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 & Wilson, 1997).

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, ξ, while the student outcome levels (the latent variable), θ, 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 $\xi^T = (\xi_1, \xi_2,\ldots, \xi_p)$, 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. D, 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 $\mathbf{A}^T = (\mathbf{a}_{11}, \mathbf{a}_{12},\ldots, \mathbf{a}_{1K_1}, \mathbf{a}_{21}, \ldots, \mathbf{a}_{2K_2},\ldots, \mathbf{a}_{IK_I})$, 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 $\theta = (\theta_1, \theta_2,\ldots, \theta_D)'$, 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 $\mathbf{B} = \left(\mathbf{B}_1^T, \mathbf{B}_2^T,\ldots, \mathbf{B}_I^T\right)^T$ for the entire test. (The score for a response in the zero category is zero, but other responses may also be scored zero.)

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

9.2

$$\Pr(\mathrm{X}_{ij} = 1; \mathrm{A}, \mathrm{B}, \xi \mid \theta) = \frac{\exp\left(\mathrm{b}_{ij}\theta + \mathrm{a}'_{ij}\xi\right)}{\sum_{k=1}^{K_i} \exp\left(\mathrm{b}_{ik}\theta + \mathrm{a}'_{ik}\xi\right)}.$$

For a response vector, we have:

9.3

$$f(\mathrm{x}; \xi \mid \theta) = \psi(\theta, \xi) \exp\left[\mathrm{x}' (\mathrm{B}\theta + \mathrm{A}\xi)\right]$$

with

9.4

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

where Ω 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, θ**.** The complete definition of the model, therefore, requires the specification of a density, $f_\theta(\theta, \alpha)$ for the latent variable, θ. Let α symbolise a set of parameters that characterise the distribution of θ**.** 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

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

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, $\mathbf{Y}_n^T\beta$, 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

$$\theta_n = Y_n^T\beta + E_n$$

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

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

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

The generalisation needs to be taken one step further to apply it to the vector-valued θ rather than the scalar-valued θ. The extension results in the multivariate population model:

9.9

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

where γ is a $u \times d$ matrix of regression coefficients, Σ is a $d \times d$ variance-covariance matrix, and $\mathbf{W}_n$ is a $u \times 1$ vector of fixed variables.

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

Combined model

In [9.10], the conditional item response model [9.3] and the population model [9.9] are combined to obtain the unconditional, or marginal, item response model:

9.10

$$f_x(x;\xi,\gamma,\Sigma) = \int_\theta f_x(x;\xi \mid \theta)\, f_\theta(\theta;\gamma,\Sigma)\,d\theta \; .$$

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 γ, Σ and ξ.

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

$$h_\theta(\theta_n; W_n, \xi, \gamma, \Sigma \mid x_n) = \frac{f_x(x_n;\xi \mid \theta_n)\; f_\theta(\theta_n; W_n, \gamma, \Sigma)}{f_x(x_n; W_n, \xi, \gamma, \Sigma)}$$

$$= \frac{f_x(x_n;\xi \mid \theta_n)\; f_\theta(\theta_n; W_n, \gamma, \Sigma)}{\int_{\theta_n} f_x(x_n;\xi \mid \theta_n)\; f_\theta(\theta_n; W_n, \gamma, \Sigma)} \; .$$

APPLICATION TO PISA

In PISA, this model 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 is used in conjunction with the population model , but conditioning variables are not used. That is, it is assumed that students have been sampled from a multivariate normal distribution.

Two five-dimensional scaling models were used in the PISA 2006 main study. The first model, made up of one reading, one science, one mathematics and two attitudinal dimensions, was used for reporting overall scores for reading, science, mathematics and two attitudinal scales. A second model, made up of one reading, one mathematics and three science dimensions, was used to generate scores for the three science scales.

The design matrix was chosen so that the partial credit model was used for items with multiple score categories and the simple logistic model was fit to the dichotomously scored items.

National calibrations

National calibrations were performed separately, country by country, using unweighted data. 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 regarded as not-administered in particular countries if it has poor psychometric characteristics in those countries but functions well in the vast majority of others. If an item is identified as behaving differently in different countries, the second option will have the same impact on inter-country comparisons.

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 (Infit Mean Square)

For each item parameter, the *ConQuest*® fit mean square statistic 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. An infit mean square greater than one is often associated with a low discrimination index, and an infit mean square less than one is often associated with a high discrimination index.

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 0.25. 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 be lower than the point-biserial correlation of 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.

National reports

After national scaling, four reports were returned to each participating country to assist in reviewing their data with the consortium.

Report 1: 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 (*see* Figure 9.1).

The first column in the table, *Label,* shows each of the possible response categories for the item. For this particular multiple-choice item, relevant categories were 1, 2, 3, 4 (the multiple-choice response categories), 8 (invalid, usually double responses) and 9 (missing).

The second column indicates the score assigned to the different categories. For this item, score 1 was allocated for the category 2 (the correct response for this multiple-choice item). Categories 1, 3, 4, 8 and 9 each received a score of 0. In this report non-reached values were treated as not administered, because this report provides information at the item calibration stage. Therefore, non-reached values are not included in this table.

The columns *Count* and *% of tot* show the number and percentage of students who responded to each category. For example, in this country, 138 students, or 38.87%, responded to *S423Q01* correctly and received score 1.

The next three columns, *Pt Bis*, *t*, and *(p)*, represent the point-biserial correlation between success on the item and a total score, the *t*-statistics associated with the point-biserial correlation and *p*-value for the *t*-statistics, respectively.

The two last columns, *PV1Avg:1* and *PV1 SD:1,* show the average ability of students responding in each category and the associated standard deviation. The average ability is calculated by domain. In this example the average ability of those students who responded correctly (category 2) is 0.12, while the average ability of those students who responded incorrectly (categories 1, 3 and 4) are –0.30, 0.07 and –0.41, respectively. Average ability of those students who selected distracter 3 for this item (0.07) is similar to the average ability of the students who selected the correct response 2. This suggests close checking of distracter three.

Figure 9.1

Example of item statistics in Report 1

```
Item:70 (S423Q01)
Cases for this item     355   Discrimination  0.13
Item Threshold(s):      0.49  Weighted MNSQ   1.17
Item Delta(s):      0.49
------------------------------------------------------------------------------
 Label     Score     Count   % of tot  Pt Bis      t  (p)    PV1Avg:1 PV1 SD:1
------------------------------------------------------------------------------
   0                   0       0.00      NA       NA (.000)     NA        NA
   1        0.00      65      18.31    -0.16    -3.02(.003)  -0.30      0.78
   2        1.00     138      38.87     0.13     2.54(.011)   0.12      0.89
   3        0.00     115      32.39     0.09     1.76(.080)   0.07      0.83
   4        0.00      26       7.32    -0.08    -1.44(.152)  -0.41      0.83
   5                   0       0.00      NA       NA (.000)     NA        NA
   6                   0       0.00      NA       NA (.000)     NA        NA
   8        0.00       4       1.13    -0.06    -1.19(.233)  -0.62      0.79
   9        0.00       7       1.97    -0.15    -2.87(.004)  -0.76      0.58
=============================================================================
```

Report 2: Summary of descriptive statistics by item

Report 2 provided descriptive statistics and comparisons of national and international parameters by item. An example of this report for the item *S478Q01* is shown in Figure 9.2.

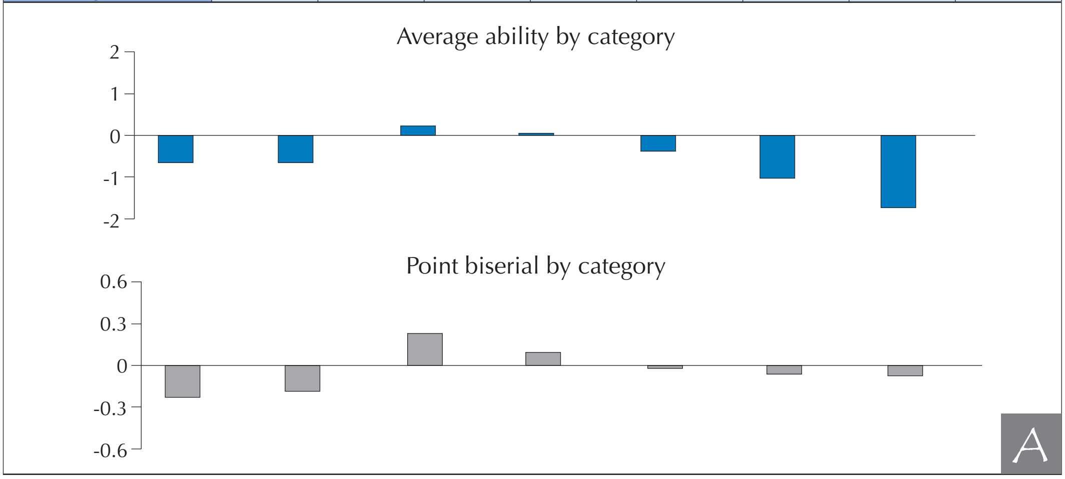
Figure 9.2

Example of item statistics in Report 2

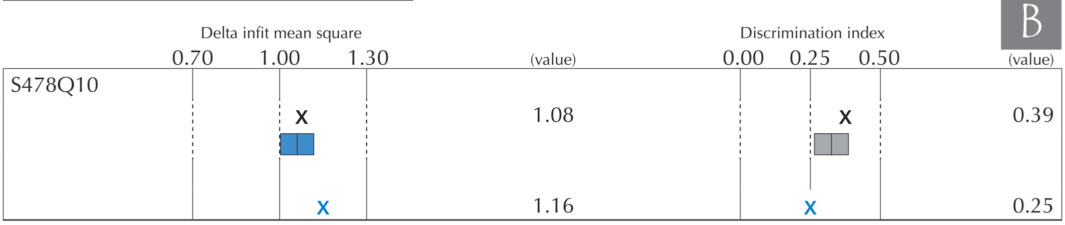
PISA 2006 Main Study: item details, Science – *S478Q01*

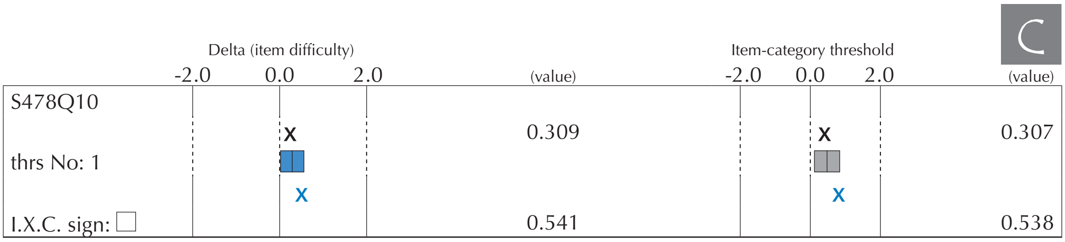
Response Frequencies

Category	**1**	**2**	**3**	**4**	**8**	**9**	**r**	**Total**
Number of students	256	145	544	467	25	9	5	1 448
Percentage	18	10	38	32	2	1	0	



ID: S478Q10	Discrimination: 0.25
Name: Antibiotics Q1	Key: 3





D

	Item by country interactions			**Discrimination**				**PISA 2003 link items**	
	Number of valid response	Easier than expected	Harder than expected	Non-key PB is positive	Key PB is negative	Low discrimination	Ability not ordered	Link items	Requires checking
S478Q10	1 443	□	□	☑	□	□	□	□	□

In this example, the graph marked with the letter A displays the statistics from Report 1 in a graphical form. The table above graph A shows the number and percentage of students in each response category, as shown in the columns *Label*, *Count* and *% of tot* in Report 1. The categories (1, 2, 3, 4, 8, 9 and *r*) are shown under each of the bar charts. An additional category, *r*, is included to indicate the number of students who did not reach the item.

The graph marked with A in Figure 9.4 facilitates the process of identifying the following anomalies:

- A non-key category has positive point-biserial or a point-biserial higher than the key category;
- A key category has a negative point-biserial;
- In the case of partial-credit items, checks can be made on whether the average ability (and the point-biserial) increases with score points.

For example, category 4 was circled by 461 students (32%) and has positive point biserial.

The initial national scaling provides the following item statistics for each country and for each item:

- Delta infit mean square;
- Discrimination index;
- Difficulty estimate (delta); and
- Thresholds.

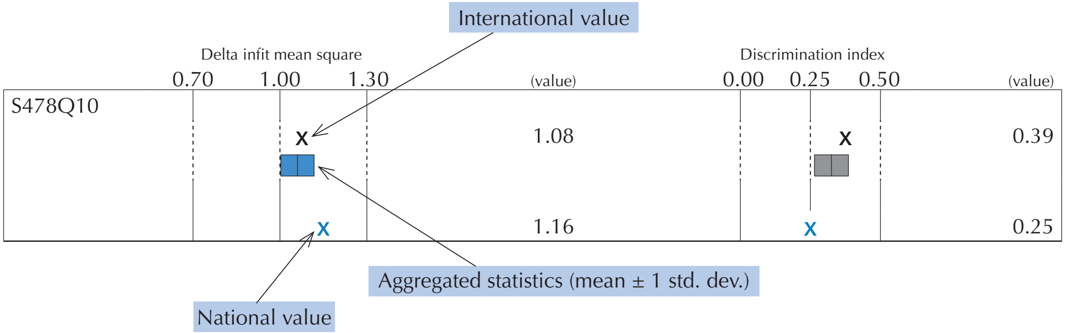
Graph B (*see* Figure 9.3) and Graph C (*see* Figure 9.4) of Report 2 present the above statistics for each item in three different forms.

- National value, calculated for country;
- Average of national values across all countries (vertical line within the shaded box);
- International value calculated for all countries scaled together.

Graph B presents a comparison of the delta infit mean square statistic and the discrimination index.

Figure 9.3

Example of item statistics shown in Graph B



Graph C presents a comparison of the item difficulty parameters and the thresholds.

Substantial differences between the national value and the international value or the national value and the mean show that the item is behaving differently in that country in comparison with all other countries. This may be an indication of a mistranslation or some other problem.

Figure 9.4
Example of item statistics shown in Graph C

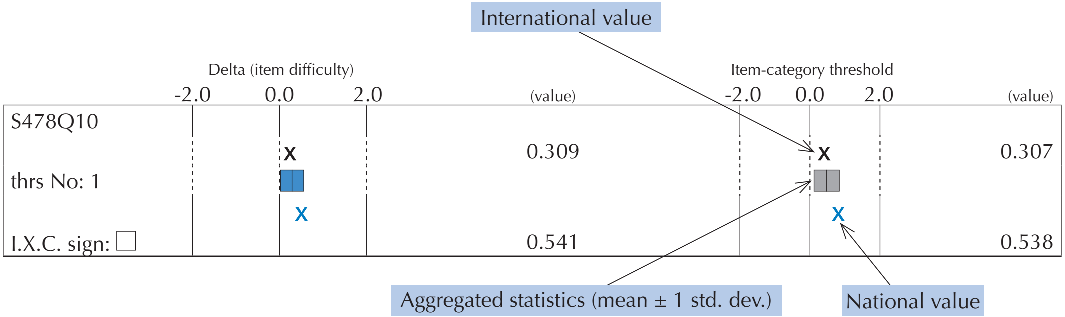


Table D (see Figure 9.5) indicates if an item is a dodgy item for the national dataset, *i.e.* an item that was flagged for one of the following reasons:

- The item difficulty is significantly lower than the average of all available countries;
- The item difficulty is significantly higher than the average of all available countries;
- 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 10 students);
- The key category point-biserial correlation is lower than –0.05;
- The item discrimination is lower than 0.2;
- The category abilities for partial credit items are not ordered;
- Link item difficulty is different from the PISA 2003 main study national item difficulty. ("Link item" box indicates if an item is a link item. "Requires checking" box is ticked when the link item performed differently in Pisa2006 main study. Only relevant to the countries that participated in both PISA cycles).

In this example item *S478Q01* was flagged as having a positive point-biserial for a non-key category.

Figure 9.5
Example of item statistics shown in Table D

	Item by country interactions			**Discrimination**				**PISA 2003 link items**	
	Number of valid response	Easier than expected	Harder than expected	Non-key PB is positive	Key PB is negative	Low discrimination	Ability not ordered	Link items	Requires checking
S478Q10	1 443	□	□	☑	□	□	□	□	□

Report 3a: national summary of dodgy items

Report 3a summarises the dodgy items for each country as listed in report 2 section D (*see* Figure 9.6).

Figure 9.6
Example of summary of dodgy items for a country in Report 3a

PISA 2006 Main Study, Report 3a: Science dodgy items

	Item by country interactions			Discrimination				PISA 2003 link items	
	Number of valid responses	Easier than expected	Harder than expected	Non-key PB is positive	Key PB is negative	Low discrimination	Ability not ordered	Link items	Requires checking
S456Q02	1 437	□	□	□	□	☑	□	□	□
S476Q01	1 482	☑	□	□	□	□	□	□	□
S477Q04	1 442	☑	□	□	□	□	□	□	□
S478Q01	1 443	□	□	☑	□	□	□	□	□
S493Q01	1 452	☑	□	□	□	□	□	□	□
S495Q01	1 442	☑	□	□	□	□	□	□	□
S495Q02	1 440	□	☑	□	□	□	□	□	□
S508Q02	1 435	□	☑	□	□	□	□	□	□
S510Q04	1 459	☑	□	□	□	□	□	□	□
S519Q01	1 438	□	☑	□	□	□	□	□	□
S524Q06	1 427	□	☑	□	□	☑	□	□	□

Report 3b: international summary of dodgy items

Report 3b (see Figure 9.7) provided a summary of dodgy items for all countries included in the analysis. If an item showed poor psychometric properties in a country but also in most of the other countries then it could most likely be explained by reasons other than mistranslation and misprint. Note that item *S478Q01* that has been used as an example in Report 1 and Report 2 was problematic in many countries. It was easier than expected in two countries, harder in three countries, had positive point-biserial for a non-key category in 27 countries and a poor discrimination in 15 out of 58 countries.

Figure 9.7
Example of summary of dodgy items in Report 3b

PISA 2006 Main Study, Report 3: Summary of Science dodgy items – Number of countries: 58

	Item by country interactions		Discrimination				Fit	
	Easier than expected	Harder than expected	Non-key PB is positive	Key PB is negative	Low discrimination	Ability not ordered	Small, high dicr. item	Large, low discr. item
S476Q02	1	8	0	0	0	0	0	0
S476Q03	3	2	2	1	1	0	0	1
S477Q01	1	0	0	0	11	0	0	0
S477Q02	1	1	0	0	0	0	0	0
S477Q03	0	1	0	0	0	0	3	0
S477Q04	4	4	0	0	2	0	0	0
S478Q01	2	3	27	0	15	0	0	10
S478Q02	1	1	0	0	1	0	0	0
S478Q03	1	0	0	0	3	0	0	4
S478Q04	1	1	0	0	0	0	1	0
S485Q02	3	4	0	0	0	0	0	2
S485Q03	0	0	0	0	0	0	0	0
S485Q04	3	0	28	0	25	0	0	3
S485Q05	3	1	0	0	0	0	0	2
S485Q08	8	8	0	0	0	0	1	0
S493Q01	7	3	0	0	6	0	0	0
S493Q03	7	6	0	0	19	0	0	4
S493Q04	10	2	0	0	2	0	1	0

International calibration

International item parameters were set by applying the conditional item response model (9) in conjunction with the multivariate population model (15), without using conditioning variables, to a sub-sample of students. This subsample of students, referred to as the international calibration sample, consisted of 15 000 students comprising 500 students drawn at random from each of the 30 participating OECD countries[1].

The allocation of each PISA item to one of the five PISA 2006 scales is given in Appendix 1.

Student score generation

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.

Plausible values

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, , 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.

M vector-valued random deviates, $\{\varphi_{mn}\}_{m=1}^{M}$, from the multivariate normal distribution, $f_\theta(\theta_n; \mathrm{W}_n, \gamma, \Sigma)$, for each case *n*.[2] These vectors are used to approximate the integral in the denominator of , using the Monte-Carlo integration

9.12

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

At the same time, the values

9.13

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

are calculated, so that we obtain the set of pairs $\left(\varphi_{mn}, {p_{mn}}/{\Im}\right)_{m=1}^{M}$, which can be used as an approximation of the posterior density [9.11]; and the probability that φ_{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; and for each random draw, the vector, φ_{ni_0}, that satisfies the condition

9.15

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

is selected as a plausible vector.

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 TIMSS (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 ID were represented by deviation contrast codes and were used as direct regressors. Each booklet was represented by one variable, except for reference booklet 11. Booklet 11 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 2003 is that with deviation contrast coding the sum of each column is zero (except for the UH booklet), whereas for simple contrast coding the sum is one. The contrast coding scheme is given in Table 0.1. In addition to the deviation contrast codes, regression coefficients between reading or mathematics and the booklet contrasts that represent booklets without mathematics 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 booklet is used as independent variable. This way, the imputation of abilities for students that did not respond to any mathematics 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 simple contrast codes for schools with the largest school as reference school. In PISA 2003 school mean performance in the major domain was used as regressor instead of contrast codes to simplify the model. The intra-class correlation was generally slightly higher in PISA 2006 than in PISA 2003, which is likely to be caused by using school dummy coding instead of school performance means. As expected, using school means slightly underestimates the between school variance.

All other categorical variables from the student, ICT and parent questionnaire were dummy coded. These dummy variables and all numeric variables (the questionnaire indices) were analysed in a principle component analysis. The details of recoding the variables before the principle component analysis are listed in Appendix 2. 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.

Table 9.1
Deviation contrast coding scheme

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	d11	d12	UH
Booklet 1	1	0	0	0	0	0	0	0	0	0	0	0	0
Booklet 2	0	1	0	0	0	0	0	0	0	0	0	0	0
Booklet 3	0	0	1	0	0	0	0	0	0	0	0	0	0
Booklet 4	0	0	0	1	0	0	0	0	0	0	0	0	0
Booklet 5	0	0	0	0	1	0	0	0	0	0	0	0	0
Booklet 6	0	0	0	0	0	1	0	0	0	0	0	0	0
Booklet 7	0	0	0	0	0	0	1	0	0	0	0	0	0
Booklet 8	0	0	0	0	0	0	0	1	0	0	0	0	0
Booklet 9	0	0	0	0	0	0	0	0	1	0	0	0	0
Booklet 10	0	0	0	0	0	0	0	0	0	1	0	0	0
Booklet 11	–1	–1	–1	–1	–1	–1	–1	–1	–1	–1	–1	–1	0
Booklet 12	0	0	0	0	0	0	0	0	0	0	1	0	0
Booklet 13	0	0	0	0	0	0	0	0	0	0	0	1	0
UH	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**

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.

Two models were run, each with five dimensions. The first model included mathematics, reading, science, interest and support. The second model included mathematics, reading and the three science scales. For each domain plausible values were drawn using the method described in the *PISA 2003 Technical Report* (OECD, 2005).

BOOKLET EFFECTS

As with PISA 2003, the PISA 2006 test design was balanced, 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 PISA 2000 and PISA 2003.

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[3].

The calibration model given above was used to estimate the international item parameters. It was estimated using the international calibration sample of 15 000 students, and not-reached items in the estimation were treated as not administered.

The booklet parameters obtained from this analysis were not used to correct for the booklet effect. Instead, a set of booklet parameters was obtained by scaling the entire data set of OECD countries using booklet as a conditioning variable and a senate weight. The students who responded to the UH booklet were excluded from the estimation. The booklet parameter estimates obtained are reported in Chapter 12. The booklet effects are the amount that must be added to the proficiencies of students who responded to each booklet.

To correct the student scores for the booklet effects, two alternatives were considered:

- To correct all students' scores using one set of the internationally estimated booklet parameters; or
- To correct 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.

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 (17). As such, plausible values contain random error variance components and are not 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).

The PISA student file contains 45 plausible values, five for each of the eight PISA 2006 scales. *PV1MATH* to *PV5MATH* are for mathematical literacy; *PV1SCIE* to *PV5SCIE* for *scientific literacy*, *PV1READ* to *PV5READ* for *reading literacy, PV1INTR* to *PV5INTR* for *interest in science* and *PV1SUPP* to *PV5SUPP* for *support for scientific inquiry*. For the three scientific literacy scales, *explaining phenomena scientifically*, *identifying scientific issues, using scientific evidence,* the plausible values variables are *PV1SCIE1* to *PV5SCIE1*, *PV1SCIE2* to *PV5SCIE2, and PV1SCIE3* to *PV5SCIE3,* respectively.

If an analysis were to be undertaken with one of these eight scales, 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 $r(\theta, \mathbf{Y})$ is a statistic that depends upon the latent variable and some other observed characteristic of each student. That is: $(\theta, \mathrm{Y}) = (\theta_1, y_1, \theta_2, y_2, \ldots, \theta_N, y_N)$ 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 $h_\theta\,(\theta_n;\, y_n, \xi, \gamma, \Sigma \mid x_n)$. If $h_\theta\,(\theta;\, \mathrm{Y}, \xi, \gamma, \Sigma \mid \mathrm{X})$ is the joint marginal posterior for $n = 1, \ldots N$ then we can compute:

9.16

$$r^*(\mathrm{X},\mathrm{Y}) = E\left[r^*(\theta,\mathrm{Y})|\mathrm{X},\mathrm{Y}\right]$$

$$= \int_\theta r(\theta,\mathrm{Y})h_\theta(\theta;\mathrm{Y},\xi,\gamma,\Sigma\,|\mathrm{X})d\theta \quad .$$

The integral in can be computed using the Monte-Carlo method. If M random vectors $(\theta_1, \theta_2, \ldots, \theta_M)$ are drawn from $h_\theta\,(\theta; Y, \xi, \gamma, \Sigma \mid X)$ is approximated by:

9.17

$$r^*\,(X,Y) \approx \frac{1}{M}\sum_{m=1}^{M} r\left(\theta_m, Y\right)$$

$$= \frac{1}{M}\sum_{m=1}^{M} \hat{r}_m$$

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

From **[9.16]** we can see that the final estimate of r is the average of the estimates computed using each plausible value in turn. If U_m is the sampling variance for $\hat{r}_m$ then the sampling variance of r^* is:

9.18

$$V = U^* + (1+M^{-1})B_M,$$

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 $r^* \pm t_v\left(\left.(1-\alpha)\right/2\right)V^{1/2}$ where $t_\upsilon(s)$ is the s- percentile of the t-distribution with ν degrees of freedom. $\upsilon = \left[\frac{f_M^2}{M-1} + \frac{(1-f_M)^2}{d}\right]^{-1}$, $f_M = (1 + M^{-1})B_M/V$ 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.

DEVELOPING COMMON SCALES FOR THE PURPOSES OF TRENDS

The reporting scales that were developed for each 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 27 OECD countries that participated in PISA 2000 that had acceptable response rates (Wu & Adams, 2002).[4]

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 – that is, the mean score in 2000 of the sampled students in the 27 OECD countries that participated in PISA 2000.[5]

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.[6]

For PISA 2006 the decision was made to report the reading on these previously developed scales. That is the reading reporting scales used for PISA2000, 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.[7]

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.

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

Linking PISA 2003 and PISA 2006 for reading and mathematics

The linking of PISA 2006 reading and mathematics to the existing scales was undertaken using standard common item equating methods.

The steps involved in linking the PISA 2003 and PISA 2006 reading and mathematics scales were as follows:

Step 1: Item parameter estimates for reading and mathematics where obtained from the PISA 2006 calibration sample.

Step 2: The above item parameters estimates where transformed through the addition of constant, so that the mean of the item parameter estimates for the common items was the same in 2006 as it was in 2003.

Step 3: The 2006 student abilities where estimated with item parameters anchored at their 2006 values.

Step 4: The above estimated students abilities where transformed with the shift estimated in step 2.

Note that this is a much simpler procedure than the 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 for both PISA 2003 and 2006.

Uncertainty in the link

In each case the transformation that equates the 2006 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 in2000 and let $\hat{\delta}_i^{2003}$ be the estimated difficulty of link i in2003, where the mean of the two sets difficulty estimates for all of the link items for a domain is set at zero. We now define the value:

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

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:

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

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 a 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

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

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

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

Further let:

$$c_{\bullet j} = \frac{1}{m_j}\sum_{i=1}^{m_j} c_{ij} \quad \text{and} \quad \overline{c} = \frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{m_j} c_{ij}$$

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

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

The link standard errors are reported in chapter 12.

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 2003 and PISA 2006) in reading resulted in an over-estimation of student scores by two points on the PISA 2003 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 2006 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 2006 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 (eg males and females) in PISA 2006 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 (eg a country) between PISA 2003 and PISA 2006 will be influenced because each student's score in PISA 2003 will be influenced by the error; and finally;
- A change in the difference in performance between two groups from PISA 2003 to PISA 2006 will not be influenced. This is because neither of the components of this comparison, which are differences in scores in 2006 and 2003 respectively, is influenced by a common error that is added to all student scores in PISA 2006.

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.

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 2006 of the performance of Canada in mathematics. The mean performance of Canada in 2003 was 532 with a standard error of 1.8, while in 2006 the mean was 527 with a standard error of 2.0. The standardised difference in the Canadian mean is -1.82, which is computed as follows: $-1.82 = (527 - 532)/\sqrt{2.0^2 + 1.8^2 + 1.4^2}$, and is not statistically significant.

Notes

[1] 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.

[2] The value *M* should be large. For PISA we have used 2000.

[3] Note that because the design was balanced the inclusion of the booklet term in the item response model did not have an appreciable effect on the item parameter estimates.

[4] Using senate weights.

[5] Again using senate weights.

[6] Again using senate weights.

[7] Again using senate weights.

Reader's Guide

Country codes – the following country codes are used in this report:

OECD countries

AUS	Australia
AUT	Austria
BEL	Belgium
BEF	Belgium (French Community)
BEN	Belgium (Flemish Community)
CAN	Canada
CAE	Canada (English Community)
CAF	Canada (French Community)
CZE	Czech Republic
DNK	Denmark
FIN	Finland
FRA	France
DEU	Germany
GRC	Greece
HUN	Hungary
ISL	Iceland
IRL	Ireland
ITA	Italy
JPN	Japan
KOR	Korea
LUX	Luxembourg
LXF	Luxembourg (French Community)
LXG	Luxembourg (German Community)
MEX	Mexico
NLD	Netherlands
NZL	New Zealand
NOR	Norway
POL	Poland
PRT	Portugal
SVK	Slovak Republic
ESP	Spain
ESB	Spain (Basque Community)
ESC	Spain (Catalonian Community)
ESS	Spain (Castillian Community)
SWE	Sweden
CHE	Switzerland
CHF	Switzerland (French Community)
CHG	Switzerland (German Community)
CHI	Switzerland (Italian Community)
TUR	Turkey
GBR	United Kingdom
IRL	Ireland
SCO	Scotland
USA	United States

Partner countries and economies

ARG	Argentina
AZE	Azerbaijan
BGR	Bulgaria
BRA	Brazil
CHL	Chile
COL	Colombia
EST	Estonia
HKG	Hong Kong-China
HRV	Croatia
IDN	Indonesia
JOR	Jordan
KGZ	Kyrgyztan
LIE	Liechtenstein
LTU	Lithuania
LVA	Latvia
LVL	Latvia (Latvian Community)
LVR	Latvia (Russian Community)
MAC	Macao-China
MNE	Montenegro
QAT	Qatar
ROU	Romania
RUS	Russian Federation
SRB	Serbia
SVN	Slovenia
TAP	Chinese Taipei
THA	Thailand
TUN	Tunisia
URY	Uruguay

References

Adams, R.J., Wilson, M. & **Wang, W.C.** (1997), The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement,* No. 21, pp. 1-23.

Adams, R.J., Wilson, M. R. & **Wu, M.L.** (1997), Multilevel item response models: An approach to errors in variables regression, *Journal of Educational and Behavioural Statistics,* No. 22 (1), pp. 46-75.

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

Bollen, K.A. & **Long, S.J.** (1993) (eds.), *Testing Structural Equation Models*, Newbury Park: London.

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

Buchmann, C. (2000), Family structure, parental perceptions and child labor in Kenya: What factors determine who is enrolled in school? *Soc. Forces,* No. 78, pp. 1349-79.

Buchmann, C. (2002), Measuring Family Background in International Studies of Education: Conceptual Issues and Methodological Challenges, in Porter, A.C. and Gamoran, A. (eds.). *Methodological Advances in Cross-National Surveys of Educational Achievement* (pp. 150-97), Washington, DC: National Academy Press.

Creemers, B.P.M. (1994), *The Effective Classroom*, London: Cassell.

Cochran, W.G. (1977), *Sampling techniques,* third edition, New York, NY: John Wiley and Sons.

Ganzeboom, H.B.G., de Graaf, P.M. & **Treiman, D.J.** (1992), A standard international socio-economic index of occupational status, *Social Science Research*, No. 21, pp. 1-56.

Ganzeboom H.B. & **Treiman, D.J.** (1996), Internationally comparable measures of occupational status for the 1988 international standard classification of occupations, *Social Science Research*, No. 25, pp. 201-239.

Grisay, A. (2003), Translation procedures in OECD/PISA 2000 international assessment, *Language Testing,* No. 20 (2), pp. 225-240.

Hambleton, R.K., Swaminathan, H. & **Rogers, H.J.** (1991), *Fundamentals of item response theory*, Newbury Park, London, New Delhi: SAGE Publications.

Hambleton, R.K., Merenda, P.F. & **Spielberger, C.D.** (2005), *Adapting Educational and Psychological Tests for Cross-Cultural Assessment,* IEA Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey.

Harkness, J.A., Van de Vijver, F.J.R. & **Mohler, P.Ph** (2003), *Cross-Cultural Survey Methods,* Wiley-Interscience, John Wiley & Sons, Inc., Hoboken, New Jersey.

Harvey-Beavis, A. (2002), Student and School Questionnaire Development, in R.J. Adams and M.L. Wu (eds.), *PISA 2000 Technical Report,* (pp. 33-38), OECD, Paris.

International Labour Organisation (ILO) (1990), *International Standard Classification of Occupations: ISCO-88.* Geneva: International Labour Office.

Jöreskog, K.G. & **Sörbom, Dag** (1993), *LISREL 8 User's Reference Guide*, Chicago: SSI.

Judkins, D.R. (1990), Fay's Method of Variance Estimation, *Journal of Official Statistics*, No. 6 (3), pp. 223-239.

Kaplan, D. (2000), *Structural equation modeling: Foundation and extensions*, Thousand Oaks: SAGE Publications.

Keyfitz, N. (1951), Sampling with probabilities proportionate to science: Adjustment for changes in probabilities, *Journal of the American Statistical Association,* No. 46, American Statistical Association, Alexandria, pp. 105-109.

Kish, L. (1992), Weighting for Unequal, *Pi. Journal of Official Statistics*, No. 8 (2), pp. 183-200.

LISREL (1993), K.G. Jöreskog & D. Sörbom, [computer software], Lincolnwood, IL: Scientific Software International, Inc.

Lohr, S.L. (1999), *Sampling: Design and Analysis*, Duxberry: Pacific Grove.

Macaskill, G., Adams, R.J. & **Wu, M.L.** (1998), Scaling methodology and procedures for the mathematics and science literacy, advanced mathematics and physics scale, in M. Martin and D.L. Kelly, Editors, *Third International Mathematics and Science Study, technical report Volume 3: Implementation and analysis*, Boston College, Chestnut Hill, MA.

Masters, G.N. & **Wright, B.D.** (1997), The Partial Credit Model, in W.J. van der Linden, & R.K. Hambleton (eds.), *Handbook of Modern Item Response Theory* (pp. 101-122), New York/Berlin/Heidelberg: Springer.

Mislevy, R.J. (1991), Randomization-based inference about latent variables from complex samples, *Psychometrika,* No. 56, pp. 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,* No. 29 (2), pp. 133-161.

Mislevy, R.J. & **Sheehan, K.M.** (1987), Marginal estimation procedures, in Beaton, A.E., Editor, 1987. *The NAEP 1983-84 technical report*, National Assessment of Educational Progress, Educational Testing Service, Princeton, pp. 293-360.

Mislevy, R.J. & **Sheehan, K.M.** (1989), Information matrices in latent-variable models, *Journal of Educational Statistics*, No. 14, pp. 335-350.

Mislevy, R.J. & **Sheehan, K.M.** (1989), The role of collateral information about examinees in item parameter estimation, *Psychometrika*, No. 54, pp. 661-679.

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

Monseur, C. (2005), An exploratory alternative approach for student non response weight adjustment, *Studies in Educational Evaluation*, No. 31 (2-3), pp. 129-144.

Muthen, B. & **L. Muthen** (1998), [computer software], *Mplus* Los Angeles, CA: Muthen & Muthen.

Muthen, B., du Toit, S.H.C. & **Spisic, D.** (1997), *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*, unpublished manuscript.

OECD (1999), *Classifying Educational Programmes. Manual for ISCED-97 Implementation in OECD Countries,* OECD, Paris.

OECD (2003), *Literacy Skills for the World of Tomorrow: Further results from PISA 2000*, OECD, Paris.

OECD (2004), *Learning for Tomorrow's World – First Results from PISA 2003*, OECD, Paris.

OECD (2005), *Technical Report for the OECD Programme for International Student Assessment 2003*, OECD, Paris.

OECD (2006), *Assessing Scientific, Reading and Mathematical Literacy: A framework for PISA 2006,* OECD, Paris.

OECD (2007), *PISA 2006: Science Competencies for Tomorrow's World*, OECD, Paris.

PISA Consortium (2006), *PISA 2006 Main Study Data Management Manual*, *https://mypisa.acer.edu.au/images/mypisadoc/opmanual/pisa2006_data_management_manual.pdf*

Rasch, G. (1960), Probabilistic models for some intelligence and attainment tests, Copenhagen: Nielsen & Lydiche.

Routitski A. & **Berezner, A.** (2006), Issues influencing the validity of cross-national comparisons of student performance. Data Entry Quality and Parameter Estimation. Paper presented at the Annual Meeting of the American Educational Research Association (AERA) in San Francisco, 7-11 April, *https://mypisa.acer.edu.au/images/mypisadoc/aera06routitsky_berezner.pdf*

Rust, K. (1985), Variance Estimation for Complex Estimators in Sample Surveys, *Journal of Official Statistics*, No. 1, pp. 381-397.

Rust, K.F. & **Rao, J.N.K.** (1996), Variance Estimation for Complex Surveys Using Replication Techniques, *Survey Methods in Medical Research*, No. 5, pp. 283-310.

Shao, J. (1996), Resampling Methods in Sample Surveys (with Discussion), *Statistics*, No. 27, pp. 203-254.

Särndal, C.-E., Swensson, B. & **Wretman, J.** (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.

SAS® CALIS (1992), W. Hartmann [computer software], Cary, NC: SAS Institute Inc.

Scheerens, J. (1990), School effectiveness and the development of process indicators of school functioning, *School effectiveness and school improvement,* No. 1, pp. 61-80.

Scheerens, J. & **Bosker, R.J.** (1997), *The Foundations of School Effectiveness*, Oxford: Pergamon.

Schulz, W. (2002), Constructing and Validating the Questionnaire composites, in R.J. Adams and M.L. Wu (eds.), *PISA 2000 Technical Report*, OECD, Paris.

Schulz, W. (2004), Mapping Student Scores to Item Responses, in W. Schulz and H. Sibberns (eds.), *IEA Civic Education Study, Technical Report* (pp. 127-132), Amsterdam: IEA.

Schulz, W. (2006a), *Testing Parameter Invariance for Questionnaire Indices using Confirmatory Factor Analysis and Item Response Theory,* Paper presented at the Annual Meetings of the American Educational Research Association (AERA) in San Francisco, 7-11 April.

Schulz, W. (2006b), *Measuring the socio-economic background of students and its effect on achievement in PISA 2000 and PISA 2003*, Paper presented at the Annual Meetings of the American Educational Research Association (AERA) in San Francisco, 7-11 April.

Thorndike, R.L. (1973), *Reading comprehension in fifteen countries,* New York, Wiley: and Stockholm: Almqvist & Wiksell.

Travers, K.J. & **Westbury, I.** (1989), *The IEA Study of Mathematics I: Analysis of Mathematics Curricula*, Oxford: Pergamon Press.

Travers, K.J., Garden R.A. & **Rosier, M.** (1989), Introduction to the Study, in Robitaille, D. A. and Garden, R. A. (eds), *The IEA Study of Mathematics II: Contexts and Outcomes of School Mathematics Curricula,* Oxford: Pergamon Press.

Verhelst, N. (2002), Coder and Marker Reliabilaity Studies, in R.J. Adams & M.L. Wu (eds.), *PISA 2000 Technical Report*. OECD, Paris.

Walberg, H.J. (1984), Improving the productivity of American schools, *Educational Leadership,* No. 41, pp. 19-27.

Walberg, H. (1986), Synthesis of research on teaching, in M. Wittrock (ed.), *Handbook of research on teaching* (pp. 214-229), New York: Macmillan.

Walker, M. (2006), *The choice of Likert or dichotomous items to measure attitudes across culturally distinct countries in international comparative educational research*. Paper presented at the Annual Meetings of the American Educational Research Association (AERA) in San Francisco, 7-11 April.

Walker, M. (2007), Ameliorating Culturally-Based Extreme Response Tendencies To Attitude items, *Journal of Applied Measurement,* No. 8, pp. 267-278.

Warm, T.A. (1989), Weighted Likelihood Estimation of Ability in Item Response Theory, *Psychometrika*, No. 54 (3), pp. 427-450.

Westat (2007), *WesVar® 5.1* Computer software and manual, Rockville, MD: Author (also see *http://www.westat.com/wesvar/*).

Wilson, M. (1994), Comparing Attitude Across Different Cultures: Two Quantitative Approaches to Construct Validity, in M. Wilson (ed.), *Objective measurement II: Theory into practice* (pp. 271-292), Norwood, NJ: Ablex.

Wolter, K.M. (2007), *Introduction to Variance Estimation*. Second edition, Springer: New York.

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

List of abbreviations – the following abbreviations are used in this report:

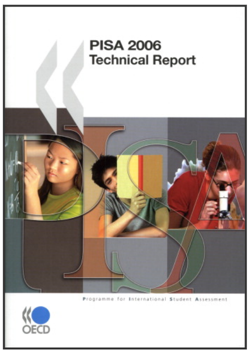
ACER	Australian Council for Educational Research
AGFI	Adjusted Goodness-of-Fit Index
BRR	Balanced Repeated Replication
CBAS	Computer Based Assessment of Science
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CITO	National Institute for Educational Measurement, The Netherlands
CIVED	Civic Education Study
DIF	Differential Item Functioning
ENR	Enrolment of 15-year-olds
ESCS	PISA Index of Economic, Social and Cultural Status
ETS	Educational Testing Service
IAEP	International Assessment of Educational Progress
I	Sampling Interval
ICR	Inter-Country Coder Reliability Study
ICT	Information Communication Technology
IEA	International Association for the Evaluation of Educational Achievement
INES	OECD Indicators of Education Systems
IRT	Item Response Theory
ISCED	International Standard Classification of Education
ISCO	International Standard Classification of Occupations
ISEI	International Socio-Economic Index
MENR	Enrolment for moderately small school
MOS	Measure of size
NCQM	National Centre Quality Monitor
NDP	National Desired Population
NEP	National Enrolled Population
NFI	Normed Fit Index
NIER	National Institute for Educational Research, Japan
NNFI	Non-Normed Fit Index
NPM	National Project Manager
OECD	Organisation for Economic Cooperation and Development
PISA	Programme for International Student Assessment
PPS	Probability Proportional to Size
PGB	PISA Governing Board
PQM	PISA Quality Monitor
PSU	Primary Sampling Units
QAS	Questionnaire Adaptations Spreadsheet
RMSEA	Root Mean Square Error of Approximation
RN	Random Number
SC	School Co-ordinator
SE	Standard Error
SD	Standard Deviation
SEM	Structural Equation Modelling
SMEG	Subject Matter Expert Group
SPT	Study Programme Table
TA	Test Administrator
TAG	Technical Advisory Group
TCS	Target Cluster Size
TIMSS	Third International Mathematics and Science Study
TIMSS-R	Third International Mathematics and Science Study – Repeat
VENR	Enrolment for very small schools
WLE	Weighted Likelihood Estimates

Table of contents

LIST OF BOXES

LIST OF FIGURES

LIST OF TABLES



From:

PISA 2006 Technical Report

Access the complete publication at:

https://doi.org/10.1787/9789264048096-en

Please cite this chapter as:

OECD (2009), “Scaling PISA Cognitive Data”, in *PISA 2006 Technical Report*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/9789264048096-10-en

This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of OECD member countries.

This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.