-0.472	0.054	0.93
ASEMS01	3501	43.32	0.600	0.024	1.01
ASSMS02	3356	59.06	-0.068	0.039	0.86
ASEMS03	3297	28.45	1.280	0.026	0.95
ASSMS04	3096	48.87	0.496	0.040	0.91
ASSMS05	3016	52.06	0.360	0.040	1.08
ASEMT01	3336	70.92	-0.734	0.042	0.85
ASEMT01	3266	40.55	0.760	0.025	0.94
ASSMT02	3328	41.17	0.827	0.039	0.94
ASSMT03	3257	45.96	0.601	0.039	0.92
ASEMT04a	3082	18.23	2.231	0.050	0.91
ASEMT04b	2984	12.40	2.771	0.059	0.89
ASSMT05	3033	62.45	-0.179	0.041	0.99

Table 7.2 Population 1 Mathematics: Item Statistics and Parameter Estimates for the International Calibration Sample (Continued 2)

Item Label	Number of Respondents in International Calibration Sample	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
ASEMU01	3483	53.37	0.209	0.023	1.19
ASSMU02	3418	51.20	0.300	0.038	1.10
ASEMU03a	3323	58.47	-0.035	0.039	0.84
ASEMU03b	3274	40.16	0.877	0.039	0.83
ASEMU03c	3250	75.75	-0.975	0.044	0.98
ASSMU04	3237	54.71	0.167	0.039	0.94
ASSMU05	3152	80.43	-1.277	0.048	0.95
ASEMV01	3486	37.74	0.872	0.026	1.04
ASSMV02	3438	41.97	0.711	0.038	0.87
ASSMV03	3347	60.86	-0.188	0.039	0.88
ASEMV04a	3305	49.26	0.390	0.030	1.16
ASEMV04b	3232	45.39	0.578	0.039	0.90
ASSMV05	3104	47.29	0.515	0.039	0.99

Figure 7.1 Empirical and Modelled Item Characteristic Curves for Mathematics Population 1 Item: ASMMB06. Fit MNSQ=1.13

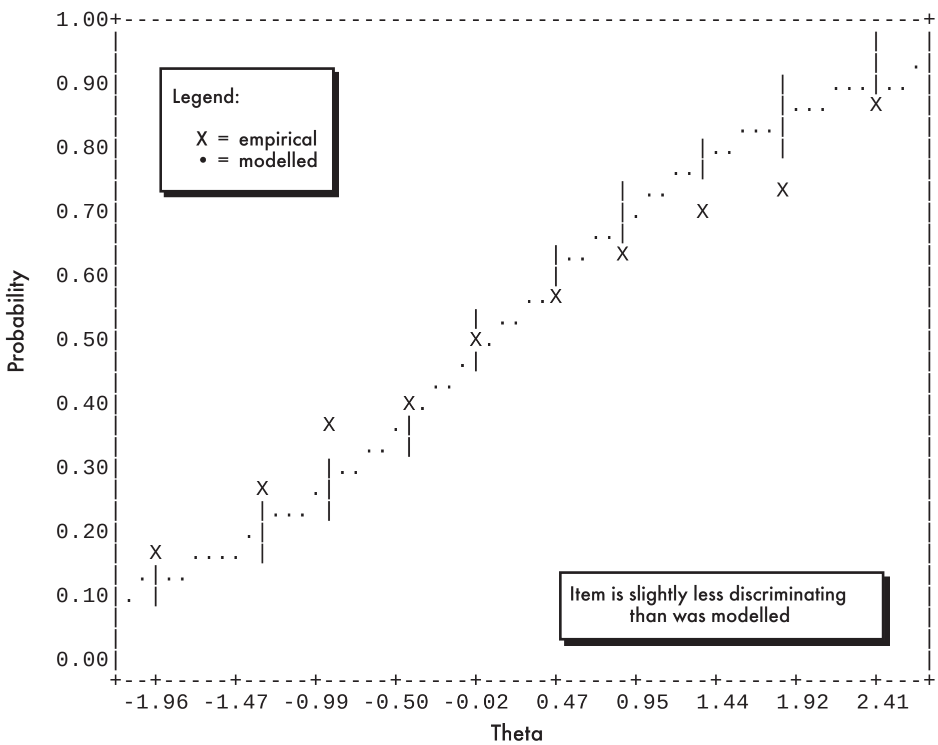


Figure 7.2 Empirical and Modelled Item Characteristic Curves for Mathematics Population 1 Item: ASEMH05. Fit MNSQ=1.15

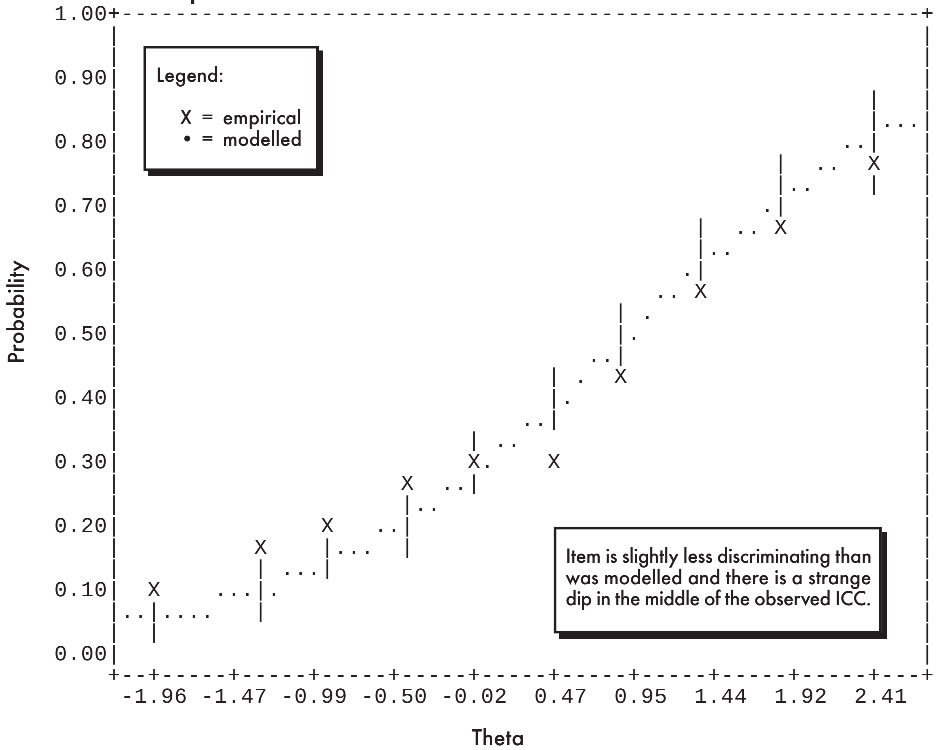
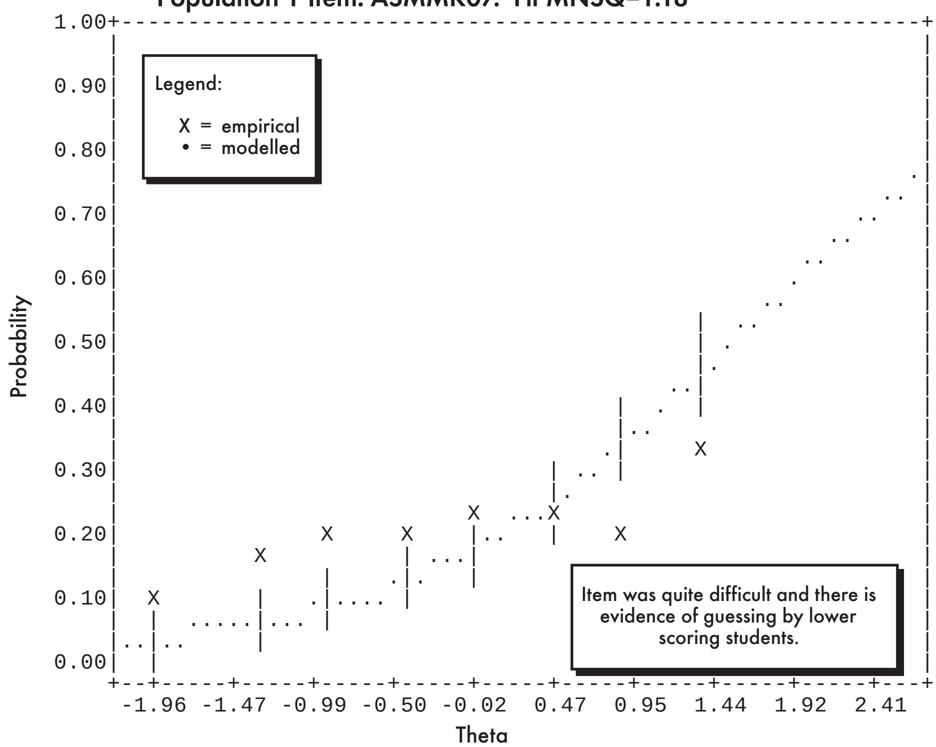


Figure 7.3 Empirical and Modelled Item Characteristic Curves for Mathematics Population 1 Item: ASMMK07. Fit MNSQ=1.18



For Population 1 science there are fewer misfitting items than for mathematics. The worst-fitting items are G08, H04, O05, and R03. Item G08 (Figure 7.4) was a difficult item and its misfit may be caused by guessing, while the misfit of H04 is caused by lower than modeled discrimination. An examination of the country-level data for these items shows that both have distracters that have positive point-biserials in a number of countries. The misfit of items O05 and R03, which is illustrated in Figures 7.5 and 7.6, is more difficult to explain – both of these items performed quite well in each of the participating countries. Item O05 does show some evidence of lower than expected discrimination, but there is also a large "blip" in the observed percentage correct for student in the second-lowest ability grouping. Item R03 has fewer than expected students at the upper achievement levels. An examination of the item-by-country interactions shows that students in the countries that had high average scores found this item more difficult than expected. Notably, this was the case in Japan, Korea, Singapore and Hong Kong, while the item was easier than expected in Slovenia, Hungary, and the Czech Republic.

Table 7.3 Population 1 Science: Item Statistics and Parameter Estimates for the International Calibration Sample

Item Label	Number of Respondents in International Calibration	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
ASMSA06	14554	84.15	-1.546	0.024	0.97
ASMSA07	14516	82.86	-1.439	0.023	0.96
ASMSA08	14501	62.02	-0.197	0.018	1.02
ASMSA09	14478	66.13	-0.404	0.019	1.00
ASMSA10	14454	74.32	-0.856	0.020	0.96
ASMSB01	7312	70.69	-0.658	0.028	0.99
ASMSB02	7298	72.06	-0.734	0.028	0.96
ASMSB03	6987	68.33	-0.524	0.028	0.98
ASMSB04	6975	57.49	0.029	0.026	1.08
ASMSC05	5429	68.06	-0.494	0.031	1.01
ASMSC06	5420	91.09	-2.248	0.049	0.97
ASMSC07	5410	43.51	0.694	0.030	1.05
ASMSC08	5387	81.60	-1.322	0.037	0.97
ASMSC09	5380	48.83	0.448	0.029	1.06
ASMSD01	5282	58.54	-0.026	0.030	0.96
ASMSD02	5483	68.19	-0.512	0.031	1.01
ASMSD03	5479	88.37	-1.944	0.044	0.95
ASMSD04	5474	72.84	-0.770	0.033	1.02
ASMSE05	5411	61.50	-0.193	0.030	0.99
ASMSE06	5401	92.33	-2.467	0.053	0.98
ASMSE07	5399	57.47	0.004	0.030	0.96
ASSSE08	5364	73.60	-0.832	0.033	0.98
ASMSE09	5349	43.30	0.678	0.030	1.07
ASMSF01	5507	84.60	-1.610	0.039	0.99
ASMSF02	5495	65.13	-0.375	0.031	1.06
ASMSF03	5491	61.28	-0.181	0.030	0.97
ASMSF04	5481	68.87	-0.569	0.031	0.93
ASMSG05	5356	79.41	-1.191	0.036	0.98
ASMSG06	5354	86.33	-1.743	0.042	1.03
ASMSG07	5346	70.61	-0.649	0.032	0.99
ASMSG08	5338	26.60	1.548	0.033	1.14
ASMSG09	5323	86.02	-1.709	0.042	0.94
ASMSH01	5460	77.34	-1.059	0.034	1.03
ASMSH02	5449	83.56	-1.506	0.039	0.93
ASSSH03	5438	39.48	0.870	0.030	0.96
ASMSH04	5434	42.69	0.717	0.030	1.21
ASMSN01	1862	38.56	0.933	0.051	1.03
ASMSN02	1860	69.68	-0.581	0.054	0.94
ASMSN03	1858	40.90	0.820	0.051	1.10
ASMSN04	1855	32.40	1.246	0.053	1.10
ASMSN05	1850	70.81	-0.642	0.055	1.05
ASMSN06	1849	40.02	0.866	0.051	1.01
ASMSN07	1842	51.09	0.346	0.050	1.03
ASMSN08	1838	58.60	-0.008	0.051	1.05
ASMSN09	1832	59.55	-0.052	0.051	0.99
ASMSO01	1829	42.81	0.676	0.051	1.03
ASMSO02	1825	38.03	0.912	0.052	1.07

Table 7.3 Population 1 Science: Item Statistics and Parameter Estimates for the International Calibration Sample (Continued 1)

Item Label	Number of Respondents in International Calibration	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
ASMSO03	1823	62.97	-0.281	0.052	0.96
ASMSO04	1820	66.76	-0.473	0.054	1.04
ASMSO05	1815	54.16	0.154	0.051	1.15
ASSSO06	1808	54.31	0.153	0.051	0.94
ASMSO07	1808	71.35	-0.708	0.056	1.08
ASMSO08	1804	51.39	0.297	0.051	1.04
ASSSO09	1783	38.70	0.909	0.052	0.93
ASMSP01	1780	83.99	-1.472	0.068	0.92
ASMSP02	1780	84.16	-1.480	0.068	0.89
ASMSP03	1777	32.86	1.254	0.054	1.07
ASSSP04	1773	31.64	1.323	0.055	1.01
ASMSP05	1768	45.76	0.630	0.052	1.00
ASMSP06	1698	33.69	1.214	0.055	1.01
ASMSP07	1765	39.60	0.929	0.052	0.99
ASMSP08	1763	53.66	0.269	0.052	0.95
ASMSP09	1759	42.30	0.807	0.052	1.02
ASMSQ01	1850	60.65	-0.071	0.051	0.95
ASMSQ02	1845	63.58	-0.211	0.052	1.07
ASMSQ03	1841	49.48	0.456	0.050	1.00
ASSSQ04	1834	58.34	0.044	0.051	0.92
ASMSQ05	1824	69.19	-0.494	0.054	1.02
ASMSQ06	1822	41.93	0.812	0.051	1.06
ASMSQ07	1819	40.74	0.870	0.051	1.03
ASSSQ08	1808	46.13	0.617	0.051	0.97
ASMSQ09	1795	52.70	0.318	0.051	1.00
ASSSR01	1830	18.80	2.106	0.063	1.03
ASMSR02	1814	39.14	0.944	0.052	1.04
ASMSR03	1805	55.62	0.173	0.051	1.15
ASMSR04	1797	73.46	-0.735	0.057	0.89
ASMSR05	1790	56.09	0.149	0.051	0.99
ASMSR06	1776	39.92	0.908	0.052	1.07
ASMSR07	1765	55.30	0.190	0.051	0.96
ASMSR08	1670	53.77	0.242	0.053	1.09
ASMSR09	1734	44.87	0.678	0.052	1.07
ASESW01	3512	55.25	0.178	0.026	1.09
ASSSW02	3431	38.41	0.989	0.038	0.95
ASSSW03	3218	26.51	1.658	0.043	1.00
ASSSW04	3143	50.81	0.458	0.038	0.88
ASESW05	2900	52.14	0.440	0.040	0.95
ASESW05	2747	36.77	1.188	0.042	0.95
ASESX01	3581	66.23	-0.721	0.031	0.90
ASSSX02	3557	73.35	-0.787	0.041	0.92
ASESX03	3471	42.64	0.736	0.025	0.97
ASSSX04	3397	82.31	-1.344	0.048	0.93
ASMSX05	3323	59.43	-0.009	0.038	1.07
ASESY01	3399	27.83	1.519	0.041	0.91
ASESY02	3258	66.73	-0.353	0.040	0.96

Table 7.3 Population 1 Science: Item Statistics and Parameter Estimates for the International Calibration Sample (Continued 2)

Item Label	Number of Respondents in International Calibration	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
ASESY02	3126	39.38	0.985	0.039	1.01
ASESY03	3021	65.54	-0.231	0.041	0.94
ASESY03	2884	45.67	0.725	0.040	0.95
ASESZ01	3479	58.47	0.026	0.037	0.94
ASESZ01	3406	20.35	1.996	0.045	1.02
ASESZ02	3390	63.54	-0.203	0.038	0.94
ASESZ03	3361	49.02	0.485	0.024	0.99

Figure 7.4 Empirical and Modelled Item Characteristic Curves for Science Population 1 Item: ASMSG08. Fit MNSQ=1.14

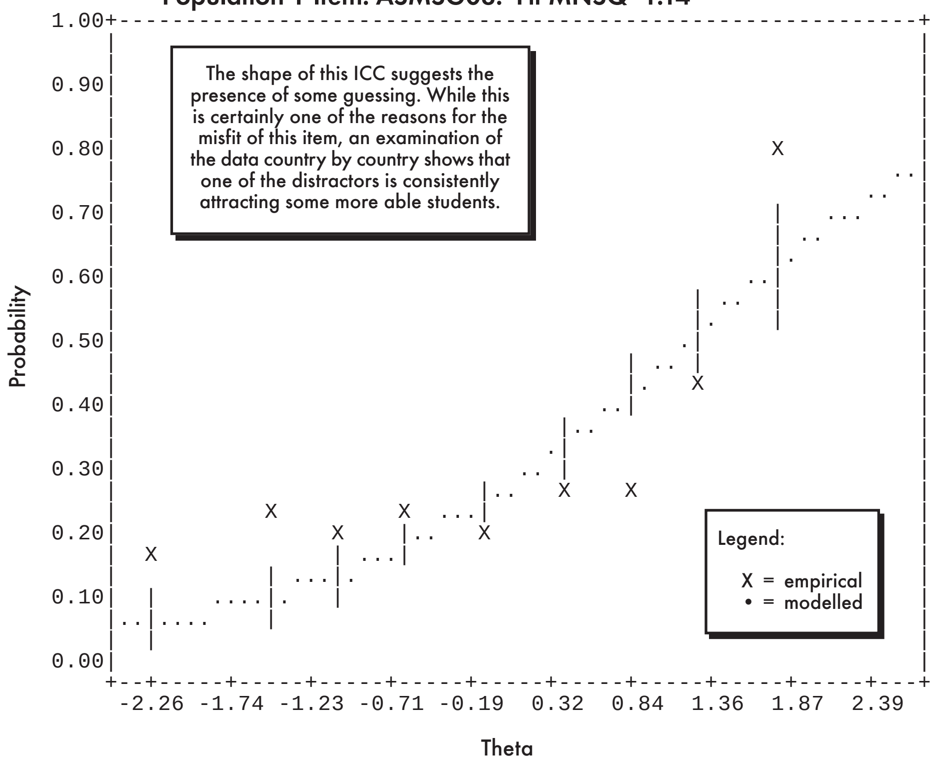


Figure 7.5 Empirical and Modelled Item Characteristic Curves for Science Population 1 Item: ASMSO05. Fit MNSQ=1.15

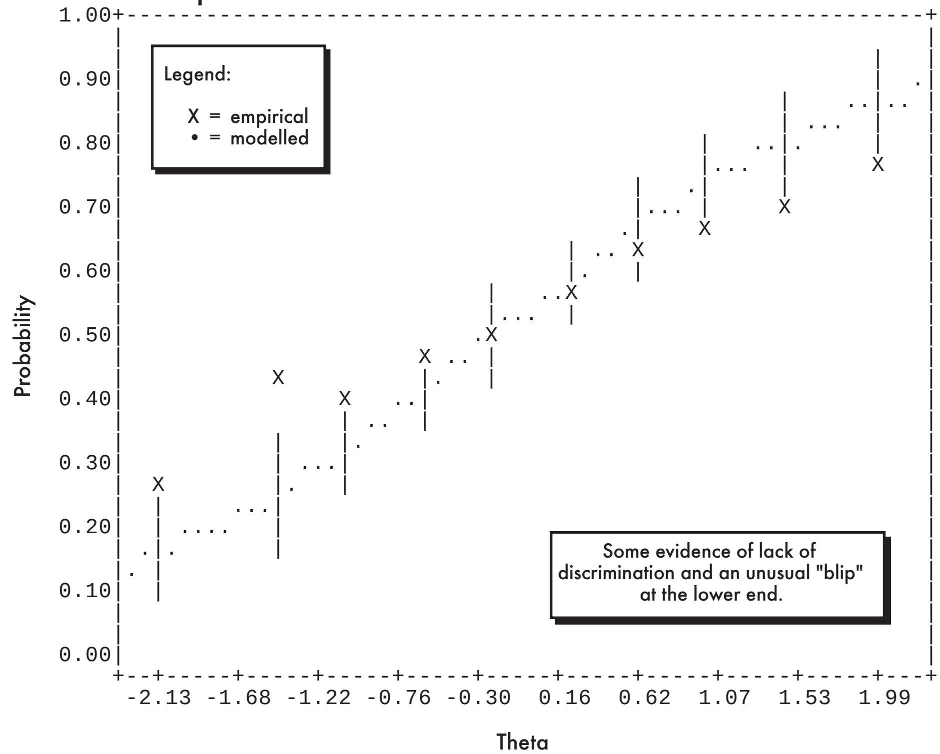
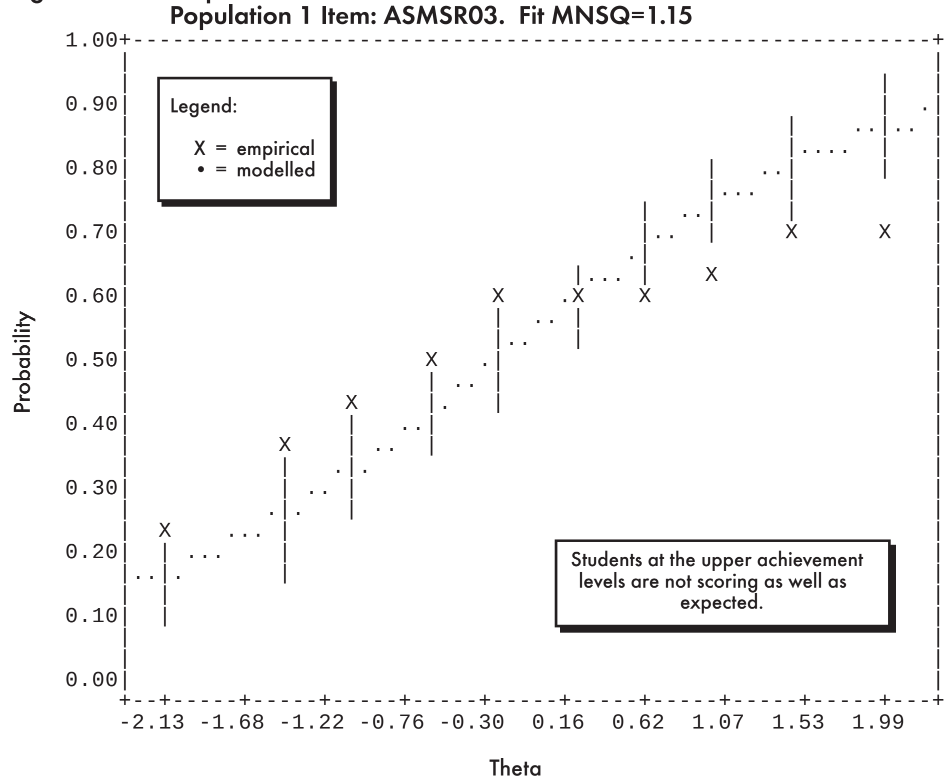


Figure 7.6 Empirical and Modelled Item Characteristic Curves for Mathematics Population 1 Item: ASMSR03. Fit MNSQ=1.15



The fit of the items for Population 2 mathematics is quite acceptable, although it is the least favorable of the four data sets. There are eight items with fit that is less than or equal to 0.85 – six of them short-answer or extended-response – and ten items with a fit greater than 1.15 – nine of them multiple-choice. For the items that have weighted fit mean squares greater than 1.15 the reason for that misfit is quite varied. Items I03, J18, L11, and N17 are all relatively difficult multiple-choice questions and exhibit evidence of guessing. As with the questions that showed elements of guessing characteristics in the Population 1 data sets, each of these items has a distracter that had a positive point-biserial in a large number of countries. Items L11 and N17 also showed bad fit in a number of countries. Item N16 is the only item with fit above 1.15, where it is reasonably clear that the misfit is due to the item having lower than modeled discrimination. The misfit for items B07, D10, and N15, which cannot be easily characterized, is illustrated in Figure 7.8, which shows the observed and expected item characteristic curves for item D10. Examining this item at the country level we note that in a number of countries it has a distracter with a positive point-biserial. This distracter has probably attracted some of the more able students, resulting in the empirical item characteristic curve being lower than the modeled curve for students toward the upper end of the achievement distribution. Plots for items B07, N15, and P09 show a similar pattern, but in examining the data we have not been able to find an explanation for the unusual shape of the observed item characteristic curve.

Table 7.4 Population 2 Mathematics: Item Statistics and Parameter Estimates for the International Calibration Sample

Item Label	Number of Respondents in International Calibration	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
BSMMA01	23039	56.91	-0.127	0.015	0.87
BSMMA02	23036	74.91	-1.120	0.017	1.02
BSMMA03	22437	61.57	-0.359	0.015	0.97
BSMMA04	23033	53.16	0.062	0.015	1.03
BSMMA05	23032	58.68	-0.217	0.015	1.07
BSMMA06	23026	76.54	-1.225	0.017	1.05
BSMMB07	11400	62.71	-0.421	0.022	1.01
BSMMB08	11389	68.82	-0.754	0.022	1.16
BSMMB09	11383	57.46	-0.148	0.021	1.08
BSMMB10	11377	49.42	0.258	0.021	0.85
BSMMB11	11372	63.32	-0.451	0.022	0.95
BSMMB12	11366	64.17	-0.496	0.022	0.91
BSMMC01	8671	57.57	-0.158	0.024	0.96
BSMMC02	8670	76.21	-1.196	0.027	0.95
BSMMC03	8667	60.61	-0.313	0.024	0.94
BSMMC04	8661	54.28	0.009	0.024	1.14
BSMMC05	8660	52.81	0.082	0.024	1.02
BSMMC06	8654	71.13	-0.883	0.026	1.03
BSMMD07	8773	60.97	-0.345	0.024	1.02
BSMMD08	8761	66.00	-0.610	0.025	1.02
BSMMD09	8756	66.71	-0.649	0.025	0.86
BSMMD10	8753	44.27	0.499	0.024	1.16
BSMMD11	8743	85.38	-1.906	0.032	1.02
BSMMD12	8740	72.15	-0.957	0.026	1.04
BSMME01	8715	69.15	-0.791	0.026	1.00
BSMME02	8710	44.32	0.492	0.024	0.99
BSMME03	8705	56.01	-0.096	0.024	1.00
BSMME04	8698	65.56	-0.590	0.025	0.95
BSMME05	8687	53.02	0.056	0.024	0.97
BSMME06	8679	40.45	0.693	0.024	0.95
BSMMF07	8615	30.47	1.243	0.026	1.03
BSMMF08	8606	59.57	-0.253	0.024	1.15
BSMMF09	8602	65.21	-0.546	0.025	0.99
BSMMF10	8596	53.52	0.052	0.024	1.01
BSMMF11	8398	45.83	0.444	0.024	0.89
BSMMF12	8583	54.44	0.007	0.024	1.07
BSMMG01	8638	52.92	0.072	0.024	1.15
BSMMG02	8633	75.98	-1.192	0.027	0.98
BSMMG03	8631	49.70	0.234	0.024	1.02
BSMMG04	8629	67.44	-0.682	0.025	0.95
BSMMG05	8622	58.37	-0.202	0.024	0.94
BSMMG06	8621	40.82	0.686	0.025	1.01
BSMMH07	8581	66.37	-0.613	0.025	1.04
BSMMH08	8575	73.84	-1.046	0.027	0.92
BSMMH09	8570	84.75	-1.837	0.032	0.96
BSMMH10	8564	43.57	0.559	0.024	0.97
BSMMH11	8549	60.51	-0.297	0.025	0.97

Table 7.4 Population 2 Mathematics: Item Statistics and Parameter Estimates for the International Calibration Sample (Continued 1)

Item Label	Number of Respondents in International Calibration	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
BSMMH12	8536	73.11	-0.996	0.027	0.94
BSMMI01	2884	34.71	1.008	0.044	1.09
BSMMI02	2883	55.64	-0.070	0.042	0.94
BSMMI03	2882	39.73	0.738	0.043	1.16
BSSMI04	2880	42.43	0.597	0.042	1.03
BSMMI05	2877	72.54	-0.981	0.046	0.95
BSSMI06	2876	76.29	-1.217	0.048	1.09
BSMMI07	2877	63.30	-0.464	0.043	1.13
BSMMI08	2871	41.14	0.666	0.043	1.12
BSMMI09	2872	64.28	-0.516	0.043	0.94
BSMMJ10	2937	40.86	0.661	0.042	0.87
BSMMJ11	2865	45.62	0.405	0.042	1.08
BSSMJ12	2929	41.62	0.624	0.042	1.03
BSSMJ13	2928	81.69	-1.576	0.051	0.99
BSMMJ14	2930	43.38	0.536	0.041	1.02
BSMMJ15	2926	63.91	-0.486	0.042	1.07
BSMMJ16	2924	51.74	0.126	0.041	0.98
BSMMJ17	2916	65.84	-0.585	0.043	1.01
BSMMJ18	2913	41.57	0.631	0.042	1.18
BSMMK01	2958	68.80	-0.804	0.044	1.05
BSSMK02	2958	64.16	-0.549	0.042	0.98
BSMMK03	2956	65.93	-0.644	0.043	1.08
BSMMK04	2957	38.35	0.772	0.042	1.12
BSSMK05	2956	36.87	0.851	0.043	0.79
BSMMK06	2956	38.94	0.741	0.042	1.08
BSMMK07	2955	50.59	0.147	0.041	1.03
BSMMK08	2955	31.78	1.134	0.044	1.05
BSMMK09	2952	47.63	0.297	0.041	0.90
BSMML08	2857	57.75	-0.168	0.042	1.10
BSMML09	2857	84.35	-1.805	0.055	0.98
BSMML10	2855	86.76	-2.024	0.059	1.00
BSMML11	2854	32.20	1.146	0.044	1.24
BSMML12	2855	72.71	-0.976	0.046	0.93
BSMML13	2854	89.66	-2.332	0.065	0.98
BSMML14	2852	22.72	1.735	0.049	1.08
BSMML15	2845	35.75	0.959	0.044	1.02
BSSML16	2844	37.48	0.868	0.043	0.93
BSMML17	2745	47.14	0.379	0.043	0.92
BSMMM01	2832	85.56	-1.887	0.057	0.95
BSMMM02	2831	62.91	-0.410	0.043	1.13
BSMMM03	2830	75.69	-1.142	0.048	0.97
BSMMM04	2830	37.77	0.868	0.043	0.87
BSMMM05	2768	48.48	0.316	0.042	1.07
BSSMM06	2828	33.80	1.084	0.044	0.90
BSMMM07	2827	71.45	-0.878	0.046	1.01
BSSMM08	2827	46.44	0.427	0.042	1.10
BSMMN11	2831	82.20	-1.600	0.053	0.94

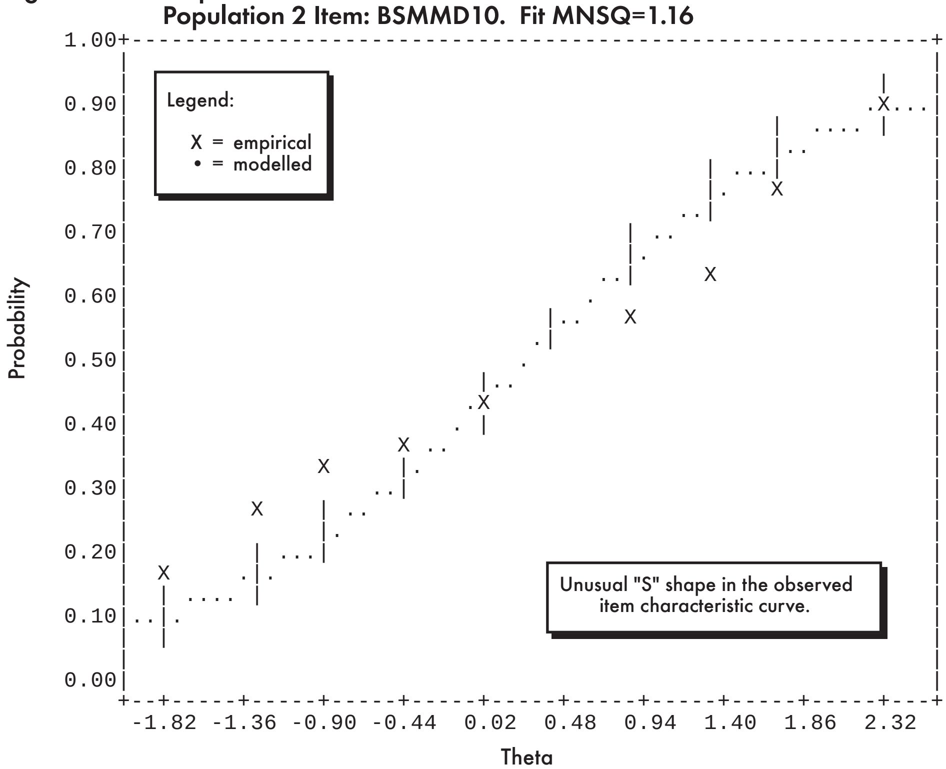
Table 7.4 Population 2 Mathematics: Item Statistics and Parameter Estimates for the International Calibration Sample (Continued 2)

Item Label	Number of Respondents in International Calibration	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
BSMMN12	2829	65.04	-0.522	0.044	1.10
BSSMN13	2825	46.27	0.439	0.042	0.90
BSMMN14	2821	66.43	-0.593	0.044	0.96
BSMMN15	2822	64.56	-0.492	0.044	1.19
BSMMN16	2746	45.59	0.482	0.043	1.19
BSMMN17	2728	38.56	0.819	0.044	1.22
BSMMN18	2799	54.31	0.048	0.042	0.99
BSSMN19	2782	50.32	0.253	0.042	0.79
BSMMO01	2881	55.22	-0.015	0.042	0.98
BSMMO02	2879	25.81	1.589	0.047	0.92
BSMMO03	2881	45.33	0.489	0.042	0.99
BSMMO04	2808	44.16	0.555	0.043	1.08
BSMMO05	2881	43.91	0.562	0.042	0.85
BSSMO06	2879	69.78	-0.793	0.045	0.95
BSMMO07	2879	68.04	-0.694	0.044	1.01
BSMMO08	2877	66.28	-0.595	0.044	1.02
BSSMO09	2876	47.46	0.383	0.042	0.84
BSMMP08	2765	54.94	-0.005	0.043	0.98
BSMMP09	2764	37.34	0.895	0.044	1.16
BSMMP10	2757	54.12	0.037	0.043	0.96
BSMMP11	2754	53.96	0.044	0.043	1.15
BSMMP12	2752	70.17	-0.809	0.046	1.00
BSMMP13	2740	67.12	-0.636	0.045	0.95
BSMMP14	2736	77.60	-1.262	0.050	0.99
BSMMP15	2730	62.78	-0.401	0.044	0.98
BSSMP16	2720	33.57	1.110	0.045	0.90
BSMMP17	2674	84.89	-1.798	0.057	1.06
BSMMQ01	2784	41.81	0.651	0.043	1.07
BSMMQ02	2778	47.70	0.353	0.043	1.09
BSMMQ03	2776	33.43	1.104	0.045	1.05
BSMMQ04	2774	84.35	-1.777	0.056	1.01
BSMMQ05	2770	65.42	-0.549	0.044	1.03
BSMMQ06	2762	38.88	0.810	0.044	1.01
BSMMQ07	2752	58.76	-0.199	0.043	0.96
BSMMQ08	2746	43.81	0.556	0.043	0.86
BSMMQ09	2683	50.09	0.238	0.043	0.99
BSSMQ10	2728	43.80	0.560	0.043	1.00
BSMMR06	2786	74.80	-1.098	0.047	1.01
BSMMR07	2785	44.34	0.516	0.043	0.99
BSMMR08	2784	48.38	0.312	0.042	1.10
BSMMR09	2783	40.14	0.734	0.043	0.94
BSMMR10	2783	51.49	0.155	0.042	1.02
BSMMR11	2783	44.05	0.529	0.043	1.04
BSMMR12	2781	86.19	-1.954	0.058	0.95
BSSMR13	2779	32.13	1.167	0.045	0.91
BSSMR14	2778	37.08	0.892	0.044	0.82
BSEMS01	2829	77.48	-1.273	0.049	1.05

Table 7.4 Population 2 Mathematics: Item Statistics and Parameter Estimates for the International Calibration Sample (Continued 3)

Item Label	Number of Respondents in International Calibration	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
BSEMS01	2739	23.80	1.722	0.050	0.97
BSEMS02	2534	65.04	-0.421	0.046	0.92
BSEMS02	2382	30.31	1.420	0.050	0.82
BSEMS02	2171	28.33	1.590	0.053	0.90
BSEMT01	5661	31.90	0.915	0.019	1.01
BSEMT01	5053	35.84	1.047	0.033	0.79
BSEMT02	4998	23.41	1.799	0.037	0.88
BSEMT02	4221	9.90	3.076	0.055	1.00
BSEMU01	5585	34.45	1.045	0.031	0.96
BSEMU01	5330	33.66	1.132	0.032	0.99
BSEMU02	5009	37.10	0.778	0.020	1.13
BSEMU02	4671	20.65	1.745	0.026	0.95
BSSMV01	5477	52.67	0.110	0.030	0.91
BSEMV02	5582	27.11	1.113	0.017	1.17
BSMMV03	5538	40.75	0.732	0.031	0.95
BSSMV04	5512	39.26	0.813	0.031	0.90

Figure 7.7 Empirical and Modelled Item Characteristic Curves for Mathematics Population 2 Item: BSMMD10. Fit MNSQ=1.16



The compatibility of the model and data for Population 2 science is better than for any of the other three data sets. There is just one item, item Q18, with a weighted mean square greater than 1.15, and there are no items with weighted mean squares as low as 0.85. Figure 7.8 is a plot of the observed and expected item characteristic curves for Q18. The plot shows evidence of guessing. Examining the behavior at the country level again reveals that there is a distracter that is positive in many countries.

As a set, the data appear to be quite compatible with the assumed Rasch scaling model. Certainly the extent of deviation from the model will have had no influence on the substantive outcomes of the study. A few isolated items that were retained in the scaling did not fit the model. The source of this misfit can generally be traced to multiple-choice item distracters that were attractive to some more able students.

Table 7.5 Population 2 Science: Item Statistics and Parameter Estimates for the International Calibration Sample

Item Label	Number of Respondents in International Calibration	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
BSMSA07	23016	67.11	-0.562	0.015	1.02
BSMSA08	23024	65.54	-0.484	0.015	1.04
BSMSA09	23015	76.66	-1.089	0.016	0.93
BSMSA10	22998	68.38	-0.626	0.015	1.03
BSMSA11	22982	57.92	-0.119	0.014	0.99
BSMSA12	22968	58.51	-0.146	0.014	0.98
BSMSB01	11410	86.92	-1.845	0.029	0.98
BSMSB02	11406	53.40	0.095	0.020	1.08
BSMSB03	11407	26.47	1.394	0.022	1.00
BSMSB04	11404	88.48	-1.998	0.030	0.95
BSMSB05	11362	48.88	0.299	0.020	1.10
BSMSB06	11402	83.44	-1.547	0.026	1.01
BSMSC07	8642	37.75	0.816	0.024	1.01
BSMSC08	8637	71.63	-0.786	0.025	0.98
BSMSC09	8633	72.95	-0.859	0.025	1.02
BSMSC10	8636	77.08	-1.102	0.027	1.03
BSMSC11	8624	45.26	0.468	0.023	0.98
BSMSC12	8618	52.70	0.131	0.023	1.09
BSMSD01	8794	40.22	0.674	0.023	1.02
BSMSD02	8787	73.39	-0.914	0.025	0.94
BSMSD03	8787	36.95	0.830	0.023	0.98
BSMSD04	8779	54.99	-0.001	0.023	1.02
BSMSD05	8769	66.27	-0.535	0.024	0.97
BSMSD06	8770	72.75	-0.877	0.025	0.97
BSMSE07	8669	41.38	0.634	0.023	1.09
BSMSE08	8666	79.17	-1.247	0.028	1.01
BSMSE09	8662	77.80	-1.158	0.027	0.96
BSMSE10	8651	53.59	0.080	0.023	0.99
BSMSE11	8642	57.30	-0.089	0.023	1.04
BSMSE12	8396	53.93	0.074	0.023	1.06
BSMSF01	8637	66.61	-0.549	0.024	0.98
BSMSF02	8634	63.19	-0.380	0.024	0.91
BSMSF03	8392	66.17	-0.515	0.024	1.08
BSMSF04	8399	68.33	-0.632	0.025	0.96
BSMSF05	8631	80.44	-1.349	0.028	0.97
BSMSF06	8630	68.81	-0.661	0.025	0.95
BSMSG07	8619	86.99	-1.856	0.033	0.98
BSMSG08	8619	59.38	-0.189	0.023	1.00
BSMSG09	8614	74.32	-0.948	0.026	1.00
BSMSG10	8612	51.64	0.166	0.023	0.98
BSMSG11	8605	49.56	0.260	0.023	1.05
BSMSG12	8596	51.14	0.189	0.023	1.06
BSMSH01	8264	69.26	-0.648	0.025	0.99
BSMSH02	8496	79.26	-1.240	0.028	1.01
BSMSH03	8494	79.15	-1.233	0.028	0.96
BSMSH04	8587	50.73	0.216	0.023	1.04
BSMSH05	8586	22.99	1.603	0.027	1.02

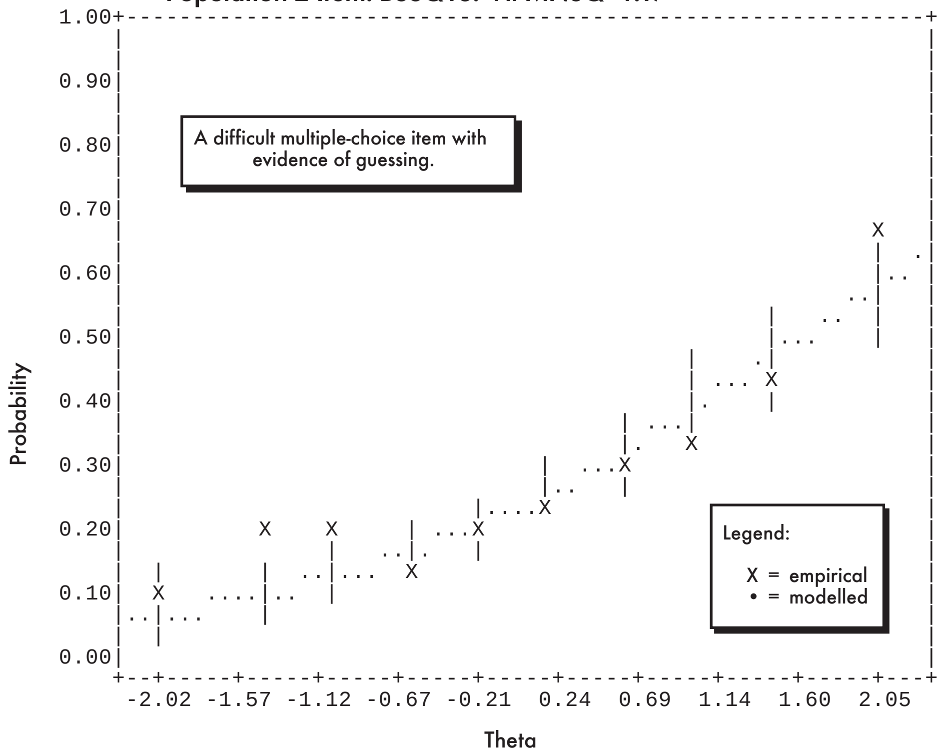
Table 7.5 Population 2 Science: Item Statistics and Parameter Estimates for the International Calibration Sample (Continued 1)

Item Label	Number of Respondents in International Calibration	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
BSMSH06	8583	51.40	0.186	0.023	0.96
BSMSI10	2871	74.54	-0.921	0.045	1.01
BSMSI11	2869	45.59	0.477	0.040	0.98
BSMSI12	2866	35.35	0.957	0.041	0.96
BSMSI13	2795	60.86	-0.217	0.041	0.94
BSMSI14	2862	54.72	0.066	0.040	1.05
BSMSI15	2859	49.11	0.320	0.040	0.97
BSMSI16	2854	85.00	-1.632	0.054	0.99
BSMSI17	2851	40.72	0.704	0.040	1.06
BSSSI18	2847	35.30	0.964	0.041	0.96
BSMSI19	2759	51.40	0.223	0.040	0.91
BSMSJ01	2784	39.22	0.752	0.041	1.04
BSMSJ02	2942	62.71	-0.369	0.040	0.97
BSSSJ03	2941	27.13	1.337	0.044	0.97
BSMSJ04	2941	41.11	0.629	0.040	1.00
BSMSJ05	2940	64.35	-0.447	0.041	0.98
BSMSJ06	2940	22.69	1.603	0.046	1.00
BSMSJ07	2873	47.16	0.362	0.040	1.00
BSMSJ08	2936	45.16	0.442	0.040	0.98
BSSSJ09	2935	76.12	-1.079	0.045	0.94
BSSSK10	2876	34.14	0.947	0.042	1.08
BSMSK11	2946	55.02	-0.012	0.039	0.96
BSMSK12	2945	51.85	0.132	0.039	0.99
BSMSK13	2944	74.01	-0.953	0.044	0.93
BSMSK14	2942	79.74	-1.306	0.048	0.94
BSMSK15	2940	59.35	-0.210	0.040	0.99
BSMSK16	2938	35.94	0.867	0.041	0.99
BSMSK17	2936	51.63	0.143	0.039	1.01
BSMSK18	2925	54.22	0.029	0.039	1.02
BSSSK19	2907	72.38	-0.852	0.044	0.97
BSMSL01	2859	47.22	0.365	0.040	1.00
BSMSL02	2858	51.68	0.163	0.040	0.99
BSMSL03	2859	68.42	-0.631	0.043	0.95
BSESL04	2858	32.96	1.041	0.042	0.94
BSMSL05	2857	64.30	-0.425	0.041	1.04
BSMSL06	2856	52.21	0.139	0.040	1.00
BSMSL07	2857	68.15	-0.618	0.042	0.96
BSMSM10	2822	46.03	0.425	0.040	1.01
BSESM11	2821	65.65	-0.483	0.042	0.92
BSSSM12	2817	52.36	0.141	0.040	0.92
BSMSM13	2813	46.82	0.391	0.040	0.96
BSSSM14	2810	71.14	-0.764	0.044	1.02
BSMSN01	2839	43.68	0.550	0.040	1.08
BSMSN02	2837	39.76	0.732	0.041	1.04
BSMSN03	2837	60.35	-0.207	0.041	0.98
BSMSN04	2834	50.53	0.241	0.040	1.07
BSMSN05	2688	31.29	1.161	0.044	1.08

Table 7.5 Population 2 Science: Item Statistics and Parameter Estimates for the International Calibration Sample (Continued 2)

Item Label	Number of Respondents in International Calibration	Percentage of Correct Responses	Difficulty Estimate in Logit Metric	Asymptotic Standard Error in Logit Metric	Mean Square Fit Statistic
BSMSN06	2833	64.03	-0.382	0.041	0.98
BSSSN07	2833	88.92	-2.017	0.061	0.91
BSMSN08	2834	70.82	-0.727	0.043	1.00
BSMSN09	2774	47.08	0.399	0.040	0.95
BSSSN10	2830	53.11	0.123	0.040	0.96
BSSSO10	2874	53.58	0.087	0.040	0.96
BSMSO11	2797	36.11	0.905	0.042	1.02
BSMSO12	2812	24.64	1.516	0.046	0.98
BSMSO13	2872	58.67	-0.147	0.040	1.01
BSESO14	2870	54.36	0.052	0.040	0.91
BSMSO15	2862	38.29	0.791	0.041	1.01
BSSSO16	2862	57.97	-0.112	0.040	0.95
BSSSO17	2803	55.26	0.025	0.040	1.04
BSMSP01	2780	82.77	-1.501	0.052	0.92
BSSSP02	2776	21.65	1.669	0.048	0.95
BSSSP03	2777	79.40	-1.264	0.049	0.91
BSMSP04	2774	54.25	0.047	0.040	0.98
BSSSP05	2770	55.96	-0.031	0.041	0.99
BSSSP06	2764	48.20	0.304	0.025	0.98
BSMSP07	2766	52.39	0.132	0.040	1.06
BSMSQ11	2713	42.87	0.569	0.041	1.03
BSSSQ12	2709	45.99	0.427	0.041	0.99
BSMSQ13	2703	59.67	-0.193	0.042	0.98
BSMSQ14	2678	45.29	0.463	0.041	1.07
BSMSQ15	2612	32.50	1.080	0.044	1.03
BSMSQ16	2597	25.53	1.444	0.047	1.05
BSSSQ17	2567	65.41	-0.460	0.044	0.99
BSSSQ18	2433	36.33	0.728	0.026	1.17
BSMSR01	2786	68.88	-0.654	0.043	1.03
BSMSR02	2786	38.30	0.771	0.041	1.05
BSESR03	2784	29.76	1.160	0.030	0.94
BSSSR04	2784	50.11	0.231	0.040	0.91
BSSSR05	2783	49.08	0.277	0.040	0.88
BSESW01	5652	80.08	-1.306	0.035	1.00
BSESW01	5408	42.77	0.607	0.029	1.05
BSESW02	5176	43.51	0.517	0.018	1.05
BSESX01	5690	22.12	1.398	0.022	0.98
BSESX02	5123	68.79	-0.609	0.032	1.03
BSESX02	4923	34.51	1.019	0.032	0.99
BSESY01	5562	6.53	3.162	0.055	0.94
BSESY02	5291	28.34	1.208	0.021	1.13
BSESZ01	2725	62.35	-0.287	0.042	0.99
BSESZ01	2449	42.63	0.642	0.043	0.90
BSESZ01	2335	28.09	1.379	0.048	0.91
BSESZ02	2341	75.22	-0.895	0.050	1.02
BSESZ02	2260	50.18	0.344	0.045	1.00

Figure 7.8 Empirical and Modelled Item Characteristic Curves for Science Population 2 Item: BSSQ18. Fit MNSQ=1.17



7.4.4 Reliability

Table 7.6 reports a variety of reliability indices for the four tests. The median Cronbach Alpha coefficients were computed by calculating the Cronbach Alpha coefficient for each test booklet within each country. The median of these values was then used as a reliability index for each country. The median of those country medians is reported in Table 7.6.

The separation reliability for the international calibration sample was computed by fitting the scaling model without the use of any conditioning variables, drawing five plausible values for each student, and then computing the median of the ten correlations between pairs of plausible values. In general, these statistics show that the science tests are slightly less reliable than the mathematics tests and that the Population 1 tests are slightly less reliable than the Population 2 tests.

Table 7.6 Unconditional Reliabilities

TIMSS Test	Median of Lower Grade National Cronbach Alpha Coefficients	Median of Upper Grade National Cronbach Alpha Coefficients	Separation Reliability in the International Calibration Sample
Mathematics Population 1	0.82	0.84	0.83
Science Population 1	0.78	0.77	0.77
Mathematics Population 2	0.86	0.89	0.89
Science Population 2	0.77	0.78	0.80

7.4.5 The Population Model for Population 2

For Population 2 it was considered expedient to proceed with a scaling that did not make extensive use of conditioning. There were two reasons for this. First, the reliability of the Population 2 data was relatively high so that the possible effect of conditioning would be ignorable. Second, the background data were not fully cleaned and checked at the time of processing, and extensive conditioning would have delayed publication of the international reports.

For each participating country the scaling was undertaken with all item parameters set at the values obtained from fitting the model to the international calibration sample. In the population model sampling weights were used, and student grade was used as a conditioning variable. Five plausible values were drawn and an EAP estimate of achievement was obtained for each student. As illustrated in Table 7.7, conditioning on grade led to little improvement in the person separation reliability. This conditioning was, however, necessary to ensure that consistent results were obtained when plausible values were used to estimate characteristics of the achievement distributions for the upper and lower grades separately.

Table 7.7 Population 2 Reliabilities For Three Countries With and Without Conditioning on Grade

Country	Mathematics		Science	
	Conditioning on Grade	No Conditioning	Conditioning on Grade	No Conditioning
Australia	0.88	0.88	0.80	0.80
Cyprus	0.86	0.86	0.78	0.77
Hong Kong	0.87	0.87	0.76	0.76

7.4.6 The Population Model for Population 1

In Population 1 conditioning was used much more extensively. For both mathematics and science the variables sex, grade, and the interaction between sex and grade were used as conditioning variables.[1] Additionally, for mathematics the mean of the mathematics score variable ASMRAWST was computed for each class, assigned to each student in that class, and then used as a conditioning variable. This variable was called ASMRAWAV. Similarly, for science the mean of the mathematics score variable ASSRAWST was computed for each class, assigned to each student in that class, and then used as a conditioning variable. This variable was called ASSRAWAV. This conditioning was undertaken so as to improve the estimation of between-class and between-school variance components that would be obtained from secondary analyses using plausible values. Each individual student's science score ASSRAWST was also used as a conditioning variable for mathematics, and in the case of science, each individual student's mathematics score ASMRAWST was used.

[1] The gender variable ASBGSEX is trichotomous (male, female, missing). When used in conditioning, this variable was replaced with two dummy coded variables.

Table 7.8 Number of Principal Components Retained In Conditioning - Population 1

Country	Number of Retained Principal Components
Australia	62
Austria	85
Canada	84
Cyprus	69
Czech Republic	84
England	51
Greece	78
Hong Kong	81
Hungary	86
Iceland	69
Iran	83
Ireland	83
Israel	62
Japan	59
Korea	78
Kuwait	88
Latvia	74
Mexico	103
Netherlands	83
New Zealand	87
Norway	75
Portugal	91
Scotland	46
Singapore	106
Slovenia	77
Thailand	73
United States	72

For both mathematics and science, the pool of over 100 student-level background variables was also represented in the conditioning. For each student-level variable a set of dummy variables was constructed from the original variables (see Appendix D). This new set of dummy variables retained all of the information in the original set of variables but made them appropriate for use in a principal components analysis. A principal components analysis of the set of dummy variables was then undertaken for each country and as many components retained as explained 90% of the variance. Scores on each of the retained components were then computed for each student. The number of retained components for each country is shown in Table 7.8.

These components, and the products of these components and ASMRAWAV (in the case of mathematics) and ASSRAWAV (in the case of science), were used as conditioning variables. Table 7.9 shows the conditioning variables that were used for mathematics and science. For some countries the total was in excess of 200. Table 7.10 illustrates for three selected countries the increase in reliability that was attained by conditioning, first by grade and then with the full set of conditioning variables.

Table 7.9 Variables Used in Conditioning - Population 1

Variables	Mathematics	Science
Grade	✓	✓
Gender	✓	✓
Gender by grade interaction	✓	✓
Mathematics score	✗	✓
Science score	✓	✗
Class mean mathematics score	✓	✗
Class mean science score	✗	✓
Principal components	✓	✓
Principal component by class mean mathematics score	✓	✗
Principal component by class mean science score	✗	✓

Table 7.10 Variables Used in Conditioning - Population 1

Country	Mathematics			Science		
	No Conditioning	Conditioning on Grade	Full Conditioning	No Conditioning	Conditioning on Grade	Full Conditioning
Australia	0.83	0.84	0.87	0.77	0.78	0.83
Cyprus	0.82	0.83	0.86	0.74	0.75	0.81
Hong Kong	0.78	0.79	0.84	0.73	0.74	0.81

REFERENCES

Adams, R.J. and Gonzalez, E.J. (1996). TIMSS test design. In M.O. Martin and D.L. Kelly (Eds.), *TIMSS technical report, volume I: Design and development.* Chestnut Hill, MA: Boston College.

Adams, R.J., Wilson, M.R., and Wang, W.C. (1997). The multidimensional random coefficients multinomial logit. *Applied Psychological Measurement*, 21, 1-24.

Adams, R.J., Wilson, M.R., and Wu, M.L. (1997). Multilevel item responses models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, 22 (1), 46-75.

Beaton, A.E. (1987). *Implementing the new design: The NAEP 1983-84 technical report.* Report No. 15-TR-20. Princeton, NJ: Educational Testing Service.

Bock, R.D. and Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika*, 46, 443-459.

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

Kelderman, H. and Rijkes, C.P.M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika*, 59, 149-176.

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

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

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

Mislevy, R.J. and Sheehan, K.M. (1987). Marginal estimation procedures. In A.E. Beaton (Ed.), *Implementing the new design: The NAEP 1983-84 technical report.* Report No. 15-TR-20. Princeton, NJ: Educational Testing Service.

Mislevy, R.J. and Sheehan, K.M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika*, 54 (4), 661-679.

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

Volodin, N. and Adams, R.J. (1997). *The estimation of polytomous item response models with many dimensions.* Paper presented at the Annual Meeting of the Psychometric Society, Gatlinburg, TN.

Wilson, M.R. (1992). The ordered partition model: An extension of the partial credit model. *Applied Psychological Measurement*, 16 (3).

Wu, M.L., Adams, R.J., and Wilson, M. (1997). *Conquest: Generalized item response modelling software – Manual.* Melbourne: Australian Council for Educational Research.

Wu, M.L. (1997). *The development and application of a fit test for use with marginal maximum estimation and generalized item response models.* Unpublished Master's dissertation, University of Melbourne.