9

Scaling PISA Cognitive Data

The mixed coefficients multinomial logit model as described by Adams, Wilson and Wang (1997) was used to scale the PISA data, and implemented by *ConQuest* software (Wu, Adams and Wilson, 1997). This chapter presents the model employed, and its application to the analysis of the PISA 2012 data.

THE MIXED COEFFICIENTS MULTINOMIAL LOGIT MODEL

The model applied to PISA is a generalised form of the Rasch model. The model is a mixed coefficients model where items are described by a fixed set of unknown parameters, $\boldsymbol{\xi}$, while the student outcome levels (the latent variable), $\boldsymbol{\theta}$, is a random effect.

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

9.1

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

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

A vector of zeroes denotes a response in category zero, making the zero category a reference category, which is necessary for model identification. Using this as the reference category is arbitrary, and does not affect the generality of the model. The $\mathbf{X}_i$ can also be collected together into the single vector $\mathbf{X}^T = (\mathbf{X}_1^T,\mathbf{X}_2^T,\ldots,\mathbf{X}_I^T)$, called the response vector (or pattern). Particular instances of each of these random variables are indicated by their lower case equivalents: x, x_i and x_{ik}.

Items are described through a vector $\boldsymbol{\xi}^T = (\xi_1,\xi_2,\ldots,\xi_p)$, 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. A set of design vectors $\mathbf{a}_{ij}$, $(i=1,\ldots,I; j=1,\ldots,K_i)$, each of length p, which can be collected to form a design matrix $\mathbf{A}^T = \left(\mathbf{a}_{11},\mathbf{a}_{12},\ldots,\mathbf{a}_{1K_1},\mathbf{a}_{21},\ldots,\mathbf{a}_{2K_2},\ldots,\mathbf{a}_{IK_I}\right)$, define these linear combinations.

The multi-dimensional form of the model assumes that a set of D traits underlies the individuals' responses. The D latent traits define a D-dimensional latent space. The vector $\boldsymbol{\theta} = (\theta_1,\theta_2,\ldots,\theta_D)'$, represents an individual's position in the D-dimensional latent space.

The model also introduces a scoring function that allows specifying the score or performance level assigned to each possible response category to each item. To do so, the notion of a response score b_{ijd} is introduced, which gives the performance level of an observed response in category j, item i, dimension d. The scores across D dimensions can be collected into a column vector $\mathbf{b}_{ik} = (b_{ik1},b_{ik2},\ldots,b_{ikD})^T$ and again collected into the scoring sub-matrix for item i, $\mathbf{B}_i = (\mathbf{b}_{i1},\mathbf{b}_{i2},\ldots,\mathbf{b}_{iD})^T$ and then into a scoring matrix $\mathbf{B} = (\mathbf{B}_1^T,\mathbf{B}_2^T,\ldots,\mathbf{B}_I^T)^T$ for the entire test. (The score for a response in the zero category is zero, but, under certain scoring schemes, other responses may also be scored zero.) The scoring matrix, $\mathbf{B}$, represents the relationships between items and dimensions, and the design matrix, $\mathbf{A}$, represents the relationships between items and the model parameters.

The probability of a response in category j of item i is modelled as

9.2

$$\Pr\left(X_{ij} = 1; \mathbf{A},\mathbf{B},\boldsymbol{\xi} \mid \boldsymbol{\theta}\right) = \frac{\exp\left(\mathbf{b}_{ij}\theta + \mathbf{a}_{ij}^T\boldsymbol{\xi}\right)}{\sum_{k=1}^{K}\exp\left(\mathbf{b}_{ik}\theta + \mathbf{a}_{ik}^T\boldsymbol{\xi}\right)}$$

There is a response vector,

9.3

$$f(\boldsymbol{x},\boldsymbol{\xi} \mid \theta) = \Psi(\theta,\boldsymbol{\xi})\exp[\boldsymbol{x}^T(\mathbf{B}\theta + \mathbf{A}\boldsymbol{\xi})]$$

with

9.4

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

where $\boldsymbol{\Omega}$ is the set of all possible response vectors.

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, $\boldsymbol{\theta}$. The complete definition of the model, therefore, requires the specification of a density, $f_{\boldsymbol{\theta}}(\boldsymbol{\theta}; \boldsymbol{\alpha})$ for the latent variable, $\boldsymbol{\theta}$. Let $\boldsymbol{\alpha}$ symbolise a set of parameters that characterise the distribution of $\boldsymbol{\theta}$**.** The most common practice, when specifying uni-dimensional marginal item response models, is to assume that students have been sampled from a normal population with mean μ and variance σ^2.That is:

9.5

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

or equivalently

9.6

$$\theta = \mu + E$$

where $E \sim N(0, \sigma^2)$

Adams, Wilson and Wu (1997) discuss how a natural extension of [9.6] is to replace the mean, μ, with the regression model, $\mathbf{Y}_n^T\boldsymbol{\beta}$, 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 or socio-economic status. Then the population model for student n becomes

9.7

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

where it is assumed that the E_n are independently and identically normally distributed with mean zero and variance σ^2 so that [9.7] is equivalent to:

9.8

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

a normal distribution with mean $\mathbf{Y}_n^T\boldsymbol{\beta}$ and variance σ^2. If [9.8] is used as the population model then the parameters to be estimated are $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\xi}$**.**

The generalisation needs to be taken one step further to apply it to the vector-valued $\boldsymbol{\theta}$ rather than the scalar-valued θ. The multi-dimensional extension results in the multivariate population model:

9.9

$$f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_n; \mathbf{w}_n, \boldsymbol{\gamma}, \boldsymbol{\Sigma}) = (2\pi)^{-d/2} \, |\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2} \left(\boldsymbol{\theta}_n - \boldsymbol{\gamma}\mathbf{w}_n\right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\theta}_n - \boldsymbol{\gamma}\mathbf{w}_n\right) \right]$$

where $\boldsymbol{\gamma}$ is a $u \times D$ matrix of regression coefficients, $\boldsymbol{\Sigma}$ is a $D \times D$ variance-covariance matrix, and $\mathbf{W}_n$ is a $u \times 1$ vector of fixed variables.

In PISA, the $\mathbf{W}_n$ variables are referred to as conditioning variables.

Combined model

In [9.10], the conditional item response model [9.2] and the population model [9.9] are combined to obtain the unconditional, or marginal, item response model:

9.10

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

It is important to recognise that under this model the locations of individuals on the latent variables are not estimated. The parameters of the model are $\boldsymbol{\gamma}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\xi}$.

The procedures used to estimate model parameters are described in Adams, Wilson and Wu (1997), Adams, Wilson and Wang (1997), and Wu, Adams and Wilson (1997).

For each individual it is possible, however, to specify a posterior distribution for the latent variable, given by:

9.11

$$\mathrm{h}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_n;\mathbf{W}_n,\boldsymbol{\xi},\boldsymbol{\gamma},\boldsymbol{\Sigma}\,|\,\mathbf{x}_n) = \frac{f_x(\mathbf{x}_n;\boldsymbol{\xi}\,|\,\boldsymbol{\theta}_n)\, f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_n;\mathbf{W}_n,\boldsymbol{\gamma},\boldsymbol{\Sigma})}{f_x(x_n;\mathbf{W}_n,\boldsymbol{\xi},\boldsymbol{\gamma},\boldsymbol{\Sigma})}$$

$$= \frac{f_x(\mathbf{x}_n;\boldsymbol{\xi}\,|\,\boldsymbol{\theta}_n)\, f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_n;\mathbf{W}_n,\boldsymbol{\gamma},\boldsymbol{\Sigma})}{\int_{\boldsymbol{\theta}_n} f_x(\mathbf{x}_n;\boldsymbol{\xi}\,|\,\boldsymbol{\theta}_n)\, f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_n;\mathbf{W}_n,\boldsymbol{\gamma},\boldsymbol{\Sigma})}$$

Plausible values

As with all item response scaling models, student proficiencies (or measures) are not observed; they are missing data that must be inferred from the observed item responses. There are several possible alternative approaches for making this inference. PISA uses the imputation methodology usually referred to as plausible values (PVs). PVs are a selection of likely proficiencies for students that attained each score. For each scale and subscale, five plausible values per student are included in the international database.

Using item parameters anchored at their estimated values from the international calibration, the plausible values are random draws from the marginal posterior of the latent distribution [9.11] for each student. For details on the uses of plausible values, see Mislevy (1991) and Mislevy et al. (1992).

In PISA, the random draws from the marginal posterior distribution are taken as follows.

Draw M vector-valued random deviates, $\{\boldsymbol{\varphi}_{mn}\}_{m=1}^{M}$, from the multivariate normal distribution, $f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_n;\mathbf{W}_n,\boldsymbol{\gamma},\boldsymbol{\Sigma})$, for each case n, these vectors are used to approximate the integral in the denominator of [9.11], using the Monte-Carlo integration:[1]

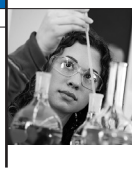
9.12

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

At the same time, the values

9.13

$$p_{mn} = f_x(x_n;\boldsymbol{\xi}\,|\,\boldsymbol{\varphi}_{mn})\, f_{\boldsymbol{\theta}}(\boldsymbol{\varphi}_{mn};\mathbf{W}_n,\boldsymbol{\gamma},\boldsymbol{\Sigma})$$



are calculated, so that we obtain the set of pairs $\left(\boldsymbol{\varphi}_{mn}, p_{mn}/\Im\right)_{m=1}^{M}$, which can be used as an approximation of the posterior density (11); and the probability that $\boldsymbol{\varphi}_{nj}$ could be drawn from this density is given by:

9.14

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

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

9.15

$$\sum_{s=1}^{i_0-1} q_{sn} < \eta_i \leq \sum_{s=1}^{i_0} q_{sn}$$

is selected as a plausible vector.

ANALYSIS OF DATA WITH PLAUSIBLE VALUES

It is very important to recognise that plausible values are not test scores and should not be treated as such. They are random numbers drawn from the distribution of scores that could be reasonably assigned to each individual - that is, the marginal posterior distribution [9.11]. As such, plausible values contain random error variance components and are not as optimal as scores for individuals. Plausible values as a set are better suited to describing the performance of the population. This approach, developed by Mislevy and Sheehan (1987, 1989) and based on the imputation theory of Rubin (1987), produces consistent estimators of population parameters. Plausible values are intermediate values provided to obtain consistent estimates of population parameters using standard statistical analysis software such as *SPSS*® and *SAS*®. As an alternative, analyses can be completed using *ConQuest* (Wu, Adams and Wilson, 1997).

If an analysis with plausible values were to be carried out, then it would ideally be undertaken five times, once with each relevant plausible values variable. The results would be averaged, and then significance tests adjusting for variation between the five sets of results computed.

More formally, suppose that $r(\boldsymbol{\theta},\mathbf{Y})$ is a statistic that depends upon the latent variable and some other observed characteristic of each student. That is: $(\boldsymbol{\theta},\mathbf{Y}) = (\theta_1,y_1,\theta_2,y_2,\ldots,\theta_N,y_N)$ where (θ_n,y_n) are the values of the latent variable and the other observed characteristic for student n. Unfortunately, θ_n is not observed, although we do observe the item responses, x_n from which we can construct for each student n, the marginal posterior $h_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_n;y_n,\boldsymbol{\xi},\boldsymbol{\gamma},\boldsymbol{\Sigma}\,|\,\mathbf{x}_n\right)$. If $h_{\boldsymbol{\theta}}\left(\boldsymbol{\theta};\mathbf{Y},\boldsymbol{\xi},\boldsymbol{\gamma},\boldsymbol{\Sigma}\,|\,\mathbf{X}\right)$ is the joint marginal posterior for $n = 1,\ldots N$ then we can compute:

9.16

$$r^*(\mathbf{X},\mathbf{Y}) = E\left[r^*(\boldsymbol{\theta},\mathbf{Y})\,|\,\mathbf{X},\mathbf{Y}\right]$$
$$= \int_{\boldsymbol{\theta}} r(\boldsymbol{\theta},\mathbf{Y})\, h_{\boldsymbol{\theta}}\left(\boldsymbol{\theta};\mathbf{Y},\boldsymbol{\xi},\boldsymbol{\gamma},\boldsymbol{\Sigma}\,|\,\mathbf{X}\right) d\boldsymbol{\theta}$$

The integral in [9.16] can be computed using the Monte-Carlo method. If M random vectors $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots,\boldsymbol{\theta}_M)$ are drawn from $h_{\boldsymbol{\theta}}\left(\boldsymbol{\theta};\mathbf{Y},\boldsymbol{\xi},\boldsymbol{\gamma},\boldsymbol{\Sigma}\,|\,\mathbf{X}\right)$ [9.16] is approximated by:

9.17

$$r^*(\mathbf{X},\mathbf{Y}) \approx \frac{1}{M}\sum_{m=1}^{M} r(\boldsymbol{\theta}_m,\mathbf{Y})$$
$$= \frac{1}{M}\sum_{m=1}^{M} \hat{r}_m$$

where $\hat{r}_m$ is the estimate of r computed using the m-th set of plausible values.

From [9.17] we can see that the final estimate of r is the average of the estimates computed using each randomly drawn vector in turn. If U_m is the sampling variance for $\hat{r}_m$ then the sampling variance of r^* is:

9.18

$$V = U^* + \left(1 + M^{-1}\right) B_M$$

where $U^* = \frac{1}{M}\sum_{m=1}^{M} U_m$ and $B_M = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{r}_m - r^*\right)^2$

An α-% confidence interval for r^* is $r^* \pm t_\nu\left((1-\alpha)/2\right)v^{1/2}$ where $t_\nu(s)$ is the s-percentile of the t-distribution with v degrees of freedom. $\nu = \left[\frac{f_M^2}{M-1} + \frac{\left(1-f_M\right)^2}{d}\right]^{-1}$, $f_M = \left(1+M^{-1}\right)B_M/V$ and d is the degree of freedom that would have applied had θ_n been observed. In PISA, d will vary by country and have a maximum possible value of 80.

APPLICATION TO PISA

In PISA, the mixed coefficients multinomial logit model described above was used in three steps: national calibrations, international scaling and student score generation.

For both the national calibrations and the international scaling, the conditional item response model [9.2] is used in conjunction with the population model [9.9], but conditioning variables are not used. That is, it is assumed that students have been sampled from a multivariate normal distribution.

The design matrix was chosen so that the partial credit model (Masters, 1982) was used for items with multiple score categories and the simple logistic model was fitted to the dichotomously scored items.

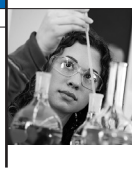
National calibrations

National calibrations were performed separately, country by country, using unweighted data. Country means were constrained to zero during the estimation process. For the countries that administered booklet sets that included the core and standard items a linear transformation was applied to the national item difficulties so that the core and standard items have a mean of zero. For the countries that have used booklets that included core and easy items a linear transformation was applied to the national item difficulties so that the core items have the same mean as the mean of the core items for the OECD calibration sample. The results of these analyses, which were used to monitor the quality of the data and to make decisions regarding national item treatment, are given in Chapter 12.

The outcomes of the national calibrations were used to make a decision about how to treat each item in each country. This means that an item may be deleted from PISA altogether if it has poor psychometric characteristics in more than ten countries (a "dodgy" item); it may be deleted from the scaling in particular countries if it has poor psychometric characteristics in those particular countries but functions well in the vast majority of others. When reviewing the national calibrations, particular attention was paid to the fit of the items to the scaling model, item discrimination and item-by-country interactions.

Item response model fit (weighted mean square MNSQ)

For each item parameter, the *ConQuest* fit mean square index (Wu, 1997) was used to provide an indication of the compatibility of the model and the data. For each student, the model describes the probability of obtaining the different item scores. It is therefore possible to compare the model prediction and what has been observed for one item across students. Accumulating comparisons across students gives an item-fit statistic. As the fit statistics compare an observed value with a predicted value, the fit is an analysis of residuals. In the case of the item infit mean square, values near one are desirable. A weighted MNSQ greater than one is associated with a low discrimination index, meaning the data exhibits more variability than expected by the model, and an infit mean square less than one is associated with a high discrimination index, meaning the data exhibits less variability than expected by the model.



Discrimination coefficients

For each item, the correlation between the students' score and aggregate score on the set for the same domain and booklet as the item of interest was used as an index of discrimination. If p_{ij} (calculated as x_{ij}/m_i) is the proportion of score levels that student i achieved on item j, and $p_i = \sum_j p_{ij}$ (where the summation is of the items from the same booklet and domain as item j) is the sum of the proportions of the maximum score achieved by student i, then the discrimination is calculated as the product-moment correlation between p_{ij} and p_i for all students. For multiple-choice and short-answer items, this index will be the usual point-biserial index of discrimination.

The point-biserial index of discrimination for a particular category of an item is a comparison of the aggregate score between students selecting that category and all other students. If the category is the correct answer, the point-biserial index of discrimination should be higher than 0.20 (Ebel and Frisbie, 1986). They set out the following recommendations regarding the index of discrimination:

Magnitude	Comment	Recommended action for item
> 0.39	Excellent	Retain
0.30 – 0.39	Good	Possibilities for improvement
0.20 – 0.29	Mediocre	Need to check/review
0.00 – 0.20	Poor	Discard or review in depth
< -0.01	Worst	Definitely discard

Non-key categories should have a negative point-biserial index of discrimination. The point-biserial index of discrimination for a partial credit item should be ordered, i.e. categories scored 0 should have a lower point-biserial correlation than the categories scored 1, and so on.

Item-by-country interaction

The national scaling provides nationally specific item parameter estimates. The consistency of item parameter estimates across countries was of particular interest. If the test measured the same latent trait per domain in all countries, then items should have the same relative difficulty or, more precisely, would fall within the interval defined by the standard error on the item parameter estimate (i.e. the confidence interval).

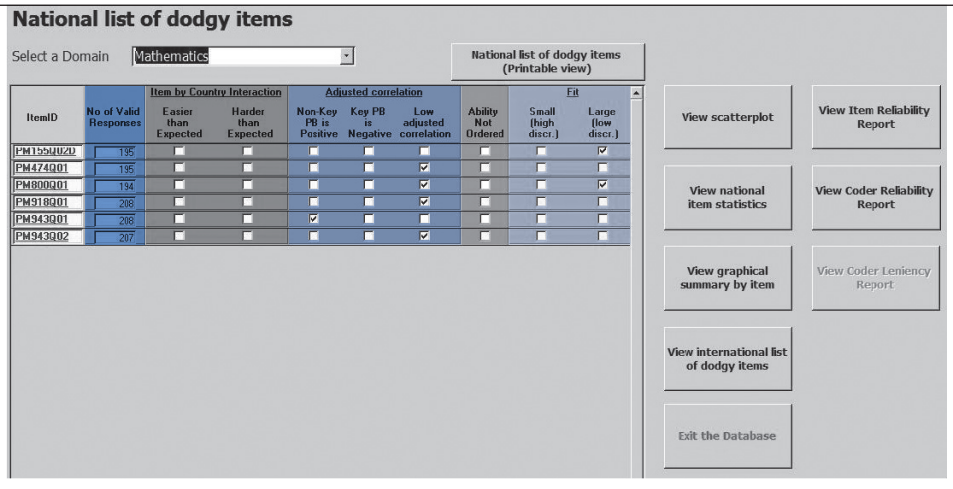
National reports

After national scaling was completed, all the available national item statistics were imported in the international item database. International level item statistics described next in this section were also included in this database. This allowed summarising national level statistics and performing the comparison to the international and aggregated item statistics. Database with national items statistics was returned to each participating country to assist in reviewing their data with the international contractor.

Figure 9.1 illustrates an interface of the national database. The main screen represents the interactive list of items by domain that are flagged as dodgy items in a country. Each column indicates a specific problem. Reliability and leniency reports will be discussed separately in Chapter 13.

■ Figure 9.1 ■

Main screen



Countries were asked to check the following statistics:

- Item by country interaction: The consistency of item parameters across countries is of particular importance in an international study. If the test measures the same underlying construct (or latent trait) the item should have similar relative difficulty in each country.
- Adjusted correlation: Adjusted correlation is a correlation between students' scores on an item and their adjusted domain scores, where the adjusted domain score is a student's total score for a domain minus the student's score for the particular item. For multiple-choice items this is equivalent to the point-biserial correlation of the correct response (key) and it should be 0.20 or higher. Otherwise it is marked as Low Adj. Correlation. If the item category is the key, the point biserial index should be positive (the same as for the item). Non-key categories (incorrect responses or distractors) should have a negative point biserial index.
- Ability not ordered: For partial credit items the student mean abilities should increase with increasing raw score; students that received score 0 should have lower mean abilities than those that had score 1 and those with score 2 should have higher mean abilities than those with 1.
- Fit: Infit Mean Square index is used to compare predicted value and observed value by analysis of residuals. Good fit should have values near one. An Infit Mean Square greater than one is associated with a low discrimination index while an Infit Mean Square lower than one is associated with a high discrimination index.

Four item reports could be generated using this database.

Report 1: Scatter plot

An example of a scatter plot report is given in Figure 9.2. This report shows the scatter plot of national and OECD/international item difficulties. Both sets of difficulties are centred on zero and are therefore referred to as relative difficulties. The vertical axis represents the national relative item difficulties and the horizontal axis the OECD or international relative item difficulties. Each dot is an item.

The scatter plot gives an overview of the behaviour of all items in a domain in one country compared to the pooled OECD set (500 students from each OECD country available at the time of analysis pooled together) or the international set (500 students from each country available at the time of analysis pooled together).

■ Figure 9.2 ■

Example of scatter plot

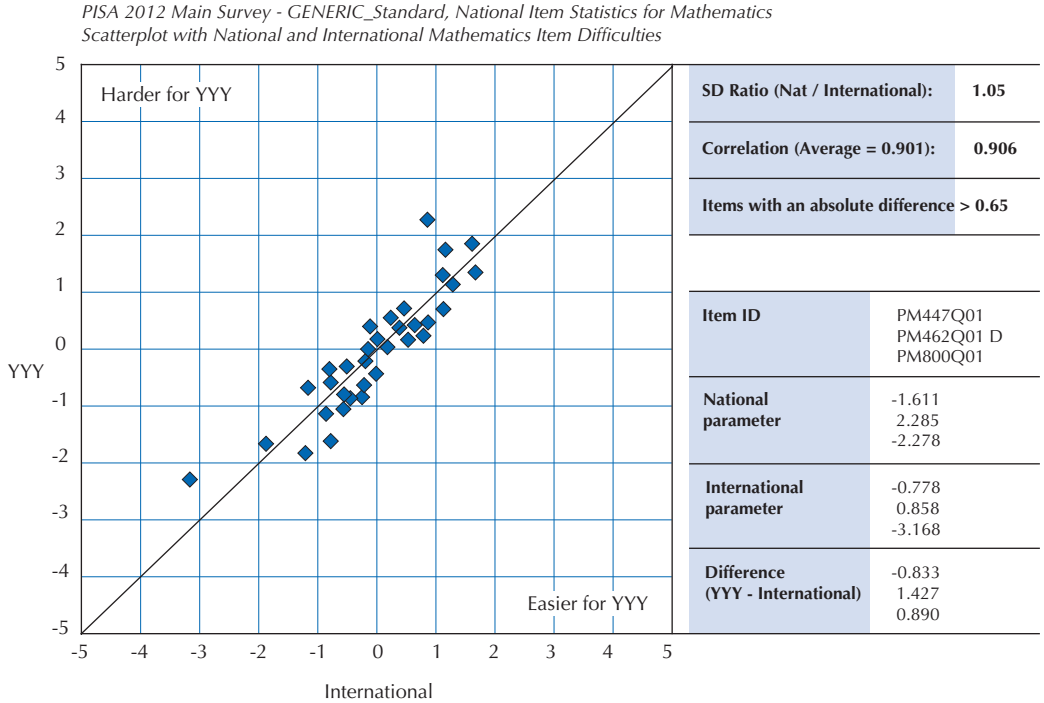


Figure 9.2 provides an illustration of the overall level of agreement and it assists in identifying outliers. Items that lie exactly on the identity line (the diagonal line) have equal national and international relative item difficulties. An outlier occurs when the relative national item difficulty is very different from the OECD/international relative item difficulty. In Figure 9.2 there are a couple of obvious outliers. This suggests that something could be wrong with these items.

The table next to the scatter plot lists all items with an absolute difference of more than 0.65. The National Centres were asked to check these items carefully for any translation or printing errors.

There are two types of summary statistics displayed in the top right box of Figure 9.2:

- Standard deviation ratio compares the spread of national item difficulties to the spread of the OECD/international item difficulties. It should be close to 1.
- Correlation should be similar to the OECD average correlation.

For this particular country both figures are satisfactory: the standard deviation ratio is sufficiently close to one and the correlation is sufficiently similar to the OECD average correlation.

Report 2: Descriptive statistics on individual items in tabular form

A detailed item-by-item report was provided in tabular form showing the basic item analysis statistics at the national level. This report provides classical item statistics for each item used in the national calibration. Summaries of item statistics are presented in a tabular form in item identifier order. If for any reason, an item is excluded from the national calibration, the item identifier will be listed at the end of the report. An example of item statistics for the fictitious item with identifier *PM999Q03* is shown in Figure 9.3.

■ Figure 9.3 ■

Example of item statistics in tabular form

Item : 50 (PM999Q03), Graphical Summary Page 50

Cases for this item 247 — Adjusted correlation: 0.18
Item Threshold(s): -0.116 — Fit (Weighted MNSQ): 1.29
Item Delta(s): -0.116

Code	Score	Count	% of Total	Pt Bis	Ability Avg	Ability SD
1	0	66	26.7	-0.10	-0.47	1.02
2	0	37	15.0	-0.07	-0.39	1.03
3	1	124	50.2	0.18	-0.06	0.92
4	0	12	4.9	0.10	0.00	1.31
8	0					
9	0	6	2.4	-0.18	-1.60	0.39
R	0	2	0.8	-0.13	-1.82	0.79

Two hundred and forty seven students have responded to this item in this country.

The national threshold and delta (difficulty) are -0.116 (for dichotomous items these two values are always the same).

The item adjusted correlation is 0.18. This is lower than 0.2 and would be reported on the interactive list of dodgy items and in the graphical summary report that is described in the next section.

The weighted mean square (MNSQ) fit statistic is 1.29. Small variations around one are expected, however, values larger than 1.2 indicate that the item discrimination is lower than assumed by the model, and values below 0.8 show that the item discrimination is higher than assumed. In this particular case the item would have a tick on the interactive screen in the 'Large Fit' column and in the graphical summary report that is described in the next section.

The first column gives the original responses. This is a multiple-choice item and therefore, the responses are: 1=A, 2=B, 3=C, 4=D, 8='invalid', 9='missing' and R='not reached'. Please note that there are no statistics for code 8. This is because there were no students in this country who gave invalid responses to this item.

The second column shows the score assigned to each response category. The correct response to this item is 3 (C).

The third and fourth columns in the table list the number and percentage of students in each category. In this country, 124 students (50.2%) gave the correct response.

The point-biserial correlations are presented in column five. This is the correlation between a response category coded as a dummy variable (a score of 1 for students that responded with the current code and a score of 0 for students in other response categories) and the total domain score. For dichotomous items the point-biserial is equal to the adjusted correlation (0.18 in Figure 9.3). Correct responses should have positive correlations with the total score, incorrect responses negative correlations. In this case one of the incorrect responses (4) has positive point-biserial (0.10). However the item would not have a tick on the interactive screen in the corresponding column for positive point biserial in non-key category, because there were fewer than 15 students who responded to distractor 4. Rather, this item would be flagged for low adjusted correlation less than 0.20.

The two last columns show the average ability of students responding in each category and the associated standard deviation. The average ability is calculated by domain. If an item is functioning well the group of students that gave the correct response should have a higher mean ability than the groups of students that provided each of the incorrect responses. This is true for categories 1 and 2. For category 4 this doesn't hold, but since the number of students is less than fifteen, this is not flagged.

Report 3: Graphical summary of descriptive statistics by item

This report provides comparisons between national and international item statistics in graphical form, one page per item.

An example of a full page for one item is given in Figure 9.4. More detailed information about each part of this report labelled A to D follows.

Part A

The top table in Figure 9.4 starts with the item code followed by the item name and item number (*CM999Q01*: Graph Example Q1).[2] For mathematics items, there is also a group identifier on the right hand side of the table. In the PISA 2012 Main Survey, the majority of mathematics items (common and link items) were administered in all participating countries. Seventeen countries used booklets that included a set of easier items. This was done to better cover the range of abilities in every country.

Item identifiers are followed by the overall item statistics, the same as in the national item statistics report described in the previous section: number of cases, adjusted correlation, weighted (infit) mean square (MNSQ), item thresholds and item difficulty (delta). In addition, item type (e.g. multiple choice) is presented. For multiple-choice items a key (correct choice) is also shown. Graph Example Q1 in Figure 9.4 is a partial credit item and therefore the key is not shown. Processing the responses to items of this type usually required manual coding.

The next section of part A contains national, international and OECD statistics by response category. The first row contains the score for each category, the second and third rows contain number of students and percentage of students in each category in the country. OECD % is the percentage of students in each category in the pooled OECD data. INT % is the percentage of students in the category in the pooled data of all countries that administered the item. Note that OECD % is not available for the set of easier mathematics items and financial literacy items.

Average ability (ability avg), standard deviation ability (ability SD), and point-biserial (pt bis) are the same national statistics as in the national item statistics report. These statistics were described in the previous section.

Part B

The displayed graphs in part B facilitate the process for identifying the possible national anomalies related to item statistics by response category.

The first graph is important for partial credit items. It helps to check whether the average ability increases with the score points, as shown in Figure 9.4. Note that categories "9" and "R" are not identified as score points.

The second graph is important for multiple-choice items. It helps to check whether:

- a non-key category has a positive point-biserial;
- a non-key category has a point-biserial higher than the key category; or
- the key category has a negative point-biserial.

■ Figure 9.4 ■

Example of graphical summary by item

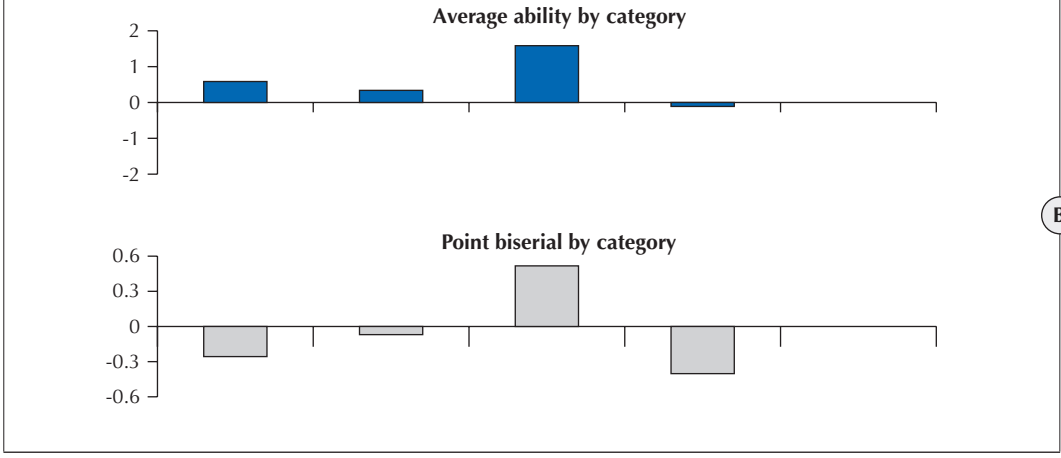
PISA MS12: Graphical presentation of item statistics for Country

CM999Q01: Graph Example Q1 (Item Format: Constructed Reponse Expert)

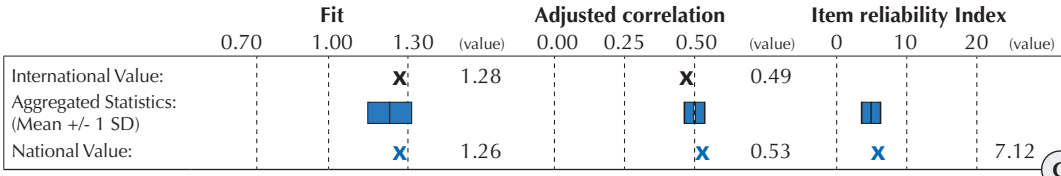
Number of Cases: 235	*Adjusted Correlation: 0.53*	*Item Threshold(s): 0.674*	*1.025*
Item Type: Partial Credit Item	*Fit (Weighted MNSQ): 1.26*	*Item Delta(s): 1.891*	*-0.192*

Response	0	1	2	9	R
Score	0	1	2	0	0
Students	70	26	115	24	
Percentage of students	29.79	11.06	48.94	10.21	
OECD %	30.42	9.91	50.06	9.54	0.06
INT %	30.73	12.45	45.12	10.59	0.12
Ability Avg	0.56	0.34	1.55	-0.1	
Ability SD	0.86	0.94	0.89	0.94	
Pt Bis	-0.27	-0.06	0.52	-0.39	

A



B



C

D

	Item by country interaction		Adjusted correlation				Fit	
No of Countries	Easier than Expected	Harder than Expected	Non-key PB is Positive	Key PB is Negative	Low Adjusted Correlation	Ability not Ordered	Small (High Discrimination Item)	Large (Low Discrimination Item)
CM999Q01	❑	❑	❑	❑	❑	❑	❑	☑
Countries: 48	12	9	0	0	0	2	0	23
OECD countries: 22	4	7	0	0	0	2	0	15
Other countries: 26	8	2	0	0	0	0	0	8

Part C

This part presents the graphical comparisons of overall item statistics at the national and OECD level.

National scaling provides for each country and item, the weighted MNSQ, adjusted correlation, delta item parameter estimate (or difficulty estimate) and threshold estimates. Item Reliability Index would also be provided if the item is a constructed response item that requires multiple coding as part of the process of evaluating the reliability of items and coders. For each item these national values will be compared with the pooled OECD value and average value for all OECD countries in the database at the time of comparison.

The black crosses at the top of each of the pictures represent the value of the coefficients computed from the pooled OECD data. The blue rectangles show the distribution of values obtained from each of available OECD country (all students). To obtain this distribution each OECD country is calibrated separately. Then the mean and standard deviation of the national estimates are computed. The rectangles are located so that their mid-point (indicated with a vertical bar) is at the mean and the left and right boundaries are located at the mean plus and minus one standard deviation respectively.

The blue crosses at the bottom of the pictures indicate the values computed only for the national dataset.

Any substantial differences between the national value and the OECD value, or the average OECD value, indicate that the item is behaving differently in that country in comparison to the other countries. This might reflect a mistranslation or printing problem. On the other hand, if the item is misbehaving in many countries, it might reflect a specific problem in the source item and not with one or more national versions of this item.

OECD statistics are not available for the easier mathematics items. Hence, the statistics for these items are calculated based on the pooled data from 17 countries.

Part D

At the bottom of the page a table with check boxes shows whether any substantial problems were found as a result of the national calibration for the particular item. The table indicates if an item was flagged for one of the following reasons:

- the relative national item difficulty is significantly higher or lower than OECD/International relative item difficulties;
- for multiple-choice items one of the non-key categories has a point-biserial correlation higher than 0.05 (only reported if the category was chosen by at least 15 students);
- for multiple-choice items the key category has a point-biserial lower than –0.05 (only reported if the category was chosen by at least 15 students);
- the adjusted correlation of the item is lower than 0.2;
- for partial credit items the category abilities are not ordered (only reported if both score categories in comparison have at least 15 students each); or
- the fit statistics are higher than 1.2 or lower than 0.8.

In the example in Figure 9.4, the box is ticked indicating large fit index. This is also shown in Part A (weighted MNSQ=1.26).

The next row below the tick boxes shows how many countries in total have a similar problem for the same item. The last two rows are the numbers of OECD countries and partner countries that have the same problem. The large fit problem, which is identified in Parts A and D, does not look problematic on the graph in Part C for this particular country. It is because out of 48 available countries, 23 countries (or 15 out of 22 available OECD countries) have the same problem (the figures are fictitious). This indicates a specific problem in the source item instead of possible mistranslation or misprint problems in the national versions.

However, if an item has at least one tick, and the number of countries below this tick is less than 10, the National Centres were strongly recommended to review the translation and printing of the item in all booklets and its appropriateness for the national context.

All flagged items are considered to be dodgy items either nationally if a problem occurs only in a particular country, or internationally if the same problem occurs in many countries (in more than 50% of cases).

Report 4: International list of dodgy items

The last report gives a summary of dodgy items for all countries included in the analysis at the time of reporting. A part of this table is given in Figure 9.5, showing items for which no indication of problems is evident, and items for which some such indication is present.

■ Figure 9.5 ■

Example of an international list of dodgy items

PISA 2012 Main Study - International counts of dodgy items **Mathematics**

		Item by Country Interaction		Adjusted correlation				Fit	
	No of Countries Included	Easier than Expected	Harder than Expected	Non-key PB is Positive	Key PB is Negative	Low adjusted correlation	Ability not Ordered	Small (high discr.)	Large (low discr.)
PM903Q01	10	1	1	0	0	0	2	0	8
PM903Q03	10	0	0	0	0	0	0	0	0
PM905Q01T	10	0	0	0	0	0	0	0	0
PM905Q02	10	0	0	0	0	0	0	2	0
PM906Q01	14	0	0	2	0	0	0	0	0
PM906Q02	14	0	0	0	0	0	0	0	0
PM909Q01	14	0	0	0	0	2	0	0	2
PM909Q02	14	0	0	0	0	0	0	0	0
PM909Q03	14	0	1	0	0	1	0	0	0
PM915Q01	14	0	1	0	0	0	0	0	0
PM915Q02	14	0	0	3	0	0	0	0	0
PM918Q01	10	0	1	0	0	0	1	0	7

If an item has poor psychometric properties in a large number of countries then it most likely should be explained by reasons other than mistranslation and misprint.

International calibration

There were new elements added to the PISA 2012 paper-based assessment of mathematics, reading and science; computer-based assessment of problem solving; digital reading and computer-based mathematics; financial literacy and reading components.

Since PISA 2000 international item parameters were set by applying the conditional item response model [9.2] in conjunction with the multivariate population model [9.9], without using conditioning variables, to a sub-sample of students. Traditionally a subsample of students referred to as an OECD calibration sample was formed using 500 students drawn at random from each of the OECD participating countries. In PISA 2009 countries with an expected mean reading score less than 450 were given the option to choose an easier set of booklets for the Main Survey. In total, 20 countries opted for the easier booklets, of which two, Mexico and Chile, were OECD member countries. This required creation of an extra calibration sample that included non-OECD countries.

In 2009, reading items required a two-step calibration process.

Step 1: The core and standard items were calibrated using the OECD calibration sample, which contained 500 students from 34 OECD countries.

Step 2: The easier items that were not included in the regular booklets were calibrated using the easy booklets calibration sample, while anchoring the core and standard items to the estimates obtained from Step 1. The easy booklets calibration sample was formed by adding subsamples of 500 students from each of the 20 countries that used the easy booklets to the international OECD calibration sample.

In PISA 2012, the use of data for calibration purposes from OECD countries only was not viable because not a sufficient number of the same OECD countries were participating in all options. Calibration of all optional parts was based on all countries that participated in those options. Consistent with the treatment of options, it was decided that the calibration is based on data from all available countries for the PISA 2012 paper-based assessment.

Similarly to PISA 2009, in PISA 2012 countries with an expected mean mathematics score less than 450 were given the option to choose an easier set of booklets for the Main Survey (see Chapter 2 for more details). In total, 17 countries opted for the easier booklets, of which two, Mexico and Chile, were OECD member countries. This subsample of students, referred to as an international calibration sample, consisted of 31 500 students comprising 500 students drawn at random from each of the 63 participating countries.[3] Not-reached items were excluded from the calibration. For model identification the average difficulty of all items in each domain was set to zero.

For the options it was decided to create a calibration sample with a similar number of responses per item as for the paper-based test. For the paper-based test sampling 500 students yields 154 responses per item, since each student responds to approximately 4/13 of all items. This resulted in the following sample sizes for each PISA 2012 option:

- Problem Solving (PS)
  - 308 (=154×8/4) students per country that implemented PS only
  - 924 (=154×24/4) students per country that implemented PS and CBA (mathematics and reading)
- Computer-Based Assessment (CBA)
  - 924 (=154×24/4) students per country for mathematics
  - 462 (=154×24/8) students per country for reading
- Financial Literacy
  - 154 (=154× 4/4) students per country

The allocation of each PISA item to one of the PISA 2012 scales and corresponding item parameters are given in Annex A.

Student score generation

Five multi-dimensional scaling models were used in the PISA 2012 Main Survey. The first model, made up of one reading, one science and one mathematics dimension, was used for reporting overall scores for reading, science and mathematics. A second model, made up of one science, one reading and four mathematics scales, was used to generate scores for the four mathematics subscales *Change and Relationships, Quantity, Space and Shape, Uncertainty and Data*. A third model, made up of one science, one reading and three mathematics scales was used to generate scores for the three mathematics subscales: *Employ, Formulate* and *Interpret*. A fourth model, made up of one reading, one science, one mathematics, one digital reading dimension, one digital mathematics and one digital problem solving dimension was used for reporting overall scores for reading, science, mathematics and computer-based mathematics, digital reading and computer problem solving scales for countries that implemented the computer-based assessment (CBA) in the PISA 2012 Main Survey. A fifth model, made up of one reading, one science, one mathematics and one digital problem solving dimension was used for reporting overall scores for reading, science, mathematics and computer problem solving scales for those countries that implemented problem solving in the PISA 2012 Main Survey as the only computer-based component.

The first three models were implemented in one step and the last two models, for countries that participated in CBA, were implemented in two steps, as it will be described later.

Sixty-five plausible values, five for each of the 13 PISA 2012 scales are included in the PISA 2012 database. *PV1MATH* to *PV5MATH* are for mathematical literacy; *PV1SCIE* to *PV5SCIE* for scientific literacy, *PV1READ* to *PV5READ* for reading literacy, *PV1CPRO* to *PV5CPRO* for computer problem solving assessment, *PV1CMAT* to *PV5CMAT* for the computer-based mathematics assessment and *PV1CREA* to *PV5CREA* for digital reading assessment. For the four mathematics content subscales, *change and relationships, quantity, space and shape, uncertainty and data,* the plausible values variables are *PV1MACC* to *PV5MACC, PV1MACQ* to *PV5MACQ, PV1MACS* to *PV5MACS, and PV1MACU* to *PV5MACU* respectively. For the three mathematics process subscales *employ, formulate and interpret*, the plausible values variables are *PV1MAPE* to *PV5MAPE, PV1MAPF* to *PV5MAPF, and PV1MAPI* to *PV5MAPI* respectively.

Constructing conditioning variables

The PISA conditioning variables are prepared using procedures based on those used in the United States National Assessment of Educational Progress (Beaton, 1987) and in IEA Third International Mathematics and Science Study

(Macaskill, Adams and Wu, 1998). All available student-level information, other than their responses to the items in the booklets, is used either as direct or indirect regressors in the conditioning model. The preparation of the variables for the conditioning proceeds as follows.

Variables for booklet identifier were represented by deviation contrast codes and were used as direct regressors. Each booklet was represented by one variable, except for reference booklet 13. Booklet 13 was chosen as reference booklet because it included items from all domains. The difference between simple contrast codes that were used in PISA 2000 and PISA 2003 is that with deviation contrast coding the sum of each column is zero (except for the UH – one hour – booklet), whereas for simple contrast coding the sum is one. The contrast coding scheme is given in Annex B. In addition to the deviation contrast codes, regression coefficients between reading or science and the booklet contrasts that represent booklets without science or reading were fixed to zero. The combination of deviation contrast codes and fixing coefficients to zero resulted in an intercept in the conditioning model that is the grand mean of all students that responded to items in a domain if only the booklet is used as independent variable. This way, the imputation of abilities for students that did not respond to any science or reading items is based on information from all booklets that have items in a domain and not only from the reference booklet as in simple contrast coding.

Other direct variables in the regression are gender (and missing gender if there are any) and deviation contrast codes for schools with the largest school as reference school, grade, mother and father *ISEI* (International Socio-Economic Index). All other categorical variables from the student, ICT (information and communication technology), ECQ (educational career questionnaire) and parent questionnaire were dummy coded. These dummy variables and all numeric variables (the questionnaire indices) were analysed in a principal component analysis. The details of recoding the variables before the principal component analysis are listed in Annex B. The number of component scores that were extracted and used in the scaling model as indirect regressors was country specific and explained 95% of the total variance in all the original variables.

The item-response model was fitted to each national data set and the national population parameters were estimated using item parameters anchored at their international location, the direct and indirect conditioning variables described above and fixed regression coefficients between booklet codes and the minor domains that were not included in the corresponding booklet.

For the countries with very large samples over 10 000 students the sample was divided into smaller data sets using either stratification variables or any other national variable that allow clearly identify distinct groups for example test of the language variable was used in some countries.

Given that the CBA reporting scale cannot influence the PISA paper-based assessment, it was suggested that the plausible values for computer-based assessment countries are drawn in two steps. The first model is a three-dimensional model with reading, mathematics and science. This model was used to estimate the regression coefficients for the background variables for three main domains. Subsequently final plausible values for all domains have been drawn from a four or six dimensional models including computer-based assessment and anchoring regression coefficients to the parameters from the three-dimensional paper-based model.

All students from schools that are sampled for computer-based assessment received plausible values for paper-based PISA and plausible values for computer-based assessment.

Five multi-dimensional scaling models described above were estimated.

BOOKLET EFFECTS

As with PISA 2003, PISA 2006 and PISA 2009, the PISA 2012 test design was balanced, so that the item parameter estimates that are obtained from scaling are not influenced by a booklet effect, as was the case in PISA 2000. However, due to the different location of domains within each of the booklets it was expected that there would still be booklet influences on the estimated proficiency distributions.

Modelling the order effect in terms of item positions in a booklet or at least in terms of cluster positions in a booklet would result in a very complex model. For the sake of simplicity in the international scaling, the effect was modelled separately for each domain at the booklet level, as in previous cycles.

To correct the student mathematics, reading and science scores for the booklet effects, two alternatives were considered:

- correcting all students' scores using one set of the internationally estimated booklet parameters; or
- correcting the students' scores using nationally estimated booklet parameters for each country.

When choosing between these two alternatives a number of issues were considered. First, it is important to recognise that the sum of the booklet correction values is zero for each domain, so the application of either of the above corrections does not change the country means or rankings. Second, if a national correction was applied then the booklet means will be the same for each domain within countries. As such, this approach would incorrectly remove a component of expected sampling and measurement error variation. Third, the booklet corrections are essentially an additional set of item parameters that capture the effect of the item locations in the booklets. In PISA all item parameters are treated as international values so that all countries are therefore treated in exactly the same way. Perhaps the following scenario best illustrates the justification for this. Suppose students in a particular country found the reading items on a particular booklet surprisingly difficult, even though those items have been deemed as central to the PISA definition of PISA literacy and have no technical flaws, such as a translation or coding error. If a national correction were used then an adjustment would be made to compensate for the greater difficulty of these items in that particular country. The outcome would be that two students from different countries who responded in the same way to these items would be given different proficiency estimates. This differential treatment of students based upon their country has not been deemed as suitable in PISA. Moreover this form of adjustment would have the effect of masking real underlying differences in literacy between students in those two countries, as indicated by those items.

Applying an international correction was therefore deemed the most desirable option from the perspective of cross-national consistency.

When estimating the item parameters, booklet effects were included in the measurement model to prevent confounding item difficulties and booklet effects. For the *ConQuest* model statement, the calibration model was:

item + item*step + booklet.

The booklet parameter, formally defined in the same way as item parameters, reflects booklet difficulty. This calibration model was used to estimate the international item parameters for mathematics, reading and science.

The booklet parameters obtained from this analysis were not used to correct for the booklet effect. Instead, a set of booklet parameters for the standard booklets was obtained by scaling the entire data set of equally weighted countries using booklet as a conditioning variable. The students who responded to the UH booklet were excluded from the estimation. A set booklet parameter for the easy booklets was obtained by scaling the entire set of equally weighted countries that opted to use an easy booklet set, using booklet as a conditioning variable.

The booklet parameter estimates obtained are reported in Chapter 12. The booklet effects are the amount that must be added to or subtracted from the proficiencies of students who responded to each booklet.

As the computer-based assessment test was balanced and only included two clusters of 20 minutes it was found that it was unnecessary to add a set of booklet parameters to the model and estimate a booklet effect.

DEVELOPING COMMON SCALES FOR THE PURPOSES OF TRENDS

The reporting scales that were developed for each of the domains of reading, mathematics and science in PISA 2000 were linear transformations of the natural logit metrics that result from the scaling as described above. The transformations were chosen so that the mean and standard deviation of the PISA 2000 scores was 500 and 100 respectively, for the equally weighted 27 OECD countries that participated in PISA 2000 that had acceptable response rates (Wu and Adams, 2002).

For PISA 2003, the decision was made to report the reading and science scores on these previously developed scales. That is, the reading and science reporting scales used for PISA 2000 and PISA 2003 are directly comparable. The value of 500, for example, has the same meaning as it did in PISA 2000.

For mathematics this was not the case, however. Mathematics, as the major domain, was the subject of major development work for PISA 2003, and the PISA 2003 mathematics assessment was much more comprehensive than the PISA 2000 mathematics assessment – the PISA 2000 assessment covered just two (*space and shape*, and *change and relationships*) of

the four areas that are covered in PISA 2003. Because of this broadening in the assessment it was deemed inappropriate to report the PISA 2003 mathematics scores on the same scale as the PISA 2000 mathematics scores. For mathematics the linear transformation of the logit metric was chosen such that the mean was 500 and standard deviation 100 for the 30 OECD countries that participated in PISA 2003. For PISA 2006 the decision was made to report the reading on these previously developed scales. That is the reading reporting scales used for PISA 2000, PISA 2003 and PISA 2006 are directly comparable. Mathematics reporting scales are directly comparable for PISA 2003 and PISA 2006. For science a new scale was established in 2006. The metric for that scale was set so that the mean was 500 and standard deviation 100 for the 30 OECD countries that participated in PISA 2006.

To permit a comparison of the PISA 2006 science results with the science results in previous data collections a science link scale was prepared. The science link scale provides results for 2003 and 2006 using only those items that were common to the two PISA studies. These results are provided in a separate database.

For PISA 2009, the decision was made to report the reading, mathematics and science scores on these previously developed scales. That is the reading scales used for PISA 2000, PISA 2003. PISA 2006 and PISA 2009 are directly comparable. PISA 2009 mathematics reporting scale is directly comparable to PISA 2003 and PISA 2006 and the science reporting scale is directly comparable to PISA 2006 scale.

Again for PISA 2012 the decision was made to report the reading, mathematics and science scores on these previously developed scales. That is the reading scales used for PISA 2000, PISA 2003. PISA 2006, PISA 2009 and PISA 2012 are directly comparable. PISA 2012 mathematics reporting scale is directly comparable to PISA 2003, PISA 2006 and PISA 2009 and the science reporting scale is directly comparable to PISA 2006 and PISA 2009 scale.

Further details on the various PISA reporting scales are given in Chapter 12.

Linking PISA 2012 for mathematics, reading, science and digital reading

The linking of PISA 2012 mathematics, reading and science to the existing scales was undertaken using standard common item equating methods.

The steps involved in linking the PISA 2009 and PISA 2012 scales were as follows:

Step 1: Item parameter estimates were obtained from the PISA 2012 calibration sample.

Step 2: A shift constant was computed to place the above item parameters estimates on the PISA 2009 scale so that the mean of the item parameter estimates for the common items was the same in 2012 as it was in 2009.

Step 3: The 2012 student abilities were estimated with item parameters anchored at their 2012 values.

Step 4: The above estimated student abilities were transformed with the shift computed in Step 2.

Note that this is a much simpler procedure than that which was employed in linking the reading and science between PISA 2003 and PISA 2000. The simpler procedure could be used on this occasion because the test design was balanced since PISA 2003 onwards.

Uncertainty in the link

In each case the transformation that equates the 2012 data with previous data depends upon the change in difficulty of each of the individual link items and as a consequence the sample of link items that have been chosen will influence the choice of transformation. This means that if an alternative set of link items had been chosen the resulting transformation would be slightly different. The consequence is an uncertainty in the transformation due to the sampling of the link items, just as there is an uncertainty in values such as country means due to the use of a sample of students.

The uncertainty that results from the link-item sampling is referred to as linking error and this error must be taken into account when making certain comparisons between the results from different PISA data collection. Just as with the error that is introduced through the process of sampling students, the exact magnitude of this linking error cannot be determined. We can, however, estimate the likely range of magnitudes for this error and take this error into account when interpreting PISA results. As with sampling errors, the likely range of magnitude for the errors is represented as a standard error.

In PISA 2003 the link error was estimated as follows.

Let $\hat{\delta}_i^{2000}$ be the estimated difficulty of link *i* in PISA 2000 and let $\hat{\delta}_i^{2003}$ be the estimated difficulty of link *i* in PISA 2003, where the mean of the two sets of difficulty estimates for all of the link items for a domain is set at zero. We now define the value:

9.19

$$c_i = \hat{\delta}_i^{2003} - \hat{\delta}_i^{2000}$$

The value c_i is the amount by which item *i* deviates from the average of all link items in terms of the transformation that is required to align the two scales. If the link items are assumed to be a random sample of all possible link items and each of the items is counted equally then the link error can be estimated as follows:

9.20

$$error_{2000,2003} = \sqrt{\frac{1}{L}\sum c_i^2}$$

Where the summation is over the link items for the domain and *L* is the number of link items.

Monseur and Berezner (2007) have shown that this approach to the link error estimation is inadequate in two regards. First, it ignores the fact that the items are sampled as units and therefore a cluster sample rather than a simple random sample of items should be assumed. Secondly, it ignores the fact that partial credit items have a greater influence on students' scores than dichotomously scored items. As such, items should be weighted by their maximum possible score when estimating the equating error.

To improve the estimation of the link error the following improved approach has been used in PISA 2006. Suppose we have *L* link items in *K* units. Use *i* to index items in a unit and *j* to index units so that $\hat{\delta}_{ij}^{y}$ is the estimated difficulty of item *i* in unit *j* for year *y*, and let

9.21

$$c_{ij} = \hat{\delta}_{ij}^{2006} - \hat{\delta}_{ij}^{2003}$$

The size (total number of score points) of unit *j* is m_j so that:

$$\sum_{j=1}^{K} m_j = L \quad \text{and} \quad \bar{m} = \frac{1}{K}\sum_{j=1}^{K} m_j$$

Further let:

9.22

$$c_{\cdot j} = \frac{1}{m_i}\sum_{i=1}^{m_j} c_{ij} \quad \text{and} \quad \bar{c} = \frac{1}{K}\sum_{j=1}^{K} c_{\cdot j}$$

and then the link error, taking into account the clustering is as follows:

9.23

$$error_{2006,2003} = \sqrt{\frac{\sum_{j=1}^{K} m_j^2 \left(c_{\cdot j} - \bar{c}\right)^2}{K(K-1)\bar{m}^2}}$$

The PISA 2006 approach for estimating the link errors was used again in PISA 2009.

The most obvious example of a situation where there is a need to use linking error is in the comparison of the mean performance for a country between two PISA data collections. For example, let us consider a comparison between 2003 and 2009 of the performance of Norway in mathematics. The mean performance of Norway in 2003 was 495 with a standard error of 2.38, while in 2009 the mean was 498 with a standard error of 2.40.

The standard error on this difference, as mentioned above, is influenced by the linking error. The standard error is therefore equal to:

9.24

$$SE = \sqrt{\sigma^2_{\bar{x}_{2003}} + \sigma^2_{\bar{x}_{2009}} + error^2_{2003,2009}}$$
$$SE = \sqrt{2.38^2 + 2.40^2 + 1.99^2} = 3.92$$

The standardised difference in the Norwegian mean is 0.71, which is computed as follows:

$$0.71 = \frac{498 - 495}{3.92}$$

and is not statistically significant (absolute values less than 1.96 are not statistically significant at the 95% level of confidence).

In PISA 2012 the same method was used for link errors involving only 2 survey administrations. If however we are considering trends over more than 2 survey administrations, we should consider in addition the covariance between link errors. For example, consider now the correlation of linking errors for 2006 and 2009 referred to 2003. Let $c_{ij} = \hat{\delta}_{ij}^{2006} - \hat{\delta}_{ij}^{2003}$ and $d_{ij} = \hat{\delta}_{ij}^{2009} - \hat{\delta}_{ij}^{2003}$ and suppose we have L scores points for link items common to all 3 cycles in K units, then

9.25

$$\text{cov\_}2003_{2006,2009} = \frac{\sum_{j=1}^{K} m_j^2 \left(c_{.j} - \bar{c}\right)\left(d_{.j} - \bar{d}\right)}{K(K-1)\bar{m}^2}$$

Suppose we are looking at the comparison of mean performance for a country between several PISA survey administrations, say for mathematics from 2003 to 2009. Consider the PISA 2006 and 2009 administrations referred back to PISA 2003. The standard error of the difference between 2006 and 2003 will be

9.26

$$SE(diff_{2003,2006}) = \sqrt{error^2_{2003,2006} + \sigma^2_{2003} + \sigma^2_{2006}}$$

where $error_{2003,2006}$ is the linking error from above and σ_{2003} is the standard error of the mean. An analogous result is obtained for the differences between 2009 and 2003. The covariance between any of these two differences will involve the covariance between the link errors as given above and the standard error of 2003, so the covariance between the differences 2006 versus 2003 and 2009 versus 2003 will be

9.27

$$Cov(diff_{2003,2006}, diff_{2003,2009}) = cov\_2003_{2006,2009} + \sigma^2_{2003}$$

with similar results for the other two covariances between the differences. Suppose we wished to test if the sizes of two consecutive trends say from 2003 to 2006, and 2006 to 2009, were significantly different. We use:

9.28

$$\delta_{2009} - \delta_{2006} = (\delta_{2009} - \delta_{2003}) - (\delta_{2006} - \delta_{2003})$$

so that the link errors are referred back to 2003:

9.29

$$\begin{aligned} Cov(diff_{2003,2006}, diff_{2006,2009}) &= E((\varepsilon_{2006} - \varepsilon_{2003} + \delta_{2006} - \delta_{2003})(\varepsilon_{2009} - \varepsilon_{2006} + \delta_{2009} - \delta_{2006})) \\ &= E((\varepsilon_{2006} - \varepsilon_{2003})(\varepsilon_{2009} - \varepsilon_{2006})) + E((\delta_{2006} - \delta_{2003})(\delta_{2009} - \delta_{2003})) - E((\delta_{2006} - \delta_{2003})(\delta_{2006} - \delta_{2003})) \\ &= -\sigma^2_{2006} + Cov\_2003_{2006,2009} - error^2_{2003,2006} \end{aligned}$$

Now the variance and standard error of the difference between the 2 trends will be

9.30

$$\begin{aligned} Var(diff_{2006,2009} - diff_{2003,2006}) &= Var(diff_{2006,2009})^2 + Var(diff_{2003,2006})^2 - 2Cov(diff_{2006,2009}, diff_{2003,2006}) \\ &= \sigma^2_{2006} + \sigma^2_{2009} + error^2_{2006,2009} + \sigma^2_{2003} + \sigma^2_{2006} + error^2_{2003,2006} \\ &\quad - 2(-\sigma^2_{2006} - error^2_{2003,2006} + \text{cov}\_2003_{2006,2009}) \\ &= 4\sigma^2_{2006} + \sigma^2_{2009} + \sigma^2_{2003} + error^2_{2006,2009} + 3error^2_{2003,2006} - 2\text{cov}\_2003_{2006,2009} \end{aligned}$$

9.31

$$SE(diff_{2006,2009} - diff_{2003,2006}) = \sqrt{4\sigma^2_{2006} + \sigma^2_{2009} + \sigma^2_{2003} + error^2_{2006,2009} + 3error^2_{2003,2006} - 2\text{cov}\_2003_{2006,2009}}$$

For example, consider Greece in 2003 to 2009, where we have for Mathematics in 2003 the mean performance 445 with standard error 3.9, 459 and 3.0 in 2006 and 466 and 3.9 in 2009.

So here the difference of the trends is (466-459)-(459-445) =-7. The standard error, using link errors and covariances from Chapter 12, will be

$$\begin{aligned} SE &= \sqrt{4(3.0)^2 + (3.9)^2 + (3.9)^2 + (1.33)^2 + 3(1.35)^2 - 2(2.340)} \\ &= \sqrt{68.97} \\ &= 8.31 \end{aligned}$$

So the test statistic in this case is -7/8.31 = .84, which is not significant at the 5% level. In this case, although the link errors reduce the size of the statistic, the standard errors of the mean are relatively so large that the result of the test would not be changed if they were ignored.

In PISA a common transformation has been estimated, from the link items, and this transformation is applied to all participating countries. It follows that any uncertainty that is introduced through the linking is common to all students and all countries. Thus, for example, suppose the *unknown* linking error (between PISA 2006 and PISA 2009) in reading resulted in an over-estimation of student scores by two points on the PISA 2006 scale. It follows that every student's score will be over-estimated by two score points. This over-estimation will have effects on certain, but not all, summary statistics computed from the PISA 2009 data. For example, consider the following:

- Each country's mean will be over-estimated by an amount equal to the link error, in our example this is two score points.
- The mean performance of any subgroup will be over-estimated by an amount equal to the link error, in our example this is two score points.
- The standard deviation of student scores will not be effected because the over-estimation of each student by a common error does not change the standard deviation.
- The difference between the mean scores of two countries in PISA 2009 will not be influenced because the over-estimation of each student by a common error will have distorted each country's mean by the same amount.
- The difference between the mean scores of two groups (e.g. males and females) in PISA 2009 will not be influenced, because the over-estimation of each student by a common error will have distorted each group's mean by the same amount.
- The difference between the performance of a group of students (e.g. a country) between PISA 2006 and PISA 2009 will be influenced because each student's score in PISA 2006 will be influenced by the error.

- A change in the difference in performance between two groups from PISA 2006 to PISA 2009 will not be influenced. This is because neither of the components of this comparison, which are differences in scores in 2009 and 2006 respectively, is influenced by a common error that is added to all student scores in PISA 2009.

In general terms, the linking error need only be considered when comparisons are being made between results from different PISA data collections, and then usually only when group means are being compared.

Link error for other types of comparisons of student performance

The link error for other comparisons of performance does not have a straightforward theoretical solution as does the link error for comparison between two PISA assessments. The link error between two PISA assessments, described above, can be used, however, to empirically estimate the magnitude of the link error for the comparison of the percentage of students in a particular proficiency level or the magnitude of the link error associated with the estimation of the annualised and curvilinear change.

The empirical estimation of these link errors uses the assumption that the magnitude of the link error follows a normal distribution with mean 0 and a standard deviation equal to the link error or comparisons of performance between PISA 2012 and previous assessments. From this distribution, 500 errors are drawn and added to the first plausible value for each assessment prior to 2012. The estimate of interest (change in the percentage of students in a particular proficiency level or the annualised change) is calculated for each of the 500 replicates. The standard deviation of these 500 estimates is then used as the link error for the annualised change, the quadratic change, and the change in the percentage of students scoring in a particular proficiency level. For further details on these link errors, see OECD (2014), *PISA 2012 Results: What Students Know and Can Do, Student Performance in Mathematics, Reading and Science (Volume I, Revised edition)*, Annex A5.

Note

1. The value *M* should be large. For PISA we have used 2000.

2. This is a fictitious item.

3.The samples used were simple random samples stratified by the explicit strata used in each country. Students who responded to the UH booklet were not included in this process.

References

Adams, R.J., M. Wilson and **W.C. Wang** (1997), "The Multidimensional Random Coefficients Multinomial Logit Model", *Applied Psychological Measurement*, No. 21, pp. 1-23.

Adams, R.J. and **M.L. Wu** (2002), *PISA 2000 Technical Report*, OECD Publishing, Paris.

Beaton, A.E. (1987), *Implementing the New Design: The NAEP 1983-84 Technical Report* (Rep. No. 15-TR-20), Educational Testing Service, Princeton, NJ.

Ebel, R.L. and **D.A. Frisbie** (1986), *Essentials of Education Measurement*, Prentice Hall, Englewood Cliffs, New Jersey.

Macaskill, G., R.J. Adams and **M.L. Wu** (1998), "Scaling Methodology and Procedures for the Mathematics and Science Literacy, Advanced Mathematics and Physics Scale", in M. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study*, *Technical Report Volume 3: Implementation and Analysis*, Boston College, Chestnut Hill, MA.

Masters, G.N. (1982), "A Rasch Model for Partial Credit Scoring", *Psychometrika*, No. 47(2), pp. 149-174.

Mislevy, R.J. (1991), "Randomization-based Inference about Latent Variables from Complex Samples", *Psychometrika*, No. 56, pp. 177-196.

Mislevy, R.J., A. Beaton, B.A. Kaplan and **K. Sheehan** (1992), "Estimating Population Characteristics from Sparse Matrix Samples of Item Responses", *Journal of Educational Measurement*, No. 29(2), pp. 133-161.

Mislevy, R.J. and **K.M. Sheehan** (1987), "Marginal Estimation Procedures", in A.E. Beaton (ed.), *The NAEP 1983-84 Technical Report, National Assessment of Educational Progress*, Educational Testing Service, Princeton, pp. 293-360.

Monseur, C. and **A. Berezner** (2007), "The Computation of Equating Errors in International Surveys in Education", *Journal of Applied Measurement*, No. 8(3), pp. 323-335.

OECD (2014), *PISA 2012 Results: What Students Know and Can Do (Volume I, Revised edition, February 2014): Student Performance in Mathematics, Reading and Science*, PISA, OECD Publishing, Paris. *http://dx.doi.org/10.1787/9789264208780-en*

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

Wu, M.L., R.J. Adams and **M.R. Wilson** (1997), *ConQuest*: *Multi-Aspect Test Software* [computer programme manual], Australian Council for Educational Research, Camberwell, Victoria.