11

Scaling the PIRLS Reading Assessment Data

Eugenio J. Gonzalez

11.1 Overview

To achieve its goal of broad coverage of the reading purposes and processes specified in the assessment framework,[1] the PIRLS 2001 assessment included a range of reading passages and items arranged into eight 40-minute assessment blocks. Each student participating in the assessment completed one student booklet made up of just two of these blocks, keeping individual student response burden to a minimum. PIRLS used a matrix-sampling design[2] to assign assessment blocks to student booklets so that a comprehensive picture of the reading achievement of fourth-grade students in each country could be assembled from the components completed by individual students. PIRLS relied on Item Response Theory (IRT) scaling to combine the student responses to provide accurate estimates of reading achievement in the student population in each country. The PIRLS IRT scaling also uses multiple imputation or "plausible values" methodology to obtain proficiency scores in reading for all students, even though each student responded to only a part of the assessment item pool.

This chapter first reviews the psychometric models and the multiple imputation or "plausible values" methodology used in scaling the PIRLS 2001 data, and then describes how this approach was applied to the PIRLS 2001 data and to the data from IEA's Trends in Reading

1 The PIRLS 2001 assessment framework is described in Campbell, Kelly, Mullis, Martin, & Sainsbury (2001).

2 The PIRLS 2001 achievement test design is described in Chapter 2.

Literacy Study. The PIRLS scaling was conducted at the PIRLS International Study Center (ISC) at Boston College, with software and psychometric support from Educational Testing Service.[3]

11.2 PIRLS 2001 Scaling Methodology[4]

The scaling approach used by PIRLS was developed originally by Educational Testing Service for use in the U.S. National Assessment of Educational Progress. It is based on psychometric models that were first used in the field of educational measurement in the 1950s, and have become popular since the 1970s for use in large-scale surveys, test construction, and computer adaptive testing.[5] This approach also has been used to scale IEA's TIMSS data.

Three distinct scaling models, depending on item type and scoring procedure, were used in the analysis of the PIRLS 2001 assessment data. Each is a "latent variable" model that describes the probability that a student will respond in a specific way to an item in terms of the respondent's proficiency, which is an unobserved or "latent" trait, and various characteristics (or "parameters") of the item. A three-parameter model was used with multiple-choice items, which were scored as correct or incorrect, and a two-parameter model for those constructed-response items with just two response options – which also were scored as correct or incorrect. Since each of these item types has just two response categories, they are known as dichotomous items. A partial credit model was used with polytomous constructed-response items (i.e., those with more than two score points).

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

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

Equation 1

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

where

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

3 PIRLS is indebted to Matthias von Davier, Ed Kulick, and John Barone of Educational Testing Service for their advice and support.

4 This section describing the PIRLS scaling methodology has been adapted with permission from the *TIMSS 1999 Technical Report* (Yamamoto & Kulick, 2000).

5 For a description of IRT scaling see Birnbaum (1968); Lord and Novick (1968); Lord (1980); Van Der Linden and Hambleton (1996). The theoretical underpinning of the imputed value methodology was developed by Rubin (1987), applied to large-scale assessment by Mislevy (1991), and studied further by Mislevy, Johnson and Muraki (1992), and Beaton and Johnson (1992). The procedures used in PIRLS have been used in several other large-scale surveys, including Trends in International Mathematics and Science Study (TIMSS), 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).

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 the location parameter for the item, characterizing its difficulty;

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

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

Equation 2

$$P_{i0} \equiv P\left(x_i = 0 \middle| \theta_k, a_i, b_i, c_i\right) = 1 - P_{i1}\left(\theta_k\right)$$

The two-parameter (2PL) model was used for the short constructed-response items that were scored as correct or incorrect. The form of the 2PL model is the same as Equation 1, with the c_i parameter fixed at zero.

11.2.2 The IRT Model for Polytomous Items

In PIRLS 2001, constructed-response items requiring an extended response were scored for partial credit (with 0, 1, 2, and 3 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 (see Equation 3), 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 the location parameter of item i, characterizing its difficulty;

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

Equation 3

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

Indeterminacy of model parameters of the polytomous model are resolved by setting $d_{i,0}$ =0, and setting the sum of the threshold parameters equal to 0.

For all of the IRT models there is a linear indeterminacy of the values of item parameters and proficiency parameters (i.e., mathematically equivalent but different values of item parameters can be estimated on an arbitrarily linearly transformed proficiency scale). This linear indeterminacy can be resolved by setting the origin and unit size of the proficiency scale to arbitrary constants, (such as a mean of 500 and a 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_k (a measure of person proficiency) and the specified parameters of the item, and are assumed unaffected by the demographic characteristics or unique experiences of the respondents, the data collection conditions, or the other items presented in the test. Under this assumption, the joint probability of a particular response pattern x across a set of n items is given by:

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

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:

$$U_{il} = \begin{matrix}1 \text{ if response } x_i \text{ is in category}_l \\ 0 \text{ otherwise}\end{matrix}$$

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. 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.

11.2.3 Proficiency Estimation Using Plausible Values

Most cognitive skills testing is concerned with accurately assessing the performance of individual respondents for the purposes of diagnosis, selection, or placement. Regardless of the measurement model used, whether classical test theory or item response theory, the accuracy of these measurements can be improved – that is, the amount of measurement error can be reduced – by increasing the number of items given to the individual. Thus, it is common to see achievement tests designed to provide information on individual students that contain more than 70 items. Since the uncertainty associated with each θ in such tests is negligible, the distribution of θ or the joint distribution of θ with other variables can be approximated using individual θ's.

For the distribution of proficiencies in large populations, however, more efficient estimates can be obtained from a matrix-sampling design like that used in PIRLS 2001. 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 is more efficiently offset by the inability to make precise statements about individuals. The uncertainty associated with individual θ estimates becomes too large to be ignored. In this situation, aggregations of individual student scores can lead to seriously biased estimates of population characteristics (Wingersky, Kaplan, & Beaton, 1987).

Plausible values methodology was developed as a way to address this issue by using all available data to estimate directly the characteristics of student populations and subpopulations, and then generating multiple imputed scores (called plausible values) from these distributions, which can be used in analyses with standard statistical software. A detailed review of plausible values methodology is given by Mislevy (1991).[6]

What follows 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. One 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:

Equation 4

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

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

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 TIMSS, NAEP, NALS, and IALLS. The value of θ for any respondent that would enter into the computation of t is thus replaced by a randomly selected value from his or her conditional distribution. Rubin (1987) proposed repeating this process several times so that the uncertainty associated with imputation can be quantified. 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 Equation 4; the

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

Equation 5

$$P\left(\theta_j \middle| x_j, y_j, \Gamma, \Sigma\right) \propto P\left(x_j \middle| \theta_j, y_j, \Gamma, \Sigma\right) P\left(\theta_j \middle| y_j, \Gamma, \Sigma\right) = P\left(x_j \middle| \theta_j\right) p\left(\theta_j \middle| y_j, \Gamma, \Sigma\right)$$

variance among them reflects uncertainty due to not observing $\underline{\theta}$. It should be noted that this variance does not include the variability of sampling from the population.

Plausible values are not test scores for individuals in the usual sense, but rather are imputed values that may be used to estimate population characteristics correctly. When the underlying 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.[7]

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 Equation 5, 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 equation.

11.2.4 Conditioning

A multivariate normal distribution was assumed for $P(\theta_j | x_j, y_j, \Gamma, \Sigma)$, with a common variance, Σ, and with a mean given by a linear model with regression parameters, Γ. Since, in large-scale studies like PIRLS, there are many hundreds of background variables, it is customary to conduct a principal components analysis to reduce the number to be used in Γ. Typically, components representing 90 percent of the variance in the data are selected. These principal components are referred to as the conditioning variables, and denoted as y^c. The following model is then fit to the data:

Equation 6

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

In Equation 6, 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.

7 For further discussion, see (Mislevy, Beaton, Kaplan, & Sheehan, 1992).

Note that, in order to be strictly correct for all functions Γ of θ, it is necessary that $P(\theta \mid y)$ be correctly specified for all background variables in the survey. Estimates of functions Γ involving background variables not conditioned on in this manner are subject to estimation error due to misspecification. The nature of these errors was discussed in detail in Mislevy (1991). In PIRLS 2001, however, principal component scores based on nearly all background variables were used. Those selected variables were chosen to reflect high relevance to policy, and to education practices. The computation of marginal means and percentile points of θ for these variables is nearly optimal.

The basic method for estimating Γ and Σ with the expectation and maximization (EM) procedure is described in Mislevy (1985) for a single scale case. The EM algorithm requires the computation of the mean θ, and variance Σ, of the posterior distribution in Equation 6.

11.2.5 Generating Proficiency Scores

After completing the EM algorithm, plausible values for all sampled students are drawn from the joint distribution of the values of Γ in a three-step process. First, a value of Γ is drawn from a normal approximation to $P(\Gamma,\Sigma \mid x_j,y_j)$ 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 6 are computed using the methods applied in the EM algorithm. In the third step, the proficiency values are drawn independently from a multivariate normal distribution with mean θ 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 are fixed. Hence, all respondents – regardless of the number of items attempted – are assigned a set of plausible values.

The plausible values can then be employed to evaluate an arbitrary statistic T as follows:

1. Using the first vector of plausible values for each respondent, evaluate T as if the plausible values were the true values of θ. Denote the result T_1.

2. As in step 1 above, evaluate the sampling variance of T, or $Var(T_1,)$, with respect to respondents' first vectors of plausible values.

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:

$$T. = \frac{\sum_{\cup} T_u}{5}$$

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

$$Var(T.) = \sum_{\cup} \frac{VAR_{ut}}{M} + \left(1 + M^{-1}\right) \frac{\Sigma\left(T_u - T.\right)^2}{M-1}$$

The first component in V_M reflects uncertainty due to sampling respondents from the population; the second reflects uncertainty due to the fact that sampled respondents' θs are not known precisely, but only indirectly through x and y.

11.2.6 Working with Plausible Values

Plausible values methodology was used in PIRLS 2001 to ensure the accuracy of estimates of the proficiency distributions for the PIRLS population as a whole, and particularly for comparisons between subpopulations. A further advantage of this method is that the variation between the five plausible values generated for each respondent reflects the uncertainty associated with proficiency estimates for individual respondents. However, retaining this component of uncertainty requires that additional analytical procedures be used to estimate respondents' proficiencies, as follows:

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

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

where d is the degrees of freedom for the complete-data statistic, and f is the proportion of total variance due to not observing θ values:

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

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 the incomplete-data statistic differs little from the reference distribution for the corresponding complete-data statistics. If, in addition, d is large, the normal approximation can be used instead of the t-distribution.

For k-dimensional t, such as the k coefficients in a multiple regression analysis, each U and U^* is a covariance matrix, and B is an average of squares and cross-products rather than simply an average of squares. In this case, the quantity $(T\text{-}t^*)V^{-1}\ (T\text{-}t^*)'$ is approximately F distributed with degrees of freedom equal to k and ν, with ν defined as above but with a matrix generalization of f_M :

$$f = \frac{\left(1 - M^{-1}\right) Trace\left(BV^{-1}\right)}{k}$$

A chi-square distribution with k degrees of freedom can be used in place of the above quantity $(T\text{-}t^*)V^{-1}\ (T\text{-}t^*)'$ 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 & Johnson (1990), Mislevy (1991), and Mislevy & Sheehan (1987). To avoid such biases, the PIRLS 2001 analyses included effectively all background variables in the conditioning.

11.3 Implementing the Scaling Procedures for the PIRLS 2001 Assessment Data

The application of IRT scaling and plausible value methodology to the PIRLS 2001 assessment data involved three major tasks: calibrating the achievement test items (estimating model parameters for each item), creating principal components from the questionnaire data for use in conditioning, and generating IRT scale scores (proficiency scores) for reading overall, and for each of two reading purposes (reading for literary experience and reading to acquire and use information).

11.3.1 Calibrating the PIRLS 2001 Test Items

As shown in Exhibit 11.1, the PIRLS achievement test design consisted of a total of eight reading blocks (a block consisting of a text passage to be read followed by a set of questions about the passage) distributed across nine student booklets and a PIRLS Reader. Each block was developed to assess one of the two reading purposes specified in the framework: reading for literary experience, or reading to acquire and use information. The literary blocks are designated L1, L2, L3, and L4 – and the information blocks I1, I2, I3, and I4. Each student booklet, as well as the Reader, contained two blocks, which were chosen according to a matrix-sampling scheme that kept the number of booklets as low as possible while maximizing the number of times blocks were paired together in a booklet. Each sampled student completed one of the nine student booklets or the Reader.

Exhibit 11.1: Distribution of Literary and Information Blocks Across Booklets*

	Reading Achievement Overall							
	Reading for Literary Experience				Reading to Acquire and Use Information			
Booklet	L1	L2	L3	L4	I1	I2	I3	I4
Booklet 1	X	X						
Booklet 2		X	X					
Booklet 3			X		X			
Booklet 4					X	X		
Booklet 5						X	X	
Booklet 6	X						X	
Booklet 7	X				X			
Booklet 8		X				X		
Booklet 9			X				X	
Reader				X				X

* An 'X' in a cell indicates that the block in that column was assigned to the booklet in that row.

The booklets and Reader were distributed among the students in each sampled class according to a scheme that ensured comparable random samples of students responded to each block. Because blocks L1 through L3 and I1 through I3 each appear in three booklets, but blocks L4 and I4 appear only in the Reader, the assignment plan ensured that the Reader was assigned after every third booklet. Effectively, this meant that each block was administered to approximately 1/4 of the student sample.

Following the PIRLS framework, IRT scales for reporting student reading achievement were constructed for reading overall (both reading purposes combined) as well as separately for reading for literary experience, and for reading to acquire information.

The first step in constructing these scales was to estimate the IRT model item parameters for each item on each of the scales. This item calibration was conducted using the commercially available Parscale software (Muraki & Bock, 1991; version 3.5). The entire PIRLS student sample (146,490 students from 35 countries) was used in the calibration runs. However, to ensure that the data from each country contributed equally to the item calibration, the student sampling weights within each country were scaled to add to 500, so that – for the purposes of item parameter estimation only – each country had a weighted sample size of 500 students.

Three separate item calibrations were run: one for the overall reading scale, and one for each of the literary and information scales. All items were included in the calibration of the overall reading scale. Interim reading scores[8] for use in generating conditioning variables were produced as a by-product of this calibration. For the calibration run for the reading for literary experience scale, only those items from the literary blocks and only those students completing a booklet containing a literary block (121,228 students) were included. Similarly, only the items from the information blocks and only the students responding to information items (121,065 students) were included in the calibration for the information scale. Exhibits D.1 through D.3 in Appendix D present the item parameters for the three calibration runs.

11.3.2 Omitted and Not-Reached Responses

Apart from missing data on items that by design were not administered to a student, missing data could also occur because a student did not answer an item – whether because the student did not know the answer, omitted it by mistake, or did not have time to attempt the item. An item was considered not reached when (within part 1 or part 2 of the booklet) the item itself and the item immediately preceding were not answered, and there were no other items completed in the remainder of the booklet.

8 Because each student responded to only a subset of the assessment item pool, these interim scores, known as expected a posterior or EAP scores, were not sufficiently reliable for reporting PIRLS results. The plausible value proficiency estimates were used for this purpose.

In PIRLS 2001, 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, because the time allotment for the PIRLS tests was generous – enough for even marginally able respondents to attempt all items – not-reached items were considered as incorrect responses when student proficiency scores were generated.

11.3.3 Evaluating Fit of IRT Models to the PIRLS 2001 Data

After the calibration runs were completed, checks were performed to verify that the item parameters obtained from Parscale adequately reproduced the observed distribution of responses across the proficiency continuum. The fit of the IRT models to the PIRLS 2001 data was examined by comparing the theoretical item response function curves generated using the item parameters estimated from the data with the empirical item response functions calculated from the posterior distributions of the θs for each respondent who received the item.

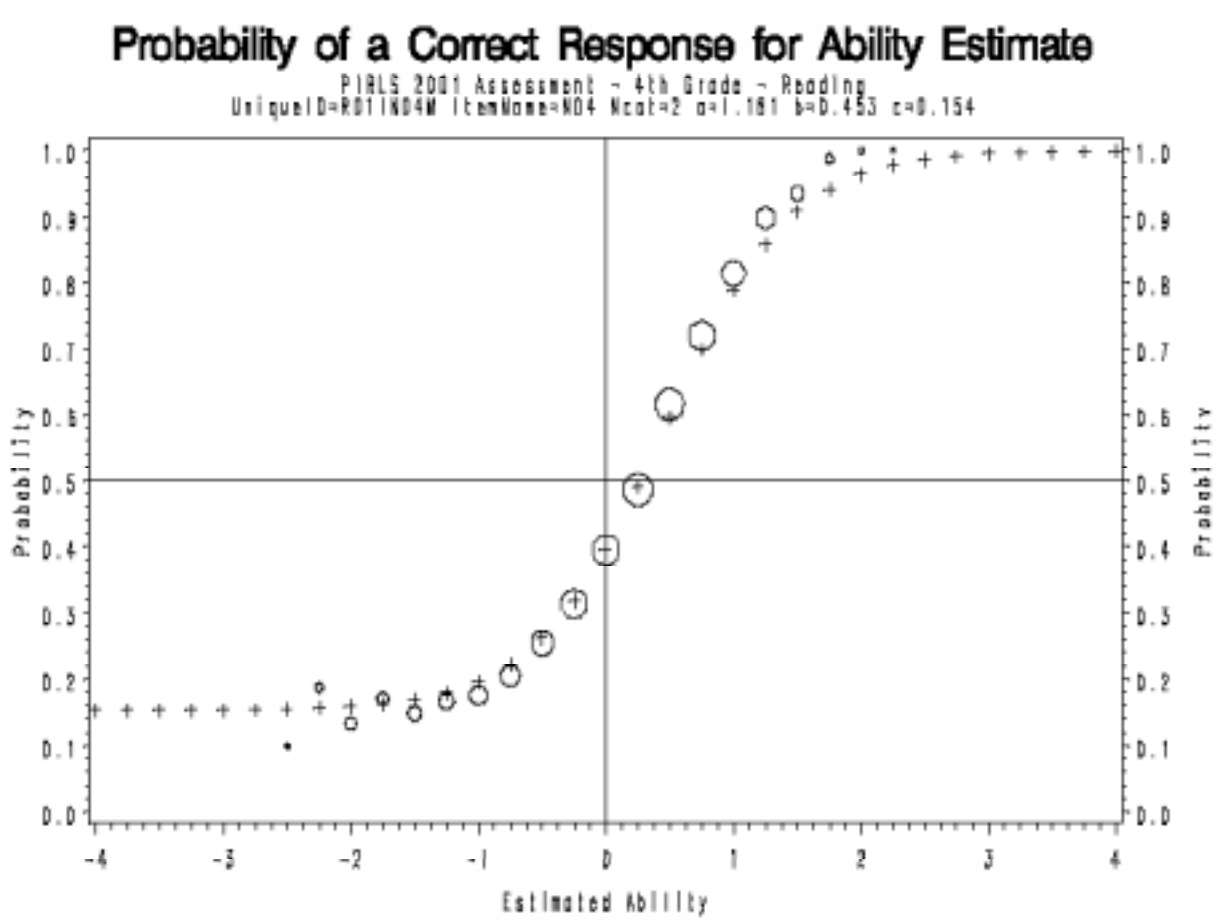
Exhibit 11.2 shows a plot of the empirical and theoretical item response functions for a dichotomous item. In the plot, the horizontal axis represents the proficiency scale, and the vertical axis represents the probability of a correct response. Values from the theoretical curve based on the estimated item parameters are shown as crosses. Empirical results are represented by circles. The centers of the circles represent the empirical proportions correct. The size of the circles is proportional to the sum of the posteriors at each point on the proficiency scale for all of those who received the item; this is related to the number of respondents contributing to the estimation of that empirical proportion correct. Exhibit 11.3 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 11.2. For items where the model fits the data well, the empirical and theoretical curves are close together.

11.3.4 Variables for Conditioning the PIRLS 2001 Data

Because there were so many background variables that could be used in conditioning, PIRLS followed the practice established in other large-scale studies of using principal components analysis to replace the original variables with a smaller number of principal components that explain most of their common variance. Principal components for the PIRLS 2001 student background data were constructed as follows:

- For categorical variables (questions with a small number of fixed response options), a "dummy coded" variable was created for each response option, with a value of one if the option was chosen and zero otherwise. If a student omitted or was

Exhibit 11.2: PIRLS 2001 Reading Assessment Example Item Response Function Dichotomous Item



not administered a particular question, all dummy coded variables associated with that question were assigned the value zero.

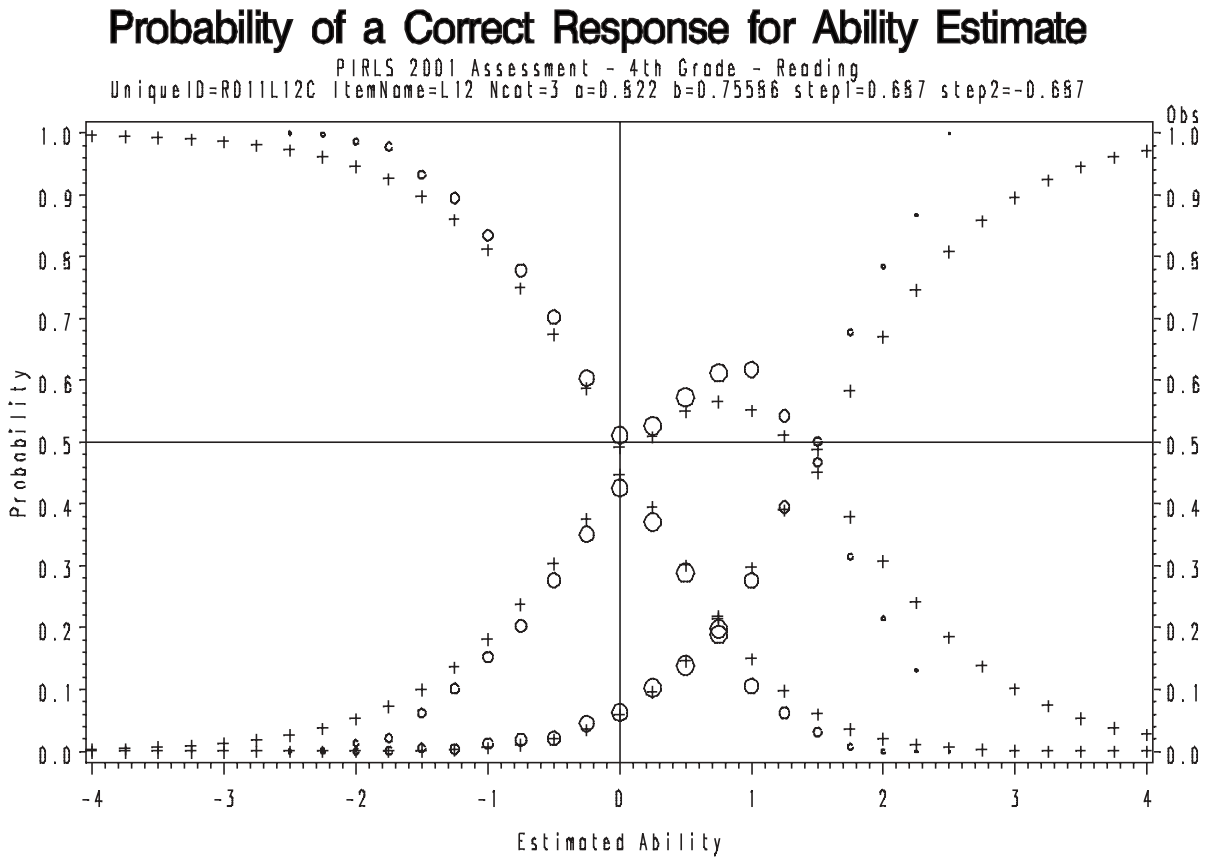
- Background variables with numerous response options (such as year of birth, or number of people who live in the home) were recoded using criterion scaling.[9] This was done by replacing each response option with the mean interim (EAP) score of the students choosing that option.

- Separately for each PIRLS country, all the dummy-coded and criterion-scaled variables were included in a principal components analysis. Those principal components accounting for 90 percent of the variance of the background variables were retained for use as conditioning variables.[10] Because the principal components analysis was performed separately for each country, the number of principal components required to account for 90 percent of the variance in the background variables varied from country to country. Exhibit 11.4 shows

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

10 Exceptions were Belize, Latvia, and Lithuania – where component accounting for only 80% of the variance were selected.

Exhibit 11.3: PIRLS 2001 Reading Assessment Example Item Response Function Polytomous Item



the total number of variables that were used in the principal component analysis and the number of principal components needed to account for 90 percent of the variance in the background variables within each country.

In addition to the principal components, student gender (dummy coded), the language of the test (dummy coded), an indicator of the classroom in the school to which the student belonged (criterion scaled), and an optional, country-specific variable (dummy coded) were included as conditioning variables.

11.3.5 Generating IRT Proficiency Scores for the PIRLS 2001 Data

Educational Testing Service's MGROUP program (ETS, 1998; version 3.1)[11] was used to generate the IRT proficiency scores. This program takes as input the students' responses to the items they were given, the item parameters estimated at the calibration stage, and the conditioning variables, then generates the plausible values that represent student proficiency in reading as output. Three MGROUP runs were conducted,

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

one for reading overall, and one each for reading for literary experience and reading to acquire and use information.

Plausible values generated by the MGROUP program are initially on the same scale as the item parameters used to estimate them. This scale metric is generally not useful for reporting purposes, since it is somewhat arbitrary, ranges between approximately -3 and +3, and has a mean of zero across all countries. For reporting PIRLS results, a scale metric was selected such that the combined proficiency distribution for fourth grade students across all PIRLS countries had a mean of 500 and a standard deviation of 100. The same metric (mean of 500 and standard deviation of 100) was also used for the literary and information scales.

Although practically all of the plausible values on the new metric were between 0 and 1000, there were a few outliers with values outside this range. These were recoded so that plausible values below 5 were set to 5, and plausible values above 995 were set to 995. This truncation did not have a noticeable effect on the distribution of the plausible values.

11.3.6 Implementing the Scaling Procedures for the Trends in IEA's Reading Literacy Study Data

In conjunction with the PIRLS 2001 assessment, IEA offered countries that had participated in the 1991 IEA Reading Literacy Study at the fourth grade the opportunity to measure trends over a ten-year period by re-administering the 1991 test at the same time as the PIRLS data collection was taking place. Nine of the 35 PIRLS countries took part in this Trends in IEA's Reading Literacy Study. The IRT scaling methodology used with the PIRLS 2001 data was applied also in scaling the trend study data. From a scaling perspective, the challenge was to place the 1991 data and the 2001

Exhibit 11.4: Number of Principal Components Selected to Account for the Variance in PIRLS 2001 Background Variables

Country	Number of Principal Components
Argentina	291
Belize	221
Bulgaria	287
Canada (O,Q)	291
Colombia	305
Cyprus	290
Czech Republic	288
England	265
France	282
Germany	305
Greece	284
Hong Kong, SAR	294
Hungary	282
Iceland	291
Iran, Islamic Rep. of	304
Israel	295
Italy	298
Kuwait	265
Latvia	223
Lithuania	272
Macedonia, Rep. of	296
Moldova	293
Morocco	160
Netherlands	280
New Zealand	280
Norway	293
Romania	287
Russia	288
Scotland	279
Singapore	303
Slovak Republic	297
Slovenia	226
Sweden	297
Turkey	287
United States	163

data on the same scale so that changes in average student reading literacy in the participating countries over the ten-year period could be accurately described.

The Reading Literacy data collected in 1991 were scaled, at that time, using a one-parameter IRT model known as the Rasch model.[12] However, the two- and three-parameter models with conditioning and plausible values used in scaling the PIRLS data were preferred also for scaling the trend data – partly for consistency with the PIRLS approach, but mainly because they were likely to be a better fit to the data (important when trying to detect possibly small changes in achievement between 1991 and 2001). Under the Rasch model, items may vary in difficulty, but are assumed to have the same discriminating power, and to not be answerable by guessing. The two- and three-parameter models feature an extra item parameter that accounts for differences among items in discriminating power, and the three-parameter model introduces a third parameter that models guessing behavior. The extra parameters mean that these models can more accurately account for the differences among items in their ability to discriminate between students of high and low ability, and are more effective than the simpler Rasch models in reducing errors due to model misspecification. Specification errors are apparent when data predicted on the basis of the model do not match the observed data. The difference between the observed data and those generated by the model is directly proportional to the degree of model misspecification.

One disadvantage of the one- and two-parameter models, compared with the one-parameter Rasch model, is that because 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 trend database is more than large enough to provide the required level of confidence.

As with the PIRLS 2001 data, the application of IRT scaling and plausible value methodology to the trend study data involved three major tasks: calibrating the items of the Reading Literacy test using the combined data from 1991 and 2001, creating principal components from the questionnaire data for use in conditioning, and creating IRT scale scores (proficiency scores) for the required reading scales.

11.3.7 Calibrating the 1991 Reading Literacy Test Items

By comparison with the PIRLS assessment, the design of the 1991 Reading Literacy test was relatively simple, consisting of a single test booklet administered to all sampled students. This test booklet contained a total of 65 test items addressing three different text types: narrative texts (22 items), expository texts (21 items), and documents (22 items).

12 The analysis of the 1991 data is described in Elley (1994).

Scales for reporting student achievement in reading literacy were constructed for reading overall (using all 65 items), and of the three text types – narrative texts, expository texts, and documents. The data from each of the nine countries consisted of student responses to the test items collected from nationally-representative samples of students at two points in time: 1991 and 2001.

The first step in constructing the trend study reading scales was to estimate the IRT model item parameters for each item on each of the scales. As with the PIRLS data, the item calibration was conducted using the commercially available Parscale software (Muraki & Bock, 1991; version 3.5). The data from 1991 and 2001 were combined for the calibration runs. A total of 59,761 student records were used in the calibration of the test items. To ensure that the data from each country contributed equally to the item calibration, and that data from 1991 and from 2001 contributed equally, the student sampling weights within each country for each data collection were scaled to add to 500 – so that, for the purposes of item parameter estimation, each country had a weighted sample size of 1000 students, 500 from 1991 and 500 from 2001.

Four separate item calibrations were run: one for the overall reading scale, and one for each of the text types – narrative texts, expository texts, and documents. All items and all students were included in the calibration of the overall reading scale. As in the PIRLS 2001 scaling, interim reading scores for use in generating conditioning variables were produced as a by-product of this calibration. Only the narrative items were included in the calibration run for the narrative scale, only the expository items for the expository scale, and only the documents items for the documents scale. Since all students responded to all items, all students were included in the calibration of each of the three scales. Exhibits D.4 through D.7 in Appendix D present the item parameters for the four calibration runs.

After the calibration was completed, checks were performed to verify that the item parameters obtained from the Parscale runs adequately reproduced the observed distribution of responses across the proficiency continuum.

11.3.8 Variables for Conditioning the Reading Literacy Trend Data

Similar to the procedure followed in conditioning the PIRLS 2001 data, principal components analysis was used to summarize the background questionnaire data collected during the 1991 and 2001 administrations of the 1991 Reading Literacy test. Identical procedures for coding the questionnaire variables prior to extracting principal components were followed. As before, those components accounting for 90 percent of the variance in the background variables were retained for conditioning.

Because the principal component analysis was performed separately for each country and for each data-collection year, the number of principal components necessary to account for 90 percent of the variance varied from country to country. Exhibit 11.5 shows the total number of variables that were used in the principal component analysis as well as the number of principal

Exhibit 11.5: Number of Principal Components Selected to Account for the Variance in Trends in IEA's Reading Literacy Study Background Variables

Country	Number of Principal Components	
	1991	2001
Greece	124	117
Hungary	122	129
Iceland	128	122
Italy	121	117
New Zealand	121	116
Singapore	123	124
Slovenia	125	120
Sweden	121	113
United States	119	119

components selected to account for 90 percent of the background variance within each country.

As with the PIRLS 2001 data, student gender (dummy coded), the language of the test (dummy coded), an indicator of the classroom in the school to which the student belonged (criterion scaled), and an optional, country-specific variable (dummy coded) were included as conditioning variables in addition to the principal components.

11.3.9 Generating IRT Proficiency Scores for the Trends in IEA's Reading Literacy Study

As with the PIRLS 2001 data, the MGROUP program (ETS, 1998; version 3.1) was used to generate the IRT proficiency scores for the trend study data. Four MGROUP runs were conducted: one for reading overall, and one each for the narrative, expository, and documents reading scales.

Because the data from 1991 and 2001 were combined during item calibration, the plausible values generated by the MGROUP program for each of the two data collections were on the same scale, and could be compared directly for purposes of analysis and reporting. To facilitate reporting, the original metric of the plausible values, which had a range of approximately from −3 to +3 with a mean of zero over all countries and across both data collections, was rescaled so that the mean of the 2001 data across all countries was 500 and the standard deviation was 100. This transformation was then applied to the 1991 data also, so that the 1991 and 2001 data had the same metric. This metric (mean of 500 and standard deviation of 100) also was used for the narrative, expository, and documents scales.

As with PIRLS 2001, practically all of the plausible values on the new metric were between 0 and 1000, with few outliers with values outside this range. Outliers were recoded so that plausible values below 5 were set to 5, and plausible values above 995 were set to 995. This truncation did not have a noticeable effect on the distribution of the plausible values.

References

Beaton, A.E. (1969). Scaling criterion of questionnaire items. *Socio-Economic Planning Sciences*, 2, 355-362.

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 (Eds.). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley Publishing.

Elley, W.B. (Ed.). (1994). *The IEA study of reading literacy: Achievement and instruction in 32 school systems*. Oxford, England: Elsevier Science Ltd.

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

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

Mislevy, R.J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, 80, 993-97.

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

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

Mislevy, R.J., Johnson, E.G. & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, 17(2), 131- 154.

Mislevy, R.J., & Sheehan, K. (1987). Marginal estimation procedures. In A. E. Beaton (Ed.), *Implementing the new design: The NAEP 1983-84 technical report* (pp. 293-360). Princeton, NJ: Educational Testing Service, 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.

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

Van Der Linden, W.J. & Hambleton, R. (1996). *Handbook of modern item response theory*. New York. Springer-Verlag.

Wingersky, M., Kaplan, B.A., & Beaton, A.E. (1987). Joint estimation procedures. In A. E. Beaton (Ed.), *Implementing the new design: The NAEP 1983-84 technical report* (pp.285-92). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Yamamoto, K., & Kulick, E. (2000). Scaling Methodology and Procedures for the TIMSS Mathematics and Science Scales. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS 1999 technical report*. Chestnut Hill, MA: Boston College.