9

Scaling PISA Cognitive Data

The mixed co-efficients multinomial logit model as described by Adams *et al.* (1997) was used to scale the PISA data, and implemented by *ConQuest* software (Wu *et al.,* 1997).

THE MIXED CO-EFFICIENTS MULTINOMIAL LOGIT MODEL

The model applied to PISA is a generalised form of the Rasch model. The model is a mixed co-efficients 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 $\mathrm{X}_i = \left(X_{i1}, X_{i2},\ldots, X_{iK_i}\right)^T$, where

$$X_{ij} = \begin{cases} 1 & \text{if response to item } i \text{ is in category } j \\ 0 & \text{otherwise} \end{cases}, \qquad (9.1)$$

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 X_i can also be collected together into the single vector $\mathrm{X}^T = \left(\mathrm{X}_1^T, \mathrm{X}_2^T,\ldots,\mathrm{X}_I^T\right)$, called the response vector (or pattern). Particular instances of each of these random variables are indicated by their lower case equivalents; x, $\mathrm{x_i}$ and x_{ik}.

Items are described through a vector $\xi^T = \left(\xi_1,\xi_2,\ldots,\xi_p\right)$, 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 $\mathrm{a}_{ij}, \left(i = 1,\ldots,I;\ j = 1,\ldots K_i\right)$, each of length p, which can be collected to form a design matrix $\mathrm{A}^T = \left(\mathrm{a}_{11},\mathrm{a}_{12},\ldots,\mathrm{a}_{1K_1},\mathrm{a}_{21},\ldots,\mathrm{a}_{2K_2},\ldots,\mathrm{a}_{IK_I}\right)$ 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 $\theta = \left(\theta_1,\theta_2,\ldots,\theta_D\right)$, 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 $\mathrm{b}_{ik} = \left(b_{ik1}, b_{ik2},\ldots,b_{ikD}\right)^T$ and again collected into the scoring sub-matrix for item i, $\mathrm{B}_i = \left(\mathrm{b}_{i1},\mathrm{b}_{i2},\ldots,\mathrm{b}_{iD}\right)^T$ and then into a scoring matrix $\mathrm{B} = \left(\mathrm{B}_1^T,\mathrm{B}_2^T,\ldots,\mathrm{B}_I^T\right)^T$ 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

$$\Pr\left(\mathrm{X}_{ij} = 1; \mathrm{A}, \mathrm{B}, \xi \mid \theta\right) = \frac{\exp\left(\mathrm{b}_{ij}\theta + \mathrm{a}'_{ij}\xi\right)}{\sum_{k=1}^{K_i}\exp\left(\mathrm{b}_{ik}\theta + \mathrm{a}'_{ik}\xi\right)}. \qquad (9.2)$$

For a response vector we have

$$f(x;\xi|\theta) = \Psi(\theta,\xi)\exp\left[x'(B\theta + A\xi)\right], \qquad (9.3)$$

with

$$\Psi(\theta,\xi) = \left\{\sum_{z\in\Omega}\exp\left[z^T(B\theta + A\xi)\right]\right\}^{-1} \qquad (9.4)$$

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:

$$f_\theta(\theta;\alpha) \equiv f_\theta(\theta;\mu,\sigma^2) = (2\pi\sigma^2)^{-1/2}\exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2}\right] \qquad (9.5)$$

or equivalently

$$\theta = \mu + E \qquad (9.6)$$

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

Adams *et al.* (1997) discuss how a natural extension of (9.6) is to replace the mean, μ with the regression model, $Y_n^T\rceil$ where Y_n is a vector of *u,* fixed and known values for student *n,* and β is the corresponding vector of regression co-efficients. For example, Y_n could be constituted of student variables such as gender or socio-economic status. Then the population model for student *n*, becomes,

$$\theta_n = Y_n^T\rceil + E_n \qquad (9.7)$$

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:

$$f_\theta(\theta_n;Y_n,b,\sigma^2) = (2\pi\sigma^2)^{-1/2}\exp\left[-\frac{1}{2\sigma^2}(\theta_n - Y_n^T\rceil)^T(\theta_n - Y_n^T\rceil)\right], \qquad (9.8)$$

a normal distribution with mean $Y_n^T\rceil$ and variance σ^2. If (9.8) 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:

$$f_\theta(\theta_n;W_n,\gamma,\Sigma) = (2\pi)^{-d/2}|\Sigma|^{-1/2}\exp\left[-\frac{1}{2}(\theta_n - \gamma W_n)^T\Sigma^{-1}(\theta_n - \gamma W_n)\right], \qquad (9.9)$$

where γ is a $u\times d$ matrix of regression co-efficients, Σ is a $d\times d$ variance-covariance matrix and W_n is a $u\times 1$ vector of fixed variables.

In PISA, the W_n variables are referred to as conditioning variables.

Combined model

In (9.10), the conditional item response model (9.4) and the population model (9.9) are combined to obtain the unconditional, or marginal, item response model:

$$f_x(x;\xi,\gamma,\Sigma)=\int_\theta f_x(x;\xi\,|\theta)\,f_\theta(\theta;\gamma,\Sigma)\,d\,\theta \qquad (9.10)$$

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 *et al.* (1997a), Adams *et al.* (1997b), and Wu *et al.* (1997).

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

$$\begin{aligned} h_\theta(\theta_n;W_n,\xi,\gamma,\Sigma\,|x_n) &= \frac{f_x(x_n;\xi\,|\theta_n)\,f_\theta(\theta_n;W_n,\gamma,\Sigma)}{f_x(x_n;W_n,\xi,\gamma,\Sigma)} \\ &= \frac{f_x(x_n;\xi\,|\theta_n)\,f_\theta(\theta_n;W_n,\gamma,\Sigma)}{\int_{\theta_n} f_x(x_n;\xi\,|\theta_n)\,f_\theta(\theta_n;W_n,\gamma,\Sigma)} \end{aligned} \qquad (9.11)$$

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 (9.3) 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.

In PISA 2003 the main scaling model was seven-dimensional, made up of one reading, one science, one problem solving and four mathematics dimensions. 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 13.

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; or an item with sound characteristics in each country but which shows substantial item-by-country interactions may be regarded as a different item (for scaling purposes) in each country (or in some subset of countries) that is, the difficulty parameter will be free to vary across countries. Both

the second and third options have the same impact on comparisons between countries. That is, if an item is identified as behaving differently in different countries, choosing either the second or third option will have the same impact on inter-country comparisons. The choice between them could, however, influence within-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 *et al.,* 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 cases gives us 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 co-efficients

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} $(= x_{ij} / m_i)$ 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, five reports were returned to each participating country to assist in reviewing their data with the consortium:

- *Report 1* presented the results of a basic item analysis in tabular form. For each item, the number of students, the percentage of students, the point-biserial correlation, and student-centred Item Response Theory (IRT) ability average were provided for each valid category.
- *Report 2* provided, for each item and for each valid category, the point-biserial correlation and the student-centred IRT ability average in graphical form.
- *Report 3* provided a graphical comparison of the item infit mean square co-efficients and the item discrimination co-efficients computed at national and international levels.
- *Report 4* provided a graphical comparison of both the item difficulty parameter and the item thresholds, computed at national and international levels.
- *Report 5* listed the items that national project managers (NPMs) needed to check for mistranslation and/or misprinting, referred to as dodgy items.
- *Report 6* provides in a graphical form a comparison of the deviation of observed scores from expected scores for each item.

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 table shows each possible response category for each item. The second column indicates the score assigned to the different categories. For each category, the number and percentage of students responding is shown, along with the point-biserial correlation and the associated *t* statistic. Note that for the item in the example the correct answer is '4', indicated by the '1' in the score column; thus the point-biserial for a response of '4' is the item's discrimination index, also shown along the top. The two last columns, *PV1Avg:1* and *PV1 SD:1*, show the average ability of students responding in each category and the standard deviation

Figure 9.1 ■ **Example of item statistics shown in Report 1**

```
Item 1
------
item:1 (M033Q01)
Cases for this item   1258   Discrimination  0.27
Item Threshold(s)    -2.06   Weighted MNSQ   1.11
--------------------------------------------------------------------------------
 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        8       0.64   -0.07    -2.32(.021) -0.67     1.25
   2       0.00       76       6.04   -0.15    -5.46(.000) -0.62     1.13
   3       0.00       94       7.47   -0.18    -6.47(.000) -0.64     1.12
   4       1.00     1069      84.98    0.27     9.91(.000)  0.17     1.12
   5                   0       0.00     NA       NA (.000)    NA       NA
   6                   0       0.00     NA       NA (.000)    NA       NA
   7                   0       0.00     NA       NA (.000)    NA       NA
   8       0.00        4       0.32   -0.06    -2.31(.021) -1.04     1.42
   9       0.00        7       0.56   -0.05    -1.88(.060) -0.66     1.21
================================================================================
```

for it. The average ability is calculated by domain. In this example the average ability of those students who responded correctly (category 4) is 0.17, while the average ability of those students who responded incorrectly (categories 1, 2, 3) is around -0.6.

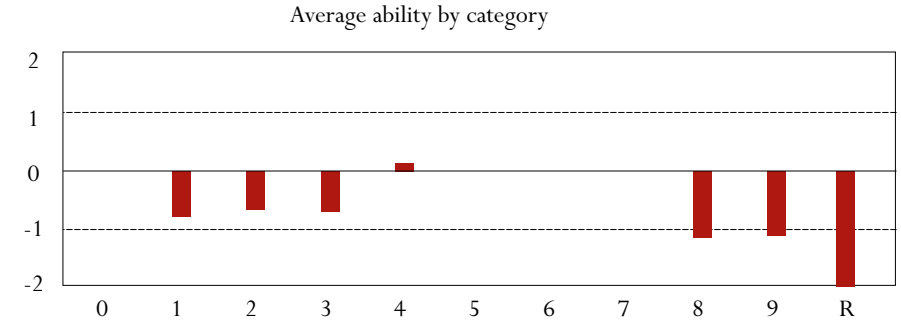
Report 2: Descriptive statistics on individual items in graphical form

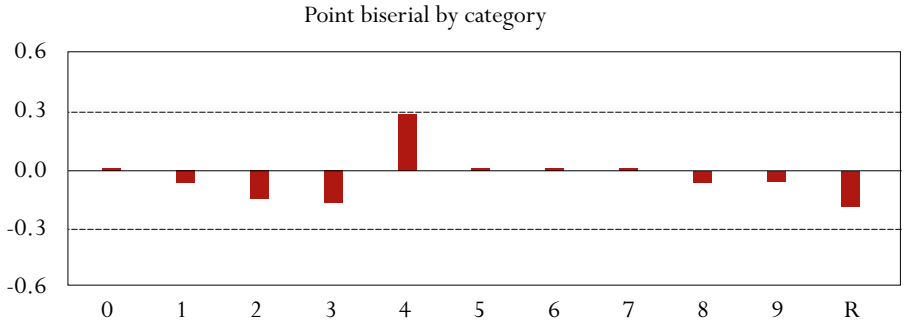
Report 2 (see Figure 9.2) graphs the ability average and the point-biserial correlation by category. Average ability by category is calculated by domain and centred for each item. This makes it easy to identify positive and negative ability categories, so that checks can be made to ensure that, for multiple-choice items, the key category has the highest average ability estimate, and for constructed-response items, the mean abilities are ordered consistently with the score levels. The displayed graphs also facilitate the process of identifying the following anomalies:

- A non-key category with a positive point-biserial or a point-biserial higher than the key category;
- A key category with a negative point-biserial; and
- For partial-credit items, average abilities (and point-biserials) not increasing with the score points.

Figure 9.2 ■ **Example of item statistics shown in Report 2**

Students	0	8	76	94	1069	0	0	0	4	7	18
%	0	1	6	7	84	0	0	0	0	1	1

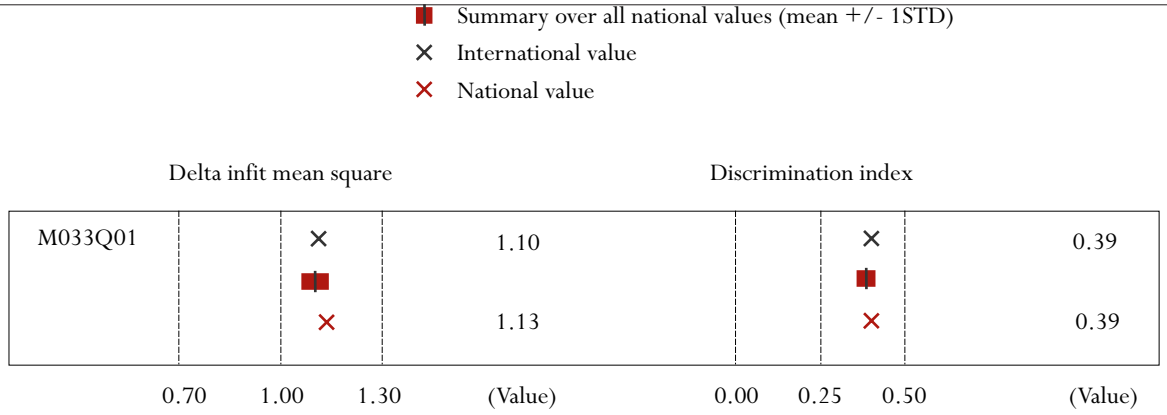




ID: M033Q01
Name: View room Q1

Discrimination: 0.27
Key: 4

Figure 9.3 ■ **Example of item statistics shown in Report 3**



Report 3: Comparison of national and international infit mean square and discrimination co-efficients

The national scaling provided the infit mean square, the point-biserial correlation, the item parameter estimate (or difficulty estimate) and the thresholds for each item in each country. Reports 3 and 4 (see Figures 9.3 and 9.4) compare the value computed for one country with those computed for all other countries and with the value computed at international level for each item.

The black crosses present the values of the co-efficients computed from the international database. Shaded boxes represent the mean plus or minus one standard deviation of these national values. Red crosses represent the values for the national data set of the country to which the report was returned.

Substantial differences between the national and international value on one or both of these indices show that the item is behaving differently in that country. This might reflect a mistranslation or another problem specific to the national version, but if the item was misbehaving in all or nearly all countries, it might reflect a specific problem in the source item and not with the national versions.

Report 4: Comparison of national and international item difficulty parameters and thresholds

Report 4 presents the item difficulty parameters and the thresholds, in the same graphic form as Report 3. Substantial differences between the national value and the international value (*i.e.* the national value mean) might be interpreted as an item-by-country interaction. Nevertheless, appropriate estimates of the item-by-country interaction are provided in Report 5.

Figure 9.4 ■ **Example of item statistics shown in Report 4**

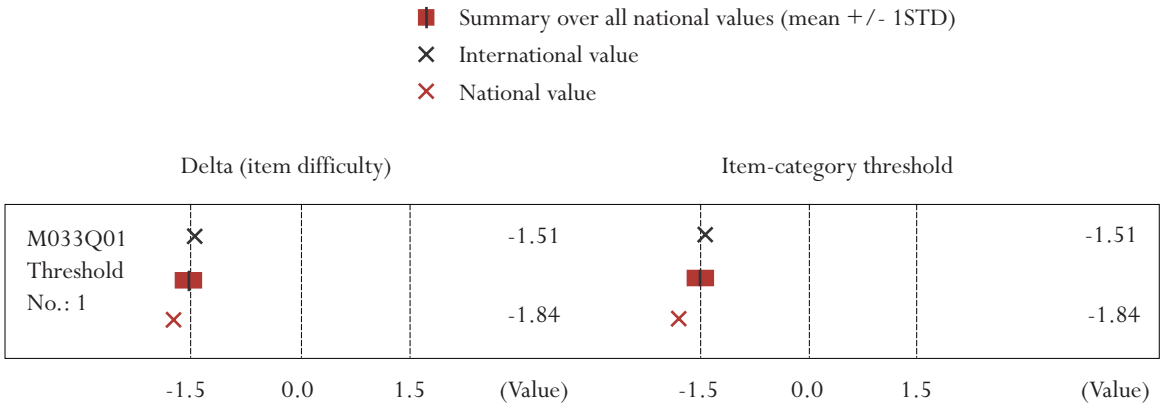


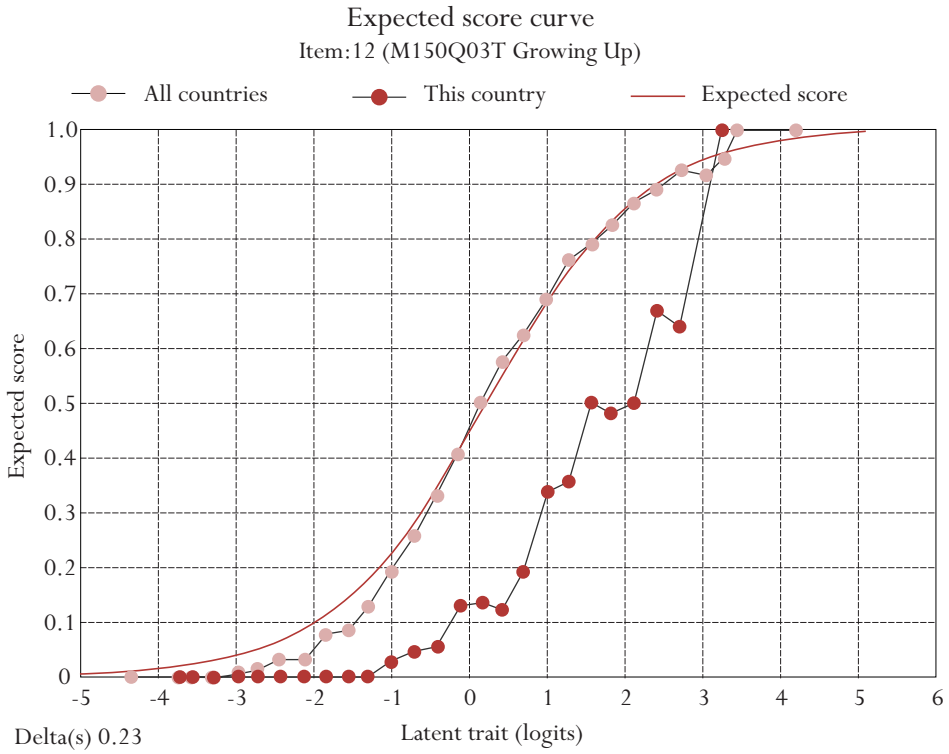
Figure 9.5 ■ **Example of item statistics shown in Report 5**

		Item by Country Interactions		Discrimination			Pisa 2000 Link Items	
	No of Valid Responses	Easier than Expected	Harder than Expected	Non-key PB is Positive	low discrimi-nation	Ability not Ordered	Link Item	Required checking
M124Q03T	1788	☐	☐	☐	☐	☐	☑	☑
M150Q03T	1601	☐	☑	☐	☐	☐	☑	☑
M155Q02T	1620	☐	☐	☐	☐	☐	☑	☑
M406Q01	1608	☑	☐	☐	☐	☐	☐	☐
M406Q02	1607	☑	☐	☐	☐	☐	☐	☐

Report 5: National dodgy item report

For each country's dodgy items, Report 5 lists where the items were flagged for one or more of the following reasons: difficulty is significantly easier or harder than average; a non-key category has a point-biserial correlation higher than 0.05 if at least 10 students selected it; the key category point-biserial correlation is lower than 0.25; the categories abilities for partial credit items are not ordered; and/or the link item difficulty was different from the PISA 2000 Main Study. An example extract is shown in Figure 9.5.

Figure 9.6 ■ **Example of item statistics shown in Report 6**



Report 6: Expected score curves

For the analysis of item performance expected score curves (ESC) were constructed and reported for each item. Report 6 provided a graphical comparison of both national and international observed scores with an expected score. Figure 9.6 is an example of the deviation of observed scores from the expected score curve. The solid line represents expected scores and the dots (connected by dotted lines) are observed scores.

International calibration

International item parameters were set by applying the conditional item response model (9.3) in conjunction with the multivariate population model (9.9), without using conditioning variables, to a sub-sample of students. This sub-sample 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 seven PISA 2003 scales is given in Appendix 12 (for mathematics), Appendix 13 (for reading), Appendix 14 (for science) and Appendix 15 (for problem solving).

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 (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.

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 (9.11), using the Monte-Carlo integration

$$\int_\theta f_x(x;\xi\,|\theta) f_\theta(\theta,\gamma,\Sigma) d\theta \approx \frac{1}{M}\sum_{m=1}^{M} f_x(x;\xi\,|\varphi_{mn}) \equiv \Im \qquad (9.12)$$

At the same time, the values

$$p_{mn} = f_x(x_n;\xi\,|\varphi_{mn}) f_\theta(\varphi_{mn};\mathrm{W}_n,\gamma,\Sigma) \qquad (9.13)$$

are calculated, so that 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) is obtained; and the probability that φ_{nj} could be drawn from this density is given by

$$q_{nj} = \frac{p_{mn}}{\sum_{m=1}^{M} p_{mn}} \qquad (9.14)$$

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

$$\sum_{s=1}^{i_0-1} q_{sn} < \eta_i \leq \sum_{s=1}^{i_0} q_{sn} \qquad (9.15)$$

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 (NAEP) (Beaton, 1987) and in TIMSS (Macaskill, Adams and Wu, 1998). The steps involved in this process are:

- *Step 1*: Five variables (booklet ID, gender, mother's occupation, father's occupations and school mean mathematics score) were prepared to be directly used as conditioning variables. The booklet ID was dummy coded so that booklet 9 was used as the reference booklet. Booklet 9 had to be chosen as the reference booklet because it is the only booklet that contains items from all four assessment domains. For mother's and father's occupation the ISEI index was used. For each student the mean mathematics achievement for that student's school was estimated using the mean of the weighted likelihood estimates for mathematics for each of the students that also attended that student's school.
- *Step 2*: Each variable in the Student Questionnaire was dummy coded. The details of this dummy coding are provided in Appendix 10.
- *Step 3*: For each country, a principal components analysis of the dummy-coded variables was performed, and component scores were produced for each student (a sufficient number of components to account for 95 per cent of the variance in the original variables).
- *Step 4*: The item-response model was fit to each national data set and the national population parameters were estimated using item parameters anchored at their international location and conditioning variables derived from the national principal components analysis and from step 1.
- *Step 5*: Five vectors of plausible values were drawn using the method described above. The vectors were of length seven, one for each of the PISA 2003 reporting scales.

As described in Chapter 2, the PISA test design is such that not all students are assessed in all four domains. In PISA 2000, the plausible values for those students who did not respond to any items from a domain were removed from the database and a set of weight adjustments were provided for dealing with the smaller data set. The assumption under this approach is that the students who did not get domain scores were missing at random. For PISA 2003, the plausible values for all domains have been retained for all students. This approach has a number of advantages. First, the database structure is simpler and analysis is simpler because the use of a weight adjustment is not necessary. Second, the missing at random assumption is loosened somewhat. The plausible value generation assumes that the relationships between the domain for which no items are observed and all other variables (both conditioning variables and the other domain) is the same for both the students who did respond to items from a domain and those that did not. Using all of this relationship information, and all available information about the student an imputation is made. Because of the amount of data that is available to make the imputation, the analysis of the full data set will produce more accurate results than will analysis of the data set that omits students who did not respond to a domain. Additionally it can be expected that, due to sampling variation, the characteristics of the students who did not

respond to a domain will be slightly different to the characteristics of those that did. These differences will be appropriately adjusted for in the imputation and the estimated characteristics of, for example, the reading proficiency distribution for all students will be slightly different to the estimated characteristics of the reading proficiency distribution for the subset of students that responded to reading items.

The one disadvantage of this approach is that the average performances on a reference booklet (booklet 9) will influence the imputations for students who did not respond to items from a domain. As we note in Chapter 13, booklet- and item-by-country interactions do result in variations across booklets in the country means. If a country has an unusually low or high performance on the reference booklet, for a particular domain, then this unusual performance will influence the imputations for all students that did not respond to that domain. The consequential effect is that the reference booklet will be given more weight than the other booklets in the assessment of national means.

ANALYSIS OF DATA WITH PLAUSIBLE VALUES

It is 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 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 *et al.*, 1997a).

The PISA student file contains 40 plausible values, five for each of the seven PISA 2003 cognitive scales and five for the combined mathematics scale. *PV1MATH* to *PV5MATH* are five for mathematical literacy; *PV1SCIE* to *PV5SCIE* for scientific literacy, *PV1READ* to *PV5READ* for reading literacy and *PV1PROB* to *PV5PROB* for problem solving. For the four mathematics literacy subscales – space and shape, change and relationship, uncertainty and quantity – the plausible values variables are *PV1MATH1* to *PV5MATH1*, *PV1MATH2* to *PV5MATH2, PV1MATH3* to *PV5MATH3,* and *PV1MATH4* to *PV5MATH4,* respectively.

If an analysis were to be undertaken with one of these seven cognitive scales, or for the combined mathematics scale, 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 $r(\theta, \mathrm{Y})$ 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 $h_\theta(\theta_n; y_n, \xi, \gamma, \Sigma \,|\mathrm{x}_n)$. If $h_\theta(\theta; \mathrm{Y}, \xi, \gamma, \Sigma \,|\mathrm{X})$ is the joint marginal posterior for $n=1,\ldots N$ then we can compute:

$$r^*(\mathrm{X}, \mathrm{Y}) = E\left[r^*(\theta, \mathrm{Y}) \middle| \mathrm{X}, \mathrm{Y}\right] = \int_\theta r(\theta, \mathrm{Y}) h_\theta(\theta; \mathrm{Y}, \xi, \gamma, \Sigma \,|\mathrm{X}) d\theta \qquad (9.16)$$

The integral (9.16) can be computed using the Monte-Carlo method. If M random vectors $(\Theta_1, \Theta_2, \ldots, \Theta_M)$ are drawn from $h_\theta(\theta; Y, \xi, \gamma, \Sigma | X)$ (9.16) is approximated by:

$$r^*(X,Y) \approx \frac{1}{M}\sum_{m=1}^{M} r(\Theta_m, Y) = \frac{1}{M}\sum_{m=1}^{M} \hat{r}_m \qquad (9.17)$$

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:

$$V = U^* + (1 + M^{-1})B_M \qquad (9.18)$$

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

An α-% confidence interval for r^* is $r^* \pm t_\upsilon\left(\frac{(1-\alpha)}{2}\right)V^{1/2}$

where $t_\upsilon(s)$ is the s-percentile of the t-distribution with V 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 (see Adams and Wu, 2002).[3]

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.[4]

For problem solving, which is a new domain for PISA 2003, and for mathematics this is 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 both problem solving and 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.[5]

Linking PISA 2000 and PISA 2003 reading and science

The PISA 2000 and PISA 2003 assessments of mathematics, reading and science are linked assessments. That is, the sets of items used to assess each of mathematics, reading and science in PISA 2000 and the sets of items used to assess each of mathematics, reading and science in PISA 2003 include a subset of items common to both sets. For mathematics 20 items were used in both assessments, for reading 28 items were used in both assessments and for science 25 items were used in both assessments (see Chapter 2). These common items are referred to as link items.

The steps involved in linking the PISA 2000 and PISA 2003 reading and science scales were:

- *Step 1*: The PISA 2000 data from each of the OECD countries were then re-scaled with full conditioning and with link items anchored at their PISA 2003 values.
- *Step 2*: The mean and standard deviation of each domain were calculated for a combined data set of 25 OECD countries[6]. Senate weights were used so that each country was given the same weight.
- *Step 3*: The mean and standard deviations computed in Step 2 were then compared with the matching means and standard deviations from the PISA 2000 scaling. Linear transformations that mapped the PISA 2003 based scores to scores that would yield a mean and standard deviation equal to the PISA 2000 results were then computed.

Linking PISA 2000 and PISA 2003 mathematics

In the case of mathematics a decision was made to produce a new scale for PISA 2003 and to undertake a retrospective mapping of the 2000 data onto this new PISA 2003 scale for each of the two areas (space and shape, and change and relationships) that were assessed both times. The steps involved were:

- *Step 1*: The PISA 2000 calibration sample was scaled with a two dimensional model, the two dimensions being the two mathematics scales included in PISA 2000. The items were anchored at their PISA 2000 values. No conditioning was used in this scaling.
- *Step 2*: Step 1 was then replicated with the items anchored at their PISA 2003 values.
- *Step 3*: For the two sets of scaling results the means and standard deviations for both dimensions were calculated for a combined data set of 25 OECD countries.[7] Senate weights were used so that each country was given the same weight.
- *Step 4*: Linear transformations that mapped the PISA 2000 based scores to scores that would yield a mean and standard deviation equal to the PISA 2003 results were then computed.

Uncertainty in the link

In each case the transformation that equates the 2000 and 2003 data depends upon the change in difficulty of each of the individual link items and as a consequence the sample of link items that has 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 PISA 2000 and PISA 2003 results. Just as with the error that is introduced through the process of sampling students, the exact magnitude of this linking error cannot be determined. The likely range of magnitudes for this error can, however, be estimated and this error can be taken into account when interpreting PISA results. As with sampling errors, the likely range of magnitude for the errors is represented as a standard error. The link standard errors are reported in Chapter 13.

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 2000 and PISA 2003) in reading resulted in an over-estimation of student scores by two points on the PISA 2000 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 2003 data. For example, consider the following:

- Each country's mean will be over-estimated by an amount equal to the link error. In this example, it is two score points.
- The mean performance of any subgroup will be over-estimated by an amount equal to the link error. In this example, it is two score points.
- The standard deviation of student scores will not be affected 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 2003 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 2003 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 2000 and PISA 2003 will be influenced because each student's score in PISA 2003 will be influenced by the error.
- A change in the difference in performance between two groups from PISA 2000 to PISA 2003 will not be influenced. This is because neither of the components of this comparison, which are differences in scores in 2000 and 2003 respectively, is influenced by a common error that is added to all student scores in PISA 2003.

In general terms, the linking error need only be considered when comparisons are being made between PISA 2000 and PISA 2003 results, 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 PISA 2000 and PISA 2003. For example, let us consider a comparison between 2000 and 2003 of the performance of Denmark in reading. The mean performance of Denmark in 2000 was 497 with a standard error of 2.4, while in 2003 the mean was 492 with a standard error of 2.8. The standardised difference in the mean for Denmark is 0.89, which is computed as follows: $0.89 = (497 - 492) / \sqrt{2.4^2 + 2.8^2 + 3.744^2}$, 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, 2000 has been used.

3 Using senate weights.

4 Using senate weights.

5 Using senate weights.

6 The Netherlands was excluded because it did not meet PISA standards in 2000. The United Kingdom was excluded because it did not meet PISA standards in 2003. Luxembourg was omitted because of a change in test administration procedures between PISA 2000 and 2003. The Slovak Republic and Turkey were excluded because they did not participate in PISA 2000.

7 See footnote 6.

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

BRA Brazil
HKG Hong Kong-China
IND Indonesia
LVA Latvia
LVL Latvia (Latvian Community)
LVR Latvia (Russian Community)
LIE Liechtenstein
MAC Macao-China
RUS Russian Federation
YUG Serbia and Montenegro (Serbia)
THA Thailand
TUN Tunisia
URY Uruguay

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
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CITO	National Institute for Educational Measurement, The Netherlands
CIVED	Civic Education Study
DIF	Differential Item Functioning
ESCS	Economic, Social and Cultural Status
ENR	Enrolment of 15-year-olds
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
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

References

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

Aiken, L. R. (1974), "Two scales of attitudes toward mathematics," *Journal for Research in Mathematics Education 5,* National Council of Teachers of Mathematics, Reston, pp. 67-71.

Andersen, Erling B. (1997), "The Rating Scale Model", in van der Linden, W. J. and R.K. Hambleton (eds.), *Handbook of Modern Item Response Theory,* Springer, New York/Berlin/Heidelberg.

Bandura, A. (1986), *Social Foundations of Thought and Action: A Social Cognitive Theory,* Prentice Hall, Englewood Cliffs, N.J.

Baumert, J. and **O. Köller** (1998), "Interest Research in Secondary Level I : An Overview", in L. Hoffmann, A. Krapp, K.A. Renninger & J. Baumert (eds.), *Interest and Learning*, IPN, Kiel.

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

Bryk, A. S. and **S.W. Raudenbush** (1992), *Hierarchical Linear Models: Applications and Data Analysis Methods,* SAGE Publications, Newbury Park.

Bollen, K.A. and **S.J. Long** (eds.) (1993), *Testing Structural Equation Models*, SAGE publications, Newbury Park.

Branden, N. (1994), *Six Pillars of Self-Esteem*. Bantam, New York.

Brennan, R.L. (1992), *Elements of Generalizability Theory*, American College Testing Program, Iowa City.

Buchmann, C. (2000), *Measuring Family Background in International Studies of Educational Achievement: Conceptual Issues and Methodological Challenges,* paper presented at a symposium convened by the Board on International Comparative Studies in Education of the National Academy of Sciences/National Research Council on 1 November, in Washington, D.C.

Cochran, W.G. (1977), *Sampling Techniques* (3rd edition), Wiley, New York.

Cronbach, L.J., G.C. Gleser, H. Nanda and **N. Rajaratnam** (1972), *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*, Wiley and Sons, New York.

Eccles, J.S. (1994), "Understanding Women's Educational and Occupational choice: Applying the Eccles *et al.* Model of Achievement-Related Choices", *Psychology of Women Quarterly 18*, Society for the Psychology of Women, Washington, D.C., pp. 585-609.

Eccles, J.S. and **A. Wigfield** (1995), "In the mind of the achiever: The structure of adolescents' academic achievement-related beliefs and self-perceptions", *Personality and Social Psychology Bulletin 21*, Sage Publications, Thousand Oaks, pp. 215-225.

Ganzeboom, H.B.G., P.M. de Graaf and **D.J. Treiman** (1992), "A standard international socio-economic index of occupational status", *Social Science Research 21*, Elsevier, pp. 1-56.

Gifi, A. (1990), *Nonlinear Multivariate Analysis*, Wiley, New York.

Greenacre, M.J. (1984), *Theory and Applications of Correspondence Analysis*, Academic Press, London.

Grisay, A. (2003), "Translation procedures in OECD/PISA 2000 international assessment", *Language Testing 20,* Holder Arnold Journals, pp. 225-240.

Gustafsson, J.E and **P.A. Stahl** (2000), *STREAMS User's Guide, Version 2.5 for Windows,* MultivariateWare, Mölndal, Sweden.

Hacket, G. and **N. Betz.** (1989), "An Exploration of the mathematics Efficacy/mathematics Performance Correspondence", *Journal of Research in Mathematics Education 20,* National Council of Teachers of Mathematics, Reston, pp. 261-273.

Harvey-Beavis, A. (2002), "Student and Questionnaire Development" in OECD, *PISA 2000 Technical Report*, OECD, Paris.

Hatcher, L. (1994), *A Step-by-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling,* SAS Institute Inc., Cary.

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

Jöreskog, K.G. and **Dag Sörbom** (1993), *LISREL 8 User's Reference Guide,* Scientific Software International, Chicago.

Judkins, D.R. (1990), "Fay's Method for Variance Estimation", *Journal of Official Statistics 6,* Statistics Sweden, Stockholm, pp. 223-239.

Kaplan, D. (2000), *Structural Equation Modeling: Foundation and Extensions,* SAGE Publications, Thousand Oaks.

Keyfitz, N. (1951), "Sampling with probabilities proportionate to science: Adjustment for changes in probabilities", *Journal of the American Statistical Association 46,* American Statistical Association, Alexandria, pp. 105-109.

Lepper, M. R. (1988), "Motivational considerations in the study of instruction", *Cognition and Instruction 5,* Lawrence Erlbaum Associates, Mahwah, pp. 289-309.

Ma, X. (1999), "A Meta-Analysis of the Relationship Between Anxiety Toward mathematics and Achievement in mathematics", *Journal for Research in Mathematics Education 30,* National Council of Teachers of Mathematics, Reston, pp. 520-540.

Macaskill, G., R.J. Adams and **M.L. Wu** (1998), "Scaling methodology and procedures for the mathematics and science literacy, advanced mathematics and physics scales", in M. Martin and D.L. Kelly (eds.) *Third International Mathematics and Science Study, Technical Report Volume 3: Implementation and Analysis,* Center for the Study of Testing, Evaluation and Educational Policy, Boston College, Chestnut Hill.

Marsh, H. W. (1990), *Self-Description Questionnaire (SDQ) II: A theoretical and Empirical Basis for the Measurement Of Multiple Dimensions of Adolescent Self-Concept: An Interim Test Manual and a Research Monograph,* The Psychological Corporation, San Antonio.

Marsh, H. W. (1994), "Confirmatory factor analysis models of factorial invariance: A multifaceted approach" *Structural Equation Modeling 1,* Lawrence Erlbaum Associates, Mahwah, pp. 5-34.

Marsh, H. W. (1999), *Evaluation of the Big-Two-Factor Theory of Motivation Orientation: Higher-order Factor Models and Age-related Changes*, paper presented at the 31.62 Symposium, Multiple Dimensions of Academic Self-Concept, Frames of Reference, Transitions, and International Perspectives: Studies From the SELF Research Centre. Sydney: University of Western Sydney.

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

Meece, J., A. Wigfield and **J. Eccles** (1990), "Predictors of Maths Anxiety and its Influence on Young Adolescents' Course Enrolment and Performance in Mathematics", *Journal of Educational Psychology 82,* American Psychological Association, Washington, D.C., pp. 60-70.

Middleton, J.A. and **P.A. Spanias** (1999), "Findings, Generalizations, and Criticisms of the Research", *Journal for Research in Mathematics Education 30*, National Council of Teachers of Mathematics, Reston, pp. 65-88.

Mislevy, R.J. (1991), "Randomization-based inference about latent variable from complex samples", *Psychometrika 56*, Psychometric Society, Greensboro, pp. 177-196.

Mislevy, R.J. and **K.M. Sheehan** (1987), "Marginal estimation procedures", in A.E. Beaton (ed.), *The NAEP 1983-1984 Technical Report* (Report No. 15-TR-20), Educational Testing Service, Princeton, N.J.

Mislevy, R.J. and **K.M. Sheehan** (1980), "Information matrices in latent-variable models", *Journal of Educational Statistics 14.4,* American Educational Research Association and American Statistical Association, Washington, D.C., and Alexandria, pp. 335-350.

Mislevy, R.J., A.E. Beaton, B. Kaplan and **K.M. Sheehan.** (1992), "Estimating population characteristics form sparse matrix samples of item responses", *Journal of Educational Measurement 29*, National Council on Measurement in Education, Washington, D.C., pp. 133-161.

Multon, K. D., S. D. Brown and **R.W. Lent** (1991), "Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation", *Journal of Counselling Psychology 38,* American Psychological Association, Washington, D.C., pp. 30-38.

Muthén, B. O., S. H. C. du Toit and **D. Spisic** (1997), "Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical outcomes", *Psychometrika*, Psychometric Society, Greensboro.

Muthen, L. and **B. Muthen** (2003), *Mplus User's Guide Version 3.1,* Muthen & Muthen, Los Angeles.

Nishisato, S. (1980), *Analysis of Categorical Data: Dual Scaling and its Applications,* University of Toronto Press, Toronto.

OECD (Organisation for Economic Co-Operation and Development) (1999), *Classifying Educational Programmes: Manual for ISCED-97 Implementation in OECD Countries,* OECD, Paris.

OECD (2001), *Knowledge and Skills for Life: First Results from PISA 2000*, OECD, Paris.

OECD (2002), *PISA 2000 Technical Report,* OECD, Paris.

OECD (2003), *Student Engagement at School: A Sense of Belonging and Participation: Results from PISA 2000,* OECD, Paris.

OECD (2004a), *The PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills*, OCED, Paris.

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

OECD (2004c), *Problem Solving for Tomorrow's World – First Measures of Cross-Curricular Competencies from PISA 2003,* OECD, Paris.

OECD (2005a), *PISA 2003 Data Analysis Manual: SAS® Users,* OECD, Paris.

OECD (2005b), *PISA 2003 Data Analysis Manual: SPSS® Users,* OECD, Paris.

Owens L. and **J. Barnes** (1992), *Learning Preference Scales*, Australian Council for Educational Research, Hawthorn.

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

Rust, K. (1985), "Variance estimation for complex estimators in sample surveys", *Journal of Official Statistics 1,* Statistics Sweden, Stockholm, pp. 381-397.

Rust, K. and **J.N.K. Rao** (1996), "Variance estimation for complex surveys using replication techniques", *Statistical Methods in Medical Research 5,* Holder Arnold Journals, pp. 283-310.

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

Schaffer, E. C., P. S. Nesselrodt and **S. Stringfield** (1994), "The Contribution of Classroom Observation to School Effectiveness Research" in Reynolds *et. al.* (eds.), *Advances in School Effectiveness Research and Practice,* Pergamon, Oxford/New York/Tokyo.

Schulz, W. (2003), *Validating Questionnaire Constructs in International Studies. Two Examples from PISA 2000,* paper presented at the Annual Meeting of the American Educational Research Association (AERA) in Chicago, 21-25 April.

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

Sirotnik, K. (1970), "An analysis of variance framework for matrix sampling", *Educational and Psychological Measurement 30,* SAGE Publications, pp. 891-908.

Slavin, R. E. (1983), "When does cooperative learning increase student achievement?" *Psychological Bulletin 94,* American Psychological Association, Washington, D.C., pp. 429-445.

Statistical Solutions (1992), *BMDP Statistical Software*, Statistical Solutions, Los Angeles.

Teddlie, C. and **D. Reynolds** (2000) (eds.), *The International Handbook of School Effectiveness Research,* Falmer Press, London/New York.

Thorndike, R.L. (1973), *Reading Comprehension Education in Fifteen Countries: An Empirical Study*, Almquist & Wiksell, Stockholm.

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

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

Verhelst, N. (2004), "Generalizability Theory", in Council of Europe, *Reference Supplement to the Preliminary Pilot version of the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, (Section E), Council of Europe (DGIV/EDU/LANG (2004) 13), Strasbourg.

Warm, T. A. (1989), "Weighted Likelihood Estimation of Ability in Item Response Theory", *Psychometrika* 54, Psychometric Society, Greensboro, pp. 427-45.

Wigfield, A., J. S. Eccles and **D. Rodriguez** (1998), "The development of children's motivation in school contexts", in P. D. Pearson. and A. Iran-Nejad (eds.), *Review of Research in Education 23,* American Educational Research Association, Washington D.C., pp. 73-118.

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

Wolter, K.M. (1985), *Introduction to Variance Estimation*, Springer-Verlag, New York.

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

Zimmerman, B.J. and **D.H. Schunk** (eds.) (1989), *Self-Regulated Learning and Academic Achievement. Theory, Research and Practice,* Springer, New York.

Table of Contents