Chapter 17: Scaling PIAAC Cognitive Data

Kentaro Yamamoto, Lale Khorramdel and Matthias von Davier, ETS

17.1 Overview

The test design for PIAAC was based on a variant of matrix sampling (using different sets of items, multistage adaptive testing, and different assessment modes) where each respondent was administered a subset of items from the total item pool. That is, different groups of respondents answered different sets of items. That makes it inappropriate to use any statistic based on the number of correct responses in reporting the survey results. Differences in total scores (or statistics based on them) among respondents who took different sets of items may be due to variations in difficulty in the adaptively administered test forms. Unless one makes very strong assumptions – for example, that the different test forms are perfectly parallel – the performance of the two groups assessed in a matrix sampling arrangement cannot be directly compared using total-score statistics. Moreover, item-by-item reporting ignores the dissimilarities of proficiencies of subgroups to which the set of items was administered. Finally, using the average percentage of items answered correctly to estimate the mean proficiency of examinees in a given subpopulation does not provide any other information about the distribution of skills within that subpopulation (e.g., variances).

The limitations of conventional scoring methods can be overcome by using IRT scaling. When a set of items requires a given skill, the response patterns should show regularities that can be modeled using the underlying commonalities among the items. This regularity can be used to characterize respondents as well as items in terms of a common scale, even if not all respondents take identical sets of items. This makes it possible to describe distributions of performance in a population or subpopulation and to estimate the relationships between proficiency and background variables.

To increase the accuracy of the cognitive measurement, PIAAC uses plausible values (PVs) – which are multiple imputations – drawn from a posteriori distribution by combining the IRT scaling of the cognitive items with a latent regression model using information from the BQ (see chapters 3 and 20) in a population model.

In the following, the population model used for PIAAC scaling (IRT analysis, latent regression model, and computation of plausible values) is described formally (see section 17.2.). Its application to the PIAAC data is then demonstrated (see section 17.3.).

17.2 The latent regression item response model

This section reviews the scaling model employed in the analyses of the PIAAC data in theory – a latent regression item response model – and explains the multiple imputation or "plausible values" methodology that aims to increase the accuracy of the estimates of the proficiency distributions for various subpopulations and the population as a whole.

Most cognitive skills tests are concerned with accurately assessing the performance of individual respondents for the purposes of diagnosis, selection or placement. The accuracy of these measurements can be improved, meaning reducing the amount of measurement error, by increasing the number of items administered to the individual. Thus, achievement tests containing more than 70 items are common. Because the uncertainty associated with each estimated proficiency θ is negligible, the distribution of proficiency or the joint distribution of proficiency with other variables can be approximated using individual proficiencies. When analyzing the distribution of proficiencies for populations or subpopulations, however, more efficient estimates can be obtained from a matrix-sampling design.

In international large-scale assessments (ILSAs) such as PIAAC, test forms are kept relatively short to minimize individuals' response burden. At the same time, ILSAs aim to achieve broad coverage of the tested constructs. The full set of items is organized into different, but linked, assessment booklets; each individual receives only one booklet. Thus, the survey solicits relatively few responses from each respondent while maintaining a wide range of content representation when responses are aggregated. The advantage of estimating population characteristics more efficiently is offset by the inability to reliably measure and make precise statements about individuals' performance. Point estimates of proficiency that are (in some sense) optimal for each respondent could lead to seriously biased estimates of population characteristics (Wingersky, Kaplan, & Beaton, 1987). The "plausible value" methodology correctly accounts for error (or uncertainty) at the individual level by using multiple imputed proficiency values (plausible values) rather than assuming that this type of uncertainty is zero. Retaining this component of uncertainty requires that additional analysis procedures be used to estimate examinee proficiencies. This is done by applying a latent regression item response model to the data.

The latent regression item response model used for PIAAC incorporated test responses (responses to the cognitive items) as well as variables measured by the BQ (e.g., academic and nonacademic activities, and attitudes), which serve as covariates, in the computation of plausible values (von Davier, Sinharay, Oranje & Beaton, 2006). This approach was carried out as follows:

1) *Item calibration based on IRT*: An IRT model was fitted to the item responses. The responses consisted of dichotomous and polytomously scored values. These responses were used to calibrate the test and provide item parameter estimates for the (cognitive) test items.
2) *Population modeling using latent regressions and PV generation*: The population model assumes that item parameters are fixed at the values obtained in the calibration stage. Once the item parameters were estimated, a latent regression model was fitted to the data to obtain regression weights (Γ) and a residual variance-covariance matrix for the latent regression ($\sum$). Next, plausible values (Mislevy & Sheehan, 1987; von Davier, Gonzalez

& Mislevy, 2009) were obtained for all examinees using the item parameter estimates from the item calibration stage and the estimates of Γ and ∑ from the latent regression model.

3) *Variance estimation*: To obtain a variance estimate for the proficiency means of each country and other statistics of interest, a replication approach (see, e.g. Johnson, 1989; Johnson & Rust, 1992) was used to estimate the sampling variability as well as the imputation variance associated with the plausible values.

The analytic procedures that establish these three modeling stages are explained further in the following sections.

17.2.1 Item response theory (item calibration)

PIAAC used the two-parameter logistic model (2PL; Birnbaum, 1968) for dichotomously scored responses and the generalized partial credit model (GPCM; Muraki, 1992) for items with more than two response categories.

The *2PL model* is a mathematical model for the probability that an individual will respond correctly to a particular item from a single domain of items. The probability of solving an item depends only on the respondent's ability or proficiency and two item parameters characterizing the properties of the item (item difficulty and item discrimination).The probability is given as a function of this person parameter and the two item parameters; it can be written as follows:

$$P(x_{ij} = 1 | \theta_j, \beta_i, \alpha_i) = \frac{\exp(\alpha_i(\theta_j - \beta_i))}{1 + \exp(\alpha_i(\theta_j - \beta_i))}$$

where

x_{ij} is the response of person j to item i, 1 if correct and 0 if incorrect;

θ_j is the proficiency of person j (note that a person with higher proficiency has a greater probability of responding correctly);

α_i is the slope parameter of item i, characterizing its sensitivity to proficiency (item discrimination);

β_i is its locator parameter, characterizing item difficulty.

Note that, for $\alpha_i > 0.0$ this is a monotone increasing function with respect to θ; that is, the conditional probability of a correct response increases as the value of θ increases. In addition, a linear indeterminacy exists with respect to the values of θ_j, α_i, and β_i for a scale defined under the 2PL model. In other words, for an arbitrary linear transformation of θ say $\theta^* = A\theta + B$, the corresponding transformations $\alpha^*_i = \alpha_i / A$ and $\beta^*_i = A\beta_i + B$ give:

$$P(x_{ij} = 1 | \theta_j^*, \beta_i^*, \alpha_i^*) = P(x_{ij} = 1 | \theta_j, \beta_i, \alpha_i)$$

A central assumption of IRT is conditional independence (sometimes also called local independence). In other words, item response probabilities depend only on θ and the specified item parameters – there is no dependence on any demographic characteristics of the examinees, or responses to any other items presented in a test, or the survey administration conditions.

Moreover, the 2PL model assumes unidimensionality, that is, a single latent variable, θ, accounts for performance on a set of items. This enables the formulation of the following joint probability of a particular response pattern $\boldsymbol{x} = (x_1, \ldots, x_n)$ across a set of n items.

$$P(\boldsymbol{x}|\theta, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \prod_{i=1}^{n} P_i(\theta)^{x_i} \left(1 - P_i(\theta)\right)^{1-x_i}$$

When replacing the hypothetical response pattern with the scored observed data, the above function can be viewed as a likelihood function that is to be maximized with respect to the item parameters. To do this, it is assumed that respondents provide their answers independently of one another and that the respondent's proficiencies are sampled from a distribution $f(\theta)$. The likelihood function is characterized as

$$P(\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\alpha}) = \prod_{j=1}^{J} \int \left(\prod_{i=1}^{n} P_i\left(\theta_j\right)^{x_{ij}} \left(1 - P_i\left(\theta_j\right)\right)^{1-x_{ij}} \right) f(\theta) d\theta$$

The item parameters obtained by maximizing this function are used in the subsequent analyses.

The GPCM (Muraki, 1992), like the 2PL, is a mathematical model for the probability that an individual will respond in a certain response category on a particular item. While the 2PL is suitable for dichotomous responses only, the GPCM can be used with polytomous and dichotomous responses. The GPCM reduces to the 2PL when applied to dichotomous responses. For an item *i* with m_i+1 ordered categories, the model equation of the GPCM can be written as:

$$P\left(x_i = k \middle| \theta_j, \alpha_i, \beta_i, \boldsymbol{d}_i\right) = \frac{\exp\left\{\sum_{r=1}^{k} 1.7\alpha_i(\theta_j - \beta_i + d_{ir})\right\}}{\sum_{u=1}^{m_i-1} \exp\left\{\sum_{r=0}^{u} 1.7\alpha_i(\theta_j - \beta_i + d_{ir})\right\}}$$

where d_i is the category threshold parameter.

Although the assumption of unidimensionality for the 2PL and GPCM may be considered a strong assumption, the use of these models is motivated by the need to summarize overall performance parsimoniously within a single domain. Hence, item parameters are estimated for each skill scale separately.

A critical part of the data analysis involves testing the assumptions of the 2PL, especially the assumption of conditional independence and the assumption of unidimensionality. Conditional independence means that respondents at a given ability level have the same probability of producing a correct response on an item regardless of their responses to other items as well as other attributes, including background variables such as citizenship, gender, immigrant status. Serious violation of the conditional independence assumption would undermine the accuracy and integrity of the results.

It is not uncommon for some items to violate this assumption. One expression of these types of model violations is differential item functioning (DIF), which means that items are either unsuitable, or much harder or easier, for a particular subpopulation compared to the other groups within the population. While the item parameters were being estimated, empirical conditional percentage-correct statistics were monitored across the samples to test for DIF in PIAAC. More

precisely, for each item, the empirical item characteristic curves (ICC) for each country were compared to the expected ICC of the item. If the empirical ICCs for a certain item differed noticeable from the expected ICC, this would be evidence of DIF. For each country, a few items were identified that showed DIF in the international calibration (see section 17.3.2) and thus, did not conform to the common (international) item parameters.

Country-specific item parameters (computing national calibrations; see section 17.3.2) for items exhibiting country-level DIF in the international calibration were estimated to reduce potential bias introduced by these deviations. This approach was favored over dropping the country-specific item responses for these items from the analysis in order to retain the information from these responses. While the items with country DIF treated in this way no longer contribute to the international set of comparable responses, they continue to contribute to the reduction of measurement uncertainty for the specific country.

The software used for calibration, *mdltm* (von Davier, 2005), was enhanced by implementation of an algorithm that monitored DIF measures and that automatically generated a suggested list of country specific item treatments. This algorithm grouped similar deviations of subgroups of countries so that unique parameters were assigned to either individual countries or country groups that showed the same level and direction of deviation.

17.2.2 Population modeling using latent regressions

The population model used for PIAAC is a combination of an IRT model and a latent regression model. In the latent regression model, the distribution of the proficiency variable (θ) is assumed to depend not only on the cognitive item responses X but also on a number of predictors Y, which are variables obtained from the BQ (e.g., gender, country of birth, education, occupation, employment status, reading practices, etc.). Both the item parameters from the calibration stage and the estimates from the regression analysis are needed to generate plausible values.

Usually, a considerable number of background variables (predictors) are collected in ILSAs, with a principal component analysis extracting the components that explain 90% of the variation for further analysis. In PIAAC it was decided to use 80% of explained variance to avoid overparameterization; (see section 17.3.4.). The use of principal components also serves to retain information for examinees with missing responses to one or more background variables. For the regression of the background variables on the proficiency variable it is assumed that:

$$\boldsymbol{\theta} \sim N(\mathbf{y}\Gamma, \Sigma)$$

The latent regression parameters Γ and Σ are estimated conditional on the previously determined item parameter estimates (from the item calibration stage). Γ is the matrix of regression coefficients and Σ is a common residual variance-covariance matrix.

The latent regression model of Θ on Y with $\Gamma = (\gamma_{sj}, s = 1,\ldots,S; l = 0,\ldots,L)$, $Y = (1, y_1, \ldots, y_L)^t$, and $\Theta = (\theta_1, \ldots, \theta_S)^t$ can be described as follows:

$$\theta_i = \gamma_{s0} + \gamma_{s1}y_1 + \ldots + \gamma_{sL}y_L + \varepsilon_s$$

where ε_i is an error term for the assessment skill s.

The residual variance-covariance matrix can then be described with the following equation:

$$\Sigma = \Theta\Theta^t - \Gamma(YY^t)\Gamma^t$$

Plausible values for each respondent *j* are drawn from the conditional distribution:

$$P(\theta_j|\boldsymbol{x}_j, \boldsymbol{y}_j, \Gamma, \Sigma)$$

Using standard rules of probability, the conditional probability of proficiency can be represented as follows:

$$P(\theta_j|\boldsymbol{x}_j, \boldsymbol{y}_j, \Gamma, \Sigma) \propto P(\boldsymbol{x}_j|\theta_j, \boldsymbol{y}_j, \Gamma, \Sigma)P(\theta_j|\boldsymbol{y}_j, \Gamma, \Sigma) \; = P(\boldsymbol{x}_j|\theta_j)P(\theta_j|\boldsymbol{y}_j, \Gamma, \Sigma)$$

where θ_j is a vector of scale values (these values correspond to performance on each of the three skills), $P(\boldsymbol{x}_j|\theta_j)$ is the product over the scales of the independent likelihoods induced by responses to items within each scale, and $P(\theta_j|\boldsymbol{y}_j, \Gamma, \Sigma)$ is the multivariate joint density of proficiencies of the scales, conditional on the observed value y_j of background responses and parameters Γ and Σ. The item parameters are fixed and regarded as population values in the computation described in this section.

The basic method for estimating Γ and Σ using the expectation-maximization (EM) algorithm is described in Mislevy (1985) for the single scale case. The EM algorithm requires the computation of the mean and variance, of the posterior distribution in (10).

After the estimation of Γ and Σ is complete, plausible values are drawn in a three-step process from the joint distribution of the values of Γ for all sampled respondents. First, a value of Γ is drawn from a normal approximation to $P(\Gamma,\Sigma|\mathbf{x}_j,\mathbf{y}_j)$ that fixes Σ at the value $\hat{\Sigma}$ (Thomas, 1993). Second, conditional on the generated value of Γ (and the fixed value of $\Sigma = \hat{\Sigma}$), the mean m_j^p, and variance Σ_j^p of the posterior distribution are computed using the same methods applied in the EM algorithm. In the third step, the θ are drawn independently from a multivariate normal distribution with mean m_j^p and variance Σ_j^p. These three steps were repeated 10 times, producing 10 imputations of θ for each sampled respondent (see section 17.3.4.).

The software DGROUP (Rogers et al., 2010) was used to estimate the latent regression model and generate plausible values. A multidimensional variant of the latent regression model was used that is based on Laplace approximation (Thomas, 1993).

17.3 Application to PIAAC

This section illustrates an application of the different steps of the population modeling described above using the PIAAC Main Study data. First, an overview of the data preparation is given. Then the national and international item calibration using the 2PL and the GPCM is described, as well as the computation of plausible values and their transformation onto the reporting scale. More specifically, the procedures utilized for the linking, with the aim to obtain equivalent scales, are described.

Scaling and analyses of the PIAAC data were carried out separately for each of the domains literacy, numeracy, and problem solving in technology-rich environments. By creating a separate scale for each, it remains possible to explore potential differences in subpopulation performance across these skills.

17.3.1 Sample size, data preparation, scoring, handling of missing values, block order effects

The following section provides an overview of the sample size, the number of items in the PIAAC assessment, the scoring and handling of missing values, and the examination of block order effects.

Sample size

PIAAC collected competency (cognitive) information through a series of assessment booklets containing literacy, numeracy and problem-solving tasks, and descriptive information through a BQ. Respondents were sampled using a stratified sampling method. Each participating country received instructions for sampling, weighting and data collection. However, each country carried out the actual design and administration of data collection activities separately.

PIAAC respondents' ages ranged from 16 to 65. Eligible participants included individuals who were living in households; institutional populations were excluded. Australia included participants younger than 16 and older than 65 in its target population, but these respondents were excluded from the PIAAC scaling process. Thus, tables comparing proficiency distributions of countries only include respondents between the ages of 16 and 65.

As with ALL, most countries used a modest monetary incentive in PIAAC. Without incentives, the participation rate may have been low enough to undermine the comparability of results.

Twenty-four countries participated in PIAAC (see Table 17.1). All 24 countries were asked to deliver their data before a certain deadline in order to allow sufficient time for analysis and reporting. Data from 331,863 respondents were received; the weighted data from 165,599 respondents between the age of 16 and 65 were available for statistical analyses (after data cleaning).

Table 17.1: Participating countries in PIAAC and sample sizes

Country	Sample Size (n)	Country	Sample Size (n)
Australia	7,430	Italy	4,621
Austria	5,130	Japan	5,278
Canada	27,285	Korea, Republic of	6,667
Canada (English)	21,374	Netherlands	5,170
Canade (French)	5,911	Norway	5,128
Cyprus[1]	5,053	Poland	9,366
Czech Republic	6,102	Russian Federation[2]	3,892
Denmark	7,328	Slovak Republic	5,723
Estonia	7,632	Spain	6,055
Finland	5,464	Sweden	4,469
Flanders (Belgium)	5,463	United Kingdom	8,892
France	6,993	*England (UK)*	5,131
Germany	5,465	*N. Ireland (UK)*	3,761
Ireland	5,983	United States of America	5,010

Assessment mode, testing time, item number and response format:

PIAAC was composed of a BQ and a core set of questions focusing on ICT applied through an interview using a computer-assisted format, and a cognitive assessment measuring the three domains. Based on the information from the BQ, the cognitive assessment was administered with either a CBA or PBA. Table 17.2 provides an overview of the frequency of selection and routing of respondents into these assessment modes.

Table 17.2: Proportion of the application of the assessment modes by domain in PIAAC

Domain	PBA (%)	CBA (%)	PBA+CBA (%)
Core	**22.8**	**73.9**	**96.7**
Literacy	**10.6**	**50.8**	**61.4**
Numeracy	**10.4**	**50.9**	**61.3**
Problem Solving	**NA**	**33.7**	**33.7**

[1] Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

[2] Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

The BQ consisted of 258 variables (measured by more than 258 items, often exceeding 400 items; different countries had a different number of BQ items due to different country specific needs) measuring demographic characteristics, educational experiences, labor market experiences, and activities related to the assessed skills. In general, these questions did not require respondents to read any materials; they were administered by an interviewer, and only those questions that are applicable to the respondents' background were presented (see also chapters 3 and 20). Thus a respondent's reading proficiency was not a primary factor in the collection of the background information. In cases where the selected respondent was unable to speak the official language, another household member was permitted to act as an interpreter between interviewer and respondent for the collection of the background information only. Responses to the background questions served two major purposes. First, they provide a way to summarize the survey results using an array of descriptive variables, such as gender, age, educational attainment and country of birth. Second, they were used in the population model to increase the accuracy of the proficiency estimates for various subpopulations as described in section 17.2.

The *ICT core and the domain-based core part* are described in more detail in Chapter 1 of this volume. These sets of core items were used in selecting the paper or computer path for the respondents as well as the level of the computer-based stages in the subsequent assessment.

The *cognitive assessment* consisted of 166 items: literacy (76 items), numeracy (76 items), and problem solving (14 items). An additional 100 items measuring reading component skills were administered in a PBA if respondents failed to succeed in the other cognitive domains, for a total of 266 items in the cognitive assessment pool. Table 17.3 provides an overview of the number of items per cognitive domain and assessment mode. The large number of items was necessary to achieve adequate content coverage for each domain.

Table 17.3: Number of cognitive items per assessment mode and domain in PIAAC

Domain (Subscale)	Assessment Mode	Number of Items
Literacy	CBA	52
	PBA	24
Numeracy	CBA	52
	PBA	24
Problem Solving	CBA	14
Reading Components	PBA	100

Note: 18 literacy and 17 numeracy items were linking items between the PBA and CBA assessment mode, meaning these items were identical; thus PIAAC contained a total of 131 unique items

Each individual assessment started with the BQ, followed by the core items, and finished with the cognitive assessment. Each survey participant spent approximately 75-100 minutes on the entire assessment:

- BQ and ICT core items: 25-40 minutes

- cognitive assessment (including core items), one booklet: 50-60 minutes (341 booklets: four paper-based booklets and 337 computer-based booklets/paths; see Chapter 1)

The cognitive items were administered using either short open-ended response formats on paper or computer-based open response formats (e.g. highlighting the correct phrase or word); responses were classified into four categories: correct, incorrect, omitted, and not presented.

Scoring and handling of missing data

The 76 literacy items, 76 numeracy items, and 100 reading component items were dichotomously scored (solved: 1, not solved: 0), while the 14 problem-solving items were dichotomously or polytomously scored (five 3-point, one 2-point, and eight dichotomously scored items). For the problem-solving items, an automated scoring algorithm was used to score the responses from the CBA. One of the innovations introduced in PIAAC was the use of the LCS algorithm (longest common subsequence); this algorithm allowed for a scoring method that is automated yet emulates the leniency shown by human scorers in cases where underlining or highlighting responses would typically be evaluated. Humans recognize with ease if a respondent highlights or underlines the correct phrase even if they carelessly error omit one or two characters at the end of the line, at the beginning, or somewhere in the middle of the text. The LCS was used in conjunction with a discrepancy measure to allow for scoring of these "almost complete" responses in a comparable way across countries. As part of this process, a country-and-language independent threshold was established for each item based on the rationale that reasonably small deviations from the completely correct underlining should be considered as correct responses (Sukkarieh, von Davier & Yamamoto, 2012).

Regarding the handling of missing data, the PIAAC design followed a similar procedure to those used in prior studies (ALL and IALS) in order to provide comparability. Because this was a voluntary survey of the adult population without direct consequence to the test taker, missing data in PIAAC has a characteristic structure that relates to the matrix sampling design and the instituted accommodation for respondents with very low literacy skills through core items. This structure is in part characterized by data missing completely at random (MCAR; within each path due to random assignment of blocks) as well as data missing at random (MAR), due to the self-assigned choice of the paper versus computer path or the selection of this path based on background data. More specifically, there are different types of missing values within the *cognitive part* of PIAAC:

1) Missing by design: items that were not presented to each respondent due to the matrix sampling design used in PIAAC (see Chapter 1). Accordingly, these structural missing data, unrelated to respondents' literacy, numeracy, and problem-solving skills, were ignored when calculating respondent proficiencies.

2) Omitted responses: missing responses that occurred when respondents chose not to perform one or more presented items, either because they were unable to do so or some other reason. Any missing response followed by a valid response (whether correct or incorrect) was defined as an omitted response. Omitted responses in the PBA were treated as wrong, because a random response to an open-ended item would almost certainly result in a wrong answer. In the case of the CBA, where it was possible to assess response times per item, nonresponses due to rapid omission were differentiated

from nonresponses after interaction with the stimuli (based on literature on response latencies; cf. Setzer & Allspach, 2007; Wise & DeMars, 2005; Wise & Kong, 2005). Thus, omitted responses were only treated as wrong if a responded spent more than five seconds on an item. If a respondent spent less than five seconds, the nonresponse was considered not attempted and treated as a missing value.

3) Not reached or not attempted responses: missing responses at the end of a block were treated as if they were not presented due to the difficulty of determining if the respondent was unable to finish these items or simply abandoned them.

Cases where respondents did not answer a sufficient number of background questions (< 5 items) were considered as incomplete cases and not used in the latent regression, and also not included in computing plausible values.

Some respondents who answered a sufficient number of background questions may not have been able to respond to the cognitive items or were unwilling to respond to the cognitive items. In these instances, the interviewers were required to document the extent to which the background questions and cognitive items were answered and to ascertain the reason for missing responses. These reasons may be categorized as:

1) Nonresponse due to refusal to participate, thus unrelated to literacy, numeracy, and problem-solving skills

2) Unable to respond due to a language difficulty or cognitive skill-related disability, thus indicating a deficiency of literacy, numeracy and problem-solving skills

3) Inability to provide a written response due to a physical disability

4) Other unspecified reasons

Only the missing responses of nonrespondents in the second category were imputed as incorrect. The rest of the missing responses were considered unrelated to cognitive skills and thus ignored.

On average across countries (based on the weighted and standardized data), 96.9% of respondents completed a BQ and responded to the cognitive items.

Respondents who correctly solved fewer than three of the six core items on the CBA, and fewer than four of the eight core items on the PBA (after the BQ and before the cognitive assessment) were not required to continue with an additional task booklet of cognitive items; their missing responses were considered incorrect for the proficiency estimation. This decision was based on the findings in the Field Test, which showed that respondents who correctly answered fewer than three of the six, or four of the eight core items, were not likely to provide a correct answer to more than 8% of items.

Treatment of respondents with fewer than five cognitive item responses

This section addresses the issue of respondents who provided background information but did not completely respond to the cognitive items. A minimum of five completed items per domain was necessary to assure sufficient information about the proficiency of respondents. On average, 1.7% of the PIAAC samples responded to fewer than five cognitive items per subscale.

Many large-scale assessment programs such as the National Assessment of Educational Progress (NAEP), the National Educational Longitudinal Study (NELS), and the 1985 Young Adult Literacy Survey (YALS) have excluded nonresponding cases from the analyses. Even though a proportion of the missing data and some of the characteristics of the missing data sample were reported, their impact on the analyses was not determined. This practice can yield both biased and inaccurate proficiency distributions for some subpopulations because of differential response rates among subpopulations. For example, individuals who were excluded based on a failure to answer core items for the 1985 YALS were predominantly Hispanic; hence, Hispanic subpopulation results were based only on those who read English, The summary table does not indicate the impact of the non-English readers within the Hispanic population. It should be emphasized again that the presence of extensive background information related to one's cognitive skill is necessary to implement any method for the imputation of proficiency scores.

In some cases, a sampled individual decided to stop the assessment. The reasons for stopping may be classified into two groups: those unable to respond to the cognitive items (i.e., for cognitive-related reasons), and those unwilling to respond (i.e., for noncognitive-related reasons). It should be noted that 2.8% of cognitive-related reasons were either "failed PBA core items" or "failed CBA core items."

PIAAC followed the ALL and IALS procedure with respect to cases with responses to fewer than five cognitive items per domain. All consecutively missing responses at the end of a block of items were treated as incorrect if the reason for not responding to the cognitive items was related to the cognitive skills (literacy, numeracy, problem solving). Otherwise, all consecutively missing responses were treated as "not reached."

This scoring method is important with regard to the latent regression population model described in section 17.2. The population model is used to estimate proficiency values based on responses to the background questions and the cognitive items. A respondent's proficiency is determined from an a posteriori distribution that is the product of two functions: a conditional distribution of proficiency given responses to the background questions, and a likelihood function of proficiency given responses to the cognitive items. The treatment of nonresponding examinees due to noncognitive-related reasons has no impact on the likelihood function of proficiency. On the other hand, there is an impact associated with the treatment for nonresponding cases due to cognitive-related reasons. In the latter case, the likelihood function will be very peaked at the lower end of the scale, which is believed to correctly represent the proficiency of those who are unable to respond to the cognitive items. With this scoring procedure, summary statistics can be produced for the entire population, including those who respond to cognitive items correctly in various degrees, as well as those who were not able to respond to cognitive items.

Furthermore, examinees with responses to fewer than five cognitive items per domain were not included in a first run of the population modeling (with regard to the regression model) to obtain unbiased Γ and Σ. In a second analysis, the regression parameters were treated as fixed to obtain plausible values for all cases, including those with fewer than five responses to cognitive items. More detailed information is provided in section 17.3.4.

Item statistics under adaptive testing

Nonadaptive large-scale population surveys such as Programme for International Student Assessment and Trends in International Mathematics and Science Study, where each block of

items are administered to randomly equivalent respondents through a type of balanced incomplete block design, the standard item statistics represent entire samples. Solely based on this randomly equivalent groups responding to every item, the item statistics are comparable across items within a country as well as across countries. In comparison, PIAAC used two levels of adaptive testing resulting in that standard item statistics represent only subsets of the entire sample and these subsets were defined through type of skills and proficiencies. Thus the standard item statistics are not comparable across items within a country or across countries.

The first level of adaptation used in PIAAC is in terms of mode of administration. Through a series of questions and responses to the CBA core items, PBA items were administered to those without ICT skills and those who were not willing to participate in the CBA. The rest of the respondents in each country (those with ICT skills who were willing to take the assessment on the computer) took CBA. The proportions of the two groups differ by country and demographic characteristics such as age and education, and also they differ by ability. PBA and CBA items were not administered to randomly equivalent group of respondents.

The second level of adaptation in PIAAC was within the CBA portion of the assessment. PIAAC used a probability-based multistage adaptive algorithm where the cognitive items for literacy and numeracy were not administered to randomly equivalent groups of respondents. In other words, more able respondents received a more difficult set of items than less able respondents. Thus item statistics of "easy items" were no longer comparable with "difficult items." Moreover, the countries differed in the distributions of skills, resulting in the distributions of administered items being different. CBA items were not administered to randomly equivalent group of respondents.

However, the comparability of item statistics across countries could be increased by standardizing the proportions of adaptive paths. Such an approach was used to evaluate block order effect in the next section.

Block order effect in the CBA

A block order effect is present when a different order of blocks of items impacts the proportion of correct item responses, that is, the item difficulty or some other characteristic of the item. Stated differently, examinee proficiency (with regard to the measured domains) and the manner in which the survey is administered influences the survey outcomes. As a precaution, the PIAAC design in the CBA was created in order to counterbalance the potential effects of item order on the difficulty of the items. In PIAAC, each respondent received two cognitive modules, where each module comprised either literacy, numeracy or problem-solving items. Each module of literacy and numeracy items appeared in two different positions within the assessment (block-order design: literacy – numeracy; numeracy – literacy, literacy – problem solving2; problem solving1 – literacy; numeracy – problem solving2; problem solving1 – numeracy; problem solving1 – problem solving2; see Chapter 1). The order of content-related blocks was examined to determine if there was any effect on the outcome of the literacy and numeracy proficiencies (note that it was not possible to examine order effects on the domain of problem solving in technology-rich environments as the different problem-solving blocks comprised different items, in contrast to the two other domains). Table 17.4 shows the average proportion correct for items in a given block for PIAAC; the average proportion is calculated from the weighted and standardized data for all participating countries. While the average proportions correct across all countries are virtually identical within 1 percentage point regardless of paired domains as long as

domain order is the same, a slight block order effect was found, 2.8% for literacy modules and 1.3% for numeracy modules.

The weighted proportion correct for an item was calculated as follows:

$$P_i = \frac{\sum_k WP_k \sum_j W_j \left(x_{ji} = 1 \middle| k \right)}{\sum_k WP_k \left(\sum_j W_j \left(x_{ji} = 1 \middle| k \right) + \sum_j W_j \left(x_{ji} = 0 \middle| k \right) + \sum_j W_j \left(x_{ji} = 2 \middle| k \right) \right)}$$

where proportion correct on item i was calculated by using standardized weights of path k WP_k, final weights for the respondent j, scores responses correct ”1”, incorrect ”0”, and omit “2”.

Table 17.4: Average proportion correct; content-related block-by-block order (PIAAC Main Study)

Country	Average of Literacy Items 1st Module		Average of Numeracy Items 1st Module		Average of Literacy Items 2nd Module		Average of Numeracy Items 2nd Module	
	LIT-NUM	LIT-PS2	NUM-LIT	NUM-PS2	NUM-LIT	PS1-LIT	LIT-NUM	PS1-NUM
Australia	56.7%	58.8%	67.8%	67.2%	53.0%	55.9%	67.2%	67.1%
Austria	61.5%	61.0%	64.5%	65.1%	58.9%	58.8%	63.4%	63.1%
Canada	58.7%	58.4%	63.8%	62.5%	54.6%	55.6%	61.9%	62.4%
Cyprus[3]	49.4%		60.7%		45.8%		60.8%	
Czech Rep.	53.5%	54.4%	68.6%	65.4%	53.9%	51.6%	64.7%	66.5%
Denmark	58.7%	57.2%	68.9%	68.2%	55.0%	55.2%	67.0%	68.1%
England/N. Ireland (UK)	58.0%	57.6%	60.5%	60.8%	52.2%	51.8%	59.9%	60.4%
Estonia	57.0%	57.1%	65.7%	65.1%	54.2%	54.7%	65.4%	66.9%
Finland	65.5%	65.2%	72.5%	74.0%	63.3%	62.6%	70.2%	67.9%
Flanders (Belgium)	60.0%	57.9%	67.2%	69.7%	57.1%	58.5%	67.3%	65.5%
France	52.1%		60.2%		48.4%		58.8%	
Germany	57.1%	56.6%	66.3%	67.5%	53.0%	51.9%	65.9%	65.3%
Ireland	56.3%	56.4%	60.7%	60.9%	52.1%	50.7%	58.9%	56.5%
Italy	47.5%		56.9%		44.2%		55.6%	
Japan	67.0%	68.9%	75.7%	76.1%	64.3%	64.1%	73.9%	74.1%
Korea	57.2%	57.1%	62.9%	63.4%	56.9%	57.8%	62.9%	60.6%
Netherlands	62.8%	62.3%	68.5%	69.3%	59.6%	61.1%	69.0%	66.8%
Norway	60.3%	61.0%	69.2%	68.2%	59.1%	57.2%	66.2%	68.9%
Poland	56.6%	55.9%	61.5%	60.8%	51.3%	54.2%	62.1%	60.2%
Russian Fed.[4]	53.7%	52.9%	56.5%	58.4%	52.5%	50.4%	57.5%	56.0%
Slovak Rep.	54.5%	55.4%	67.2%	66.9%	53.8%	53.9%	67.0%	66.7%
Spain	48.4%		55.7%		44.8%		55.4%	
Sweden	62.4%	64.7%	69.7%	70.6%	58.5%	61.9%	67.0%	68.9%
United States	57.8%	56.7%	56.9%	58.8%	52.1%	54.9%	56.8%	55.0%
Average$_1$	58.8%	58.8%	65.7%	65.9%	55.8%	56.1%	64.7%	64.3%
Average$_2$	49.4%		58.4%		45.8%		57.7%	

Average$_1$ is based on the countries that participated in the problem solving domain.
Average$_2$ is based on the countries that did not participated in the problem solving domain.

17.3.2 National and international item calibration

Item calibration is the first step in population modeling and provides the item parameters for the cognitive items that are needed as one of the inputs for the population model used to calculate

[3] Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

[4] Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

the plausible values (see section 17.2.). All cognitive items were calibrated using the 2PL or the GPCM model using *mdltm* (von Davier, 2005) for multidimensional discrete latent traits models. The software provides marginal maximum likelihood estimates (MML) obtained using customary expectation-maximization methods (EM), with optional acceleration. Both IRT models are described in detail in section 17.2.

Of the 166 items used for PIAAC, 18 literacy and 17 numeracy items were used as linking items between PBA and CBA (this means those items were identical between PBA and CBA); therefore, PIAAC contained 131 unique items. In other words, 166 items were described by 131 sets of item parameters. The 131 unique items were calibrated together with 132 unique items from IALS and ALL (263 unique items in total; see Table 17.5). The 100 reading component items were not used for the IRT calibration; for those items, descriptive statistics were provided such as percentage of correct responses, as well as overall timing of the reading component test (only 23.5% of the tested population received the reading component assessment). The 76 literacy items (described by 58 sets of item parameters), and the 76 numeracy items (described by 59 sets of item parameters) were scored dichotomously and calibrated using the 2PL in separate unidimensional IRT analyses. The 14 problem-solving items were scored dichotomously or polytomously and were calibrated using the 2PL and GPCM.

The item calibration also comprised a combined analysis using the IALS and ALL data for the purpose of producing linked scale for trend measurement (see section 17.4.2 and the IALS/ALL technical report for more details). Table 17.5 provides an overview of the distribution of the 263 unique cognitive items across the different surveys (ALL, IALS, PIAAC) and assessment modes (PBA, CBA).

Table 17.5: Distribution of the 263 unique cognitive items across surveys and assessment modes by domain used in PIAAC item calibration (Main Study)

		IALS only	IALS + ALL	IALS + PIAAC	IALS + ALL + PIAAC	ALL only	ALL + PIAAC	PIAAC only	Total items in calib-ration
Literacy	PBA	42	30	0	0	45	0	6	123
	CBA	0	0	1	5	0	6	22	34
	PBA+ CBA	0	0	0	3	0	15	0	18
Numeracy	PBA	0	0	0	0	12	0	10	22
	CBA	0	0	0	0	0	13	22	35
	PBA+ CBA	0	0	0	0	0	17	0	17
Problem solving	CBA	0	0	0	0	0	0	14	14
Total items in calibration		42	30	1	8	57	51	74	**263**

Note: Linking items are counted to avoid duplication.

Two out of the 24 countries participating in PIAAC (France and Russia[5]) were unable to meet the data delivery deadline due to organizational reasons. The data for these countries were not included in the item calibration to obtain the international item parameters. However, the data for these countries – after they were received – went through the same quality assurance and national item calibration (to provide national item parameters for items which showed deviation with regard to the international item parameters). Altogether, data from 154,714 PIAAC respondents were used for the international IRT calibration. During the item calibration, sample weights standardized to represent each country equally were used.

As the samples for each assessment (PIAAC, IALS, ALL) came from somewhat different populations with different characteristics, the calibration procedure needed to take into account the possibility of any systematic interaction between the samples and the items that were used to produce estimates of the item parameters and sample distributions. For this reason, a multiple-group IRT model was estimated using a mixture of normal population distributions (one for each sample) where item parameters were generally constrained to be equal across countries with a unique mean and variance for each country. The moments of these distributions were updated at each iteration during IRT calibration.

The item calibration was completed in two consecutive steps: First, the data were analyzed in an international calibration under the assumption that the common data (including the data from all participating countries) were comparable for all items in the assessment. This step was used to obtain estimates of the international (or common) item parameters, which were equal for all countries. In the subsequent step, national (or unique) item parameters were estimated in order to account for national deviations for a small subset of items. This involved a close monitoring of the IRT scaling for item-by-country interactions and allowing country-specific item parameters only in instances where substantial deviations were identified. An algorithmic approach that automatically identified those country-by-item combinations requiring national parameters based on DIF detection was applied. Items not exhibiting appropriate fit using an international parameter received a country-specific parameter. However, if more than one country exhibited a deviation from the international parameters, an algorithm was applied that ensured parsimony in the parameterization. For example, if two countries showed poor item fit for the same item in the international calibration, and in the same direction, both countries received the same unique item parameter estimated for these two countries (note that the term “national item parameters” in this report is used for both cases: one country that receives a unique country-specific item parameter, and more than one country that receive the same unique item parameter which is different from the international item parameter).

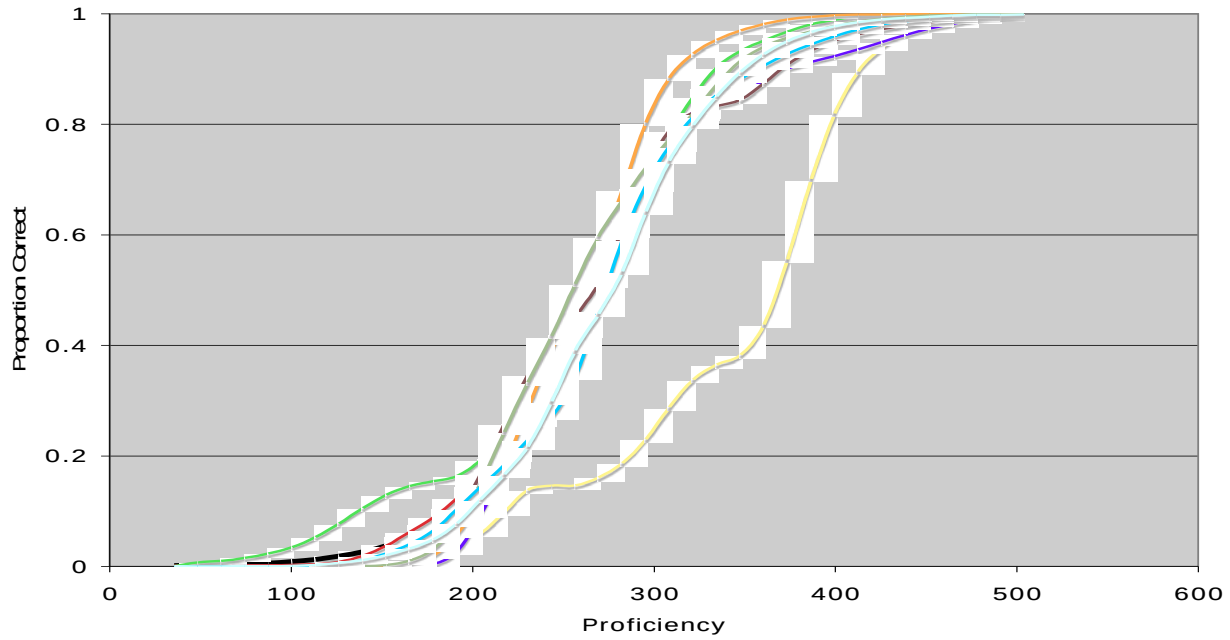
To identify misfitting items, fit statistics were estimated using the mean deviation (MD) and the root mean square deviation (RMSD). The MD is most sensitive to the difficulties of items and can represent a magnitude of shift of observed data from the estimated ICC. The RMSD is a standardized index of the discrepancy between the observed ICC and the model-based ICC; it is sensitive to measure the deviation of the observed item characteristics from the estimated ICC both in terms of slope and location of the item response function. Poorly fitting item characteristic curves were revealed using a RMSD > 0.1 criterion (a value of 0 indicates no discrepancy; in other words, a perfect fit of the model). The identification of poor fitting items and the replacement of international item parameters with country-specific (unique) parameters

[5] Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

was carried out using an automatic algorithm in *mdltm*. Thus, the international and national calibrations were conducted simultaneously for all countries, that is, all estimated item parameters (international and national) are located on one common scale.

In most cases, the item responses across countries were accurately described by the international (common) item parameters. For some items, there was evidence that the estimated parameters did not fit as well for a certain assessment sample from a few countries as compared to the others. However, this pattern was not consistent for any one particular country. Given this estimation and optimization approach, no item was dropped from the analysis in PIAAC. For those items with item functions showing substantial deviation from the international item parameters (poor fitting items), national (unique) item parameters were estimated. If an item showed poor fit but had the same kind of poor fit in multiple countries, an additional country-group specific parameter besides the international or common item parameter was used for this item. If an item showed poor fit in one or two countries only or showed item fit to a different extent in different countries (unique deviation), the unique country-specific item parameters were used for further analysis. Thus, PIAAC allowed for different sets of item parameters to improve model fit and optimize the comparability of countries. Figure 17.1 shows a typical plot of a case (for the 2PL) to illustrate how the data from one country might not support the use of international item parameters.

Figure 17.1: Item response curve for an item where the international item parameter is not appropriate for one country



The solid black line is the fitted two-parameter logistic item response curve that corresponds to the international item parameters; the other lines are observed proportions of correct responses at various points along the proficiency scale for the data from each subpopulation. The horizontal axis represents the proficiency scale. This plot indicates that the observed proportions of correct responses, given the proficiency, are quite similar for most countries. However, the data for one country indicated by the yellow line shows a noticeable departure from the common ICC. This

item is far more difficult in that particular country than expected given the responses on other items. Thus, a unique set of item parameters was estimated for that country.

Table 17.6 provides an overview of the number of country-specific (national) item parameters per country (see also Appendix 17.1 for detailed information), which were used together with the international parameters for the remainder of the items to calculate plausible values in PIAAC. For literacy, country-specific item parameters were estimated for only 8% of the items due to item-by-country interactions. For numeracy, 7% of the items necessitated country-specific parameters, and for problem solving, 3% of unique item parameters were used. (Unique item parameters for Russia[6] were determined after the reduction of the Russian sample by more than 1,200 cases due to issues in those data.)

Table 17.6: Number of national item parameters for each country and proficiency scale

Country	Number of Country-Specific Item Parameters	Number of Country-Specific Item Parameters	Number of Country-Specific Item Parameters
	Literacy (76 items)	Numeracy (76 items)	Problem Solving (14 items)
Australia	2	2	0
Austria	5	1	0
Canada (English)	2	1	0
Canada (French)	6	3	0
Cyprus[7]	13	3	NA
Czech Republic	8	5	1
Denmark	3	5	0
England/N. Ireland (UK)	3	3	0
Estonia	4	4	1
Finland	6	7	0
Flanders (Belgium)	5	5	0
France	8	3	NA
Germany	5	2	0
Ireland	2	2	0
Italy	5	3	NA
Japan	14	16	1
Korea	15	16	2
Netherlands	2	5	1
Norway	6	9	0
Poland	6	6	0
Russian Federation[8]	12	21	3
Slovak Republic	9	3	2
Spain	4	3	NA
Sweden	6	5	0
United States	4	9	0

[6] Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

[7] Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

[8] Please refer to the above note regarding the Russian Federation.

17.3.3 National reports

For the purposes of secondary analyses and transparency, every participating country received the prepared data files including plausible values for the international data, and the country-specific data, respectively. The reported values are based on the international calibration providing a common, comparable scale, with the potential adjustment of utilizing country specific item parameters to improve model fit and reduce bias. National reporting is supported by supplying these databases to each country, and additionally providing a set of tools for further analysis.

17.3.4 Generating plausible values

Plausible values are multiple imputed proficiency values based on information from the test items (the actual PIAAC literacy, numeracy, and problem solving tests) and information provided by the respondent in the BQ. Plausible values are used to obtain more accurate estimates of group proficiency than would be obtained through an aggregation of point estimates. A more detailed description is given in section 17.2 as well as in Mislevy (1991), Thomas (2002), and von Davier, Sinharay, Oranje & Beaton (2006).

In PIAAC, the computation of group-level reporting statistics involving scores in the three domains is based on 10 independently drawn plausible values for each scale assigned to each respondent. Each set of plausible values is equally well designed to estimate population parameters, however, multiple plausible values are required to represent the uncertainty in the domain measures appropriately (von Davier, Gonzalez & Mislevy, 2009). As mentioned earlier, the statistics based on scores are always computed at population or subpopulation levels. They should never be used to draw inferences at the individual level (see also section 18.4). Detailed information on the computation of plausible values in PIAAC is given in section 17.2.2.

For the population modeling and the calculation of plausible values for the scales of PIAAC, the computer program DGROUP (Rogers et al., 2006)[9] was used.

In the analyses of PIAAC, a normal multivariate distribution was assumed for $P(\theta_j|x_j, y_j, \Gamma, \Sigma)$, with a common variance, Σ, and with a mean given by a linear model with slope parameters, Γ, based on the principal components of several hundred selected main effects from the vector of background variables.

The item parameters for the cognitive items were obtained from the concurrent item calibration (see section 17.3.2) using the data from IALS, ALL and PIAAC as described above. The result of the concurrent calibration is a scale that provides comparable results across IALS, ALL and PIAAC. To calculate the plausible values for PIAAC only, the item parameters for the 166 PIAAC items (from the concurrent item calibration) were used in the population modeling.

The background variables included demographic information, educational experiences, occupational experiences and skill use, among others. A description of the different sections of the background data can be found in Chapter 3 of this report. All variables in the BQ were contrast coded before they were processed further in the population model. Contrast coding allows the inclusion of codes for refused responses as well as codes for responses that were not

[9] The statistical program DGROUP can be obtained from ETS on demand.

collected by means of routing and avoiding the necessity of linear coding. The increased number of variables obtained through contrast coding is substantial. To capture most of the common variance in the contrast-coded background questions with a reduced set of variables, a principal component analysis was conducted. Because each population can have unique associations among the background variables, a single set of principal components was not sufficient for all countries included in PIAAC. Therefore, the extraction of principal components was carried out separately by country. In PIAAC each set of principal components y^c (or conditioning variables) was selected to include 80 percent of the variance with the aim of explaining as much variance as possible while at the same time avoiding overparameterization.

Principal component scores based on nearly all background variables were used in PIAAC including international variables (collected by every participating country) as well as national background variables (country specific variables in addition to the international variables). Note, that the principal component analysis and the population modeling were calculated separately for each country in order to take into account the differences in associations between the background variables and the cognitive skills.

A small subset of respondents did not attempt the cognitive items or responded to fewer than five cognitive items for an inability to read or write in the language of assessment, a physical disability, a mental disability, or a refusal to participate in the survey. If these respondents had been excluded from the survey, the proficiency scores of some subpopulations in the PIAAC survey would have been systematically overestimated and the picture of the nation's cognitive skills would have been distorted. Those respondents with an insufficient number of responses (<5) to the cognitive items were excluded from the estimation of the latent regression. In a subsequent step, however, the latent linear regression estimated on the sample for examinees with sufficient numbers of responses was fixed and plausible values were drawn for all respondents. That is, in the second run all cases were included in the analysis but Γ and Σ were fixed to the values of the first run. Hence, a set of plausible values for the cognitive scales were calculated for all respondents regardless of the number of items attempted. The reason for this procedure is that sufficient information about the proficiency cannot be obtained for cases with fewer than five responses to cognitive items. Including these cases could influence the regression analysis, which aims to link background variables and (sufficiently accurate) proficiency estimates with the aim of predicting proficiency. For 2,616 cases across the 23 countries did not receive plausible values because of insufficient information due to literacy-related nonresponse.

17.4 Linking scales across delivery modes and surveys

PIAAC followed two aims with regard to the linking design:

1) Linking the different booklets containing different sets of items administered through different assessment (delivery) modes to each other in order to get comparable cognitive measures;
2) Linking the different ILSA adult surveys (IALS, ALL, PIAAC) to each other to provide trend measures.

17.4.1 Linking different booklets and assessment modes within PIAAC

To obtain comparable test results in all three cognitive domains for all sample groups, it was important that all items (in a given domain) were calibrated on one common scale. However, this was not easy to achieve given the complex test design in PIAAC. As illustrated in Chapter 1, PIAAC used a matrix sampling design where different items from the total item pool were administered to different test takers or groups by using different test booklets. Furthermore, items were administered through a version of adaptive testing, and by using different assessment modes, which made the design even more complex.

To establish a common scale for all items in a given domain, the items had to be linked together across test booklets (subset of items) and assessment modes. This was achieved by using common sets of items in the different booklets and assessment modes. Thus, certain items were administered in both the PBA and CBA (note that this pertains to literacy and numeracy items, as problem solving was only available for the CBA) as well as in different booklets (across different assessment modes). Out of 52 literacy and 52 numeracy items in the CBA, 18 literacy and 20 numeracy items were used to link the computer- and paper-based instruments. Within the CBA, all items were linked together in the booklet design. According to the distribution of the linking items, it was considered that the different item contexts (such as education, personal, work and everyday life), different item contents (such as data and chance, dimension and shape, quantity and number) and different cognitive processes or types of responses (such as integrate and interpret, evaluate and reflect, identify, and locate or access) were present within the linking items. In other words, the linking items were selected with the aim of being representative of the total item pool.

Through these linking items it was possible to calibrate items answered by different respondents in different booklets and assessment modes on one common scale for each cognitive domain. This was done within the item calibration (see section 17.3.2.). Deviations of item-by-country interactions were identified using a measure of MD and RMSD. Results for the PIAAC linking across assessment modes in the Main Study are presented in section 18.4.

17.4.2 Linking previous international adult assessments with PIAAC

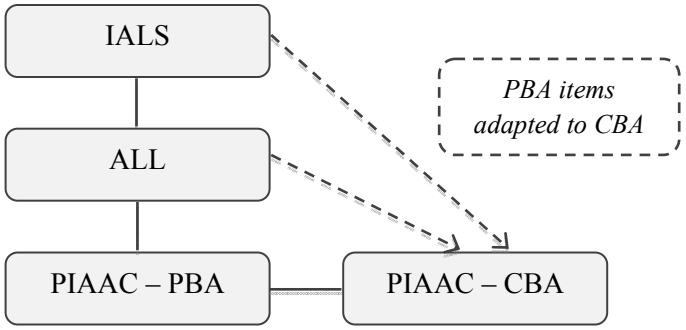
As the intent of PIAAC was to have its results linked to previous international adult assessments, 60 items of the literacy and numeracy items administered in PIAAC came from ALL and IALS. Seventy-four new items were developed for the literacy and numeracy domains, and new measures were developed for the reading components and problem solving domains (based on their respective frameworks) and tested in the PIAAC Field Test. Table 17.5 gives an overview of the item numbers per survey, domain and assessment mode.

The equivalence of item parameters among linking items from IALS and ALL to PIAAC was again evaluated through item calibration by applying IRT models (similar to the evaluation of the link between PBA and CBA in PIAAC).

Entire literacy items, including those unique to a particular survey as well as linking to multiple surveys, were reestimated using the entire aggregate data of IALS and ALL because the literacy scale in PIAAC is a joint scale of prose and document literacy scales (in IALS and ALL). These new parameters were used for the subsequent analyses. The numeracy scale was introduced in the ALL survey, and subsequent analyses used ALL numeracy item parameters.

Equivalence of item characteristics among the literacy and numeracy items common to IALS and ALL on the PBA was examined. As some IALS and ALL items (which used PBA only) were adapted to the CBA in PIAAC (see Figure 17.2), the equivalence of these adapted items to the appropriate IALS/ALL items was evaluated as well in the Field Test. Results for the PIAAC linking across surveys in the Main Study are presented in section 18.4.

Figure 17.2: Linking different international adult assessments and assessment modes (PIAAC)



To place the IALS and ALL items on the same scale as the PIAAC items, the item calibration (and thereby the linking) was used for the items and data from all three surveys. Therefore, the new estimates had to be transformed in order to be comparable to the old estimates, thus allowing the measurement of trend.

After the joint item calibration for all surveys was carried out, a linear transformation of the group means was conducted. The group means and standard deviations of the weighted scores obtained from the old item calibration of the IALS and ALL data were used to transform the new group means and standard deviations from the new joint item calibration (for IALS, ALL and PIAAC). An example of such a transformation is given in Table 17.7.

Table 17.7: Example of a transformation of IRT-based means of a set of old and new countries, calibrated together to find a transformation of the "new" countries' scores to the original scale

Old Countries	Original Mean	IRT New Calibration Based Mean	Transformed New Mean
A	240	0.3	240
B	250	0.4	250
C	260	0.5	260
D	270	0.6	270
E	280	0.7	280
New Countries	**Not Tested**		
F	-	0.3	240
G	-	0.5	260
H	-	0.7	280
I	-	0.55	265

For the trend measure, the transformed means of the weighted scores obtained from the item calibration were used for further analysis. The plausible values were influenced by this transformation as well but are not used for measuring trends.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Johnson, E. G. (1989). Considerations and techniques for the analysis of NAEP data. *Journal of Educational Statistics, 14(4)*, 303-334.

Johnson, E. G., & Rust, K. F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics, 17*, 175-190.

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

Mislevy, R. J. (1985). Estimation of latent group effects. J*ournal of the American Statistical Association, 80*(392), 993–997.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177–196.

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

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–177.

Rogers, A., Tang, C., Lin, M.-J., & Kandathil, M. (2006). DGROUP (computer software). Princeton, NJ: Educational Testing Service.

Setzer, J. C., & Allspach, J. R. (2007, October). *Studying the effect of rapid guessing on a low-stakes test: An application of the effort-moderated IRT model.* Paper presented at the annual meeting of the Northeastern Educational Research Association, Rocky Hill, CT. http://www.psyc.jmu.edu/assessment/research/pdfs/SetzerAllspach_NERA07.pdf

Sukkarieh, J., von Davier, M. & Yamamoto, K. (2012). *From biology to education: Scoring and clustering multilingual text sequences and other sequential tasks* (Research Report Series. ETS RR–12-25

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics, 2*, 309–322.

Thomas, N. (2002). The role of secondary covariates when estimating latent trait population distributions. *Psychometrika, 67*(1), 33-48.

von Davier, M. (2005). *A general diagnostic model applied to language testing data.* Research Report RR-05-16). Princeton, NJ: Educational Testing Service.

von Davier, M., Gonzalez, E. & Mislevy, R. (2009) What are plausible values and why are they useful? *In: IERI Monograph Series: Issues and Methodologies in Large Scale*

Assessments, Vol. 2. Retrieved from IERI website: http://www.ierinstitute.org/IERI_Monograph_Volume_02_Chapter_01.pdf

von Davier, M. Sinharay, S., Oranje, A. & Beaton, A. (2006) Statistical procedures used in the National Assessment of Educational Progress (NAEP): Recent developments and future directions. In C. R. Rao and S. Sinharay (Eds.), *Handbook of Statistics (Vol. 26): Psychometrics*. Amsterdam: Elsevier.

Wingersky, M., Kaplan, B., & Beaton, A. E. (1987). Joint estimation procedures. In A. E. Beaton (Ed.), Implementing the new design: The NAEP 1983-84 technical report (pp. 285-292). Princeton, NJ: ETS.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1-17.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*, 163-183.

Appendixes

Appendix 17.1: Items per country that received country-specific item parameters in the population modeling

Item	Australia	Austria	Flanders (Belgium)	Canada (Eng.)	Canada (Fr.)	Cyprus	Czech Rep.	Germany	Denmark	Spain	Estonia	Finland	France	England/N. Ireland (UK)	Ireland	Italy	Japan	Korea	Netherlands	Norway	Poland	Russia	Slovak Rep.	Sweden	United States
LITERACY																									
C301C05S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
C300C02S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
D302C02S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
D311701S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
E321001S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
E321002S	*	*	*	*	*	*	**X**	*	*	*	*	*	*	*	*	*	*	*	*	*	Δ	*	**X**	*	*
C308117S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
C308119S	*	*	*	*	Δ	*	*	**X**	*	*	*	*	*	*	*	**X**	**O**	*	*	*	*	*	*	*	*
C308120S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
C308121S	*	*	*	*	*	*	*	*	*	**X**	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
C305215S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
C305218S	*	**X**	*	*	Δ	**X**	*	*	*	*	*	*	**U**	*	*	*	*	**O**	*	**V**	**O**	*	*	*	*
D315512S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	Δ	*	*	*	*	**X**	*	*
C308118S	*	*	**X**	*	*	*	*	**X**	*	*	*	*	*	*	*	*	*	*	*	*	Δ	**O**	Δ	**U**	*
D304710S	*	*	*	*	*	*	**X**	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
D304711S	*	*	*	*	*	**X**	Δ	*	*	*	*	*	**O**	*	*	*	*	**X**	*	*	*	**U**	Δ	*	*
C308116S	*	*	**X**	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
E327001S	*	*	*	*	*	**X**	*	*	*	*	*	*	*	*	*	*	Δ	**O**	*	*	*	**U**	*	*	*
E327002S	*	*	*	*	*	*	Δ	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**X**	*	*
E327003S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**X**	*	*	Δ	**O**	*
E327004S	*	*	*	*	*	*	*	*	*	**X**	*	*	*	*	*	*	*	*	*	*	*	*	*	Δ	*

Item	Australia	Austria	Flanders (Belgium)	Canada (Eng.)	Canada (Fr.)	Cyprus	Czech Rep.	Germany	Denmark	Spain	Estonia	Finland	France	England/N. Ireland (UK)	Ireland	Italy	Japan	Korea	Netherlands	Norway	Poland	Russia	Slovak Rep.	Sweden	United States
D307401S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
D307402S	*	*	*	*	*	X	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
C309319S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*	*	*	*	*
C309320S	*	*	*	*	*	*	X	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
C309321S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
C309322S	*	*	X	*	*	*	*	*	*	*	*	Δ	O	*	*	*	*	U	*	*	*	*	*	*	*
E322001S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
E322002S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*	*	*	*	*	*
E322005S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*
C313412S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
C313414S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
E322003S	X	*	*	*	Δ	Δ	*	*	*	*	*	*	O	X	X	*	*	*	*	*	*	*	*	*	*
C310406S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
C310407S	*	*	*	*	*	X	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
E320001S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
E320003S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*	*	*	*	*
E320004S	*	*	*	*	*	*	X	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
E322004S	*	*	*	*	X	*	*	*	*	*	*	*	Δ	*	*	*	O	*	*	*	*	*	*	*	*
D306110S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
D306111S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	*	X	*	*	*	*	*	*	*
C313410S	*	*	*	*	*	X	*	*	*	*	*	*	*	*	*	*	*	*	Δ	*	*	*	*	*	*
C313411S	*	*	*	*	*	X	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
C313413S	*	*	*	*	*	*	*	*	*	*	X	*	*	*	*	*	*	*	*	*	*	*	*	*	*
E323003S	*	*	*	*	*	X	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
E323004S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*	*	*	Δ	*	*	*

Item	Australia	Austria	Flanders (Belgium)	Canada (Eng.)	Canada (Fr.)	Cyprus	Czech Rep.	Germany	Denmark	Spain	Estonia	Finland	France	England/N. Ireland (UK)	Ireland	Italy	Japan	Korea	Netherlands	Norway	Poland	Russia	Slovak Rep.	Sweden	United States
E318001S	*	*	*	*	*	*	*	*	X	*	Δ	*	*	*	*	*	O	*	*	U	V	W	*	*	*
E318003S	*	*	*	*	*	X	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
E329002S	X	*	*	Δ	*	*	*	*	*	*	*	O	*	X	X	*	U	*	*	*	V	*	*	*	Δ
E329003S	*	*	*	*	*	*	*	*	Δ	*	O	*	*	*	*	*	*	X	*	X	O	U	*	V	*
E323002S	*	*	*	*	X	*	*	*	*	*	*	*	*	*	*	*	X	Δ	*	*	*	*	*	*	*
E323005S	*	*	*	*	*	*	X	*	*	*	*	*	*	*	*	*	X	*	*	*	*	*	*	*	*
M301C05S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
P330001S	*	X	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	Δ	*	*	*	*	*	O	*
N302C02S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
M300C02S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
N306110S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*	*	*	*	*
N306111S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*	*	*	*	*	*
M313410S	*	X	Δ	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
M313411S	*	*	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*	*	*	*	*	Δ	*	*	*
M313412S	*	*	*	*	*	*	*	X	*	*	*	*	*	*	*	*	Δ	*	*	*	*	*	*	*	*
M313413S	*	*	X	*	*	*	*	*	*	*	*	Δ	*	*	*	*	O	X	*	*	*	*	U	*	X
M313414S	*	*	*	*	*	*	*	*	X	*	*	*	*	*	*	*	Δ	*	*	*	*	*	*	*	*
P324002S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	O	*	*	*	*	*	*	*	*
P324003S	*	*	*	*	*	X	*	*	*	Δ	*	O	*	*	*	*	U	*	*	*	*	*	*	*	*
M305215S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
M305218S	*	*	*	*	*	*	*	*	*	*	X	Δ	O	*	*	*	*	*	*	Δ	*	U	X	Δ	*

Item	Australia	Austria	Flanders (Belgium)	Canada (Eng.)	Canada (Fr.)	Cyprus	Czech Rep.	Germany	Denmark	Spain	Estonia	Finland	France	England/N. Ireland (UK)	Ireland	Italy	Japan	Korea	Netherlands	Norway	Poland	Russia	Slovak Rep.	Sweden	United States
P317001S	*	X	*	Δ	Δ	*	*	*	*	*	*	X	*	*	*	*	*	O	*	*	*	*	*	*	*
P317002S	*	*	*	*	*	*	*	*	*	X	*	*	*	*	*	*	*	Δ	*	*	*	O	X	*	*
P317003S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
M310406S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
M310407S	*	*	*	*	*	X	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
M309319S	*	X	*	*	*	*	*	X	*	*	*	*	O	*	*	Δ	*	*	X	Δ	*	U	*	*	*
M309320S	*	*	*	*	*	*	X	Δ	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X
M309321S	*	*	*	*	*	X	*	*	*	*	*	*	Δ	*	*	X	*	*	*	*	*	O	*	*	U
M309322S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
NUMERACY																									
C600C04S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
C601C06S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
E645001S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
C615602S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*	*	*	*	*
C615603S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	Δ	*	*	*	*	*	*	*
C624619S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X
C624620S	*	*	*	*	*	*	*	*	*	*	X	*	*	*	*	*	Δ	*	*	*	X	O	*	*	*
C604505S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
C605506S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*
C605507S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*	Δ	*	*	*
C605508S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*
E650001S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*	*	Δ	*	*	*

Item	Australia	Austria	Flanders (Belgium)	Canada (Eng.)	Canada (Fr.)	Cyprus	Czech Rep.	Germany	Denmark	Spain	Estonia	Finland	France	England/N. Ireland (UK)	Ireland	Italy	Japan	Korea	Netherlands	Norway	Poland	Russia	Slovak Rep.	Sweden	United States
C623616S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
C623617S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**X**	*	*	*	*	*	*	*
E657001S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**X**	Δ	*	*	Δ	**O**	*	*	*
C619609S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**X**	*	*	*	*	*	*	*
E632001S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
E632002S	*	*	*	*	*	*	**X**	*	*	*	*	*	*	*	*	*	*	*	*	*	Δ	*	**X**	*	*
E646002S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**X**	*	*	*
C620610S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
C620612S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**X**	*	*	*	*	*	*	*	*
C613520S	*	*	*	*	*	**X**	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
C614601S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
C618607S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
C618608S	*	*	*	*	*	*	*	*	**X**	*	*	*	*	*	*	*	Δ	*	*	*	*	#	*	*	*
E635001S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
C607510S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**X**	*	*	*
E655001S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**X**
C602502S	*	*	*	*	*	*	**X**	*	*	*	*	*	*	*	*	*	*	Δ	*	*	*	**O**	*	*	*
C602503S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**X**	**O**	*	*	*	Δ	*	*	*
C608513S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
C602501S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
C606509S	*	*	*	*	*	*	*	*	*	**X**	Δ	*	**O**	*	*	*	**U**	*	*	*	*	*	*	*	*
C611516S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**X**	*	*	*	*	Δ	*	*	*
C611517S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
C622615S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**X**	*	*	*	*	*	*	*	*
E665001S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**X**	*	*	*

Item	Australia	Austria	Flanders (Belgium)	Canada (Eng.)	Canada (Fr.)	Cyprus	Czech Rep.	Germany	Denmark	Spain	Estonia	Finland	France	England/N. Ireland (UK)	Ireland	Italy	Japan	Korea	Netherlands	Norway	Poland	Russia	Slovak Rep.	Sweden	United States
E665002S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
E636001S	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*	*	*	*	*	*	*	Δ	*	*	*
C617605S	X	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	Δ	*	*
C617606S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*
E660003S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*	*	*	*	*	*	*
E660004S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
E641001S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*	Δ	*	*	*
E661001S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
E661002S	*	*	X	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	Δ	*	*	*	*
C612518S	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*	*	*	*	*	*	*	Δ	*	X	X
E651002S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	O	*	*	*	*	*	*	*
E664001S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*	Δ	O	*	*	*
E634001S	*	*	*	*	*	*	*	X	*	*	X	*	*	*	*	*	*	*	X	Δ	*	*	*	*	O
E634002S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X
E644002S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	Δ	*	*	*	*	*	*	*
M600C04S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*	*	X
P601C06S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*	*	*	*	*
P614601S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X
P645001S	*	*	*	X	X	*	*	*	*	*	*	X	*	*	*	*	*	Δ	*	*	*	*	*	*	Δ
M615602S	*	*	*	*	*	*	X	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
M615603S	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*	*	Δ	O	*	*	*	*	*	*	*
P640001S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*	*	*	*
M620610S	*	X	Δ	*	*	X	*	*	*	Δ	*	*	*	*	Δ	X	*	*	O	O	*	*	O	*	*
M620612S	*	*	*	*	*	*	*	*	*	*	*	*	X	*	*	Δ	*	*	*	*	*	*	*	*	*
P666001S	X	*	*	*	U	Δ	*	V	*	*	*	*	*	Δ	O	*	W	*	X	*	X	Z	*	*	*

Item	Australia	Austria	Flanders (Belgium)	Canada (Eng.)	Canada (Fr.)	Cyprus	Czech Rep.	Germany	Denmark	Spain	Estonia	Finland	France	England/N. Ireland (UK)	Ireland	Italy	Japan	Korea	Netherlands	Norway	Poland	Russia	Slovak Rep.	Sweden	United States
M623616S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
M623617S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*	*	Δ	*	*	*
M623618S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*
M624619S	*	*	X	*	*	*	*	*	*	*	*	O	*	*	*	*	*	*	Δ	*	*	*	*	*	*
M624620S	*	*	*	*	*	*	*	*	X	*	*	Δ	*	X	*	*	*	Δ	*	X	*	*	*	*	*
M618607S	*	*	*	*	X	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	Δ	*	O	*
M618608S	*	*	X	*	*	*	*	*	Δ	O	O	Δ	U	O	*	*	*	*	*	O	*	*	*	Δ	*
M604505S	*	*	*	*	*	*	X	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
M610515S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X	*	*	*	*	*
P664001S	*	*	X	*	*	*	*	*	X	*	*	*	*	*	*	*	*	*	*	X	*	*	*	X	*
M602501S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
M602502S	*	*	*	*	*	*	X	*	*	*	*	*	*	*	*	*	*	*	*	Δ	*	*	*	O	*
M602503S	*	*	*	*	*	*	*	*	X	*	*	*	*	*	*	*	Δ	U	*	O	*	*	*	*	*
P655001S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	X
PSTRE																									
U01A000S	*	*	*	*	*	#	*	*	*	#	*	*	#	*	*	#	*	*	*	*	*	*	*	*	*
U01B000S	*	*	*	*	*	#	*	*	*	#	*	*	#	*	*	#	*	*	*	*	*	*	*	*	*
U03A000S	*	*	*	*	*	#	*	*	*	#	*	*	#	*	*	#	*	*	*	*	*	X	Δ	*	*
U06A000S	*	*	*	*	*	#	*	*	*	#	X	*	#	*	*	#	X	*	*	*	*	Δ	*	*	*
U06B000S	*	*	*	*	*	#	X	*	*	#	*	*	#	*	*	#	*	*	*	*	*	*	Δ	*	*
U21X000S	*	*	*	*	*	#	*	*	*	#	*	*	#	*	*	#	*	*	*	*	*	*	*	*	*
U04A000S	*	*	*	*	*	#	*	*	*	#	*	*	#	*	*	#	*	*	*	*	*	*	*	*	*
U19A000S	*	*	*	*	*	#	*	*	*	#	*	*	#	*	*	#	*	X	*	*	*	*	*	*	*
U19B000S	*	*	*	*	*	#	*	*	*	#	*	*	#	*	*	#	*	*	X	*	*	*	*	*	*
U07X000S	*	*	*	*	*	#	*	*	*	#	*	*	#	*	*	#	*	*	*	*	*	X	*	*	*

Item	Australia	Austria	Flanders (Belgium)	Canada (Eng.)	Canada (Fr.)	Cyprus	Czech Rep.	Germany	Denmark	Spain	Estonia	Finland	France	England/N. Ireland (UK)	Ireland	Italy	Japan	Korea	Netherlands	Norway	Poland	Russia	Slovak Rep.	Sweden	United States
U02X000S	*	*	*	*	*	#	*	*	*	#	*	*	#	*	*	#	*	*	*	*	*	*	*	*	*
U16X000S	*	*	*	*	*	#	*	*	*	#	*	*	#	*	*	#	*	*	*	*	*	*	*	*	*
U11B000S	*	*	*	*	*	#	*	*	*	#	*	*	#	*	*	#	*	**X**	*	*	*	*	*	*	*
U23X000S	*	*	*	*	*	#	*	*	*	#	*	*	#	*	*	#	*	*	*	*	*	*	*	*	*

Note: * denotes international item parameters; all other symbols and letters (**X**, Δ, **O**, **U**, **V**, **W**, **Z**) denote country-specific item parameters; identical symbols/letters in the same row (or for the same item) for different countries denote identical item parameters for the specific item in these countries (identical symbols/letters in different rows/items do not); # denotes items that were not presented in a country or excluded during item calibration (this was the case for one item in one country) – typically this symbol will be found for countries that optioned out of the assessment of PSTRE.