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).

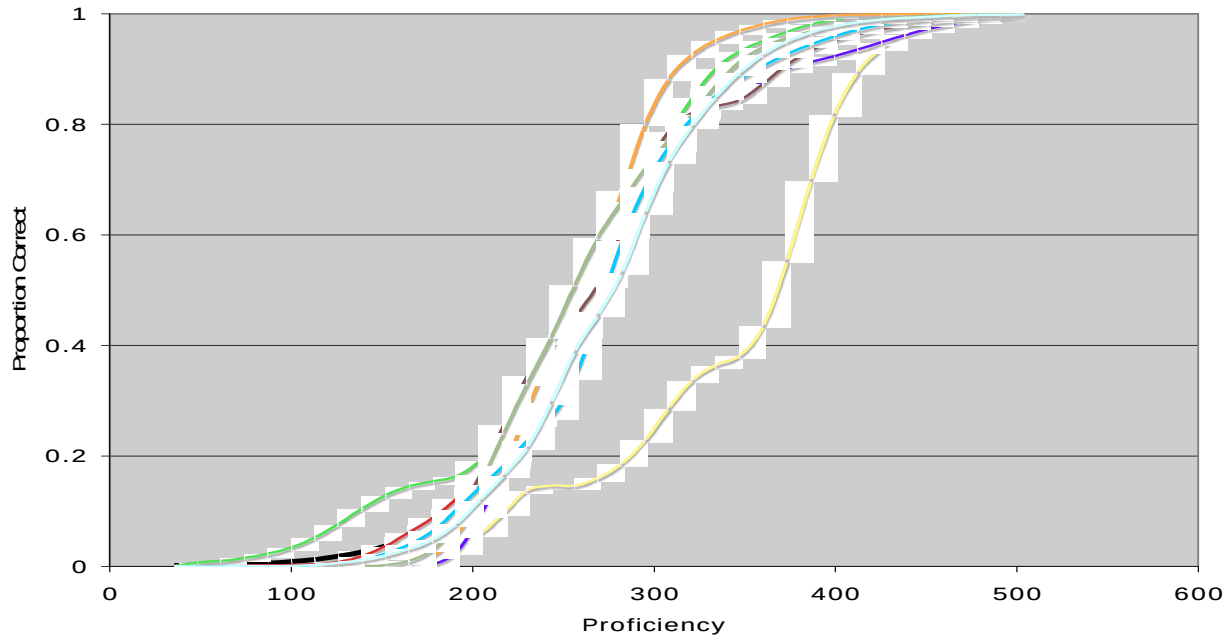
To identify misfitting items, fit statistics were estimated using the mean deviation (MD) and the root mean square deviation (RMSD). The MD is most sensitive to the difficulties of items and can represent a magnitude of shift of observed data from the estimated ICC. The RMSD is a standardized index of 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

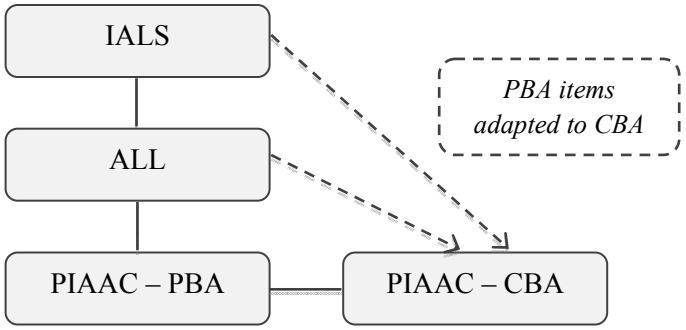
As the intent of PIAAC was to have its results linked to previous international adult assessments, 60 items of the literacy and numeracy items administered in PIAAC came from ALL and IALS. Seventy-four new items were developed for the literacy and numeracy domains, and new measures were developed for 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.

Chapter 18: Scaling Outcomes

Kentaro Yamamoto, Lale Khorramdel and Matthias von Davier

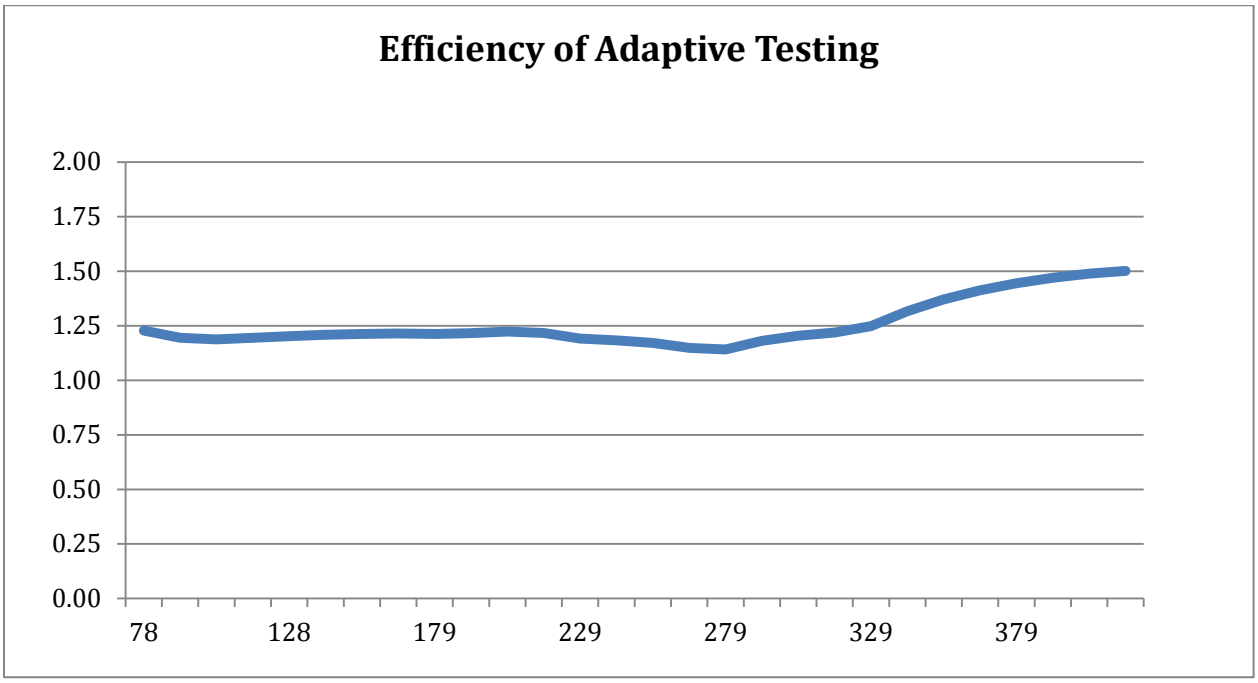
18.1 International characteristics of the PIAAC item pool and scales

18.1.1 Test information and evaluation of adaptive testing

The PIAAC multistage adaptive testing design for the CBA was developed to match a respondent's background profile and ability while maintaining a degree of randomness of assignment to ensure broad coverage of the domain in all proficiency levels. This made it possible to match respondents' abilities with the booklets' difficulties in a fair manner. Moreover, it was possible to control the exposure rates for all booklets (cf. Chen, Yamamoto & von Davier, in press). The aim of adaptive testing is to increase efficiency, validity and accuracy of the cognitive measurement. The multistage adaptive testing design may also increase engagement and test motivation, and hence reduce nonresponse and random responding.

The graph in Figure 18.1 shows the efficiency of the PIAAC multistage adaptive assessment for the literacy scale over averaged (expected) test information of the nonadaptive assessment, defined as the ratio of the conditional maximum test information of the 12 adaptive tests (note that one test consists of two clusters of items: stage 1 and stage 2) over the average test information of nonadaptive tests. The ratio of the two test information curves is shown on the vertical axis whereas the literacy scale is shown on the horizontal axis. Between the literacy proficiency values 100 and 400, the adaptive assessment was 15% to 47% more efficient than the average nonadaptive assessment based on the identical item set. Increased efficiency of adaptive testing means that the same amount of test information was obtained from the adaptive test as would be a nonadaptive test with 15-47% more items (or restated, the adaptive test required 13-32% fewer items).

Figure 18.1: Efficiency of the PIAAC multistage adaptive assessment for the scale of literacy over averaged (expected) test information of the nonadaptive assessment



18.1.2 Testing time

Each block of items for the domains of literacy, numeracy and PSTRE in the CBA was expected to take 30 minutes on average, including orientations. However, it turned out that in most cases, respondents took less than the expected amount of time (see Table 18.1). Table 18.4 shows the average time per item and block for the cognitive domains in PIAAC; the information in the tables does not include the average time spent for orientations. The reading components, which were expected to take 10 minutes on average, took less time as well (see Table 18.2).

Table 18.1: Average minutes per block of items in the CBA with regard to domain

Literacy		Numeracy		PSTRE	
Block	**Min (average)**	**Block**	**Min (average)**	**Block**	**Min (average)**
Core	1.19	Core	1.76	PS1 Block	20.65
CBA Block	22.52	CBA Block	21.89	PS2 Block	18.32

Table 18.2: Average time (in minutes) per block of items for the reading components domain

Block	Minutes (average)	SD
Vocabulary	2.48	1.86
Sentence	2.89	1.82
Passage 1, 2, 3, 4	6.30	3.64

* Vocabulary: 34 items. Sentence: 22 items. Passages 1, 2, 3 and 4: 44 items)

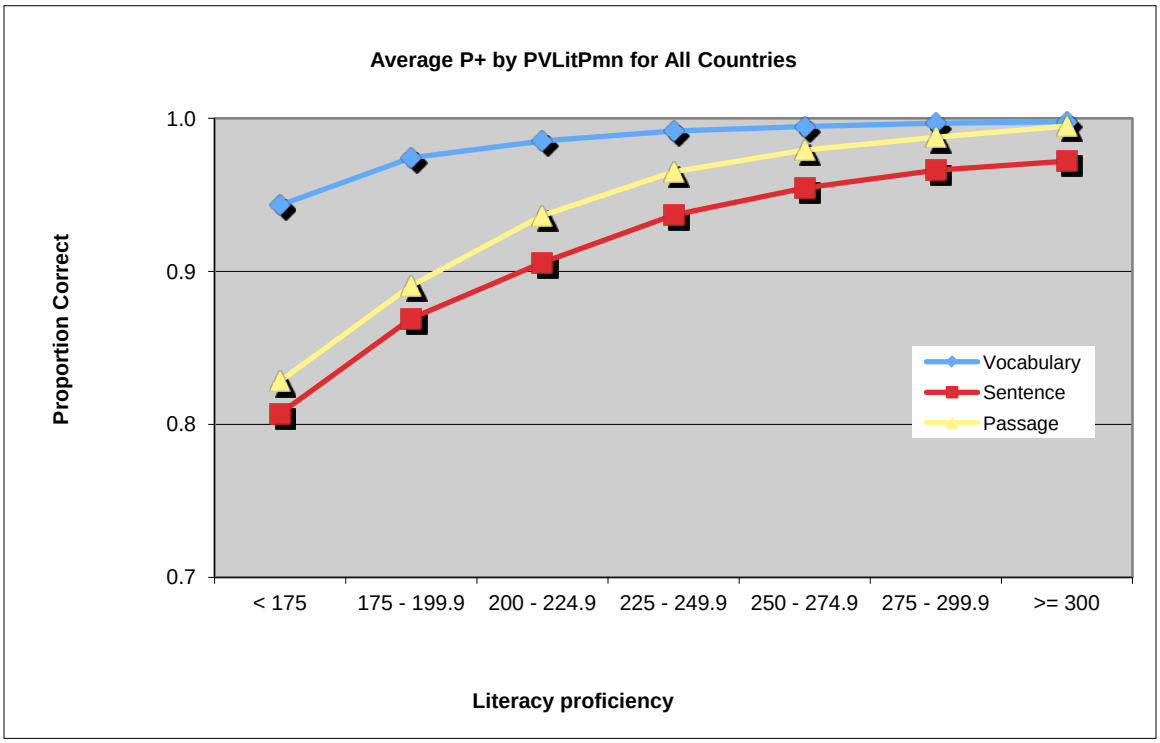
For the reading components, both response time and proportions correct had predictable relationships with literacy proficiencies (see Table 18.3). Results show a high proportion of correct responses as expected even among least able respondents of less than 175 (vocabulary: P+ = .94; sentence processing: P+ = .81; basic passage comprehension: P+ = .83), meaning the reading components were easy for every respondent. While high response accuracy was even among least able respondents, response fluency represented by the average response time indicate that less able respondents took 2.5 times longer to answer reading component items.

Table 18.3 Reading components average proportions correct and average response time by literacy posterior means

		Literacy Posterior Means						
		< 175	**175 - 199.9**	**200 - 224.9**	**225 - 249.9**	**250 - 274.9**	**275 - 299.9**	**>= 300**
Vocabulary	Average proportions correct P+	0.94	0.97	0.99	0.99	0.99	1.00	1.00
	Average response time per item (sec)	7.62	5.97	4.95	4.40	3.95	3.69	3.33
Sentence	Average proportions correct P+	0.81	0.87	0.91	0.94	0.95	0.97	0.97
	Average response time per item (sec)	14.45	10.99	9.23	8.19	7.24	6.69	6.06
Passage	Average proportions correct P+	0.83	0.89	0.94	0.97	0.98	0.99	0.99
	Average response time per item (sec)	16.63	13.29	10.93	9.26	8.18	7.33	6.38

Figure 18.2 shows the proportion of correct responses for the reading components scale projected onto the literacy proficiency scale.

Figure 18.2: Accuracy (discrimination by means of conditional P+) of the PIAAC scale for reading components projected onto the literacy proficiency scale, Main Study



18.1.3 Test reliability and accuracy

As different sets of items were administered to different respondents in the Main Study, it is not reasonable to calculate marginal reliabilities for each cognitive domain. In order to get an indication of test reliability, the explained variance for each cognitive domain (see Table 18.4) was computed based on the weighted posteriori variance. The explained variance shows how much variance is explained by the model; it is computed using the 10 plausible values as follows: 1 – (*expected error variance/total variance*). The weighted posteriori variance is an expression of the posterior measurement error and is obtained through the population modeling. The expected error variance is the weighted average of the posteriori variance. This term was estimated using the weighted average of the variance of the plausible values (the posteriori variance is the variance across the 10 plausible values). The total variance was estimated using a resampling approach (Efron, 1982). It was estimated for each country depending on the country-specific proficiency distributions for each cognitive domain.

Table 18.4: Test reliability for literacy, numeracy, and PSTRE

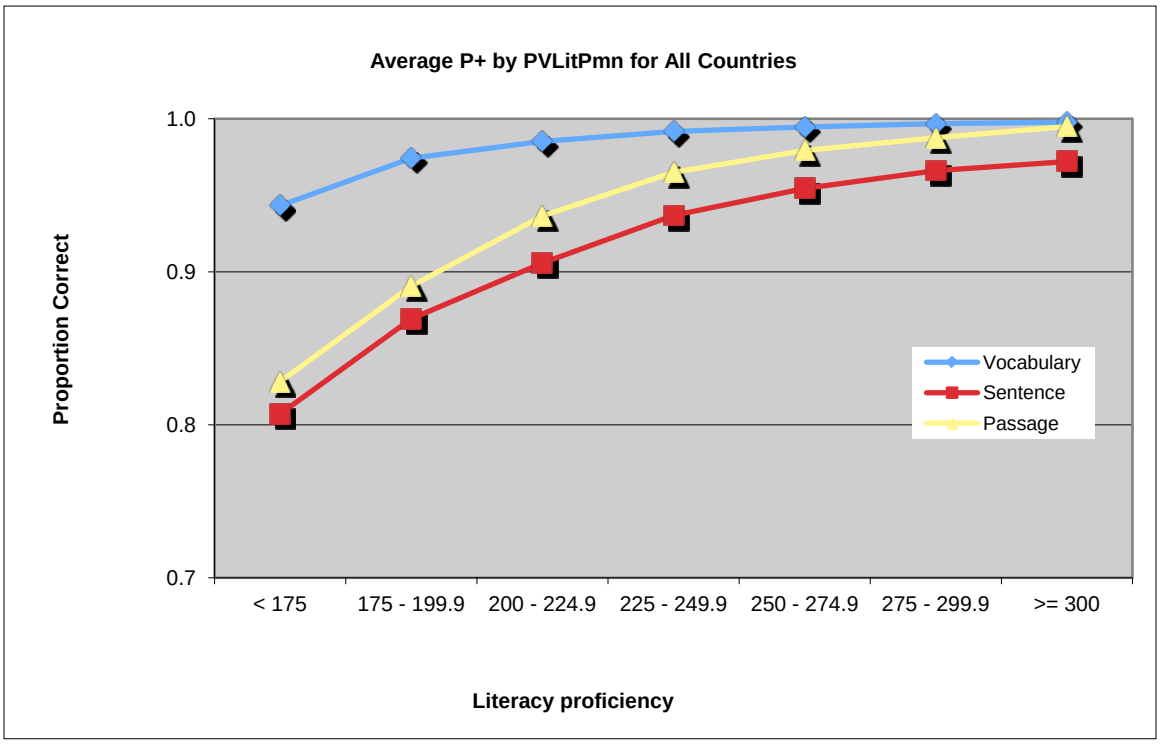
Countries	Literacy	Numeracy	PSTRE
Australia	0.879	0.875	0.834
Austria	0.865	0.860	0.844
Canada	0.878	0.874	0.847
Cyprus[1]	0.846	0.860	---
Czech Republic	0.856	0.862	0.869
Denmark	0.886	0.874	0.860
England/N. Ireland (UK)	0.879	0.896	0.876
Estonia	0.844	0.844	0.852
Finland	0.882	0.866	0.854
Flanders (Belgium)	0.881	0.868	0.846
France	0.890	0.902	---
Germany	0.887	0.889	0.864
Ireland	0.875	0.874	0.844
Italy	0.858	0.871	---
Japan	0.841	0.839	0.824
Korea	0.854	0.856	0.828
Netherlands	0.889	0.888	0.849
Norway	0.887	0.892	0.871
Poland	0.854	0.852	0.845
Russian Federation[2]	0.844	0.839	0.887
Slovak Republic	0.839	0.858	0.800
Spain	0.891	0.895	---
Sweden	0.903	0.903	0.886
United States	0.898	0.907	0.866

The table above shows that the explained variance by the combined IRT and latent regression model is at a comparable level across countries. While the joint model including background and item response data reaches levels of around 0.85, it is important to keep in mind that this is not to be confused with a classical reliability coefficient, as it is based on more than the item responses. Comparisons among individual respondents are not appropriate, because the apparent accuracy of the measures is obtained by statistically adjusting the estimates based on background data. This approach does provide improved behavior of subgroup estimates, while the plausible values obtained using this methodology are not suitable for comparisons of individuals (e.g., Mislevy & Sheehan, 1987; von Davier, Sinharay, Oranje, & Beaton, 2006).

The accuracy of the reading components in PIAAC was good as well. Results show a high proportion of correct responses as expected (vocabulary: P+ = 97.4; sentence processing: P+ = 91.4; basic passage comprehension: P+ = 92.7), meaning the reading components were easy for every respondent. Figure 18.3 shows the proportion of correct responses for the reading components scale projected onto the literacy proficiency scale.

[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.

Figure 18.3: Accuracy (discrimination by means of conditional P+) of the PIAAC scale for reading components projected onto the literacy proficiency scale, Main Study



18.1.4 Domain intercorrelations

The estimated correlations (corrected for attenuation) between the three PIAAC domains per country range from .737 to .875 (see Table 18.5). The correlations are rather high, as expected, but still show there is some distinction between each of the domains.

Table 18.5: Estimated average intercorrelations of the domains of literacy, numeracy and PSTRE by country, based on plausible values

Countries	Literacy with Numeracy	Literacy with PSTRE	Numeracy with PSTRE
Australia	0.890	0.801	0.729
Austria	0.863	0.791	0.714
Canada	0.868	0.813	0.740
Cyprus[3]	0.813	---	---
Czech Republic	0.798	0.768	0.697
Denmark	0.876	0.816	0.762
England/N. Ireland (UK)	0.875	0.773	0.769
Estonia	0.833	0.801	0.750
Finland	0.864	0.809	0.714
Flanders (Belgium)	0.873	0.811	0.734
France	0.863	---	---
Germany	0.872	0.806	0.753
Grand Total	0.861	0.781	0.725
Ireland	0.871	0.770	0.703
Italy	0.827	---	---
Japan	0.855	0.717	0.668
Korea	0.882	0.766	0.696
Netherlands	0.886	0.824	0.767
Norway	0.895	0.801	0.763
Poland	0.852	0.749	0.682
Russian Federation[4]	0.790	0.685	0.694
Slovak Republic	0.854	0.716	0.662
Spain	0.887	---	---
Sweden	0.893	0.791	0.746
United States	0.888	0.813	0.759

[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.

18.2 Scaling outcomes

18.2.1 Conditioning (population modeling)

As described in sections 17.2, and 17.3, the conditioning (population modeling) combining IRT models and latent regression models utilized all the background data for each of the PIAAC countries. Analyses were carried out by country to allow country-specific latent regression models. The resulting estimates were then used to generate plausible values.

To assure the conditioning worked well across countries, examinations of convergence efficiency, residual variances and correlations based on the 10 plausible values were conducted for each country and cognitive scale (correlations were computed with each plausible value, then the average calculated). Results showed comparable correlations among scales (see Table 18.5), comparable levels of reliability (see Table 18.4), and reasonable correlations with skill use self-reports (see Table 18.6 for selected correlations and Appendix 18.1 for detailed information).

Table 18.6: Marginal correlations per country of the respective domains with selected scales of the BQ, based on the 10 plausible values obtained from the population modeling (conditioning)*

Countries	LIT – Use of reading skills at home	LIT – Use of reading skills at work	LIT – Use of writing skills at home	LIT – Use of writing skills at work	NUM – Use of NUM skills at home	NUM – Use of NUM skills at work	PSTRE – Use of ICT skills at home	PSTRE – Use of ICT skills at work
Australia	0.335	0.211	0.280	0.210	0.319	0.200	0.307	0.218
Austria	0.327	0.290	0.257	0.208	0.276	0.260	0.350	0.255
Canada	0.337	0.214	0.229	0.175	0.273	0.194	0.318	0.19
Cyprus[5]	0.168	0.099	0.094	0.121	0.144	0.146	----	----
Czech Rep.	0.324	0.196	0.195	0.172	0.230	0.200	0.295	0.178
Denmark	0.328	0.220	0.255	0.152	0.251	0.235	0.323	0.224
England/N. Ireland (UK)	0.314	0.271	0.253	0.209	0.269	0.205	0.368	0.289
Estonia	0.315	0.197	0.229	0.155	0.270	0.219	0.403	0.240
Finland	0.313	0.207	0.279	0.171	0.315	0.253	0.384	0.216
Flanders (Belgium)	0.343	0.322	0.238	0.218	0.265	0.269	0.376	0.288
France	0.395	0.344	0.272	0.230	0.338	0.279	----	----
Germany	0.368	0.262	0.230	0.158	0.329	0.252	0.373	0.235
Ireland	0.320	0.243	0.245	0.192	0.243	0.215	0.355	0.273
Italy	0.362	0.281	0.167	0.210	0.239	0.258	----	----
Japan	0.271	0.153	0.052	0.095	0.179	0.256	0.267	0.246
Korea	0.370	0.252	0.153	0.163	0.298	0.211	0.309	0.208
Netherlands	0.336	0.252	0.280	0.188	0.278	0.229	0.365	0.208
Norway	0.283	0.236	0.176	0.178	0.236	0.227	0.335	0.257
Poland	0.384	0.226	0.264	0.13	0.314	0.223	0.317	0.149

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

Table 18.6 (cont.): Marginal correlations per country of the respective domains with selected scales of the BQ, based on the 10 plausible values obtained from the population modeling (conditioning)*

Countries	LIT – Use of reading skills at home	LIT – Use of reading skills at work	LIT – Use of writing skills at home	LIT – Use of writing skills at work	NUM – Use of NUM skills at home	NUM – Use of NUM skills at work	PSTRE – Use of ICT skills at home	PSTRE – Use of ICT skills at work
Russian Fed.[6]	0.273	0.098	0.084	0.088	0.294	0.119	0.263	0.122
Slovak Rep.	0.370	0.193	0.135	0.139	0.278	0.173	0.188	0.150
Spain	0.393	0.273	0.261	0.224	0.291	0.215	----	----
Sweden	0.279	0.184	0.203	0.156	0.223	0.243	0.377	0.269
United States	0.245	0.184	0.200	0.137	0.235	0.141	0.333	0.220

Note: The correlations for the ICT scales might be underestimated as not every respondent received the ICT items according to the path of the adaptive testing.
*LIT = Literacy, NUM = Numeracy, PSTRE = Problem-solving in technology-rich environments

The conditioning model estimations converged without any apparent issues, the between-scale correlations across countries are similar, and the correlations of direct assessed proficiency data and self-reported skill use are in a range that is comparable to prior assessments. Given these results, and the successful link across PIAAC and two prior surveys – IALS and ALL – the PIAAC database can be considered a source for consistent and valid comparisons across countries and subpopulations within countries. Good comparability was achieved over time and across assessment modes.

18.2.2 Classification of items into different proficiency levels

After estimation of the item parameters and respondents' proficiencies (person parameters) in the item calibration stage, items were classified into different proficiency levels separately for each cognitive domain. The purpose of classifying items into different levels is to provide more descriptive information about group proficiencies. That is, the different item levels provide information about the underlying or latent characteristics of an item; the higher the latent characteristic (which reflects our understanding of literacy skills), the higher the level. This item classification into different levels is done by selecting a response probability (RP) value (which defines a point on the scale for that the item function has a certain probability) to predict the probability of correctly responding to a group of items that share characteristics and then to use the selected RP value to assign items to the different proficiency levels. Each level is defined by certain score boundaries for each domain.

While the definitions of the score boundaries for the literacy and numeracy domains are the similar, the score boundaries for PSTRE are different. As there were fewer problem-solving items (14 items) than items from the other domains (2 x 76 items), and the problem-solving items were more difficult, only three levels were defined for this domain. Table 18.7 shows the score boundaries used in PIAAC for literacy and numeracy, and Table 18.8 shows the score boundaries for PSTRE. The decision for the score boundaries was based on expert judgment utilizing the distribution of item difficulties.

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

Table 18.7: Score boundaries for item classification for the domains of literacy and numeracy

Level	Literacy - Score	Numeracy - Score
below level 1	0-175	0-175
1	176-225	176-225
2	226-275	226-275
3	276-325	276-325
4	326-375	326-375
5	376-500	376-500

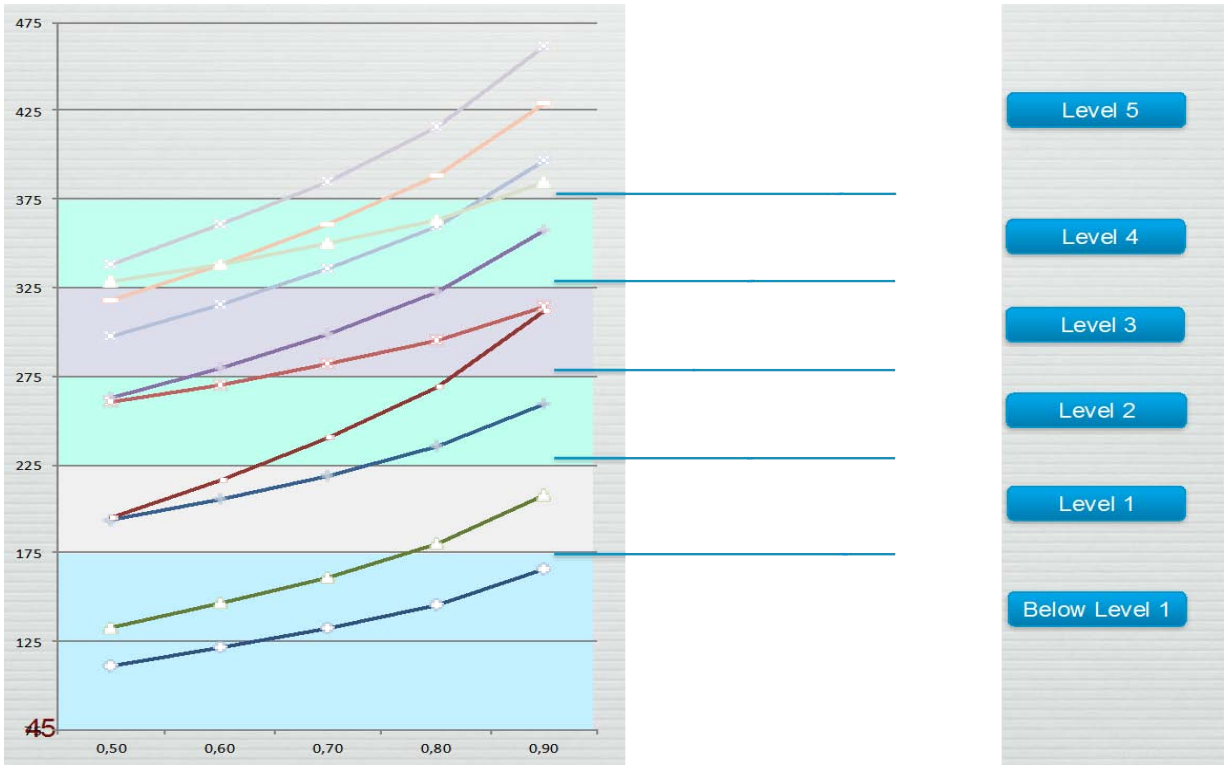
Table 18.8: Score boundaries for item classification for the domain of PSTRE

Level	PSTRE – Score
below level 1	0-240
1	241-290
2	291-340
3	341-500

So far, there is no generally agreed upon rule in the research literature that has been used to characterize items along a proficiency scale. RP values around .65 have been used in most school-based surveys, while values as high as .80 have been used in some adult surveys including IALS and ALL. More recently, however, the US National Academy of Sciences recommended that the National Assessment of Adult Literacy (NAAL) survey (the most recent US survey of adults) use an RP value more closely aligned with school-based surveys. For PIAAC it was decided to use an RP value of .67; some countries received an additional RP value of .80 at their request for the purpose of a better comparability with prior surveys (IALS, ALL). Items are assigned to different proficiency levels due to the selected RP value.

As shown in Figure 18.4, the selection of the RP value impacts where a particular item is classified along the scale. While the selection of an RP value can impact the level in which an item is located, the selection of an RP value has no impact on the proficiency distribution or the percentage of respondents who fall within a particular level (see Figure 18.4).

Figure 18.4: Example for the impact of selected RP values on the placement of items along a scale



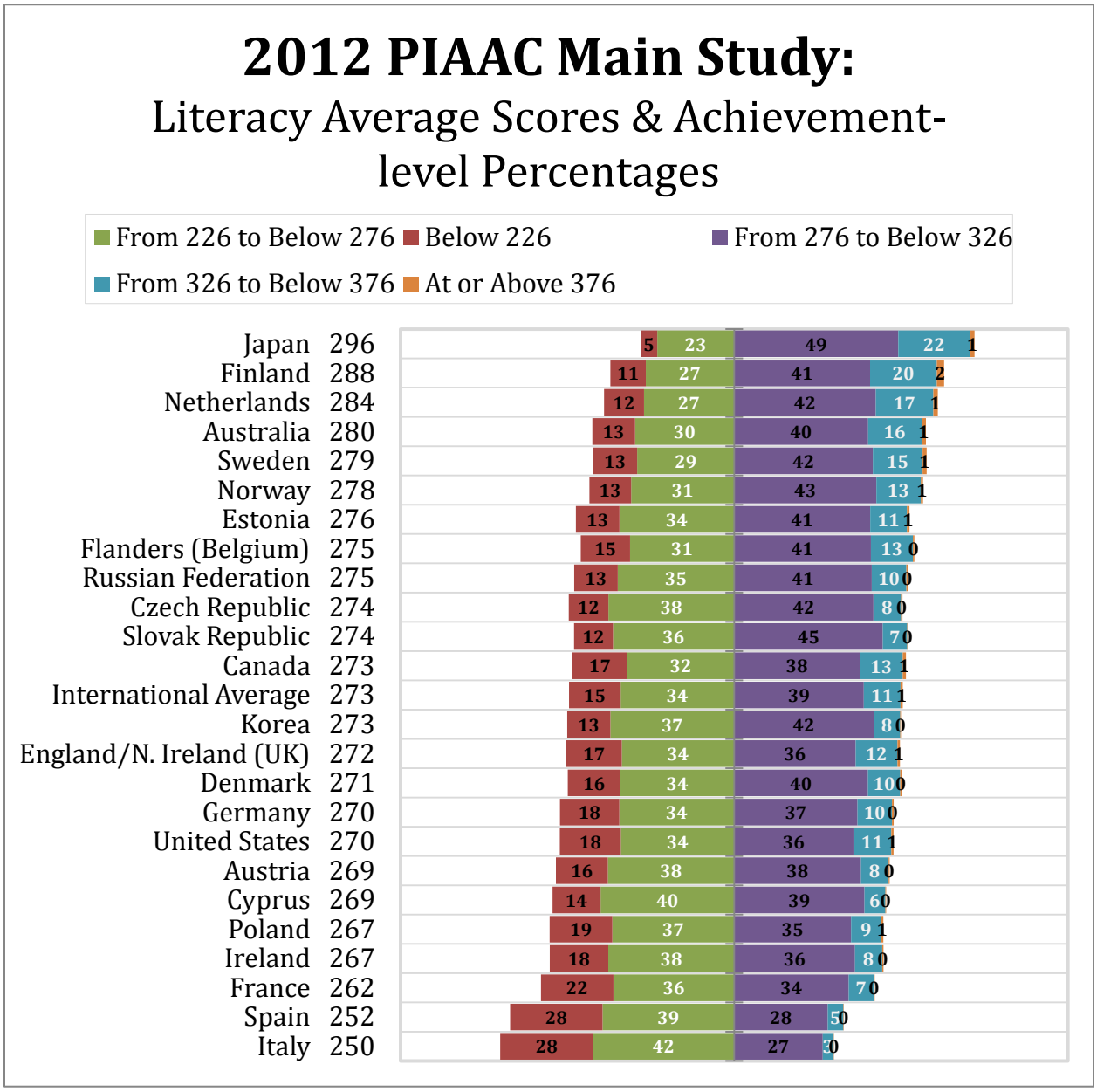
Note: The x-axis in the left hand side exhibit is the response probability (RP) and the y-axis denotes the scale score of the domain.

It is also important to keep in mind that the precision of measurement along a scale is not impacted by the RP value. The same items define the underlying scale regardless of which RP value is selected. Finally, it is important to note that the RP value does not decide on which item measures in which level: All items contribute to the measurement precision in all levels of proficiency, the RP value is one point on the item function graph at which a certain probability is reached. Respondents with a proficiency located below this point have a lower probability (but not 0.0) than the RP value chosen, and respondents with a probability above this point have a higher probability (but do not solve the item with certainty) of solving an item. That means that an item that was located in level 4 using an RP value of 0.67 will also provide information on respondents that are located in levels 3 or 5. The location of an item at a certain level simply implies that (for the chosen RP value) this item is most representative of that particular level.

Chapter 21 describes the content definition for each proficiency level per cognitive domain. Figures 18.5 to 18.7 show the percentage of respondents per country at each level of proficiency for each cognitive domain (note that France and the Russian Federation[7] are not included in the figures as their data were not received in time).

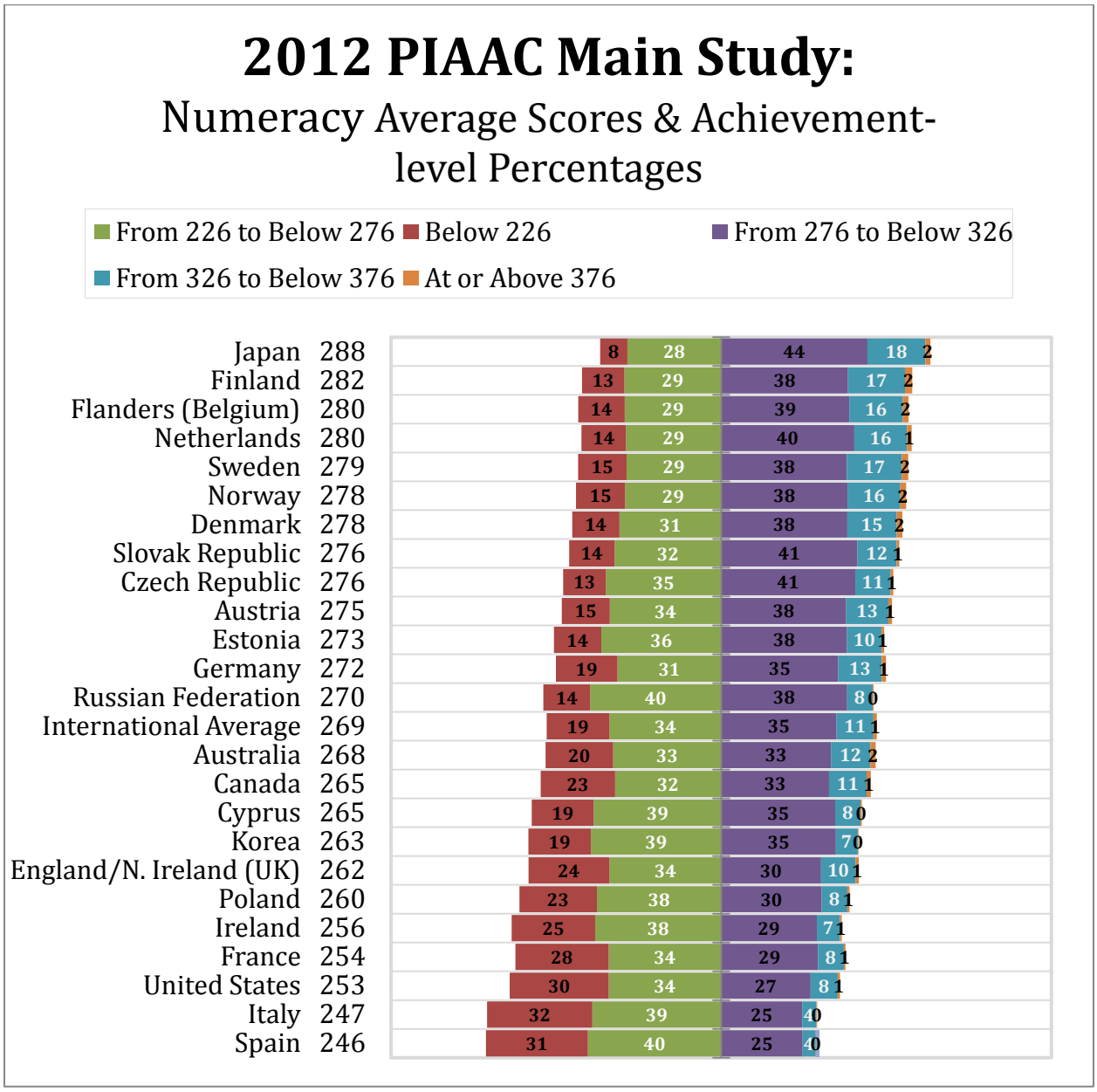
[7] Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

Figure 18.5: Percentage of respondents per country[8] at each level of proficiency for the domain of literacy



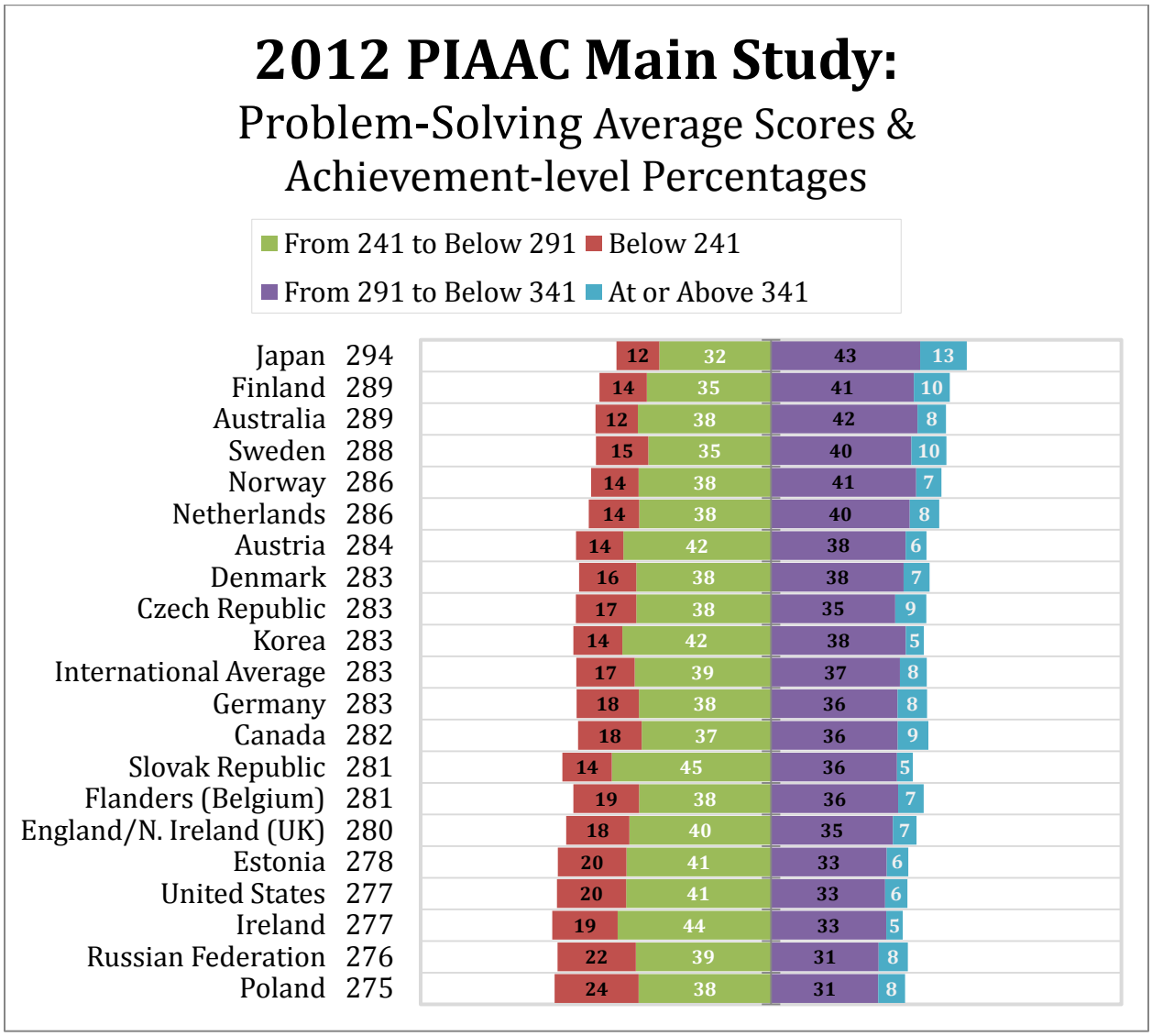
[8] Please refer to the note regarding the Russian Federation, and notes A and B regarding Cyprus in the *Note to Readers* section of this report.

Figure 18.6: Percentage of respondents per country[9] at each level of proficiency for the domain of numeracy



[9] Please refer to the note regarding the Russian Federation, and notes A and B regarding Cyprus in the *Note to Readers* section of this report.

Figure 18.7: Percentage of respondents per country[10] at each level of proficiency for the domain of PSTRE



18.2.3 Transforming the plausible values to PIAAC scales

The plausible values (derived from the population modeling) were transformed using a linear transformation to form a scale that is linked through anchor items to IALS and ALL for literacy and numeracy. This scale can be used to compare the overall performance of countries or subgroups within a country. It can also be used to compare performance along the scale based on statistical criteria such as percentiles.

The linear transformation is based on a concurrent calibration of the literacy and numeracy scales across all countries participating in PIAAC, and also includes data from countries that

[10] Please refer to the note regarding the Russian Federation, and notes A and B regarding Cyprus in the *Note to Readers* section of this report.

participated in IALS and ALL. The reported country distributions from IALS and ALL were used to align the IRT-based country distributions for PIAAC, IALS and ALL to ensure comparability between the three assessments.

To compare the proficiency estimates of the different countries with regard to the cognitive domains, the weighted mean of each of the 10 plausible values per country, and then the average of these 10 means was calculated. Table 18.9 shows the average plausible values for each cognitive domain per country as well as the resampling-based standard errors.

Table 18.9: Average plausible values and resampling-based standard errors per country for the PIAAC domains of literacy, numeracy, and PSTRE

	Literacy		Numeracy		PSTRE	
Country	**Average Plausible Values**	**Standard Error**	**Average Plausible Values**	**Standard Error**	**Average Plausible Values**	**Standard Error**
Australia	280	0.9	268	0.9	289	0.9
Austria	269	0.7	275	0.9	284	0.7
Canada	273	0.6	265	0.7	282	0.7
Cyprus[11]	269	0.8	265	0.8	---	---
Czech Rep.	274	1.0	276	0.9	283	1.1
Denmark	271	0.6	278	0.7	283	0.7
England/N. Ireland (UK)	272	1.0	262	1.1	280	0.9
Estonia	276	0.7	273	0.5	278	1.0
Finland	288	0.7	282	0.7	289	0.8
Flanders (Belgium)	275	0.8	280	0.8	281	0.8
France	262	0.6	254	0.6	---	---
Germany	270	0.9	272	1.0	283	1.0
International Avg. (OECD)	273	0.2	269	0.2	283	0.2
Ireland	267	0.9	256	1.0	277	1.0
Italy	250	1.1	247	1.1	---	---
Japan	296	0.7	288	0.7	294	1.2
Korea	273	0.6	263	0.7	283	0.8
Netherlands	284	0.7	280	0.7	286	0.8
Norway	278	0.6	278	0.8	286	0.6
Poland	267	0.6	260	0.8	275	1.3
Russian Fed.[12]	275	2.7	270	2.7	276	4.3

[11] Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.
[12] Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

Table 18.9 (cont.): Average plausible values and resampling-based standard errors per country for the PIAAC domains of literacy, numeracy, and PSTRE

	Literacy		Numeracy		PSTRE	
Country	**Average Plausible Values**	**Standard Error**	**Average Plausible Values**	**Standard Error**	**Average Plausible Values**	**Standard Error**
Slovak Rep.	274	0.6	276	0.8	281	0.8
Spain	252	0.7	246	0.6	---	---
Sweden	279	0.7	279	0.8	288	0.6
United States	270	1.0	253	1.2	277	1.1

18.3 Analysis of data with plausible values

If the scale proficiency values (θ) were known for all respondents, it would be possible to directly compute any statistic t(θ,y), for example, a scale or composite subpopulation sample mean, a sample percentile point, or a sample regression coefficient to estimate a corresponding population quantity T.

Because the scaling models are latent variable models, θ values are not observed. To overcome this problem, we follow the approach taken by Rubin (1987) and treating θ as "missing." data The value t(θ,y) is approximated by its expectation given (x,y), the data actually observed, as follows:

$$t^*(\bar{x},\bar{y}) = E[t(\bar{\theta},\bar{y})|\bar{x},\bar{y}] = \int t(\bar{\theta},\bar{y})p(\bar{\theta}|\bar{x},\bar{y})d\theta$$

It is possible to approximate t^* using plausible values (also referred to as imputations) instead of the unobserved θ values. Plausible values are random draws from the conditional distribution of the scale proficiencies given the item responses x_j, background variables y_j, and model parameters (see section 17.2.). For any respondent, the value of θ used in the computation of t is replaced by a randomly selected value from the respondent's conditional distribution. Rubin (1987) argues that this process should be repeated several times so that the uncertainty associated with imputation can be quantified. For example, the average of multiple estimates of t, each computed from a different set of plausible values, is a numerical approximation of t^* in the above equation; the variance among them reflects uncertainty due to not observing θ. It should be noted that this variance does not include any variability due to sampling from the population.

It cannot be emphasized too strongly that the plausible values are not a substitute for test scores for individuals. Plausible values incorporate responses to test items and information about the background of responses and can therefore not be used to compare individual test takers in the usual sense. Plausible values are only intermediary computations in the calculation of the integrals in the above equation in order to estimate population characteristics such as subgroup means and standard deviations. When the underlying model is correctly specified, plausible values will provide consistent estimates of population characteristics, even though they are not

generally unbiased estimates of the proficiencies of the individuals with whom they are associated (von Davier, Gonzalez & Mislevy, 2009), provide examples and a more detailed explanation). The key idea lies in a contrast between plausible values and the more familiar ability estimates of educational measurement that are in a sense optimal for each respondent (e.g., bias corrected maximum likelihood estimates, which are consistent estimates of a respondent's proficiency θ, and Bayesian estimates, which provide minimum mean-squared errors with respect to a reference population). Point estimates that are optimal for individual respondents have distributions that can produce decidedly nonoptimal (inconsistent) estimates of population characteristics (Little & Rubin, 1983). Plausible values, on the other hand, are constructed explicitly to provide consistent estimates of population effects. For further discussion, see Mislevy, Beaton, Kaplan, and Sheehan (1992).

After obtaining the plausible values from the posteriori distribution, they can be employed to evaluate the previous equation for an arbitrary function T as follows:

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

2) In the same manner as in step 1 above, evaluate the sampling variance of T, or Var(T_1,), with respect to respondents' first vectors of plausible values. Denote the result Var_1.

3) Carry out steps 1 and 2 for the second through all 10 vectors of plausible values, thus obtaining T_u and Var_u for u=2, . . .,10.

4) The best estimate of T obtainable from the plausible values is the average of the 10 values obtained from the different sets of plausible values:

   $T. = \frac{\sum_u T_u}{10}$

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

$$Var(T.) = \frac{\sum_u Var_u}{10} + (1 + \frac{1}{10}) \frac{\sum_u (T_u - T.)^2}{10 - 1}$$

The first component in *Var(T.)* reflects uncertainty due to sampling from the population; the second component reflects uncertainty because the respondents' proficiencies θ are only indirectly observed through *x* and *y*.

Example for partitioning the estimated error variance:

The following example illustrates the use of plausible values (PV) for partitioning the error variance. Tables 18.10a-c present data for nine subgroups of respondents with differing employment status (variable C_Q07: 1 = full-time employed or self-employed; 2 = part-time employed or self-employed; 3 = unemployed; 4 = pupil or student; 5 = apprentice or internship; 6 = in retirement or early retirement; 7 = permanently disabled; 9 = fulfilling domestic tasks of looking after family; 10 = other). Ten plausible values were calculated for each respondent for each scale (domain). Each column in these tables presents the means of these 10 plausible values.

Table 18.10a: Example for use of plausible values to partitioning the error – PVs 1 to 5

		Plausible Value									
		1		2		3		4		5	
C_Q07	N	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
1	2532	276.14	1.51	276.22	1.59	275.51	1.52	275.82	1.41	275.20	1.57
2	602	267.38	7.18	267.67	6.05	268.15	7.34	265.97	6.85	266.94	5.56
3	414	248.64	6.92	249.27	5.74	249.86	5.59	250.40	7.07	250.87	6.14
4	442	278.88	5.86	279.50	7.00	278.95	7.60	277.38	5.81	279.51	5.36
5	14	261.22	115.05	278.57	75.31	277.95	75.11	266.04	137.08	273.69	128.94
6	203	266.33	13.80	266.51	13.62	268.66	12.41	271.01	12.60	266.97	12.87
7	270	229.81	8.12	231.01	8.32	228.63	10.57	229.45	8.47	230.05	7.54
9	281	269.96	10.06	267.22	13.44	268.92	11.81	270.63	10.01	269.02	11.24
10	137	272.87	29.97	273.99	26.47	269.86	38.14	273.93	32.09	270.52	30.41

Table 18.10b: Example for use of plausible values to partitioning the error – PVs 6 to 10

		Plausible Value									
		6		7		8		9		10	
C_Q07	N	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
1	2532	275.74	1.65	275.60	1.50	275.66	1.58	274.70	1.53	275.65	1.53
2	602	269.03	5.00	266.45	6.58	267.41	6.25	268.85	6.85	266.45	6.38
3	414	250.63	6.21	249.98	5.78	249.53	7.05	248.78	6.27	251.82	6.97
4	442	279.61	5.84	278.27	6.78	279.04	6.07	282.19	5.37	279.11	5.62
5	14	284.81	46.85	272.05	162.29	296.01	59.46	267.64	159.43	280.77	71.99
6	203	267.92	17.15	268.15	18.38	265.38	13.60	268.05	17.40	267.07	14.06
7	270	230.91	9.76	228.51	9.89	229.83	8.29	230.72	9.81	230.06	8.73
9	281	268.73	13.24	266.79	11.60	268.63	14.09	270.38	11.23	269.29	15.18
10	137	272.85	32.07	270.49	34.29	273.31	31.97	275.00	29.46	272.68	35.37

Table 18.10c: Example for use of plausible values to partitioning the error – sample error, measurement error, and standard error based on the 10 PVs

C_Q07	N	Mean of 10 PVs	Sampling Error	Measurement Error	Standard Error
1	2532	275.62	1.24	0.46	1.32
2	602	267.43	2.53	1.08	2.75
3	414	249.98	2.52	1.03	2.73
4	442	279.24	2.48	1.29	2.79
5	14	275.88	10.16	10.60	14.68
6	203	267.61	3.82	1.63	4.15
7	270	229.90	2.99	0.91	3.13
9	281	268.96	3.49	1.30	3.73
10	137	272.55	5.66	1.79	5.94

The error variance, or squared standard error, of the mean plausible values differs greatly for the subgroups. The error variance reflects a component of error with regard to the lack of precision of the measurement instrument and a component of error with regard to sampling. The variance can be reduced by either increasing the precision of the measurement instrument (for example, increasing the number of items) or increasing the sample size. The resampling method was used to estimate the variance due to sampling using the each set of imputed values. This component of variance is similar across the 10 plausible values; the size is influenced by the homogeneity of proficiencies among respondents in a subgroup but not by the sample size or by the precision of the survey instruments. The sampling error is smaller when the subgroup consists of respondents with similar proficiencies. The total error variance can be calculated as the summation of "sampling error" and measurement error."

The last column presents the standard error of the subpopulation mean, which is equal to the square root of the sum of the two components' variance. Pairwise differences can be evaluated using these standard errors. However, multiple comparisons such as the six possible pairwise comparisons of this example need to consider an adjustment of significance level such as Hochberg Stagewise Procedure (HSP), described in Hochberg (1988).

Hochberg developed a method for multiple comparisons that utilizes the order of significance levels among all comparisons. HSP begins by placing the comparisons in an increasing order of significance levels, i.e., $P_1 \leq P_2 \leq \ldots \leq P_3 \leq \ldots P_M$. It proceeds to sequentially evaluate P_j with adjusted critical significance level of $\alpha/(m-j+1)$ where α is the target significance level. If P_j is smaller than the critical significance level then the process continues until a non-significance comparison is found. All preceding comparisons before the first nonsignificant comparison are declared significant and all subsequent comparisons are declared nonsignificant. Both the Bonferoni method and the HSP control the Type 1 error of false discovery of significant comparison when in fact it is nonsignificant. The False Discovery Rate (FDR) procedure (Benjamini & Hochberg, 1995) controls the expected proportion of falsely rejected hypotheses, finding the comparison nonsignificant when in fact it is significant. The procedure is very similar to HSP for ordering the comparisons by the significance level, then using the critical significance level of $\alpha*j/m$ for j-th in the comparisons. The determination of the significance of comparisons is identical to the HSP.

The standard errors of mean proficiencies, percentages, and percentiles play an important role in interpreting subpopulation results and in comparing the performances of two or more subpopulations. The resampling standard errors reported by PIAAC are statistics whose quality depends on certain features of the samples from which the estimates are obtained. In certain cases, primarily when the standard error is based on a small number of respondents, the mean squared error associated with the estimated standard errors may be quite large.

18.4 Developing common scales across modes of administration and for the purpose of trends

As described in section 17.4, the linking design for PIAAC aims to link items and booklets across different assessment modes as well as to the IALS and ALL surveys to provide trend measures. PIAAC items were linked between PBA and CBA and to items from IALS and ALL. Common scales were obtained through item calibration using an IRT analysis (2PL, GPCM) and

a linear transformation of the new estimates using the group means and standard deviations obtained from the previous IRT estimates in IALS and ALL (see section 17.4.2). The comparability of item parameters across countries, modes of assessment, and time (different surveys) were evaluated. If a large deviation from the common item parameters was observed for one or more countries, unique item parameters were estimated for the deviant item.

The following section present the results of this linking design for the PIAAC Main Study.

18.4.1 Linking outcomes

The linking design for the PIAAC Main Study was aimed at establishing comparability across countries with regard to both PBA and CBAs as well as the link between PIAAC and the IALS and ALL surveys. This pertains especially to the paper- and computer-based items in numeracy, and the paper-based items in literacy; deviations were limited to a few items and countries. The PIAAC item parameters for a few computer-based literacy items (which were adapted from IALS and ALL paper-based items) were not comparable with the item parameters of IALS and ALL, and with item parameters of the paper-based PIAAC assessment. By estimating new item parameters – that is, parameters were estimated for the CBA only – for those computer-based literacy items, comparability improved to the level of numeracy. The majority of linking items shared the common item parameters, that is, parameters were estimated for the data of the PBA and the CBA together.

The proportion of respondents who received the 12 different adaptive paths for the literacy scale varied between 5.0 to 13.5% across countries. For the numeracy scale, the proportions varied between 2.9% to 16.7% among paths and countries. The following two tables (18.11, 18.12) present the distribution of the 12 routing paths for literacy and numeracy scales by country, showing that the distributions are comparable between countries. A note on notation: L13 means that literacy testlets 1 and 3 were administered for the stage 1 and 2.

Table 18.11: Distribution of routing paths for the literacy module by country

Country	CBA Core	Literacy Routing Path											
		L11	L12	L13	L14	L21	L22	L23	L24	L31	L32	L33	L34
Australia	0.01	0.06	0.08	0.08	0.08	0.06	0.08	0.08	0.09	0.06	0.11	0.09	0.12
Austria	0.01	0.08	0.08	0.08	0.07	0.07	0.08	0.08	0.09	0.07	0.11	0.09	0.10
Canada	0.01	0.07	0.08	0.08	0.08	0.07	0.09	0.08	0.09	0.06	0.10	0.08	0.12
Cyprus[13]	0.01	0.07	0.09	0.08	0.07	0.07	0.08	0.08	0.08	0.09	0.10	0.09	0.08
Czech Rep.	0.01	0.07	0.09	0.08	0.08	0.07	0.08	0.09	0.08	0.07	0.09	0.08	0.10
Denmark	0.01	0.07	0.08	0.08	0.08	0.07	0.09	0.08	0.09	0.07	0.10	0.08	0.11
England/N. Ireland (UK)	0.02	0.08	0.09	0.08	0.09	0.07	0.09	0.08	0.08	0.06	0.09	0.08	0.11
Estonia	0.01	0.07	0.07	0.09	0.08	0.07	0.08	0.08	0.09	0.07	0.09	0.09	0.11
Finland	0.01	0.06	0.07	0.08	0.09	0.07	0.09	0.07	0.10	0.06	0.10	0.09	0.11
Flanders (Belgium)	0.02	0.06	0.07	0.09	0.09	0.06	0.08	0.09	0.09	0.06	0.10	0.08	0.11
France	0.02	0.07	0.07	0.09	0.08	0.07	0.09	0.08	0.08	0.06	0.10	0.08	0.10
Germany	0.01	0.07	0.08	0.09	0.08	0.07	0.08	0.07	0.08	0.07	0.09	0.09	0.11
Ireland	0.01	0.07	0.08	0.07	0.08	0.06	0.09	0.08	0.10	0.06	0.10	0.09	0.11
Italy	0.01	0.09	0.08	0.08	0.08	0.08	0.09	0.08	0.07	0.07	0.10	0.08	0.09
Japan	0.00	0.06	0.08	0.07	0.09	0.05	0.08	0.08	0.10	0.06	0.09	0.10	0.14
Korea	0.01	0.06	0.08	0.07	0.09	0.06	0.09	0.08	0.08	0.08	0.10	0.10	0.11
Netherlands	0.01	0.07	0.08	0.08	0.08	0.07	0.09	0.08	0.09	0.06	0.10	0.09	0.11
Norway	0.02	0.06	0.07	0.08	0.09	0.07	0.09	0.09	0.09	0.05	0.09	0.09	0.11
Poland	0.01	0.07	0.08	0.08	0.09	0.07	0.09	0.08	0.09	0.06	0.09	0.08	0.10
Russian Fed.[14]	0.00	0.07	0.07	0.07	0.09	0.06	0.10	0.08	0.11	0.06	0.10	0.09	0.10
Slovak Rep.	0.01	0.07	0.08	0.08	0.08	0.07	0.08	0.08	0.09	0.09	0.09	0.09	0.09
Spain	0.01	0.08	0.09	0.07	0.07	0.08	0.08	0.09	0.08	0.08	0.09	0.09	0.09
Sweden	0.01	0.08	0.07	0.08	0.08	0.07	0.09	0.09	0.09	0.06	0.09	0.09	0.11
United States	0.02	0.07	0.08	0.08	0.08	0.06	0.08	0.09	0.08	0.06	0.09	0.08	0.12

[13] Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.
[14] Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

Table 18.12: Distribution of routing paths for the numeracy module by country

Country	CBA core	Numeracy Routing Path											
		N11	N12	N13	N14	N21	N22	N23	N24	N31	N32	N33	N34
Australia	0.01	0.06	0.08	0.07	0.08	0.06	0.09	0.09	0.11	0.04	0.09	0.09	0.13
Austria	0.01	0.06	0.07	0.07	0.09	0.05	0.08	0.08	0.11	0.05	0.10	0.09	0.13
Canada	0.01	0.07	0.08	0.08	0.09	0.06	0.08	0.08	0.10	0.05	0.09	0.09	0.13
Cyprus[15]	0.01	0.06	0.09	0.08	0.08	0.05	0.08	0.08	0.09	0.05	0.10	0.10	0.13
Czech Rep.	0.01	0.07	0.08	0.08	0.09	0.05	0.09	0.09	0.10	0.04	0.08	0.10	0.12
Denmark	0.01	0.06	0.08	0.07	0.09	0.05	0.08	0.08	0.11	0.05	0.10	0.10	0.13
England/N. Ireland (UK)	0.02	0.07	0.09	0.09	0.06	0.07	0.09	0.08	0.09	0.05	0.08	0.08	0.13
Estonia	0.01	0.06	0.08	0.08	0.09	0.05	0.08	0.09	0.11	0.04	0.09	0.10	0.12
Finland	0.01	0.07	0.08	0.09	0.09	0.05	0.08	0.09	0.10	0.04	0.08	0.09	0.13
Flanders (Belgium)	0.02	0.06	0.08	0.07	0.09	0.05	0.08	0.08	0.10	0.04	0.10	0.09	0.14
France	0.02	0.07	0.07	0.08	0.08	0.06	0.08	0.08	0.10	0.05	0.09	0.09	0.12
Germany	0.01	0.07	0.08	0.08	0.09	0.06	0.08	0.07	0.11	0.04	0.09	0.09	0.13
Ireland	0.01	0.07	0.08	0.08	0.08	0.06	0.09	0.08	0.09	0.06	0.10	0.10	0.12
Italy	0.01	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.11	0.05	0.09	0.08	0.10
Japan	0.00	0.06	0.07	0.08	0.09	0.04	0.08	0.09	0.11	0.03	0.10	0.09	0.17
Korea	0.01	0.06	0.08	0.08	0.07	0.05	0.08	0.09	0.11	0.05	0.10	0.09	0.14
Netherlands	0.01	0.07	0.08	0.08	0.09	0.05	0.08	0.09	0.10	0.05	0.08	0.09	0.13
Norway	0.02	0.07	0.08	0.08	0.09	0.05	0.08	0.08	0.09	0.04	0.09	0.10	0.13
Poland	0.01	0.06	0.07	0.09	0.09	0.05	0.08	0.08	0.10	0.05	0.09	0.09	0.13
Russian Fed.[16]	0.00	0.08	0.09	0.08	0.09	0.07	0.08	0.09	0.08	0.05	0.09	0.08	0.12
Slovak Rep.	0.01	0.06	0.07	0.08	0.09	0.06	0.09	0.08	0.10	0.05	0.09	0.08	0.13
Spain	0.01	0.07	0.09	0.09	0.07	0.07	0.09	0.09	0.09	0.05	0.09	0.09	0.11
Sweden	0.01	0.07	0.08	0.08	0.10	0.05	0.08	0.08	0.10	0.04	0.08	0.10	0.12
United States	0.02	0.07	0.08	0.07	0.07	0.06	0.09	0.09	0.08	0.05	0.09	0.09	0.13

[15] Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.
[16] Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological) 57,* 289–300.

Chen, H., Yamamoto, K., & von Davier, M. (in press). Controlling MST exposure rates in international large-scale assessments. In D. Yan, D., A. A. von Davier, & C. Lewes (Eds.), *Computerized Multistage Testing –Theory and Applications*. Taylor & Francis.

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. *Society of Industrial and Applied Mathematics CBMS-NSF Monographs,* 38.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75,* 800-802.

Little, R. J. A. & Rubin, D. B. (1983). On jointly estimating parameters and missing data. *American Statistician, 37, 218-220.*

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

Mislevy, R. J. & Sheehan, K. (1987) Marginal estimation procedures. In A. E. Beaton, *Implementing the new design: The NAEP 1983-84 technical report (pp. 293-360). (No. 15-TR-20)* Princeton, NJ: Educational Testing Service

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley & Sons.

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.ieriinstitute.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.

Appendix 18.1: Marginal correlations (Pearson) per country of the cognitive domains literacy (LIT), numeracy (NUM) and problem solving in technology rich environments (PSL), respectively, with scales of the BQ, based on the 10 plausible values obtained from the population modeling (conditioning)

	Learning at work	Readiness to learn	Use of ICT skills at home	Use of ICT skills at work	Use of influencing skills at work	Use of num skills at home	Use of num skills at work	Use of planning skills at work	Use of reading skills at home	Use of reading skills at work	Use of task discretion at work	Use of writing skills at home	Use of writing skills at work
Australia													
LIT	0.010	0.323	0.313	0.214	0.183	0.313	0.151	0.116	0.335	0.211	0.146	0.28	0.210
NUM	0.003	0.284	0.277	0.188	0.169	0.319	0.200	0.118	0.310	0.207	0.127	0.253	0.195
PSTRE	0.015	0.204	0.307	0.218	0.086	0.217	0.121	0.045	0.122	0.103	0.094	0.171	0.133
Austria													
LIT	0.073	0.267	0.315	0.259	0.180	0.266	0.230	0.108	0.327	0.290	0.095	0.257	0.208
NUM	0.055	0.250	0.267	0.226	0.197	0.276	0.260	0.116	0.311	0.288	0.120	0.222	0.194
PSTRE	0.095	0.175	0.350	0.255	0.080	0.241	0.190	0.022	0.235	0.167	0.024	0.218	0.130
Canada													
LIT	-0.016	0.271	0.291	0.195	0.168	0.264	0.154	0.107	0.337	0.214	0.177	0.229	0.175
NUM	-0.031	0.237	0.253	0.175	0.139	0.273	0.194	0.098	0.296	0.193	0.158	0.193	0.148
PSTRE	-0.007	0.191	0.318	0.190	0.093	0.249	0.114	0.041	0.220	0.108	0.161	0.230	0.098
Cyprus[17]													
LIT	0.000	0.108	0.080	0.047	0.040	0.075	0.062	-0.023	0.168	0.099	-0.006	0.094	0.121
NUM	0.034	0.146	0.078	0.107	0.083	0.144	0.146	0.018	0.209	0.180	0.026	0.141	0.159
PSTRE													
Czech Rep.													
LIT	0.056	0.212	0.248	0.131	0.189	0.255	0.162	0.112	0.324	0.196	0.079	0.195	0.172
NUM	0.063	0.188	0.202	0.158	0.200	0.230	0.200	0.136	0.303	0.240	0.118	0.140	0.164
PSTRE	0.075	0.210	0.295	0.178	0.158	0.307	0.217	0.095	0.288	0.154	0.106	0.215	0.159

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

Appendix 18.1: Marginal correlations (Pearson) per country of the cognitive domains literacy (LIT), numeracy (NUM) and problem solving in technology rich environments (PSL), respectively, with scales of the BQ, based on the 10 plausible values obtained from the population modeling (conditioning)

	Learning at work	Readiness to learn	Use of ICT skills at home	Use of ICT skills at work	Use of influencing skills at work	Use of num skills at home	Use of num skills at work	Use of planning skills at work	Use of reading skills at home	Use of reading skills at work	Use of task discretion at work	Use of writing skills at home	Use of writing skills at work
Denmark													
LIT	0.057	0.209	0.307	0.228	0.184	0.269	0.180	0.067	0.328	0.22	0.066	0.255	0.152
NUM	0.030	0.155	0.246	0.233	0.181	0.251	0.235	0.087	0.284	0.231	0.074	0.193	0.141
PSTRE	0.076	0.151	0.323	0.224	0.099	0.291	0.203	-0.009	0.212	0.119	0.004	0.263	0.110
England/N. Ireland (UK)													
LIT	0.041	0.277	0.291	0.218	0.207	0.250	0.187	0.145	0.314	0.271	0.161	0.253	0.209
NUM	0.018	0.280	0.269	0.187	0.175	0.269	0.205	0.133	0.302	0.261	0.152	0.221	0.189
PSTRE	0.049	0.275	0.368	0.289	0.171	0.262	0.226	0.098	0.254	0.222	0.152	0.269	0.195
Estonia													
LIT	0.016	0.277	0.278	0.217	0.125	0.273	0.169	0.056	0.315	0.197	0.163	0.229	0.155
NUM	0.026	0.274	0.246	0.217	0.150	0.270	0.219	0.096	0.313	0.220	0.184	0.204	0.161
PSTRE	0.042	0.276	0.403	0.24	0.130	0.322	0.184	0.045	0.255	0.183	0.187	0.270	0.143
Finland													
LIT	-0.014	0.176	0.322	0.219	0.182	0.303	0.167	0.102	0.313	0.207	0.066	0.279	0.171
NUM	-0.020	0.133	0.280	0.235	0.131	0.315	0.253	0.071	0.283	0.204	0.081	0.220	0.157
PSTRE	0.036	0.148	0.384	0.216	0.090	0.315	0.152	0.039	0.236	0.084	0.057	0.265	0.074
Flanders (Belgium)													
LIT	0.100	0.272	0.329	0.274	0.219	0.267	0.24	0.103	0.343	0.322	0.129	0.238	0.218
NUM	0.082	0.261	0.299	0.235	0.230	0.265	0.269	0.130	0.320	0.316	0.142	0.219	0.221
PSTRE	0.136	0.192	0.376	0.288	0.146	0.306	0.226	0.092	0.234	0.243	0.101	0.202	0.188

Appendix 18.1: Marginal correlations (Pearson) per country of the cognitive domains literacy (LIT), numeracy (NUM) and problem solving in technology rich environments (PSL), respectively, with scales of the BQ, based on the 10 plausible values obtained from the population modeling (conditioning)

	Learning at work	Readiness to learn	Use of ICT skills at home	Use of ICT skills at work	Use of influencing skills at work	Use of num skills at home	Use of num skills at work	Use of planning skills at work	Use of reading skills at home	Use of reading skills at work	Use of task discretion at work	Use of writing skills at home	Use of writing skills at work
France													
LIT	0.175	0.274	0.318	0.210	0.222	0.332	0.230	0.113	0.395	0.344	0.146	0.272	0.230
NUM	0.154	0.268	0.297	0.226	0.232	0.338	0.279	0.136	0.392	0.376	0.148	0.257	0.243
Germany													
LIT	0.036	0.256	0.335	0.217	0.179	0.327	0.207	0.076	0.368	0.262	0.101	0.230	0.158
NUM	0.031	0.247	0.291	0.214	0.196	0.329	0.252	0.088	0.339	0.250	0.123	0.201	0.148
PSTRE	0.033	0.172	0.373	0.235	0.087	0.331	0.211	0.002	0.265	0.157	0.035	0.224	0.120
Ireland													
LIT	0.044	0.242	0.262	0.194	0.148	0.226	0.169	0.129	0.320	0.243	0.136	0.245	0.192
NUM	0.050	0.225	0.236	0.176	0.133	0.243	0.215	0.106	0.306	0.242	0.160	0.227	0.168
PSTRE	0.018	0.155	0.355	0.273	0.075	0.229	0.181	0.059	0.232	0.144	0.144	0.239	0.131
Italy													
LIT	0.030	0.220	0.241	0.133	0.20	0.214	0.204	0.110	0.362	0.281	0.085	0.167	0.210
NUM	0.043	0.219	0.238	0.145	0.193	0.239	0.258	0.113	0.371	0.263	0.123	0.159	0.213
Japan													
LIT	0.073	0.243	0.208	0.174	0.09	0.143	0.170	0.001	0.271	0.153	0.015	0.052	0.095
NUM	0.041	0.250	0.194	0.216	0.161	0.179	0.256	0.048	0.279	0.205	0.078	0.061	0.119
PSTRE	0.023	0.178	0.267	0.246	0.03	0.151	0.195	-0.036	0.149	0.088	0.011	0.023	0.064
Korea													
LIT	0.092	0.327	0.304	0.183	0.200	0.317	0.220	0.080	0.370	0.252	0.035	0.153	0.163
NUM	0.073	0.300	0.271	0.154	0.197	0.298	0.211	0.073	0.351	0.223	0.033	0.132	0.149
PSTRE	0.092	0.164	0.309	0.208	0.045	0.203	0.119	-0.051	0.129	0.063	-0.049	0.125	0.108

Appendix 18.1: Marginal correlations (Pearson) per country of the cognitive domains literacy (LIT), numeracy (NUM) and problem solving in technology rich environments (PSL), respectively, with scales of the BQ, based on the 10 plausible values obtained from the population modeling (conditioning)

	Learning at work	Readiness to learn	Use of ICT skills at home	Use of ICT skills at work	Use of influencing skills at work	Use of num skills at home	Use of num skills at work	Use of planning skills at work	Use of reading skills at home	Use of reading skills at work	Use of task discretion at work	Use of writing skills at home	Use of writing skills at work
Netherlands													
LIT	0.055	0.339	0.372	0.241	0.185	0.278	0.178	0.102	0.336	0.252	0.175	0.280	0.188
NUM	0.050	0.301	0.330	0.218	0.185	0.278	0.229	0.110	0.320	0.235	0.170	0.249	0.170
PSTRE	0.071	0.274	0.365	0.208	0.152	0.304	0.158	0.049	0.276	0.157	0.141	0.240	0.142
Norway													
LIT	0.043	0.203	0.279	0.273	0.212	0.238	0.183	0.141	0.283	0.236	0.128	0.176	0.178
NUM	-0.011	0.150	0.235	0.282	0.175	0.236	0.227	0.138	0.243	0.227	0.116	0.142	0.169
PSTRE	0.066	0.184	0.335	0.257	0.128	0.295	0.224	0.048	0.243	0.131	0.050	0.197	0.134
Poland													
LIT	0.039	0.253	0.301	0.182	0.151	0.306	0.206	0.092	0.384	0.226	0.080	0.264	0.130
NUM	0.016	0.235	0.267	0.158	0.133	0.314	0.223	0.101	0.338	0.199	0.077	0.222	0.136
PSTRE	0.004	0.126	0.317	0.149	0.072	0.255	0.124	0.011	0.260	0.107	0.077	0.207	0.064
Russian Federation[18]													
LIT	0.052	0.180	0.177	0.080	0.014	0.225	0.088	0.032	0.273	0.098	0.070	0.084	0.088
NUM	0.110	0.180	0.160	0.054	0.048	0.249	0.119	0.019	0.259	0.122	0.073	0.066	0.103
PSTRE	0.092	0.194	0.263	0.122	0.049	0.225	0.139	0.057	0.279	0.166	0.072	0.112	0.140
Slovak Rep.													
LIT	0.120	0.293	0.121	0.087	0.098	0.256	0.125	0.061	0.370	0.193	0.082	0.135	0.139
NUM	0.133	0.319	0.169	0.121	0.115	0.278	0.173	0.101	0.388	0.226	0.119	0.165	0.156
PSTRE	0.007	0.131	0.188	0.15	0.055	0.131	0.125	0.036	0.151	0.095	0.081	0.089	0.129

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

Appendix 18.1: Marginal correlations (Pearson) per country of the cognitive domains literacy (LIT), numeracy (NUM) and problem solving in technology rich environments (PSL), respectively, with scales of the BQ, based on the 10 plausible values obtained from the population modeling (conditioning)

	Learning at work	Readiness to learn	Use of ICT skills at home	Use of ICT skills at work	Use of influencing skills at work	Use of num skills at home	Use of num skills at work	Use of planning skills at work	Use of reading skills at home	Use of reading skills at work	Use of task discretion at work	Use of writing skills at home	Use of writing skills at work
Spain													
LIT	0.067	0.227	0.300	0.190	0.212	0.305	0.180	0.138	0.393	0.273	0.074	0.261	0.224
NUM	0.054	0.227	0.252	0.185	0.207	0.291	0.215	0.143	0.375	0.268	0.072	0.242	0.234
Sweden													
LIT	-0.013	0.201	0.276	0.253	0.171	0.219	0.207	0.038	0.279	0.184	0.042	0.203	0.156
NUM	0.014	0.178	0.222	0.227	0.16	0.223	0.243	0.063	0.244	0.181	0.065	0.152	0.127
PSTRE	0.070	0.204	0.377	0.269	0.107	0.276	0.253	-0.001	0.266	0.111	-0.018	0.264	0.114
United States													
LIT	-0.064	0.205	0.267	0.183	0.119	0.240	0.111	0.116	0.245	0.184	0.174	0.20	0.137
NUM	-0.036	0.179	0.245	0.199	0.118	0.235	0.141	0.112	0.220	0.186	0.152	0.164	0.131
PSTRE	-0.041	0.147	0.333	0.22	0.081	0.222	0.105	0.08	0.138	0.099	0.158	0.177	0.126

Note: The correlations for the ICT scales might be underestimated as not every respondent received the ICT items according to the path of the adaptive testing.

Chapter 19: Proficiency Scale Construction

Kentaro Yamamoto, Lale Khorramdel and Matthias von Davier, ETS

19.1 Overview

In this chapter we describe and illustrate the development of scales and items (based on respective frameworks) for the cognitive part of the PIAAC survey as well as the evaluation of the items and the instrument through a field test.

The Field Test addressed three main areas: a) operational (in terms of feasibility of implementation), b) instrumentation, and c) scaling and psychometric characteristics, and was important for the successful implementation of the Main Study. The fact that the results of PIAAC had to be linked to previous assessments, while also being implemented in both PBA and CBA modes (including an adaptive aspect), added to that importance.

Results of the Field Test provided information and guidance with regard to the sampling, data collection, refinement of scoring procedures for the CBA items, inference strategies, and analysis methods for the Main Study.

19.2 Development of the described scales

In the following, we will refer to the term "task" as an umbrella term for "item" as well as "item group associated with a common stem." A task can have a more complex structure compared to an item representing the construct or scale of interest, while an item is a question referring to a common stem or stimulus. Thus, one task can have one or multiple items. In the context of the description of the frameworks and scale developments, we refer to "tasks"; in the context of data analyses, we refer to "items."

19.2.1 Stage 1: Identifying possible scales

The identification and definition of scales (domains) to be measured in international large-scale assessments are important as they provide a foundation for the design of the assessment and set the boundaries for what will be included. PIAAC 2012 assessed the three main domains of literacy, numeracy, and problem solving and thus has to give definitions for them. All three domains are multifaceted constructs referring to complex competencies. The following section provides an overview of these definitions, explains on which prior definitions and assessments they are based, and explains to which extent prior definitions were expanded to meet new opportunities and changes in society.

Literacy scale:

The definition of the *literacy* scale in PIAAC is based on the previous adult literacy assessments known as IALS and ALL, but extends these assessments because adults have faced new literacy opportunities since (e.g., the use of email and other digital media) those assessments were created. Therefore, it was necessary to broaden the literacy construct to include new modes of text. PIAAC also provides an opportunity to deepen our understanding of the cognitive skills that underlie adult literacy and the role that engagement plays in literacy. While in IALS and ALL the literacy scale was divided into the scales *prose literacy* (continuous texts) and *document literacy* (noncontinuous texts), PIAAC joins them into one *literacy* scale. On the one hand, the concept of literacy in PIAAC was defined to support a link to the IALS and ALL assessments to enable the analysis of trends. On the other, it was expanded in three ways:

1) The range of texts to be considered should be broader than in previous assessments; in particular, the definition should include those texts often identified as electronic texts.
2) The type of cognitive activities identified should go beyond simply using text, to enable a deeper understanding of literacy ability.
3) The concept of literacy should also include engagement in literacy practices.

The Literacy Expert Group defines the PIAAC literacy scale as follows: "*Literacy is understanding, evaluating, using and engaging with written texts to participate in society, to achieve one's goals, and to develop one's knowledge and potential.*"

Excursus:

The following definitions and explanations provide a deeper understanding of the literacy definition that is used for PIAAC 2012:

- *Written text:* PIAAC aims to expand the range of texts which were assessed by IALS and ALL (informative texts of both continuous and noncontinuous form) to include a greater variety of text types, such as narrative and interactive texts, and a greater variety of media (computer, PDA, Blackberry or iPhone, etc.). Including electronic text opens the assessment to new types of text and content. Some of these novel form/content combinations include interactive texts, such as exchanges in comments sections of blogs or in email response threads, multiple texts, whether displayed at the same time on a screen or linked through hypertext, and expandable texts, where a summary can be linked to more detailed information if the user chooses.

- *Understanding:* Understanding means the construction of meaning (large and small, literal and implicit) from text. This can be as basic as understanding the meaning of the words, or as complex as comprehending the underlying theme of a lengthy argument or narrative. PIAAC aims to provide a more direct measure of understanding (not just an indirect one). While the assessment of reading components provides the construct to support basic understanding, the assessment of literacy in PIAAC includes tasks that explicitly tap more complex understanding, such as the relation(s) between different parts of the text, the gist of the text as a whole, and insight into the author's intent. Readers also have to understand the social function of each text and the way this influences structure and content.

- *Evaluating:* Readers continually (have to) make judgments about a text and evaluate information in terms of accuracy, reliability and timeliness. This is particularly important with online material as, in contrast to published print information, online information is more varied, ranging from authoritative sources to postings with unknown or uncertain authenticity.

- *Using:* Using means that the reader approaches the text with a specific task in mind, that is, reading is directed toward applying the information and ideas in a text to an immediate task or to reinforce or change beliefs. In some cases, using a text in this way requires just minimal understanding – getting the meaning of the words with some elementary recognition of structure. In others, it requires using both syntactic and more complex structural understanding to extract the information.

- *Engaging with:* Adults differ in how engaged they are with reading texts and how much a role reading plays in their lives (reading because it is required versus reading for pleasure). Studies have found that engagement with reading is an important correlate with the direct cognitive measures.

- *Participate in society:* Adults use text as a way to engage with their social surroundings, to learn about and to actively contribute to life in their community, close to home and more broadly. For many adults, literacy is essential to their participation in the labor force. Thus, literacy has a social aspect. It is a part of the interactions between and among individuals.

- *Achieve one's goals:* Literacy is increasingly complicit in meeting those needs, whether simply finding one's way through shopping, or negotiating complex bureaucracies whose rules are commonly available only in written texts. It is also important in meeting adult needs for sociability, entertainment and leisure, and work.

- *Develop one's potential:* Surveys suggest that many adults engage in some kind of learning throughout their life, much of it self-directed and informal. Much of this learning requires some use of text, and as individuals want to improve their life, whether at work or outside, they need to understand, use, and engage with printed and electronic materials.

In PIAAC texts are organized in three ways:

1) *Medium (print and digital):* A major development of PIAAC over previous adult surveys is the inclusion of digital (or electronic) texts. Because some texts that are applied electronically are just simple copies of printed texts, digital texts are not distinguished by the medium in which they occur, but by whether they make use of text navigation and display features found only through digital devices. Any text that could appear on a printed page exactly as it appears on a screen is considered a *print* text; any text that could not appear on a printed page with all its features intact is considered a *digital* text.

2) *Format (continuous and noncontinuous):* In IALS and ALL, texts were classified as

continuous (prose literacy) or noncontinuous (document literacy). This is an important distinction, as each requires different text knowledge and a different approach to text processing. At the same time, many actual texts involve some elements that are continuous and some that are noncontinuous. Thus, the distinction is better made on the basis of what type(s) of text a task requires.

a. Continuous: This type of text is conventionally made up of sentences formed into paragraphs. Some continuous texts include typographic features, such as indenting and headings, that signal the organization of the text, but many do not. Examples of continuous texts include newspaper and magazine articles, brochures, manuals, emails, and many web pages.

b. Noncontinuous: This type of text uses explicit typographic features, rather than paragraphs, to organize information. While there may be full sentences in some noncontinuous texts, most consist of words or phrases organized by some kind of matrix arrangement. Tables, graphs, charts and forms are all examples of noncontinuous texts.

c. Combined: This type of text has both continuous and noncontinuous elements. Examples of mixed texts include web pages with a list of links, newspaper articles that incorporate line graphs or pie charts, and brochures with attached order forms.

d. Multiple: Multiple texts consist of texts that have been generated and which make sense independent of each other. The texts are juxtaposed or loosely linked for a particular purpose. The relationships among the component texts need not be obvious. The texts may be contradictory or complementary. Such texts are common in digital settings, but are also found in print environments.

3) *Type (rhetorical stance of the text):* The IALS and ALL frameworks are classified as continuous texts by their rhetorical stance, because all share the same structure, but noncontinuous texts also share the same rhetorical stances. Therefore, in PIAAC, the stances of all types of text were identified using the six categories employed in the IALS and ALL assessments (the text type "hypertext" was eliminated in PIAAC because it is not a rhetorical category but a structural type which will be included under electronic text for PIAAC). The point of having rhetorical stance as a variable is not due to evidence that difficulty is affected by it, but as a way of ensuring that a variety of texts are included on the assessment. The six types of rhetorical stance for PIAAC are as follows:

a. Description: This is the type of text where the information refers to properties of objects in space. A page of a manual that identifies the parts of some device, such as a Cuisinart, is a description, as is a verbal depiction of a piece of art.

b. Narration: This is the type of text where the information refers to properties of objects

in time. Stories recounted to make a point, such as fables, are narrations, as are texts about the steps an individual took to solve a problem.

c. Exposition: In this type of text, information is presented as composite concepts or mental constructs, or those elements into which concepts or mental constructs can be analyzed. The text provides an explanation of how the component elements interrelate in a meaningful whole. A text that explains the nature of some health problem or one that tells about the election process in the United States would be an exposition.

d. Argumentation: This type of text presents propositions as to the relationship among concepts or other propositions. An important subclassification of argument texts is persuasive texts. Newspaper editorials are one example, as are advertisements.

e. Instruction (sometimes called injunction): This type of text provides directions on what to do. Most equipment manuals contain instruction texts, but so do other guides, such as those about first aid or a leisure activity.

f. Records: Records are texts that are designed to standardize, present and conserve information without embedding in other stances. A table of standings in a sports league is an example of a record, as is a graph of the changes in oil prices. The minutes of a meeting constitute another type of record.

More detailed information about how to classify noncontinuous texts (Matrix Documents, Graphic Documents, Locative Documents, Entry Documents, Combination Documents) and electronic texts (Hypertext, Index-like, Interactive) is given in the PIAAC literacy framework (OECD, 2012).

Because both the motivation to read and the interpretation of the content may be influenced by the *context*, a fair assessment must include material from a broad range of settings in order to include some material that would be familiar to any participant. PIAAC tried to include the following contexts (or content areas):

- Home and family
- Health and safety
- Community and citizenship
- Consumer economics
- Work
- Leisure and recreation
- Education and training

Furthermore, the following three *cognitive operations with text* can be identified that are needed when working on items or tasks:

1) Access and identify information in the text
2) Integrate and interpret (relate parts of text to each other)
3) Evaluate and reflect (understanding of the text as a whole)

As a supplement to the main literacy assessment, PIAAC includes an additional assessment of *reading components*. This assessment aims to provide information on the reading abilities of adults with poor skills in order to get a proper understanding of their difficulties. The following five reading components were identified:

- Alphanumeric perceptual knowledge and familiarity
- Word recognition
- Word knowledge (vocabulary)
- Sentence processing
- Passage fluency

More detailed information about contexts, cognitive operations, and further points that influence the difficulty of items (such as the transparency of items, semantic complexity, amount of information needed, prominence of information, and competing information), as well as more information about the reading components, is given in the PIAAC literacy framework (OECD, 2012).

Numeracy scale:

Basic computational or mathematical knowledge has always been considered part of the fundamental skills that adults need to function well and be able to accomplish various goals in their everyday, work and social life. Societies now present increasing amounts and wider ranges of information of a quantitative nature to citizens from all walks of life in diverse contexts. As workplaces are becoming more concerned with involving all workers in improving efficiency and quality, the importance of numeracy skills is growing. Numeracy involves, among other things, the handling of arithmetical processes, understanding of proportions and probabilistic ideas, understanding of numerical, geometric and graphical types, and representations of quantitative information, critical interpretation of statistical or mathematical messages, and ability to solve various types of quantitative problems.

The Numeracy Expert Group defines the PIAAC numeracy scale as follows: "*Numeracy is the ability to access, use, apply, interpret, and communicate mathematical information and ideas in order to effectively manage and respond to the mathematical demands of diverse situations in the information age.*"

The conceptualization of numeracy is based on the previous adult literacy assessments IALS and ALL, as well as on a review of scholarly literature and research findings (with regard to IALS, the numeracy scale in PIAAC is most closely related to the scales of document literacy and quantitative literacy). Numeracy operates on two levels:

> It relates to numeracy as a construct describing a competence as defined above, and to numerate behavior, which is the way a person's numeracy is manifested in the face of situations or contexts, which have mathematical elements or carry information of a quantitative nature. In this way, inferences about a person's numeracy are possible through analysis of performance on assessment tasks designed to elicit numerate behavior.
>
> In congruence to the view of numeracy as a competence, numeracy will be described as comprising both cognitive elements (i.e., various knowledge bases and skills) as well as noncognitive or semicognitive elements (i.e., attitudes, beliefs, habits of mind, and other dispositions) which together shape a person's numerate behavior.

The Numeracy Expert Group gives the following definition for numerate behavior: "Numerate behavior involves managing a situation or solving a problem in a real context, by responding to mathematical content/information/ideas represented in multiple ways."

Thus, numerate behavior comprises four facets:

a) contexts (everyday life, work, societal, further learning)
b) responses (identify, locate or access; act upon, use; order, count, estimate, compute, measure, model; interpret; evaluate/solve; communicate)
c) mathematical content/information/ideas (quantity and number; dimension and shape; pattern, relationships, change; data and chance)
d) representations of mathematical information (objects and pictures; numbers and mathematical symbols; formulae; diagrams and maps, graphs, tables; texts; technology-based displays)

A more detailed definition of these four facets is given in the PIAAC numeracy framework (OECD, 2012).

Numeracy is required so people can effectively cope with or respond to a range of situations that are embedded in the course of life with real, personal meaning to them. Three key types of situations are given below to illustrate the range of numeracy demands placed on adults:

- *Generative situations*: These demand that people count, quantify, compute, or otherwise manipulate numbers, concrete objects, visual elements, and so forth, to create/generate new numbers or estimates (e.g., calculating the total price of products while shopping, finding the number of boxes in a crate, measuring the area of a room to be painted in order to calculate the amount of materials needed to do the job, reading a menu and computing the cost of a specified meal, filling out an order form for a product, figuring out travel times between train stations based on a timetable, etc.). The numerical information in many types of generative situations may be evident in the situation itself (e.g., real objects to be arranged, sorted, counted, or measured; a graph on a computer

display) or may also be communicated through text or embedded in different types of text; hence, such situations may also involve language skills to varying degrees.

- *Interpretive situations*: These demand that people make sense, and grasp the implications of, messages that contain information of a mathematical or statistical nature but that *do not involve direct manipulation of numbers* (e.g., deciding whether a generalization stated in a newspaper article about results from a recent opinion poll is valid; other examples can be added where references to proportions, averages, samples, bias, correlation, risk, or causality are discussed or implied, such as in the context of genetic or medical counseling, or understanding of statistical process control displays).

- *Decision situations*: These demand that people locate and consider multiple pieces of information in order to determine a course of action, typically in the presence of conflicting goals, constraints or uncertainty. Two key subtypes here are *optimization tasks* (identification of optimal ways to use resources such as money or supplies, or schedule personnel or time) and *choice tasks* (making choices among alternatives, such as which of several apartments to rent, which pension or health insurance plan to join, whether to undergo a surgical medical procedure that has known probabilities of certain side effects, etc.). It is important to note that optimization and choice tasks can be part of a broader problem-solving process, where alternatives have to be generated and then evaluated. Thus, what is being termed here a decision situation at times also can be viewed as a problem-solving situation.

The three types of numeracy situations described above are not mutually exclusive, and other cases may exist, possibly of a hybrid nature. Moreover, it is important to keep in mind the impact of evolving technologies (Internet- or technology-based resources).

While it is possible to define numeracy in general terms without invoking literacy, the structure of the tasks and demands in adults' lives show these areas cannot be considered mutually exclusive. Mathematical or statistical information is carried by or embedded in text in some, but certainly not all, contexts in which adults have to function. To the extent this happens, one's performance on numeracy tasks will depend not only on formal mathematical or statistical knowledge but possibly also on literacy related factors such as vocabulary, reading comprehension, reading strategies, or prior literacy experiences.

Problem solving scale:

The aim of PIAAC to assess problem solving in technology-rich environments (PS-TRE) was based on the fact that digital technologies have deeply transformed the way individuals learn, communicate, work, and, more generally, the way they function in societies. Microcomputers, laptops, mobile phones, and the Internet have provided users with powerful tools to search for and make use of immense repertories of information and services. Increasingly versatile mobile technologies allow users to stay connected almost regardless of where they are and what they are doing. And the integration of digital tools in homes, cars and appliances potentially increases the safety, flexibility, and effectiveness of many activities of everyday life.

Yet using computers or other digital devices to perform personal or work-related activities and to solve problems often presents a challenge for the everyday user. People often have trouble

installing, setting up, and learning how to use new digital devices and software applications. Users often confine themselves to a few basic, but ineffective, procedures. Then, even routine computer use for mundane tasks is often prone to errors, delays and incidents. Tools and technologies are normally meant to facilitate the resolution of a problem. They may, however, also contribute to making a problem more difficult, especially when a person has limited knowledge and experience with those tools and technologies.

Therefore, PIAAC aimed to analyze the problem-solving skills involved in the uses of digital technologies, thus concentrating on problems people deal with when using ICT. Those problems share the following characteristics:

- The existence of the problem is primarily a *consequence of the availability of new technologies*. One example relates to the vast amount of information now available on the Web. This gives rise to problems related to locating and evaluating information for quality and credibility, for example, when seeking advice about legal issues or medical conditions. Other examples include the increasing capacity of electronic storage devices, with the subsequent problems of organizing and sorting large numbers of files; or the growing practice of social communication on the Web, with the subsequent problem of learning and making use of new social norms as far as private vs. public information.

- The problem solution *requires the use of computer-based artifacts* (tools, representational formats, computational procedures) that were not available previously, or at least not available to the general public. An example is the management of personal finance by using spreadsheets, statistical packages, and graphical tools. Here the problem itself may not be new (i.e., keeping spending in balance with income), but the new artifacts modify the distribution of work across social agents (professional vs. laypersons) and deeply transform the procedures and steps required to solve the problem.

- The problems are *related to the handling and maintenance of technology-rich environments* themselves (e.g., how to operate a computer, how to fix a settings problem, how to use the Internet browser in a technical sense).

Understanding and evaluating meaningful information available in technology-rich environments is central to the construct of problem solving. Most of the problems require one to handle vast amounts of symbolic information and, thus, the ability to deal with semantic content or meaning (e.g,. understanding command names in dropdown menus, naming of files and folders, hits in a search engine, or links in a Web page). Furthermore, many problems require the person to read and understand electronic texts, graphics and numerical data.

The Problem Solving Expert Group defines the PIAAC problem-solving scale as follows: "*Problem solving in technology-rich environments (PS-TRE) involves using digital technology, communication tools and networks to acquire and evaluate information, communicate with others and perform practical tasks. PIAAC 2012 will focused on the abilities to solve problems for personal, work and civic purposes by setting up appropriate goals and plans, accessing and making use of information through computers and computer networks.*"

More information and specific comments on the words and phrases used in this definition is given in the PIAAC problem-solving framework (OECD, 2012).

The PIAAC domain of problem solving may be organized along three key dimensions:

> *Cognitive dimensions:* the mental structures and processes by which a person actually performs problem solving (goal setting and monitoring progress; planning; locating and evaluating information; and selecting, organizing, and transforming information)
> *Technologies:* the devices, applications and functionalities through which problem solving is conducted (hardware devices; simulated software applications; commands and functions; representations such as text and graphics, etc.)
> *Tasks or problem statements:* elements of a situation that trigger a condition for problem solving (scenario and task directions presented to test takers; specific material conditions in which the test is organized)

More detailed information and examples of the different key dimensions of the PIAAC problem-solving scale are given in the PIAAC problem-solving framework (OECD, 2012).

Even if the domains of literacy, numeracy, and problem solving rely on the same "core" cognitive processes (e.g., the ability to decode printed symbols, working memory capacity), there are aspects that distinguish problem solving from the other two domains:

- As problem solving specifically assesses goal setting, monitoring, and planning in technology-rich environments, problem-solving tasks emphasize the processes of problem finding and problem shaping that are typical of problem solving. Problem-solving tasks also focus on the kinds of problems that are associated with these environments (e.g., problems associated with Web-based texts that are not well defined and the need for logical operators to search for information).

- Problem-solving tasks were carried out in environments that involve multiple, complex sources of information. Some of the tasks even required the test taker to use multiple environments and to shift across them. Thus, problem solving assessed decision making as far as information sources to be used (e.g., the act of choosing which environment to use or whether or not to go to another website). Evaluation was included as a critical underlying part of problem solving. Additionally, selecting appropriate devices or tools took a more prominent role for this domain.

- In terms of information processing, problem solving is a specific construct in that: a) it focuses on the pragmatic evaluation of sources in terms of reliability and the adequacy of information relative to the problem statement as opposed to mere topical relevance, which is more applicable for literacy; b) it focuses on the integration of information across sources, especially in cases where the sources provide inconsistent information.

19.2.2 Stage 2: Design principles and constraints (selecting items for the assessment)

During the item development process and the assignment of items per PIAAC domain (scale) to the assessment, the following principles were taken into consideration:

a) *Items should cover as many aspects as possible with regard to the different text types, contexts and processes of literacy, the different facets and contexts of numeracy, and the different cognitive dimensions and contexts of problem solving.* Items should require the activation of a broad range of skills and knowledge included in these constructs, as portrayed in the conceptual frameworks.
b) *Items should aspire to maximal authenticity and cultural appropriateness.* Items should be derived from real-life stimuli and pertain to situations that can be expected to be of importance or relevant in different contexts in at least some of the countries participating in PIAAC. Item content and questions should appear purposeful to respondents across cultures, even if they are not necessarily familiar to all adults in all countries.
c) *Items should have a free-response format, to the extent feasible by the computer platform used for administering the direct assessments in PIAAC.* Items should be structured to include a stimulus (e.g., a picture, drawing, visual display) and one or more questions, the answers to which the respondent communicates via the modes available on the computer, primarily: entry, click, highlight a region of the stimulus, usage of various pull-down menus. (Text entry is limited to very specific words or sometimes a simple number due to the concerns listed above regarding the inability to score text entries with keying/typing errors, and the presence of multiple ways to express the same content in words and/or numbers).
d) *Items should spread over different levels of ability.* Items should span the range of ability levels anticipated within PIAAC participants, from low-skilled individuals (which are of interest in countries where policies and educational programs may be earmarked for low-skill populations) all the way to those with advanced competencies. The need to reduce the number of items to be administered in any one domain has led to the practice (in previous assessments as well as in PIAAC) of including few very easy items (i.e., items at level 1) and few very hard items (i.e., items at Level 5). Respondents will be classified at Level 1 if they could not do well on Level 2 tasks. Likewise, those classified at Level 5 will be those who performed well on Level 4 items and on the few real Level 5 items.
   It follows that a more detailed assessment of the specific skills of Level 1 respondents requires a separate diagnostic assessment. Therefore, the reading components assessment was conducted in PIAAC. To enable the adaptive testing process and thus reach an efficient estimation of respondents' ability levels, the following distribution of items at the different difficulty levels was sought for constructing the item pool for literacy and numeracy (there was no adaptive testing for problem solving in technology-rich environments) for the main PIAAC assessment, based on the results of the Field Test (pilot test) in 2010: 5% Level 1 items, 25% Level 2 items, 40% Level 3 items, 25% Level 4 items, and 5% Level 5 items.
e) *Items should vary in the degree to which the task is embedded in text.* Some items should be embedded in or include relatively rich texts, while others should use little or no text. This distribution aimed to reflect the different levels of text involvement in real-world

numeracy tasks, as well as reduce overlap with the literacy scale.

f) *Items should be efficient.* To allow for coverage of many key facets of the literacy, numeracy, and problem-solving competencies, the inclusion of a large number of diverse stimuli and questions was needed. However, in light of testing time constraints, the use of short items was necessitated, precluding items that could simulate extended problem-solving processes or require a lengthy open-ended response.

g) *Items should be adaptable to unit systems across participating countries.* Items should be designed so that their underlying literacy/mathematical/problem solving demands are as consistent as possible across countries regarding language and conventions. For example, items were designed so that different currency systems or different systems of measurement (metric or imperial) could be applied to the numbers or figures used. Items should retain equivalency with respect to their literacy/mathematical/problem solving or cognitive demands after being translated.

In addition to the above listed principles, the assignment of items to the PIAAC assessment design had to address two further points: the linking between PIAAC and previous surveys, and the link between the CBA and PBA. To enable the linking among PIAAC, IALS and ALL, a part of the PIAAC item pool came from the IALS and ALL surveys (approximately 60%), while the other part consisted of new items that were developed for PIAAC. With regard to the literacy scale, the newly developed PIAAC items had to be assigned either to the subscale "prose literacy" or the subscale "document literacy" as the scale "literacy" was divided into these subscales in IALS and ALL. To enable the link between the PBA and the new CBA, a portion of the IALS and ALL items, which were all paper-based, had to be redesignated to be administered within the CBA. Furthermore, a portion of the newly developed items had to be assigned to both modes of assessment. Altogether, a larger portion of the IALS and ALL items as well as the newly developed PIAAC items was used for the CBA, while a smaller portion was used for the PBA. The latter procedure had not only the aim of enabling the linking design but also to provide a reliable and valid assessment for adults who were unfamiliar or uncomfortable with computers.

Due to the limited testing time (only 60-70 minutes for the core part, the cognitive adaptive assessment, and the BQ), it was decided to use a larger number of short tasks for the scales of literacy and numeracy (in order to cover all relevant contexts and facets) instead of a smaller number of more complex tasks, although it is recognized that ability to solve complex or extended literacy and numeracy problems is an inherent part of these competencies.

PIAAC also aimed to include open-ended response formats, with the limitation that the computer system (TAO) in the current stage of development could not accept most types of free-form text-based answers because of the huge possible diversity in how respondents may enter answers. The limitations stem from the difficulty of automatically coding the responses in dozens of languages while accommodating various grammatical and syntactical structures, as well as overcoming typing mistakes, which are naturally expected when people type text into a computer. Some workarounds were implemented to capture selected types of open-ended responses and circumvent the text-processing limitation to some extent, for example, by using multiple pull-down menus that allow a respondent to "construct" a response from predesigned elements or

response ranges. Maybe in future cycles of PIAAC, some of the current technical limitations will be resolved, allowing for better coverage of more aspects of the assessed constructs.

19.2.3 Stage 3: Field Test – Aims, design, and data collection

After developing new items (for literacy and numeracy) and new measures (for reading components and problem solving in technology-rich environments), and assigning old and new items to the cognitive domains based on their respective frameworks, the quality of the developed instrument had to be tested and evaluated. More precisely, the scaling and psychometric characteristics of the items had to be evaluated before using the items for the PIAAC Main Study in 2012. Furthermore, it was necessary to evaluate if the linking design was working and providing reliable trend measures, and if the computer delivery platform (for the CBA) was stable and reliable. Thus, a Field Test trial was designed and data analyzed in 2010 to yield adequate information relating to these questions. Moreover, standardized procedures and quality mechanisms were tested in the Field Test; they were embedded into various phases of PIAAC including survey development, implementation, and analysis and reporting of the data. The outcomes of the Field Test were used to assemble the final instruments that were used in the Main Study, and operational issues were modified and refined based on the Field Test. In summary the following areas were evaluated:

- Evaluation of survey operations procedures (data collection procedures, response rates for various subpopulations, data processing including scoring, recoding, and data transmission)
- Quality of the instrument: scaling and psychometric characteristics
- Equivalence of assessment modes: CBA vs. PBA
- Comparability of results between countries
- Trend measure: link between IALS, ALL and PIAAC

The PIAAC Field Test was designed to measures the domains of literacy and reading components, numeracy, and problem solving in technology-rich environments across two modes of administration (paper and pencil and computer delivered), while also offering participating countries both core and optional components. As mentioned earlier, 60 percent of the literacy and numeracy items came from the ALL and the IALS surveys to allow a link to these assessments and provide trend measures.

The full Field Test design assumed 40-45 minutes of administration time for the BQ and JRA and 60 minutes for the direct assessment. The design was based on the sample yield of 1,500 respondents per country/per language (i.e., completed cases) between the ages of 16 to 65: 1,100 for the CBA and 400 for the PBA (with a later modification of a maximum of 200 ICT-core failed samples to be routed to the PBA). On average across 23 countries, 209 respondents failed the ICT-core items, 1,426 completed the BQ section, 830 completed the CBA, and 505 completed the PBA.

Equivalence of scoring standard across countries

Achieving the goal of comparability depends on the equivalence of scoring within and between countries. Scoring was required to determine whether respondents have correctly answered the questions in the paper-based cognitive instruments. Rescoring was conducted as a quality assurance measure to determine whether the scoring rubrics have been applied consistently by every scorer within the country and without bias across the countries.

During the Field Test, participating countries checked the consistency of scoring by having a second scorer rescore 100% of the instruments. Additionally, item-level reliability was conducted to identify items that the scorers had difficulty in scoring consistently. Items with low interrater reliability have been further examined for possible ways to improve scoring accuracy through improved translation, instruction, and/or training for the Main Study.

a) Inter-country scoring reliability (equivalence of scoring of anchor booklets)

In order to evaluate scoring standard across countries, anchor booklets were produced in English (60 anchor booklets for the core part, 60 for literacy, and 60 for numeracy in both the Field Test and the Main Study). This common set of booklets was prepared by test developers and distributed to all countries. Item responses in these booklets were based on actual responses collected in the field as well as responses that reflected key points on which scorers were trained. Because responses were provided in English, scoring teams in each country designated two bilingual scorers responsible for the double-scoring process. Countries were required to follow a specified design to ensure that each booklet was scored twice and that each scorer functioned both as first and second scorer across all the booklets. Scoring results of both scorers were evaluated by the Consortium for consistency between the scorers as well as accuracy against the master scores as designed.

The unit of analysis implemented to evaluate agreement was the number of items multiplied by the number of countries, i.e., (38x23=)836 for the literacy scale and (35x23=)770 for numeracy. Average percentage agreement over items within a country averaged across all countries was 95.7% for literacy items and 95.6% for numeracy items. The variance of average agreements was 2.42 for literacy and 0.06 for numeracy. The number of item by country pairs showing less than 85% agreement was 24 for literacy and 14 for numeracy. Out of those lower agreements, two items were responsible for 12 of 24 for literacy items, and also two items accounted for eight of 14 lower agreements for numeracy. Regarding disagreements per country, there were two countries with more disagreements than the rest of countries. These two countries accounted for 15 of the total of 38 lower agreement item*country pairs.

Altogether, the rescoring of anchor booklets indicated very clearly that scoring of printed cognitive items is accurate, consistent, and without evidence of bias.

b) Intra-country scoring reliability

While reliable scores of anchor booklets ensure comparability of scoring standard across countries, reliability of scoring within a country indicates how accurately such a scoring standard was applied consistently among multiple scorers within a country. Countries followed rescoring instructions that were provided for three-, four- and five-scorer situations.

The unit of analysis implemented to evaluate agreement is identical to one used for the anchor booklets rescoring. Average percentage agreement over items within a country averaged across all countries was 96.1% for literacy items and 97.3% for numeracy items. The variance of average agreements was 10.35 for literacy and 4.00 for numeracy. Number of item by country pairs with less than 85% agreement was 43 for literacy and 13 for numeracy. Out of those lower agreements, three items were responsible for 18 of 43 for literacy items, and one item accounted for eight of 13 lower agreements for numeracy. In terms of country, there were three countries with more disagreements than the rest of the countries. These three countries accounted for 32 out of 52 lower agreement item by country pairs.

Altogether, a small number of items and countries have shown some difficulty in attaining high score reliability. As a consequence, some recommendations were given to optimize the scorer training for the PIAAC Main Study as well as the data capture, operational issues, data transmission, and quality assurance mechanisms.

Instrumentation:

The Field Test addressed the following issues related to instrumentation:

- The accuracy and comparability of survey instruments were reviewed, including translation and scoring guides and all related manuals. These activities resulted in a number of corrections and clarifications.

- The timing and flow of questions in the BQ was evaluated. (The researchers in GESIS performed this task, resulting in the reports included in the summary of BQ instruments.)

- The appropriateness of questions across participating countries was evaluated.

- The response distribution in all categories of the BQ was examined.

The timing information from the Field Test was used to make sure that the Main Study wouldn't be too long. The Field Test showed that the majority of respondents needed one hour to complete the assessment, and that they were much faster in completing the reading components than expected. Therefore, more items could be included for the Main Study from the existing item pool (one more reading component passage was used in the Main Study than originally planned).

Computer delivery platform:

To evaluate the CBA in PIAAC, the Field Test was delivered on a laptop computer to respondents in their homes. A computer-delivery platform (TAO) integrated with the CAPI tool was used for the administration of the BQ, the JRA and the cognitive instruments. The Field Test addressed the following issues related to the computer-delivery platform:

- The functioning of the cognitive portion of the delivery platform was tested and evaluated (emphasizing response capturing and automatic scoring).

- The functioning of the CAPI system was tested and evaluated (emphasizing the flow of questions and efficiency of the system in capturing information).

- The accuracy of the interviewer's instructions was evaluated.
- The effectiveness of the system during the interview was tested.
- The integration of the PIAAC platform with national survey management systems was verified.

The Field Test showed that no major architectural changes were necessary for the platform, but some system freeze occurred during the test administration that had to be fixed. After addressing this issue in later updates, the Main Study instruments became very stable.

Scaling, psychometric characteristics, equivalence between modes and assessments:

The Field Test data were used to examine scaling methodologies in order to determine the psychometric characteristics of items and scales. This included the evaluation of the equivalence of item parameter estimates among linking items from IALS and ALL to PIAAC, and the equivalence of the estimates between the PBAs and the CBAs. To identify deviations of item-by-country interactions, two measures of mean deviation (MD) and root mean squared deviations (RMSD) were used (see section 17.3.2. for detailed information about the MD and RMSD).

Furthermore, the Field Test was also an opportunity to examine the role of computer familiarity and determine the standards for branching respondents with regard to the adaptive test design of the Main Study. The Field Test provided initial IRT item parameter estimates that were used to construct the adaptive testing algorithm, which was implemented in the Main Study. Thus, the Field Test had to address the following issues with respect to IRT scaling and psychometric characteristics:

- Literacy items were re-estimated using the entire aggregate data of IALS/ALL because the literacy scale is a joint scale of prose and document literacy scales. These new parameter estimates were used for the subsequent analyses. The numeracy scale was introduced in ALL, and subsequent analyses used ALL numeracy estimates.
- In order to examine equivalence of item characteristics across countries, a common set of item parameter estimates of the two-parameter logistic (2PL) model and the general partial credit model (GPCM) was estimated and found to fit quite well to all countries, for all three scales, and in both PBA and CBA. Deviation was fairly small and almost all countries and items were found to be conforming to the international parameter estimates. The sample size in the Field Test was too small for each country to estimate country-specific item parameters.
- Equivalence of item characteristics among the literacy and numeracy items common to IALS and ALL on the paper-and-pencil version was examined. Equivalence of IALS/ALL item parameter estimates to the CBA items adapted from IALS and ALL were also evaluated. Previously estimated IALS and ALL item parameters on PBA fit very well to the PBA items adapted for both scales of literacy and numeracy. For the IALS/ALL items adapted for the PIAAC CBA, previously estimated item parameters fit quite well for the numeracy scales with a few items showing noticeable deviation from the IALS/ALL estimates. For the literacy scales, more items showed clear deviation from

the IALS/ALL estimates. Equivalence of item characteristics of literacy and numeracy items common to PBA and CBA was examined. Several items were freed to estimate CBA only item parameters, while the majority of linking items shared common item parameters between PBA and CBA items.

- Items among the literacy, numeracy and problem-solving items were identified to be assembled into the core assessment.

- The expected proportions of subsamples routed to the different assessment modes and the different stages of the CBA based on preliminary background information and the core were examined. As working with various countries with various ability distributions makes it critical to have a sufficient number of responses for every item, simulation studies were calculated to evaluate item exposure under adaptive procedure.

- The overall psychometric characteristics and quality of the Field Test items were evaluated to guide the selection of items for the Main Study.

The Field Test design

The PIAAC Field Test design provided good item level information on the full range of direct assessment measures and was useful in addressing the other operational and psychometric issues identified above. The BQ and a core set of questions focusing on information and ICT was designed to ensure that respondents who have no familiarity with computers are routed to the PBA. Because the number of respondents without ICT skills could have been numerous, a limitation on the maximum number of respondents was placed at 200 so that the CBA item parameter estimation would not be jeopardized. The limit of 200 respondents was placed to avoid such a scenario. However, most of the countries never reached this limit during the data collection in the Field Test. In order to link the PBA and the CBA, the remaining adults (the majority of adults in each country who are expected to pass the core) were randomly assigned to either one of them. The Field Test design (see Figure 19.1) comprised the following steps and procedures:

Step 1, BQ: The BQ was designed to take 30-40 minutes, and was delivered by the interviewer using a computer-assisted format with respondents taking one of three variable sections (a 20-minute core set of items and one of three, 10-minute subsets that would be administered along with the cognitive instruments). Compared to the original design for the Main Study, the BQ required some modifications to accommodate the large number of questions that go beyond 30-40 minutes (implemented by rotating some of the questions). Moreover, not every respondent answered every question because appropriate questions were presented based on the answer to the previous question(s).

Step 2a, PBA: The PBA was designed to comprise a 10-minute core of either literacy or numeracy skills (each with six items), followed by two 20-minute blocks of literacy or numeracy (totaling 29 items), and a final 10-minute cluster of reading components. Thus, the total testing time was estimated to be 60 minutes. Four paper booklets with varied

(balanced) block orders were constructed to control for possible order effects (see Figure 19.1). In the Field Test (as well as the Main Study), every respondent in the PBA took the reading components (in case the international option to assess reading components was chosen by the respective country; see below). But while there was no link between the CBA and the reading components in the Field Test (only respondents working on the PBA also worked on the reading components), respondents who performed poorly on the core and literacy items of the CBA based survey were transferred to the reading components as well.

Step 2b, CBA: The CBA was designed to included twenty-one 60-minute booklets consisting of two 30-minute blocks of items in each booklet. While the items of the CBA in the Main Study were administered adaptively, this was not the case for the Field Test: The block order was balanced, but the item order within each block was fixed. As reflected in this design (see Figure 19.1), each of the computer-delivered booklets contained literacy-only tasks, numeracy-only tasks, literacy and problem-solving tasks, numeracy and problem-solving tasks, or problem solving-only tasks. Overall, for the Field Test, there were thirteen 30-minute blocks that were grouped to form the 21 booklets: four blocks of literacy tasks, four blocks of numeracy tasks, and five blocks of problem-solving tasks.

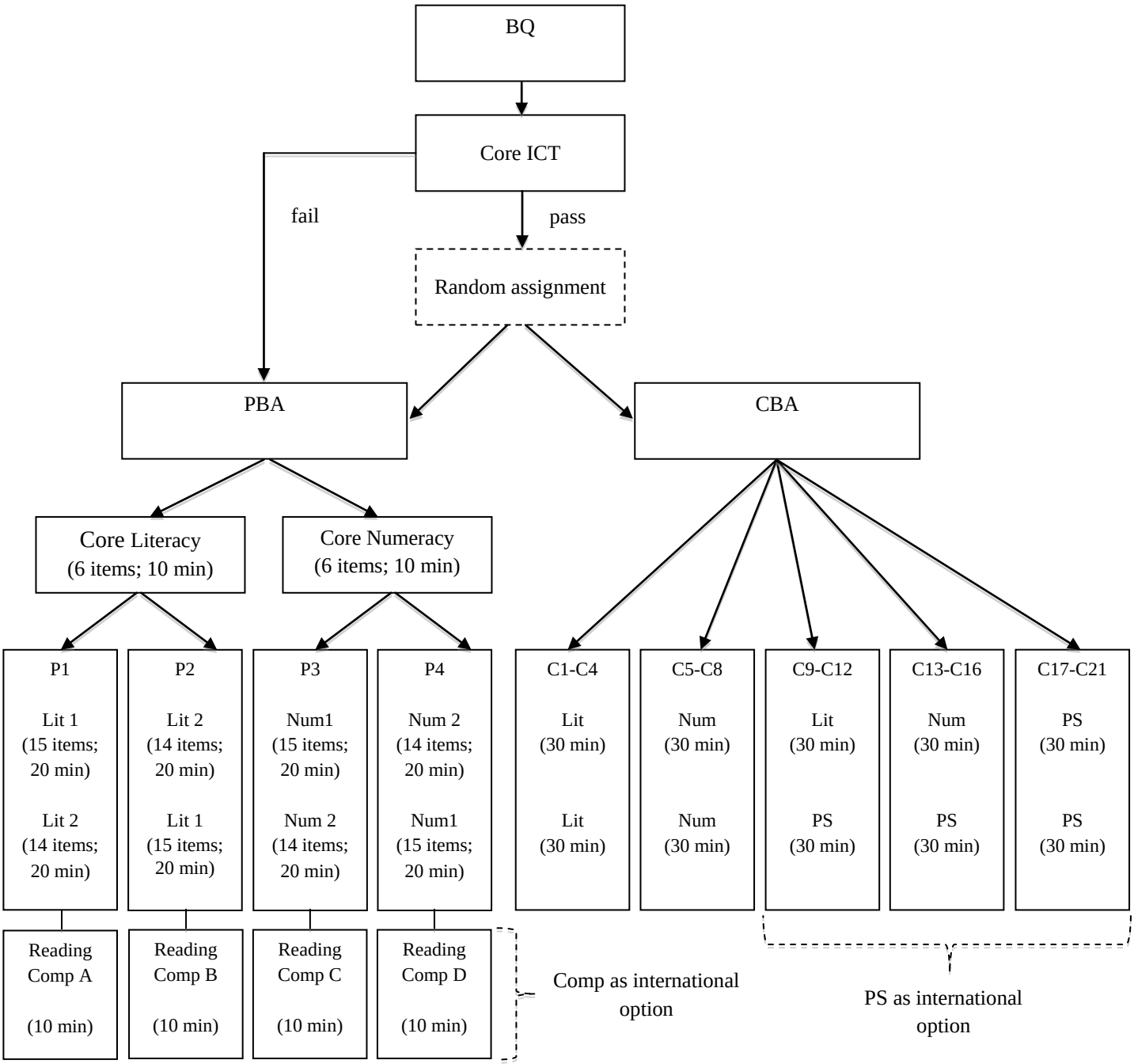
International options in the PIAAC Field Test: The Field Test offered the participating countries the option to assess reading components and problem solving, or not to assess one of them.

The reading components were optional to the participating countries, that is, each country could decide whether to include them in the assessment. Countries choosing the option not to include the reading components measures expected to save about 10 minutes in the overall assessment time and reduced their sample size by a total of 100 adults. The decision not to assess reading components had only minimal impact on the overall Field Test design.

The international option to include reading components but not to assess problem solving had a significant impact on both the sample size needed for the Field Test and the number of computer-based booklets. To compensate for the lack of covariance information between the different domains, the number of respondents per item was increased for the domains of literacy and numeracy, but the overall sample size was smaller (note that the main focus of the Field Test was not on the domain covariance but on the item parameter estimation for each single domain). In this design, assessment time per individual remains at 60 minutes, and each item is answered by 200 adults and based on an estimate of 1,200 respondents per country/per language (i.e., completed cases): 800 who respond to the computer-delivered measures and 400 who respond to the paper-and-pencil items.

In this Field Test design, the direct assessment time was 60 minutes, each item was to be answered by a minimum of 150 adults, and it was based on an estimate of 1,500 respondents per country/per language (i.e., completed cases): 1,100 for the computer-delivered test and 400 for the paper-and-pencil test. Although most countries never reached these numbers, many came close, thus allowing us to carry out the planned analyses.

Figure 19.1: Test design for the PIAAC Field Test 2010



19.2.4 Stage 4: Analyzing Field Test data

Analyses of the Field Test data were carried out to produce overall results as well as results by each participating country. The smallest unit of analysis was language by country data. For the cognitive data, the Field Test analysis included a range of descriptive analyses at both the national and international levels:

- Classical item analysis as well as analyses of collections of items using modern testing methodologies such as IRT
- Analyses of item by survey interaction for common items
- Analyses of item by mode of presentation interaction
- Analyses of item by language within a country
- Selection and rationale supporting the identification of core items, including cut points
- Development of branching rules to be used in the multistage adaptive branching of examinees into different paths of the assessment
- Evaluation of comparability of scoring standard and procedures within and between countries
- Evaluation of anchor booklets (as this was done for the first time in an international large-scale assessment)

The analysis of the Field Test data provided answers to questions related to the finalization of the design of the main assessment as well as the item selection for the main assessment. These questions include the development of the core that, in combination with items from the BQ, guided respondents to the PBA or the CBA and the assembly of booklets and design parameters for the multistage (or adaptive) testing.

The Field Test data were used to examine the comparability of the literacy and numeracy scales for PIAAC against the scales used in IALS and ALL (based on common items across the various surveys). These data were also used to evaluate the stability of the item parameters across the two modes of administration (PBA, CBA). Items that were comparable across the PBA and CBA were used to establish this important link for PIAAC. Field Test data were also used to reveal any item by country interactions and helped quantify these effects, as well as provided information on how they might be reduced (e.g., translation of display issues that can be easily identified and corrected). Results showed several issues associated with clear differences between scoring procedures of PBA and CBA. These findings were incorporated into the improved online scoring during the Main Study and the development of programs to harvest such information from nearly exhaustive log files.

Data on response time were examined as this allowed the Consortium to determine the comparability of time taken on each task across languages/countries, and whether the intended

timeframe established in the cognitive labs and previous tryouts hold up as feasible in the Field Test. In addition, the timing of the various blocks and booklets were reviewed and modified.

Item parameters estimated with the Field Test data using IRT analysis were fixed for the adaptive aspect of the Main Study.

19.2.5 Stage 5: Item selection for the Main Study based on the Field Test

The goal of the PIAAC Field Test was to provide new items to cover new domains and extensions of existing frameworks as well as linking items to establish a link among PIAAC, IALS and ALL as well as between PBA and CBA. In order to meet these target goals, it was necessary to develop and assess a larger pool of items for the Field Test compared to the Main Study. The PBA of the Field Test needed a total of 70 items – 35 literacy and 35 numeracy items (while 24+24=48 items were selected for the Main Study). The CBA of the Field Test needed 72 items for each domain (52 items were selected for the Main Study). Of these items, 42 were used to evaluate their utility as linking items for the CBA, while a subset of 25 was used to evaluate their utility for linking the PBA and CBA.

In the Field Test, on average, the respondents from most countries took less time to answer questions than anticipated by nearly 30%. It was decided to lengthen the test by about 10% for Literacy and Numeracy CBA booklets. The reason for not lengthening a full 30% was to reduce the number of respondents going over 60 minutes.

The selection of items for the Main Study was based on three main considerations:

- Measurement construct representations
- Survey design constraints
- Psychometric characteristics of an item as well as a set of items together

The assessment of problem solving in technology-rich environments involved scenarios of varying levels of complexities. Scenarios were designed to take between five and 15 minutes on average to complete. Overall, 150 minutes of testing material was developed for the Field Test (approximately 16 scenarios of varying lengths) with some 75 minutes of problem solving in technology-rich environment tasks selected for inclusion in the Main Study (approximately eight scenarios of varying lengths). The scenarios finally selected for the Main Study were organized into two 25-minute blocks.

With regard to the assessment of reading components, respondents worked through the items more quickly than expected by 2.25 minutes. However, for among least able respondents (below the 17th percentile), the average time was 9.87 minutes. The most able groups of respondents in every country converged to about 3 seconds per item for vocabulary tasks. The proportion correct (P+) differentiated reading components skills of PIAAC respondents rather well for respondents with low skills. For the Main Study, a total of 20 minutes was allotted to measure several of these skills, with final measures assembled from 40 minutes worth of Field Test data.

References

OECD (2012), Literacy, Numeracy and Problem Solving in Technology-Rich Environments: Framework for the OECD Survey of Adult Skills, OECD Publishing. http://dx.doi.org/10.1787/9789264128859-en

Chapter 20: Creating Simple and Complex Derived Variables and Validation of Background Questionnaire Data

Matthias von Davier, Jonathan Weeks and Henry Chen, ETS;
Jim Allen and Rolf van der Velden, ROA

20.1 Overview

The complex structure of the PIAAC BQ enabled the collection of variables from a diverse population of adults. But not all variables could be reasonably collected for all respondents (e.g., Loeys, Moerkerke, De Smet, & Buysse, 2012). Some were only appropriate for respondents in the workforce, while others were suitable only for those in training. Still another set was used for respondents who belonged to the group of recently unemployed. The need to adapt the BQ in order to provide appropriate sections for a diverse population can be best understood by examining the following examples:

- Current industry and occupation, as well as skill use at work, could be meaningfully asked only of those who were either employed or self-employed at the time of the interview, because respondents who are out of the labor force or never had paid work cannot reasonably be asked whether they use their literacy skills at work.

- For ICT skill use, questions assessing the domain were not presented to those without any previous contact with computers. In contrast, reading, writing and numeracy skills used at home were assessed for all respondents, and the corresponding scales for skills used at work were applied for those respondents who were part of the labor force and the recently (less than 12 months) unemployed.

- Earnings were only asked for those at work. Questions on earnings do not provide meaningful information when respondents are no longer part of the labor force or never had paid work. The same holds true for questions addressing those who were in education or training at the time.

At the same time, a host of other questions in sections addressing general domains are available for practically all respondents who completed the BQ. This is true for skills used at home, education history, questions about health, civil engagement, and approaches to learning, as well as socio-demographic information, among other things. The computer-based routing of respondents to those sections that were appropriate for respondents to answer led to an extremely high item-level response rate overall, as documented in the corresponding section of this chapter.

Clearly, care needs to be taken when analyzing these data. The sections below will provide an overview of some of the key areas for which the Consortium derived variables for use in secondary analyses. The next section presents an overview of those variables that PIAAC shares with previous large-scale assessments of adult populations. The following section discusses the assessment and derivation of earnings variables, and the final section discusses the derivation of variables related to self-reports of literacy skill use, job requirements and learning.

20.2 Overview of the BQ sections

The BQ collected data on a large variety of work-related, education-related and general domains such as socioeconomic variables, health-related questions and attitudinal variables that can be related to the cognitive assessment of literacy skills.

The BQ is too complex to try to reproduce all domains in great detail that were assessed in the instrument. Further information on the development and the content of the BQ is available in Chapter 3 of this report.

A PDF file that provides a linear representation of the international variables collected in the PIAAC BQ can be found at http://www.oecd.org/edu/48442549.pdf.

A framework that outlines the rationale of the selections made in the construction of the different sections of the BQ can be found at http://www.oecd.org/site/piaac/PIAAC(2011_11)MS_BQ_ConceptualFramework_1%20Dec%202011.pdf.

The sections of the BQ broadly covered the following domains relevant for assessing contexts of work, education, skill utilization, and demographics:

A: General information

B: Past education and current education and training

C: Current status and work history

D: Current work (if applicable)

E: Last job (past 12 months if no current job)

F: Skills used at work (JRA)

G: Literacy, numeracy, ICT at work

H: Literacy, numeracy, ICT at home

I: About yourself

J: Background

As stated above, the path through the BQ was an adaptive one, as different sections were appropriate for respondents who were employed, unemployed, out of the labor force, or still in school or training. Altogether, there were over 400 questions (without national

adaptation), so it becomes virtually impossible to report in detail on each of the questions. Instead, we provide Table 20.1, which shows the rate of response by country for those adaptively routed question paths presented to respondents. That is, only respondents that received questions are counted in terms of response or nonresponse.

Table 20.1: Response rate in detail

Country	BQ Rate	MIN	MAX	MEAN	Median	MIN Item	No Response	Below 50%	50% to 90%
Australia	88.2%	88.0%	99.5%	98.9%	99.4%	J_Q07b	37	0	2
Austria	50.3%	6.1%	100.0%	98.8%	99.8%	B_D01d	0	3	0
Canada	80.6%	9.9%	99.5%	97.9%	98.7%	B_D01d	0	2	0
Cyprus[1]	86.6%	7.8%	100.0%	98.7%	99.5%	B_D01d	0	2	0
Czech	40.5%	9.7%	100.0%	99.1%	99.9%	B_D01d	0	2	0
Denmark	46.3%	11.8%	99.6%	98.4%	99.1%	B_D01d	1	2	0
England/ N. Ireland	68.5%	15.0%	100.0%	99.1%	99.9%	B_D03d	0	2	2
Estonia	58.4%	7.5%	100.0%	98.9%	99.7%	B_D01d	0	2	0
Finland	70.7%	8.8%	99.9%	94.7%	95.3%	B_D01d	0	2	1
Flanders (Belgium)	54.2%	6.8%	100.0%	99.0%	99.9%	B_D01d	0	2	1
France	72.5%	8.3%	100.0%	90.2%	90.9%	B_D01d	0	2	18
Germany	52.5%	9.8%	100.0%	98.9%	99.9%	B_D01d	0	2	1
Ireland	92.7%	9.2%	99.9%	99.0%	99.8%	B_D01d	0	2	1
Italy	62.7%	4.5%	100.0%	98.4%	99.2%	B_D01d	0	2	0
Japan	47.0%	5.2%	100.0%	99.0%	99.9%	B_D01d	0	2	1
Korea	78.4%	9.9%	99.1%	97.7%	98.4%	B_D01d	0	2	0
Netherlands	50.3%	7.9%	100.0%	99.1%	99.8%	B_D01d	0	2	0
Norway	58.9%	12.6%	99.9%	98.9%	99.8%	B_D01d	0	2	1
Poland	51.0%	20.5%	98.0%	97.0%	97.7%	B_D01d	0	2	0
Russian Fed.[2]	99.5%	15.0%	100.0%	98.9%	99.8%	B_D01d	0	2	2
Slovak	61.6%	5.2%	100.0%	99.0%	99.8%	B_D01d	0	2	0
Spain	41.5%	11.5%	100.0%	99.1%	99.8%	B_D01d	0	2	0
Sweden	44.8%	10.5%	100.0%	99.0%	99.8%	B_D01d	0	2	0
United States	80.6%	10.0%	99.9%	98.7%	99.6%	B_D01d	0	2	2

[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.

Table 20.1 shows the mean and median percentage response rate, as well as the minimum and maximum, along with the item for which the minimum was observed. The last three columns provide an overview of the number of items without any responses, and those with responses below 50% and between 50% and 90%. It can be seen that due to confidentiality deletions, a few countries exhibit nonzero counts for items without any responses.

20.3 Overview of BQ trend variable domains

One of the major tasks of international assessments is providing trend information. For that reason, the PIAAC Consortium tried to collect and derive variables that can be viewed as comparable over three adult assessments: IALS, ALL and PIAAC. In order to achieve this, the Consortium developed a number of derived variables (DVs) based on the raw BQ variables collected during the computer-based background interview.

Table 20.2 shows an overview of these indices. It can be seen that several variables have a direct correspondence among the three assessments, while there are also variables that required derivation for one or more of the three assessments in order to arrive at a comparable definition across all three assessments. In order to do so, some variables had to be coarsened for the purpose of defining a variable that allows quantitative comparisons across assessments based on groupings using trend variables.

Table 20.2: Domains with available trend variables*

	Domain
1	Date of birth
2	Gender
3	Respondent's origin
4	Educational background - formal education
5	Language background
6	Respondent's mother's background
7	Respondent's father's background
8	Respondent's employment status
9	Work history - past 12 months
10	Job information - current job or last (past 12 months) job held
11	Education or training which the respondent has taken in the past 12 months
12	Education or training wanted but not taken in the past 12 months
13	Reading and writing in respondents' daily life
14	Civic participation - volunteer work
15	Health
16	Use of information technologies - computer use
17	Respondents' children's education

* Detailed information about matching variables in the BQ across IALS, ALL and PIAAC instruments is given in Appendix 4

While not all domains include trend variables, many of the central reporting variables were able to be matched. If no direct match could be achieved, questions that largely agree were identified. Appendix 4 provides the details on questions that are matched between IALS, ALL and PIAAC.

20.4 Development of derived earnings variables

20.4.1 Introduction

The BQ deployed two key innovations designed to make it easier for respondents to report their earnings, and thereby to improve the quality of the available earnings data and reduce item nonresponse. The first was for respondents to choose among reporting their earnings per hour, day, week, two weeks, month or year, or by piece rate. By removing the necessity for respondents to convert from their own preferred payment period to a predetermined standard, the aim was to improve the data quality and remove potential barriers to response. Furthermore, this approach automatically takes into account country differences in the payment period that are typically applied in most cases.

The second key innovation was an additional option for those who were still unwilling or unable to report their earnings as a precise amount. In this case, respondents were invited to report their earnings in broad categories. Again, in this case the categories were expressed per hour, per day, per week, per two weeks, or per month or per year according to the respondent's preference. This option was expected to be attractive for respondents who had only a rough idea of how much they earn per period and for those reluctant to reveal their precise earnings due to concerns such as privacy.

In addition to these key innovations, earnings were asked separately for wage and salary earners and for the self-employed, and there was a separate question for wage and salary earners in which they could report annual bonuses they may have received. Earnings of self-employed were asked per year, unless respondents had been in their current business for less than a year, in which case they were asked per month. For earnings of both wage and salary earners and the self-employed, as well as for annual bonuses, the option to report in broad categories was offered for those who were unwilling or unable to report directly.

Although the design of the set of questions was expected to yield significant advantages in terms of interview flow, item response and data quality, these advantages come at a price—there is no direct measure of earnings that ensues directly from the data. It was necessary to devise a fairly elaborate set of conversion rules to go from earnings as reported to the derived earnings variables used in the data. The first step is a fairly straightforward conversion of directly reported earnings from the earnings period option chosen by the respondent into every available alternative (e.g., from hourly to monthly, from yearly to daily, etc.). The second, and by far most complex, step comprised the conversion of earnings reported in broad categories into an equivalent direct amount. A third step comprised the construction of a set of standard variables that formed the basis for the earnings derived variables (DVs) to be included in the public data file. A fourth step involved a purchasing power parity (PPP) correction so that all earnings variables were expressed in terms of real disposable earnings in a fixed currency (in this case given in US dollars). Finally all earnings indicators were converted into deciles.

20.4.2 Conversion of directly reported earnings into all possible reporting periods

As stated above, this step was quite straightforward and involved using a set of fixed conversion rules from each reporting period into every other reporting period; earnings reported as a piece rate were first converted into an hourly rate based on an additional question regarding the usual number of hours per piece as estimated by the respondent. This conversion makes use of the number of hours worked per week, using rules on the ratio between the different reporting periods. Most of these variables are not intended for inclusion in the final data, which only include earnings expressed in hourly or monthly amounts. The reason for creating all of these variables is that they are needed as input for the following step, the conversion of earnings reported in broad categories into an equivalent direct amount.

20.4.3 Converting broad categories into equivalent direct amounts

As stated above, any respondents who were unable or unwilling to report their earnings precisely were given the option of reporting in broad categories. These categories were provided by each participating country on the basis of their national earnings distribution. For regular earnings of wage and salary earners, six broad bands were used, with the bands divided roughly along the 10^{th}, 25^{th}, 50^{th}, 75^{th} and 90^{th} percentiles of the national distribution, provided separately per hour, day, week, two weeks, month or year. For self-employed, the same bands were applied, but only per year or month, depending on whether the respondent had been in the current business at the time of the survey for at least a year or less than a year. For annual bonuses, three broad bands were used, with bands divided at roughly 5% and 10% of the median of national annual gross earnings.

Convenient as this option may have been for some respondents, it does not yield a unique earnings amount that can be directly compared with the direct earnings reported by the majority of respondents. Several alternative approaches were considered for dealing with this problem:

- Replacing the bands by a fixed amount, for example, the midpoint of the band or some other value considered to be the most likely value. This option was rejected for a number of reasons. The most important reason is that this would give rise to unwanted "lumpiness" in the data, which is not only a problem in its own right but leads to unavoidable and unsolvable problems when converting final earnings into deciles in a later step. Conversion into six discrete amounts inevitably means that all earnings reported in broad categories would be included in just six of the deciles. A further complication of this approach was caused by the fact that the broad bands were not usually strictly comparable across reporting periods. This was because countries usually rounded the dividing points into round amounts, for example, 6 Euros rather than, say, 5.78 Euros, which might be a strict conversion from the equivalent dividing point in terms of monthly earnings.

- Converting direct earnings into the six broad bands. This option was rejected for reasons similar to the previous option. In addition to the above-mentioned discrepancies between the different reporting periods within each country, there was the additional problem that there are non-negligible differences between the manner in which the bands were defined per country, which would negatively affect comparability. Finally, it was observed that

the bands used in the BQ were never intended to be used in this way, and in fact represent a highly unusual way in which to express earnings.

- Leaving the data unconverted, allowing users of the data to make their own conversions as they see fit. This option was not seriously considered, because it would essentially render the earnings data as included in the public data file for this group of respondents unusable.

Taking into account the serious limitations of the alternatives considered above, it was decided that a precise earnings amount would be imputed for every respondent who reported in broad categories. The imputation method comprised matching each of these respondents with a respondent who reported earnings directly, meaning the person was considered "most likely" to resemble him or her in terms of earnings, and assigning the precise amount reported by that respondent. The basis for this matching was predicted earnings on the basis of a regression model using key indicators such as highest education, skill level, age, gender and so on as predictors.

In somewhat more detail, the imputation process followed the following steps:

1. Precise earnings of wage and salary earners were converted into the same broad ranges as used in the BQ.
2. Earnings regressions were run on directly reported earnings, separately for hourly, daily, weekly, biweekly, monthly and yearly earnings, in each case also separately for low, medium and high earnings (earnings bands 1-2, 3-4 and 5-6 respectively).
3. Predicted earnings were saved in each case, both for those who reported earnings directly and those who reported in broad categories.
4. Cases that reported in broad categories were matched to their "nearest neighbor" in terms of predicted earnings among those who reported directly. This matching was conducted separately for each of the broad earnings ranges, thus ensuring that each case would always be matched with a "mate" who fell into the same broad category.
5. Based on this matching, each broad category case was assigned the actual directly reported hourly and monthly earnings value of its "mate." Note that this assignment always takes place based on the matching based on the reporting category actually used by respondents. For example, those who reported earnings based on an hourly rate were always matched on the basis of predicted hourly earnings. In this way, we ensured that the matching was as precise as possible, and removed any possible bias that might occur because the dividing amounts for the different reporting periods were not always strictly equivalent.
6. An equivalent process was used to derive imputed values for additional payments.
7. The imputed hourly and monthly earnings, as well as the imputed additional payments, were combined with directly reported earnings to form single hourly and monthly earnings variables.
8. A flag variable was created to indicate whether earnings were imputed or directly reported.

It should be noted that it did not prove possible to derive imputed earnings for the self-employed using such a methodology. The primary reason is the unusual earnings distribution for self-employed, and in particular the fact that a significant proportion of the self-employed had zero or negative earnings (in both cases reported as zero in the data). We were not successful in developing sufficiently reliable and robust regression models that were able to account for the unusual composition of this category of the self-employed, in terms of education, skills and other factors.

20.4.4 Construction of a set of standard variables

Starting with the above mentioned variables for wage and salary earners, combining actual and imputed earnings (hourly and monthly earnings, additional payments) and the direct monthly earnings measure for the self-employed, we then constructed a set of standard variables that formed the basis for the earnings DVs to be included in the public data file. The first two of these were hourly and monthly earnings of wage and salary earners, excluding bonuses. By adding additional payments to these (of course, with the necessary conversion to the payment period concerned), we then constructed two variables comprising hourly and monthly earnings of wage and salary earners including bonuses. By combining monthly earnings of wage and salary earners including bonuses with monthly earnings of the self-employed, we obtained an overall measure of total monthly earnings of wage and salary earners and self-employed.

20.4.5 Purchasing power parity (PPP) conversion

The next step involved a PPP correction, so that all earnings variables were expressed in terms of real disposable earnings in a fixed currency (in this case, US dollars). This is simply a multiplication by a constant value per country, based on data on purchasing power parity per country supplied by the OECD.

20.4.6 Conversion into deciles

Finally all earnings indicators were converted into deciles. This involved dividing the data for each earnings variable into 10 equally sized groups per country strictly based on the position in the distribution of earnings according to that variable. Where there were multiple cases at the cutoff points, these respondents were assigned to the higher and lower earnings group in the numbers required to produce groups of equal size, with individuals being randomly sorted into the higher and lower groups.

20.5 Derivation of variables related to self-reports of literacy skill use, job requirements and learning

20.5.1 Overview

In PIAAC, the skills of a population are not only measured directly through the cognitive instruments but also indirectly through the BQ by asking respondents to report on their use of skills both inside and outside of work. The frequency and type of activities associated with reading, writing, numeracy and information technology were targeted in the BQ using multiple items that were similarly worded to apply to activities both in and out of work. In addition, other areas, particularly those involving intrapersonal, interpersonal and other generic "soft" skills, not included in the direct assessment, were also addressed through a set of self-reported questions.

This set of questions makes up a module within the questionnaire that has been specifically developed for the PIAAC project: the JRA module.

Altogether, we constructed 13 scales based on a cross-country analysis of comparability, reliability, and convergent as well as discriminant validity. These scales were constructed using IRT, more specifically, the generalized partial credit model (GPCM), and person-specific levels of skill use were estimated using weighted likelihood estimation (WLE). Scale values were derived for all respondents who reported at least some limited activities in each of these domains. Those who reported no skill use in each of the 13 areas were not represented on any of these 13 scales; nevertheless, they provide important information with respect to the percentage of people in each participating country who do not use particular type of skills either in or outside of work. The data also allow users to examine the characteristics of these individuals.

20.5.2 Models and methods

The PIAAC BQ contains scales, that is, collections of questions around a topic, with regards to the domains of skill use, activities at work, and approaches to learning. These scales are mainly found in sections F, G, H and I of the BQ. The skill use scales are arranged around domains that relate to the literacy domains assessed in the cognitive part (the test) of the PIAAC. More specifically, questions around activities involving reading, writing, numeracy, and the use of technology were administered and respondents were asked to rate how often they perform these activities either at work (section G) or outside of work (section H).

Based on the arrangement of these questions within sections, and based on the topics covered therein, there are a number of groupings for these BQ items that a reviewer may come up with. Upon OECD's request, the PIAAC Consortium tested a set of 30 potential scales. The question is whether a given set of items can provide reliable, nonredundant measures of skill use (and other behavioral indicator) to justify the reporting of these results as a derived indicator. Three principal criteria were used to determine if a specific scale should be retained: average internal consistency reliabilities (as measured by Cronbach's alpha) across countries greater than or equal to 0.6, mean subscale (total score) correlations across countries less than 0.7, and ignorable country misfit as characterized by weighted root mean squared differences between empirical and expected response probabilities across countries. In all cases, the number of items associated with each potential scale is quite small (two to eight items); hence, the ability estimator must also be considered so as to minimize bias. The most common estimators of latent ability are maximum likelihood (ML) and expected a posteriori (EAP). The former does not incorporate any bias correction whereas the latter is a Bayesian approach that shrinks estimates toward the mean as a function of score reliability. In order to minimize bias without reducing the variability of the scores considerably, a weighted maximum likelihood estimation (WLE) approach was used.

20.5.2.1 Item parameter estimation

The skill use items as well as the items used in the approaches to learning and the job requirement analyses are measured using a five-point Likert scale. The items for each potential scale were fitted using the generalized partial credit model (GPCM; Muraki, 1992).[3]

[3] For the potential scales with only two items, a partial credit model (Masters, 1982) was used where the item slopes were constrained to be the same for both items.

$$P_{ij}(\theta) = \frac{\exp\left[\sum_{c=0}^{j} Da_i(\theta - b_{ic})\right]}{\sum_{c=0}^{J} \exp\left[\sum_{k=0}^{c} Da_i(\theta - b_{ik})\right]} \quad (1)$$

where the probability of responding in a given category, k, is modeled as a function of examinee ability, θ, and estimated item parameters. For the GPCM, a_i is the discrimination (slope) for item i, b_{ik} is a step parameter. D is a scaling constant equal to 1.7, not an estimated parameter, which is included in the estimation for reasons that relate the logistic models to probit models (e.g. Cramer, 2004). The item parameters were estimated using the PARSCALE software, implementing a multiple-group concurrent calibration with countries serving as the different groups, and countries equally weighted by means of standardizing the sampling weights to a constant sum per country. The estimation utilizes marginal maximum likelihood without any priors specified for the item parameters and is run in two phases. In the first phase the mean and standard deviation of the ability scale are fixed to 0 and 1 respectively, across countries, and the ability distribution is constrained to be normal. In the second phase the item parameter estimates from the first phase are used as starting values and the estimation proceeds after relaxing the normality constraint.

Scale Exclusion Criteria

With the item parameters estimated, the BQ items were checked for misfit and each of the potential scales was considered for elimination. Three primary exclusion criteria were used to identify items/scales that were problematic and/or provided redundant information:

Criterion 1: Scale Reliability – When reporting subscale results, it is important that the scores have sufficient reliability to allow for defensible inferences to be made on the basis of the scores. For cognitive measures, reliabilities of 0.70 or higher are generally preferred. If this criterion were used, nearly two-thirds of the potential scales would be flagged for possible exclusion. As such, a slightly relaxed criterion was used. In order to be considered for exclusion, the mean reliability across countries had to be less than 0.6, as characterized by Cronbach's alpha.

Criterion 2: Scale Correlations – In addition to being reliable, subscores should provide unique information about the measured background characteristics. Scales that provide redundant information may be of little utility; hence, the correlation between scales was considered. Potential scales with a mean correlation across countries greater than or equal to 0.7 were flagged for possible exclusion.

Criterion 3: Between Country Differences – When item parameters are estimated for measures administered across countries, there is a strong potential for item-by-country interactions which may lead to item misfit. Stated differently, the empirical response curves across countries may differ appreciably from the expected curves based on international item parameters. These differences may occur for individual items or all/most items in a subscale. To summarize these differences, a weighted root mean squared difference (WRMSD)

$$WRMSD_i = \sqrt{\sum_{C}\sum_{X} \frac{\omega_c(X)[p_{ijc}(X) - P_{ij}(X)]^2}{J}} \quad (2)$$

can be computed for each BQ item, *i*, where $P_{ij}(X)$ is the expected probability of responding in category *j* for a given ability, *X*, $p_{ijc}(X)$ is the proportion of examinees in country *C* responding in category *j*, and $\omega_c(X)$ is a set of weights corresponding to the expected proportion of examinees in country *C* at ability *X* for the given subscale. Items with a WRMSD greater than 0.25 logits were flagged for possible exclusion. Additionally, scales where more than half of the items had WRMSDs greater than 0.25 were flagged for possible exclusion.

20.5.2.2 Skill use level estimation

Once the final set of subscales was identified, examinee skill use levels for each scale were estimated. All of the potential scales have very few items; hence, there is an increased potential for bias in estimates of ability. The most common estimators of latent ability are ML and EAP. As mentioned earlier, the former does not incorporate any bias correction whereas the latter is a Bayesian approach that shrinks estimates toward the mean as a function of score reliability. As an alternative to EAPs, Warm (1989) proposed a weighted likelihood estimator for dichotomously scored responses that essentially serves as a bias-corrected ML estimator. Penfield and Bergeron (2005) extended this methodology to GPCM items.

The maximum likelihood estimate of θ for a given individual is equal to the value of θ that maximizes the log likelihood, *L*, of the associated response pattern given a fixed set of item parameters. This estimate is obtained iteratively through the use of the Newton-Raphson algorithm where the estimate at iteration *t* is equal to

$$\hat{\theta}_t = \hat{\theta}_{t-1} - \frac{L'}{L''} \tag{3}$$

In Equation 3, L' and L'' are given by

$$L' = \sum_{i=1}^{N} \sum_{j=0}^{J} u_{ij} D a_i (j - \lambda_1), \tag{4}$$

$$L'' = -\sum_{i=1}^{N} D^2 a_i^2 (\lambda_2 - \lambda_1^2), \tag{5}$$

where $\lambda_k = \sum_{j=0}^{J} j^k P_{ij}$ and P_{ij} is the expected probability from (1). Under this formulation $\lambda_1 = \sum_{j=0}^{J} j P_{ij}$ and $\lambda_2 = \sum_{j=0}^{J} j^2 P_{ij}$. The standard error for $\hat{\theta}$ is equal to

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I}} \tag{6}$$

where I is the information of the test at θ, and is computed as

$$I = \sum_{i=1}^{N} D^2 a_i^2 (\lambda_2 - \lambda_1^2) \tag{7}$$

Extending this approach, the weighted likelihood estimator of θ at iteration t is equal to

$$\hat{\theta}_t = \hat{\theta}_{t-1} - \frac{W'}{W''} = \hat{\theta}_{t-1} - \frac{L' + B'}{L'' + B''} \tag{8}$$

where W is the weighted log likelihood (i.e., the bias corrected log-likelihood) and B' and B'' are given by

$$B' = \frac{\sum_{i=1}^{N} D a_i^3 (\lambda_3 - 3\lambda_1\lambda_2 + 2\lambda_1^3)}{2\sum_{i=1}^{N} a_i^2 (\lambda_2 - \lambda_1^2)} \tag{9}$$

$$B'' = \frac{AB - 2C^2}{B^2} \tag{10}$$

where

$$A = \sum_{i=1}^{N} D^2 a_i^4 (\lambda_4 - 4\lambda_1\lambda_3 - 3\lambda_2^2 + 12\lambda_1^2\lambda_2 - 6\lambda_1^4) \tag{11}$$

$$B = 2\sum_{i=1}^{N} a_i^2 (\lambda_2 - \lambda_1^2) \tag{12}$$

$$C = \sum_{i=1}^{N} D a_i^3 (\lambda_3 - 3\lambda_1\lambda_2 + 2\lambda_1^3) \tag{13}$$

Because B is proportional to the likelihood, it cannot be estimated directly (Warm, 1989). As such, the Newton-Raphson method is required. The standard error is the same as that obtained for the ML estimate.

20.5.3 Potential scales

By clustering related BQ items, 30 potential scales were identified by OECD analysts and the Consortium was asked to evaluate these scales. This list of scales included 18 non-nested scales and 12 nested scales (comprising subsets of items from four non-nested scales). In the list below, the values in the parentheses indicate the number of items associated with each scale.

Non-nested scales:

- Cooperation (2)
- ICT at home (7)
- ICT at work (7)
- Influence (7)
- Learning at work (3)
- Numeracy at home (6)

- Numeracy at work (6)
- Physical (2)
- Planning (3)
- Problem solving (2)
- Reading at home (8)
- Reading at work (8)
- Readiness to learn (6)
- Self-organization (2)
- Task discretion (4)
- Trust (2)
- Writing at home (4)
- Writing at work (4)

Nested Scales:

- Numeracy at home: Basic (3), Advanced (3)
- Numeracy at work: Basic (3), Advanced (3)
- Reading at home: Basic (3), Advanced (5); Documents (4), Prose (4)
- Reading at work: Basic (3), Advanced (5); Documents (4), Prose (4)

20.5.4 Results

In an effort to provide only scale-based derived variables that meet a sufficient level of psychometric quality, all proposed scales were analyzed first for each of the participating countries separately, and then jointly for consistency across countries. While scales with two items are viewed with well-grounded concern (Eisinga, Te Grotenhuis, & Pelzer, 2013), they were included in this first round of analyses in order to ensure that all of the proposed scales would be checked as requested.

20.5.4.1 Scale reliabilities

Table 20.3 presents the mean, standard deviation, minimum, and maximum of the country-level reliabilities for each potential scale. The mean reliabilities ranged from 0.50 to 0.84 for the non-nested scales and 0.50 to 0.78 for the nested scales. Using the criterion of alpha values less than 0.6, three non-nested scales and four nested scales were flagged for possible exclusion. Three of these scales had mean alpha values substantively below 0.6: physical (r = 0.49), reading at home: basic (r = 0.50), and writing at home (r = 0.51). The other four scales had mean reliabilities at or slightly below 0.6: cooperation (r = 0.59), reading documents at home (r = 0.60), reading prose at home (r = 0.58), and reading documents at work (r = 0.60).

Table 20.3: Reliability summary statistics for potential subscales

		Mean	SD	Min	Max
Non-Nested Scales	Cooperation	**0.59**	0.07	0.48	0.70
	ICT at home	0.69	0.03	0.64	0.76
	ICT at work	0.77	0.02	0.73	0.81
	Influence	0.79	0.02	0.74	0.82
	Learning at work	0.69	0.05	0.59	0.80
	Numeracy at home	0.77	0.03	0.72	0.82
	Numeracy at work	0.81	0.02	0.77	0.84
	Physical	**0.49**	0.22	-0.26	0.71
	Planning	0.71	0.04	0.62	0.77
	Problem solving	0.68	0.04	0.57	0.74
	Reading at home	0.73	0.04	0.66	0.80
	Reading at work	0.81	0.03	0.75	0.85
	Readiness to learn	0.84	0.03	0.80	0.91
	Self organization	0.79	0.08	0.53	0.88
	Task discretion	0.80	0.04	0.73	0.92
	Trust	0.66	0.07	0.46	0.80
	Writing at home	**0.50**	0.05	0.37	0.63
	Writing at work	0.62	0.06	0.51	0.77
Nested Scales	Numeracy at home: Basic	0.68	0.04	0.61	0.73
	Numeracy at home: Adv	0.72	0.08	0.56	0.81
	Numeracy at work Basic	0.79	0.04	0.67	0.84
	Numeracy at work Adv	0.68	0.07	0.53	0.76
	Reading at home: Basic	**0.50**	0.09	0.36	0.67
	Reading at home: Adv	0.62	0.04	0.56	0.69
	Reading at home: Docs	**0.60**	0.04	0.53	0.66
	Reading at home: Prose	**0.58**	0.07	0.46	0.71
	Reading at work: Basic	0.68	0.05	0.57	0.77
	Reading at work: Adv	0.67	0.04	0.59	0.76
	Reading at work: Docs	**0.60**	0.05	0.54	0.69
	Reading at work: Prose	0.78	0.03	0.70	0.83

20.5.4.2 Scale correlations

Table 20.4 presents the mean raw-score correlation between the potential subscales. Using the criterion of correlations greater than or equal to 0.7, there are three sets of scales that appear to provide redundant information. These sets correspond primarily to the non-nested scales. The only exception is for the subscales for *self-organization* and *planning,* which were strongly correlated across all countries (mean r = 0.91).

The scales for *reading skills at home* for both document and prose type texts, and the scales for basic and advanced literacy skills at home (i.e., the nested scales for reading at home) generally had high moderate to high correlations across all countries (range of mean correlations: 0.79 – 0.93). Similarly, the scales for *reading skills at work* for both document and prose type texts, and the scales for basic and advanced literacy skills at work (i.e., the nested scales for reading at work) generally had moderate to high correlations across all countries (range of mean correlations: 0.70 – 0.94).

The subscale for *numeracy at home* was strongly correlated with both basic and advanced numeracy at home across all countries (range of mean correlations: 0.84 – 0.91), yet basic and advanced numeracy at home were only moderately correlated (r = 0.53). The subscale for *numeracy at work* was strongly correlated with both basic and advanced numeracy at work across all countries (range of mean correlations: 0.83 – 0.92), but basic and advanced numeracy at work were only moderately correlated (r = 0.56).

Table 20.4: Subscale correlations, averaged across countries

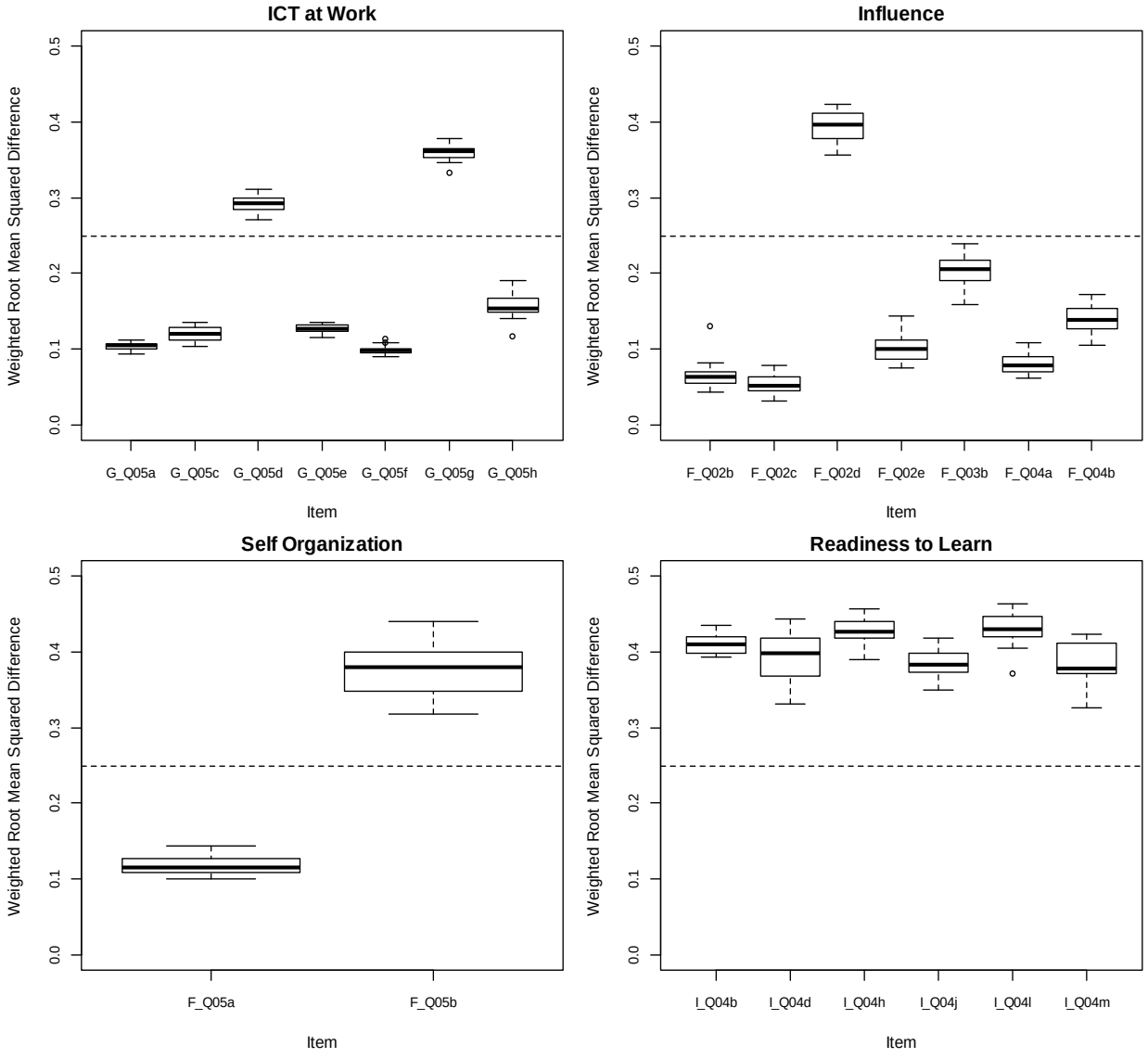
	COOPERATION	ICTHOME	ICTWORK	INFLUENCE	LERNATWORK	NUMHOME	NUMHOMEADV	NUMHOMEBAS	NUMWORK	NUMWORKADV	NUMWORKBAS	PHYSICAL	PLANNING	PROBWORK	READHOME	READHOMEADV	READHOMEBAS	READHOMEDOC	READHOMEPRO	READWORK	READWORKADV	READWORKBAS	READWORKDOC	READWORKPRO	READYTOLERN	SELFORGANISE	TASKDISC	TRUST	WRITHOME
ICTHOME	0.05																												
ICTWORK	0.02	0.37																											
INFLUENCE	0.25	0.18	0.34																										
LERNATWORK	0.27	0.17	0.21	0.36																									
NUMHOME	0.03	0.45	0.19	0.18	0.15																								
NUMHOMEADV	0.02	0.41	0.17	0.11	0.10	0.84																							
NUMHOMEBAS	0.04	0.38	0.17	0.18	0.15	0.91	0.53																						
NUMWORK	0.08	0.24	0.54	0.48	0.26	0.35	0.28	0.32																					
NUMWORKADV	0.09	0.26	0.50	0.37	0.22	0.36	0.36	0.28	0.83																				
NUMWORKBAS	0.40	0.18	0.44	0.47	0.24	0.27	0.17	0.28	0.92	0.56																			
PHYSICAL	0.10	-0.12	-0.32	-0.10	-0.02	-0.04	-0.06	-0.02	-0.21	-0.21	-0.17																		
PLANNING	0.14	0.11	0.36	0.62	0.25	0.12	0.06	0.13	0.39	0.33	0.36	-0.09																	
PROBWORK	0.21	0.18	0.35	0.47	0.37	0.17	0.11	0.17	0.39	0.35	0.35	-0.11	0.41																
READHOME	0.07	0.51	0.28	0.31	0.23	0.51	0.39	0.49	0.32	0.31	0.26	-0.11	0.23	0.28															
READHOMEADV	0.05	0.44	0.26	0.30	0.21	0.49	0.39	0.47	0.30	0.31	0.24	-0.10	0.23	0.26	0.93														
READHOMEBAS	0.08	0.45	0.22	0.24	0.20	0.40	0.28	0.40	0.25	0.23	0.21	-0.08	0.18	0.23	0.83	0.57													
READHOMEDOC	0.06	0.42	0.21	0.25	0.20	0.53	0.40	0.51	0.30	0.29	0.24	-0.04	0.19	0.24	0.86	0.83	0.67												
READHOMEPRO	0.06	0.45	0.26	0.29	0.21	0.38	0.29	0.37	0.26	0.26	0.22	-0.15	0.22	0.25	0.89	0.80	0.79	0.54											
READWORK	0.14	0.26	0.61	0.58	0.40	0.24	0.16	0.24	0.61	0.54	0.54	-0.23	0.50	0.51	0.46	0.43	0.37	0.38	0.41										
READWORKADV	0.10	0.25	0.55	0.55	0.37	0.25	0.18	0.25	0.61	0.55	0.53	-0.19	0.47	0.47	0.44	0.44	0.33	0.39	0.39	0.94									
READWORKBAS	0.16	0.21	0.55	0.52	0.37	0.18	0.10	0.19	0.50	0.44	0.45	-0.23	0.45	0.47	0.39	0.35	0.35	0.31	0.37	0.90	0.70								
READWORKDOC	0.15	0.22	0.48	0.50	0.37	0.24	0.16	0.25	0.61	0.52	0.55	-0.13	0.42	0.47	0.39	0.37	0.31	0.38	0.31	0.89	0.88	0.74							
READWORKPRO	0.09	0.24	0.57	0.54	0.35	0.19	0.12	0.19	0.50	0.46	0.43	-0.27	0.47	0.45	0.43	0.41	0.35	0.32	0.43	0.91	0.82	0.87	0.62						
READYTOLERN	0.09	0.34	0.25	0.31	0.27	0.31	0.23	0.30	0.27	0.26	0.23	-0.08	0.25	0.28	0.44	0.42	0.35	0.36	0.40	0.34	0.33	0.29	0.28	0.32					
SELFORGANISE	0.05	0.09	0.34	0.44	0.21	0.10	0.04	0.12	0.35	0.28	0.32	-0.10	0.91	0.36	0.20	0.19	0.16	0.16	0.20	0.43	0.40	0.40	0.37	0.41	0.22				
TASKDISC	-0.09	0.12	0.32	0.30	0.13	0.11	0.09	0.10	0.32	0.23	0.32	-0.16	0.43	0.21	0.16	0.16	0.12	0.13	0.15	0.32	0.31	0.26	0.25	0.31	0.23	0.45			
TRUST	0.02	0.08	0.11	0.10	0.05	0.06	0.06	0.04	0.08	0.10	0.06	-0.14	0.10	0.09	0.15	0.15	0.10	0.09	0.16	0.14	0.13	0.13	0.08	0.17	0.07	0.09	0.08		
WRITHOME	0.05	0.57	0.26	0.25	0.19	0.48	0.41	0.43	0.25	0.25	0.20	-0.11	0.16	0.20	0.60	0.53	0.54	0.52	0.54	0.31	0.30	0.27	0.26	0.30	0.36	0.13	0.12	0.11	
WRITWORK	0.16	0.24	0.54	0.49	0.30	0.18	0.12	0.19	0.51	0.47	0.43	-0.24	0.43	0.45	0.36	0.33	0.30	0.28	0.34	0.69	0.61	0.67	0.60	0.64	0.28	0.37	0.22	0.13	0.33

20.5.5 Between-country differences

Figure 20.1 presents box-and-whiskers plots of the WRMSDs for each item for four of the potential subscales. These types of plots were used to visually identify potentially problematic items/subscales. Out of the full set of BQ items, there were 20 items with WRMSD values greater than 0.25 logits. These results point to items that function differentially across countries. In most instances, these were single items on a given scale; however, there were five cases where all, or the majority, of items associated with a given scale had WRMSD values greater than or equal to the criterion. These scales (with mean WRMSDs in the parentheses) include:

cooperation (0.29), *physical* (0.31), *problem solving* (0.25), *readiness to learn* (0.41) and *trust* (0.44). Most of these are two-item scales.

Figure 20.1: Subscale weighted root mean squared differences



20.5.6 Subscale retention determinations

Based on the results of these analyses, a decision was made to exclude all two-item and nested scales. A total of 13 subscales were retained. Each of the two-item scales was flagged for exclusion based on one or more criteria. *Cooperation* and *physical* were flagged as problematic both for low reliability and between country differences. *Problem solving* and *trust* were flagged as problematic due to between-country differences, and *self-organization* was strongly correlated with the three-item scale *planning*. All four of the nested scales were highly correlated with the corresponding non-nested scales, indicating that the nested subscales provided redundant information relative to the associated non-nested scales. The subscales for reading

documents/prose at home and work also had low reliabilities. In addition to the exclusion of these scales, two items were eliminated due to large between-country differences, G_Q05g and F_Q02d, on the *ICT at work* and *influence* scales respectively. In general, any subscale flagged for exclusion was removed from the set of reported scales; however, there were two scales that were retained in spite of the exclusion flag. The *writing at home* scale had a low reliability, but it was retained to maintain consistency with the reporting of at home/at work variables. The *readiness to learn* scale did have notable between country differences, but it was also fairly reliable (0.85).

The following scales were retained:

- ICT at home (7 items)
- ICT at work (7 items)
- Influence (7 items)
- Learning at work (3 items)
- Numeracy at home (6 items)
- Numeracy at work (6 items)
- Planning (3 items)
- Reading at home (8 items)
- Reading at work (8 items)
- Readiness to learn (6 items)
- Task discretion (4 items)
- Writing at home (4 items)
- Writing at work (4 items)

20.5.7 Summary of weighted likelihood estimates

For each of the retained subscales, there are a notable number of examinees with the lowest possible score (these are not always the same examinees). For the remainder of the examinees, the distributions of WLE are unimodal and appear to be approximately normal. Figure 20.2 illustrates this pattern for two of the subscales, *ICT at work* and *numeracy at work*. In these examples, if examinees do not use ICT and/or numeracy at work, there is little justification for providing scores on these subscales. As such, a decision was made to recode these values for each scale as missing. This decision is grounded in the fact that for many self-report scales of activities, zero-inflated counts are found for those respondents for which the questions are not applicable (e.g. Goodman, 1975; Dayton & Macready, 1980; Yamamoto, 1989; Loeys et al., 2012). Also note that for *ICT at home*, no such phenomenon is found as respondents were not given the *ICT at home* questions if they responded that they never used a computer before. Table 20.5 presents the means and standard deviations of each subscale when these estimates are included/excluded. By excluding these estimates in the computation of the descriptive statistics, the mean for each scale increases by 0.07 to 0.79 logits (mean = 0.33), while the standard deviations decrease by 0.06 to 0.78 logits (mean = -0.45).

Figure 20.2: WLE distributions

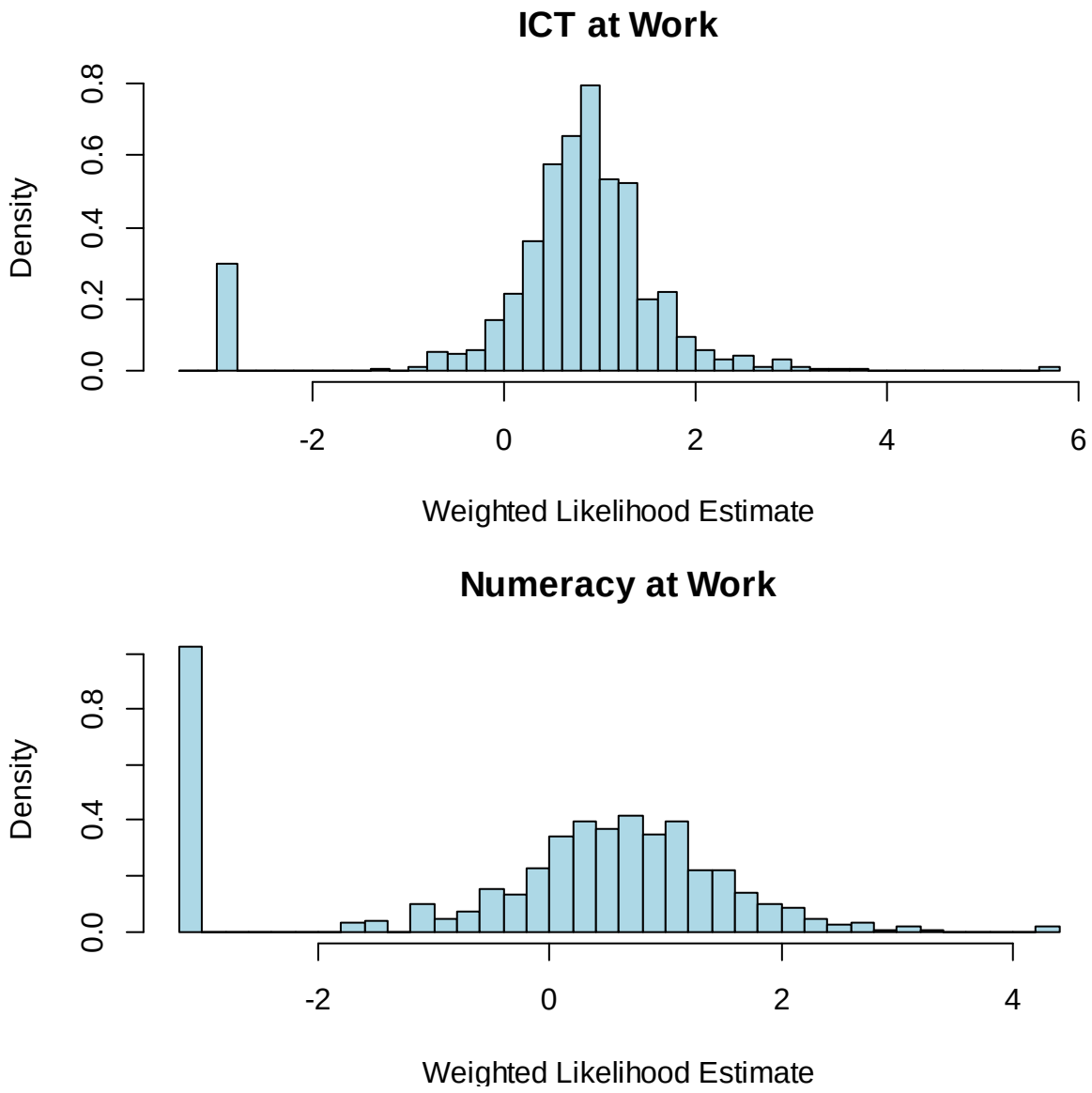


Table 20.5: WLE descriptive statistics

	Lowest Values Included		Lowest Values Excluded	
	Mean	SD	Mean	SD
ICT at Home	0.08	1.25	0.15	1.02
ICT at Work	0.15	2.00	0.54	1.37
Influence	-0.56	2.07	-0.06	1.55
Learning at Work	0.47	1.02	0.56	0.84
Numeracy at Home	-0.51	1.18	-0.21	0.76
Numeracy at Work	-0.97	1.85	-0.18	1.07
Planning	0.49	1.83	0.96	1.30
Reading at Home	0.05	1.53	0.15	1.24
Reading at Work	-0.67	2.42	-0.17	1.56
Readiness to Learn	0.81	0.58	0.83	0.52
Task Discretion	0.44	0.84	0.54	0.64
Writing at Home	-0.71	1.08	-0.43	0.69
Writing at Work	-1.00	2.01	-0.29	1.25

20.6 Item level nonresponse rates

Table 20.6 summarizes the mean proportion of missing responses across countries for all items on each scale. The standard deviation in the table indicates the variability in missing responses across countries. For most of the scales, approximately 20% to 30% of the responses were missing. With the exception of *reading at home* (SD = 19%), the variability in missing responses is fairly consistent across scales.

Table 20.6: Proportion of missing responses by scale

	Mean	SD
ICT at Home	0.01	0.03
ICT at Work	0.25	0.07
Influence	0.26	0.08
Learning at Work	0.34	0.09
Numeracy at Home	0.01	0.03
Numeracy at Work	0.25	0.07
Planning	0.26	0.07
Reading at Home	0.42	0.19
Reading at Work	0.26	0.07
Readiness to Learn	0.20	0.10
Task Discretion	0.33	0.08
Writing at Home	0.01	0.03
Writing at Work	0.25	0.07

References

Cramer, J. S. (2004). The early origins of the logit model. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 35*(4), 613-626.

Dayton, C. M., & Macready, G. B. (1980). A scaling model with response errors and intrinsically unscalable respondents. *Psychometrika, 45,* 343-356.

Goodman, L. A. (1975). A new model for scaling response patterns: An application of the quasi-independence concept. *Journal of the American Statistical Association, 70,* 755-768.

Eisinga, R., Te Grotenhuis, M., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach or Spearman-Brown? *International Journal of Public Health. 58*(4): 637-642. doi:10.1007/s00038-012-0416-3

Loeys, T., Moerkerke, B., De Smet, O., & Buysse, A. (2012). The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression. *British Journal of Mathematical and Statistical Psychology, 65,* 163-180. doi: 10.1111/j.2044-8317.2011.02031.x

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

Yamamoto, K. (1989). *HYBRID model of IRT and latent class models.* (Research Report No. RR-89-41). Princeton, NJ: Educational Testing Service.

Chapter 21: PIAAC Proficiency Scales

Claudia Tamassia and Mary Louise Lennon, ETS

21.1 Introduction

To adequately measure the skills of adults with differing educational backgrounds and life experiences, PIAAC includes tasks that range from very easy to very challenging. As described in Chapter 2, these tasks were developed to measure the range of skills and abilities defined in the frameworks for the three assessment domains – literacy, numeracy and problem solving in technology-rich environments (PSTRE). Results from the assessment are reported along three proficiency scales, each ranging from 0 to 500 with tasks at the lower end of the scale being easier than those at the higher end.

Reporting that one task falls at 215 on a scale while another falls at 345 provides some information – namely that the first task is easier than the second – but it does not tell us much about the underlying skills and knowledge each requires. To provide a richer report of the PIAAC results, described proficiency scales were developed for each of the domains, describing what performance at various points along those scales means. To create these described proficiency scales, the expert groups in each domain met with psychometricians and test developers to review the Main Study data, look at the tasks as they were distributed along the 500-point scales, and articulate how the requisite skills and knowledge to complete those tasks progressively increased along the scale. Defining clusters of tasks which required similar skills and knowledge and differentiating them from other clusters which were more or less difficult allowed the experts to define the levels of performance along the proficiency scale for each of the PIAAC domains.

21.2 Defining the proficiency levels

The IRT scaling procedures used in PIAAC constitute a statistical solution to the challenge of establishing a scale for a set of tasks with an order of difficulty that is essentially the same for everyone.

First, the response data collected from each participating country was used to estimate item parameters for each scale using a particular IRT model. In PIAAC, a two-parameter model was used that models the probability of a response based on the difficulty of an item and how well it discriminates, in combination with the person's ability or proficiency. This information was summarized in the form of item characteristic curves which show the probability of successfully completing an item at a given level of ability. Next, item parameters along with other information were used to estimate the ability distributions for each participating country along a scale with an overall mean and standard deviation. This scale can then be used to compare the

overall performance of countries or subgroups within a country. It can also be used to compare performance along the scale based on statistical criteria such as percentiles.

The IRT analysis summarizes how well the sample of individuals who responded to the pool of tasks performed. The tasks in this pool constitute a sample of the universe or "population" of tasks representing the construct that is measured (in the case of PIAAC, literacy, numeracy and PSTRE as defined by the relevant framework documents). Thus, the goal is to make inferences concerning the proficiency of respondents with respect to the population of tasks that represent the construct – that is, to make inferences about how well respondents performed on items used in the assessment as well as items having similar characteristics that also represent the construct but were not included in this particular assessment. As the items used in the survey represent a sample of tasks, it is important that any description of skills closely align to the framework used to define and construct them.

The use of IRT makes it possible not only to summarize results for various subpopulations of adults but also determine the relative difficulty of the tasks. In other words, just as individuals receive a specific score along a scale according to their performance on the assessment tasks, each task receives a specific value on a scale according to its difficulty, as determined by the performance of adults across the various countries that participated in the assessment (Kirsch et al., 2002). As tasks used in PIAAC vary widely in terms of task requirements and levels of complexity, it is possible to capture the range of difficulty of task through an item map which places items along a scale based on a selected response probability.[1]

Test items do not discriminate perfectly and each person has a chance (however small) of responding correctly to any given item. Consequently, a value representing the probability of correctly responding to an item must be selected in order to place an item on a proficiency scale. In theory, any value greater than zero and less than one can be chosen to place items on a proficiency scale, and a range of RP values are used in large-scale assessments. A value of 0.62 is used in PISA (OECD, 2009). Trends in International Mathematics and Science Study (TIMSS) uses different values for constructed responses (0.50) and multiple choice items (0.65) (TIMSS, 2007). The US National Assessment of Educational Progress (NAEP) uses an RP of 0.74 for multiple-choice items and 0.65 for open-ended items (National Center for Education Statistics, 2011). The IALS and ALL surveys used an RP of 0.80. The US National Assessment of Adult Literacy (NAAL) used an RP of 0.80 in reporting its 1992 survey and 0.67 in reporting results from its 2002 survey (Hauser, Edler, Koenig, & Elliott, 2005).

In PIAAC, the OECD Secretariat and participating countries agreed on an RP value of 0.67, similar to the approach used in PISA, to ensure that the description of what it means to be performing at a particular level of proficiency is consistent between the two surveys. There are potential risks for the credibility of both studies if being at a particular level of proficiency meant something different in each survey. While the RP value used in PIAAC and PISA will not be identical,[2] the interpretation of what it means to be at a level of proficiency will be the same.

[1] The RP section of this chapter was based on a PIAAC BPC document, Proficiency Levels in PIAAC [Doc. Ref.: COM/DELSA/EDU/PIAAC(2011)14], and written by Irwin Kirsch and Kentaro Yamamoto.

[2] This is a result of the different widths of the proficiency bands used.

Within any given scale, except for those at the lowest level, a person would be expected to pass a test made up of items from the level at which he or she performed. For example, using RP67, a person at the bottom of Level 3 on the literacy scale would be expected to successfully complete items of Level 3 difficulty approximately 50 percent of the time, a person at the top of the level would be expected get such items correct around 80 percent of the time, and a person at the middle of the level would do so 67 percent of the time. The probability of success on Level 3 items of persons at the top, bottom and middle of Level 3 based on RP80 is approximately 60, 80 and 90 percent, respectively. It is important to note that for both RP values, a person at the middle of a level would be likely to get most items at a lower level correct as well as a reasonable proportion of items at the next highest level. It is also important to note that the selection of a response probability is independent from the estimation of both item parameters and ability. The choice of an RP value has no impact on either the statistical characteristics of the items or the estimation of ability along the scale. In addition, the precision of measurement along a scale is not affected by the RP value. The same items define the underlying scale regardless of which RP value is selected.

As RP80 was used in IALS and ALL, in order to ensure that countries that wish to do so can map the change from RP80 to RP67, the OECD Secretariat provided item maps for literacy and numeracy under both the PISA approach (RP67) and the RP80 assumption in an appendix to the international report.

21.3 Interpreting the proficiency levels

As explained in the previous section, the proficiency scales range from 0 to 500 and are designed so the scores represent degrees of proficiency in a particular aspect of the domain. There are easier and harder tasks for each proficiency scale.[3] Each scale is divided into proficiency levels based on the knowledge and skills required to complete the tasks within those levels.

The purpose of described proficiency scales is to facilitate the interpretation of the scores assigned to respondents. That is, respondents at a particular level not only demonstrate knowledge and skills associated with that level but also the proficiencies required at lower levels. Thus, respondents scoring at Level 2 are also proficient at Level 1, with all respondents expected to answer at least half of the items at that level correctly.

The PIAAC proficiency scales and item descriptions were part of the work done by the PIAAC Expert Groups in December 2012 and January 2013. For a complete list of experts in these groups, please see Appendix 6.

21.3.1 Literacy

As described in Chapter 2 of this report, the PIAAC literacy items were developed and selected to represent three major aspects of processing continuous and noncontinuous texts and documents: accessing and identifying, integrating and interpreting, and reflecting on and evaluating information.

- *Access and identify* tasks require the reader to locate information in a text or document. While some tasks can be relatively straightforward because the information requested in

[3] See Appendix 1 for the complete list of Main Study PIAAC items in each domain organized by difficulty.

the question matches clearly with information that is easily located in the text, not all tasks in this category are necessarily easy. Inferences may need to be made and rhetorical understanding may be required.

- *Integrate and interpret* tasks require the reader to relate different parts of the text to each other. Requiring respondents to compare and contrast, understand problems and solutions, and identify cause/effect relationships are examples of this task type. These relationships may be explicitly signaled (e.g., the text states that "the cause of X is Y") or may require the reader to make inferences. The text components to be related may be contiguous and therefore easier to locate and integrate or may be found in different paragraphs in the same text or in separate documents.

- *Evaluate and reflect* tasks require the reader to draw on knowledge, ideas or values external to the text. The reader must assess the relevance, credibility, argumentation and truthfulness of the information presented in the text within a context of information that is not present in the text. The reader may also evaluate the purposefulness, register, structure or reader awareness of the text, or the success with which the author uses evidence and language to argue or persuade. Tasks of this type were judged to be particularly important to include in the context of PIAAC's digital texts, where it is readers must be alert to a text's accuracy, reliability and timeliness.

The PIAAC literacy framework defined features of stimulus texts and tasks that were anticipated to impact the difficulty of tasks included in the assessment.[4] These included the following:

- transparency of information in the text as it relates to the presented task or question
- degree of complexity necessary to make required inferences
- semantic and syntactic complexity of the text and/or question
- amount of text that must be processed
- prominence of needed information in the text
- competing information in the text
- text features that facilitate or hinder understanding relationships among parts of the text

The literacy proficiency scale is defined in terms of six levels. In all, the literacy scale includes 58 tasks with that ranged in difficulty from an RP67 of 75 to 376. Those tasks are distributed by level as follows:

- Below Level 1 (1 – 175): 4 tasks
- Level 1 (176 – 225): 3 tasks

[4] For the full text of the PIAAC Literacy Framework, see Chapter 3 of OECD (2012).

- Level 2 (226 – 275): 15 tasks
- Level 3 (276 – 325): 24 tasks
- Level 4 (326 – 375): 11 tasks
- Level 5 (376 – 500): 1 task

Each of the six proficiency levels is defined below and one or more representative tasks are described to illustrate the key information-processing skills at each level.

Literacy Below Level 1 **0 to 175**

The tasks at this level require the respondent to read brief texts on familiar topics to locate a single piece of specific information. Only basic vocabulary knowledge is required, and the reader is not required to understand the structure of sentences or paragraphs or make use of other text features. There is seldom any competing information in the text and the requested information is identical in form to information in the question or directive. While the texts can be continuous, the information can be located as if the text were noncontinuous. Tasks below Level 1 do not make use of any features specific to digital texts.

SGIH (C301AC05)
Difficulty: 75
In this task, respondents are asked to identify a telephone number in a very short advertisement. The question explicitly refers to literal information in a simple text with little competing information. The information is prominently located on a single line in the advertisement, labeled by an abbreviation for the word "telephone." These features of the text and question combine to make this the easiest task on the PIAAC literacy scale.

Election Results (C302BC02)
Difficulty: 162
Respondents are asked to use a notice providing results from a union election to identify the candidate with the fewest number of votes. Although the notice contains several paragraphs of information, the respondent only needs to use a very short table with three numbers and associated names within the text to answer the question. The key word ("votes") appears in both the prompt and the text making the relevant information very transparent. There is no competing information as the word "votes" appears nowhere else in the text. To locate the answer, the respondent needs to compare the three numbers (the word "fewest" in the prompt indicates the answer will involve a number), and once that is determined, locate the name associated with that number.

Literacy Level 1 **176 to225**

Most of the tasks at this level require the respondent to read relatively short digital or print continuous, noncontinuous or mixed texts to locate a single piece of information which is identical to or synonymous with the information given in the question or directive. Some tasks may require the respondent to enter personal information into a document, in the case of some noncontinuous texts. Little, if any, competing information is present. Some tasks may require simple cycling through more than one piece of information. Knowledge and skill in recognizing basic vocabulary, evaluating the meaning of sentences, and reading of paragraph text is expected.

Dutch Women (C311B701)
Difficulty: 201
This task asks the respondent to find the percentage of women in the teaching profession in Greece based on a graphically presented table showing that information for 10 countries. There is a single instance of the word "Greece" in the stimulus and a single instance of a percentage associated with that word, making the task relatively simple. There are other percentages in the text that might serve as distractors or cause the respondent to misread the table, which makes this more difficult than the Below Level 1 tasks, but the explicit connection between the question wording and information in the stimulus makes this a relatively simple task.

Generic Medicine (C309A321)
Difficulty: 219
This stimulus consists of a short newspaper article focusing on the limited use of generic medicines in Switzerland. The article includes a simple two-column table showing the market share for generic medications in 15 countries. The Level 1 item associated with this stimulus asks the respondent to identify the number of countries where generic medicines account for more than 10% of drug sales. While the phrase "drug sales" does not appear in the text, the only place a list of countries and percentages appears is in the table included in the article. The phrase "market share" is in the title of this table and might be regarded as a synonym for "drug sales," but most respondents would not need this additional information. The respondent's task is then to simply count the number of percentages that are greater than 10%, a task made simpler as the percentages are ordered from large to small.

Literacy Level 2 **226 to 275**

At this level, the complexity of text increases. The medium of texts may be digital or printed, and texts may comprise continuous, noncontinuous or mixed types. Tasks in this level require respondents to make matches between the text and information, and may require paraphrase or low-level inferences. Some competing pieces of information may be present. Some tasks require the respondent to

- cycle through or integrate two or more pieces of information based on criteria,
- compare and contrast or reason about information requested in the question, or
- navigate within digital texts to access and identify information from various parts of a document.

Lakeside Fun Run (C322P002)
Difficulty: 240
This unit is based on a Web page with information about a community relay race and walking event. The tasks associated with the unit require some understanding of Web conventions. This task, the easiest in the unit, asks respondents to identify the link they would use to find the phone number for one of the event organizers. The correct response, a link labeled "Contact Us," is one of several on the home page of this digital text. While using this link might be apparent to respondents familiar with Web-based texts, less familiar respondents need to make some inferences in order to know where to navigate to find the information.

Generic Medicines (C310A406)
Difficulty: 272
This task uses the same stimulus as that described in Level 1 above but requires the respondent to use the text of the newspaper article. Here the respondent is asked to identify two reasons given in the text for the limited use of generic medicines. Previous research has shown that tasks requiring multiple responses tend to be more difficult as respondents must search through the text more than once. While the reasons are explicitly stated in the text, they are not specifically labeled as reasons. Respondents must make an inference based on a semantic cue in the text – the single word "Why?" which signals that reasons will follow. There are other instances of "reasons" in the text (such as why generic medicines are less expensive, signaled by the explicit "because") that might serve as distractors for less able respondents.

Literacy Level 3 **276 to 325**

Texts at this level are often dense or lengthy, including continuous, noncontinuous, mixed or multiple pages. Understanding text and rhetorical structures become more central to successfully completing tasks, especially in navigation of complex digital texts. Tasks require the respondent to identify, interpret or evaluate one or more pieces of information and often require varying levels of inferencing. Many tasks require the respondent construct meaning across larger chunks of text or perform multistep operations in order to identify and formulate responses. Often tasks also demand that the respondent disregard irrelevant or inappropriate text content to answer accurately. Competing information is often present, but it is not more prominent than the correct information.

Lakeside Fun Run (C322P001)
Difficulty: 283
This question in the "Lakeside Fun Run" unit asks the respondent to identify information in the Web page that explains how this year's race differs from last year's. Not only does the task require the respondent to understand a contrast – a more difficult semantic construct – but the contrast is only indirectly signaled in the text, which says, "The popular walk will continue, but this year…"

Lakeside Fun Run (C322P004)
Difficulty: 293
A more difficult task from the "Lakeside Fun Run" task requires the respondent to understand a common convention in digital texts – a FAQ (frequently asked questions) link – and be able to

use it to navigate through the text. The respondent is asked to identify the date by which a race participants must notify organizers they want to change their race distances. In order to find the requested information, the respondent must click on the FAQ link on the home page. Once the respondent has successfully navigated to the FAQ page, the information on the page is relatively easy to find, as there is a near synonymous match between the task statement and the text.

Literacy Level 4 **326 to 375**

Tasks at this level often require respondents to perform multiple-step operations to integrate, interpret, or synthesize information from complex or lengthy continuous, noncontinuous, mixed, or multiple type texts. Complex inferences and application of background knowledge may be needed to perform successfully. Many tasks require identifying and understanding one or more specific, noncentral ideas in the text in order to interpret or evaluate subtle evidence claim or persuasive discourse relationships. Conditional information is frequently present in tasks at this level and must be taken into consideration by the respondent. Competing information is present and sometimes seemingly as prominent as correct information.

Library Search (C323P004)
Difficulty: 329
The stimulus for this unit consists of two pages from a library website listing results for a search on "genetically modified food." This task asks the reader to find two books that argue against genetically modified foods, requiring the respondent to examine the brief descriptions of all the books and decide which best meet that criterion. The respondent must scroll through the full list, using both pages on the website, to make inferences and compare the descriptions in the 10 entries. As the task asks for two books, the respondent must cycle through the text twice to locate both responses.

Library Search (C323P002)
Difficulty: 348
The same "Library Search" unit includes another example of a Level 4 task that is harder than the task above. The task asks the respondent to find the single book that suggests that the claims both for and against genetically modified foods are unreliable. The information in the text that the respondent uses to find the answer is "manufactured propaganda," which the respondent has to infer is meant to be synonymous with the word "unreliable" that is in the prompt. The task requires the careful respondent to examine all the entries.

Literacy Level 5 **376 to 500**

At this level, tasks may require the respondent to search for and integrate information across multiple, dense texts; construct syntheses of similar and contrasting ideas or points of view; or evaluate evidence-based arguments. Application and evaluation of logical and conceptual models of ideas may be required to accomplish tasks. Evaluating reliability of evidentiary sources and selecting key information is frequently a key requirement. Tasks often require respondents to be aware of subtle, rhetorical cues and to make high-level inferences or use specialized background knowledge.

Library Search (C323P005)
Difficulty: 376
One of the most difficult literacy tasks in PIAAC is also associated with the "Library Search" unit. The respondent is asked to identify the book likely to be least useful in providing more information about genetically modified food. As mentioned in the framework, negative phasing is more complex than affirmative, so evaluating the 10 books in terms of which is *least* useful for the defined purpose is expected to be difficult. The fact that the correct selection is located at the end of the second page of results also increases the difficulty of the task. The respondent must read and evaluate each of the choices in order to make a correct selection.

21.3.2 Numeracy

The PIAAC numeracy framework includes a definition of the domain as well as a description of numerate behavior.[5] Numeracy tasks were developed to cover a range of difficulty as a result of combining variables that include:

- the kind and degree of interpretation and reflection required by the problem,
- the kind of representation skills required,
- the kind and level of mathematical skill required (e.g., single-step vs. multistep problems, or more advanced mathematical knowledge, complex decision making, and problem-solving and modeling skills),
- the kind and degree of mathematical argumentation required,
- the degree of familiarity with the context, and
- the extent to which tasks require reproduction of known procedures and steps or present novel situations requiring nonroutine and perhaps more creative responses.

The numeracy proficiency scale is defined in terms of six levels and includes 56 tasks with difficulty values ranging from 129 to 375. Based on RP67, these tasks are distributed by level as follows:

- Below Level 1 (1 – 175): 3 tasks
- Level 1 (176 – 225): 6 tasks
- Level 2 (226 – 275): 21 tasks
- Level 3 (276 – 325): 20 tasks
- Level 4 (326 – 375): 6 tasks

[5] For the full text of the PIAAC Numeracy Framework, see Chapter 4 of OECD (2012).

- Level 5 (376 – 500): 0 tasks

Each of the six proficiency levels is defined below and one or more representative tasks are described to illustrate the key skills and knowledge at each level.

Numeracy Below Level 1 **0 to 175**

Tasks at this level are set in concrete, familiar contexts where the mathematical content is explicit with little or no text or distractors and that require only simple processes such as counting, sorting, performing basic arithmetic operations with whole numbers or money, or recognizing common spatial representations.

Bottles (C601AC06)
Difficulty: 129

The easiest task on the numeracy scale, with difficulty level of 129, requires respondents to look at a photograph containing two cases of water bottles. They are asked to find the total number of bottles in the two full cases being shown. Part of what makes this task easy is that content is drawn from everyday life and objects of this kind are relatively familiar to most people. Second, what respondents are asked to do is apparent and explicit – this task uses a photograph depicting concrete objects and containing no text to be read. A third contributing factor is that respondents can approach the task in a variety of ways that differ in sophistication, such as by multiplying rows and columns, but also by simple counting. This task requires that adults make a conjecture using spatial visualization because the full set of bottles in the lower case is not visible, but as can be seen from the low difficulty level of the task, this feature did not present a problem for the vast majority of adults in participating countries.

Numeracy Level 1 **176 to225**

Tasks in this level require the respondent to carry out basic mathematical processes in common, concrete contexts where the mathematical content is explicit with little text and minimal distractors. Tasks usually require simple one-step or two-step processes involving, for example, performing basic arithmetic operations; understanding simple percents such as 50%; or locating, identifying and using elements of simple or common graphical or spatial representations.

Tea Candles (C615A602)
Difficulty: 221

An example of a Level 1 task is Tea Candles Q1. The stimulus for this item consists of a photo of a box containing tea light candles. The packaging identifies the product (tea light candles), the number of candles in the box (105 candles) and its weight. While the packaging partially covers the top layer of candles, it can be seen that the candles are packed in five rows of seven candles each. The instructions inform the respondent that there are 105 candles in a box and asks him or her to calculate how many layers of tea candles are packed in the box.

Numeracy Level 2 **226 to 275**

Tasks in this level require the respondent to identify and act upon mathematical information and ideas embedded in a range of common contexts where the mathematical content is fairly explicit or visual with relatively few distractors. Tasks tend to require the application of two or more steps or processes involving, for example, calculation with whole numbers and common decimals, percents and fractions; simple measurement and spatial representation; estimation; and interpretation of relatively simple data and statistics in texts, tables and graphs.

Gas Gauge (C604A505)
Difficulty: 228

This is a somewhat more complex numeracy task falling in the lower end of Level 2. A gauge is presented that has three lines or ticks on it: one showing an "F," one showing an "E" and one in the middle of the others. A line on the gauge, representing the gauge's needle, shows a level that is roughly halfway between the middle tick and the tick indicating "F," suggesting that the tank is about three-quarters full. The task states that the tank holds 48 gallons and asks the respondent to determine "how many gallons remain in the tank." This task is drawn from an everyday context and requires an adult to interpret a display that conveys quantitative information but carries virtually no text or numbers. No mathematical information is present other than what is given in the question. What makes this task more difficult than the previous ones is that adults must first estimate the level of gas remaining in the tank by converting the placement of the needle to a fraction. Then they need to determine how many gallons this represents from the 48-gallon capacity stated in the question. Thus, this task requires adults to apply multiple operations or procedures to arrive at a correct response without specifying what the operations may be. Nonetheless, this task, like many everyday numeracy tasks, does not require an exact computation but allows an approximation that should fall within reasonable boundaries.

Cooper Test (C601AC06)
Difficulty: 234

This Level 2 item engages the respondent with moderately complex tables of numerical and textual data relating to a common measure of physical fitness – the Cooper Test – from which they have to read off the level of fitness of a 43-year-old male who runs 1,100 meters in 12 minutes. This task is drawn from everyday life and involves interpreting the headings and numerical information in the table correctly in order to locate the 40-49 age table row and the appropriate cell in this row for a male who runs 1,100 meters in the requisite 12 minutes. There is no calculation involved, but number bands for both age and distance need to be understood. However, it is a type of task many adults, particularly those who use the Internet regularly, would have experienced.

Numeracy Level 3 **276 to 325**

Tasks in this level require the respondent to understand mathematical information which may be less explicit, embedded in contexts that are not always familiar, and represented in more complex ways. Tasks require several steps and may involve the choice of problem-solving strategies and relevant processes. Tasks tend to require the application of, for example, number sense and spatial sense; recognizing and working with mathematical relationships, patterns, and proportions expressed in verbal or numerical form; and interpretation and basic analysis of data and statistics in texts, tables and graphs.

Tiles (C619A609)
Difficulty: 282

This Level 3 item presents the respondent with a plan of a kitchen floor to be tiled with nine of the proposed square tiles placed in a corner, with the plan drawn on a squared grid. It asks the respondent to use this information to find out how many tiles are needed to cover the entire floor. The task is a familiar one drawn from everyday life and, using the most obvious method an adult would choose, would require several operations to arrive at the correct answer. First, the area in terms of the number of larger grid squares in the kitchen floor plan is calculated by counting or otherwise. Then the number of tiles in each larger square is calculated by counting or multiplication. The last step involves multiplying the number of larger squares by the number of tiles per larger square to get the total number of tiles required to cover the kitchen floor. Respondents need to use their spatial reasoning ability in organizing the information in the first two steps in this task. The task could also be done using a combination of spatial visualization and counting all the small squares (tiles), but this method would be more prone to error.

Orchestra Tickets (C664P001)
Difficulty: 307

This task has a difficulty around the middle of Level 3. It presents the respondent with a table of numerical data on ticket price categories for single and multiple events (Season Ticket). The respondent has to discern the pattern in the data and identify the formula, probably in verbal or numerical terms (e.g., multiply by 4½), for calculating the cost of a season ticket from the cost of a single ticket for different seating categories to an event, and use it to calculate the cost of a season ticket for a new entry category – a student season ticket. The task requires adults to use a range of reasoning strategies, including algebraic reasoning (i.e., reasoning with variables and generalizing from specific values) and computational procedures.

Numeracy Level 4 **326 to 375**

Tasks in this level require the respondent to understand a broad range of mathematical information that may be complex, abstract or embedded in unfamiliar contexts. These tasks involve undertaking multiple steps and choosing relevant problem-solving strategies and processes. Tasks tend to require analysis and more complex reasoning about, for example, quantities and data; statistics and chance; spatial relationships; change; proportions; and formulas. Tasks in this level may also require comprehending arguments or communicating well-reasoned explanations for answers or choices.

Cooper Test (C665P002)
Difficulty: 326

This task is based on the same stimulus as the Level 2 task described above but was considerably more difficult for adults in participating countries. It requires respondents to go beyond interpreting the information in the tables to calculate the percent increase needed in the distance run by a female in 12 minutes for her fitness level to be in the "Good" category. To arrive at a correct response, respondents have to locate the "Good" band for a 27-year-old female and use the difference between the runner's current 12-minute distance and the minimum distance for the "Good" band to calculate the percent increase in distance run by her to qualify for that band. There is considerable use of reasoning and knowledge and understanding of percentages in carrying out this task.

Compound Interest (P610A515)
Difficulty: 348

This is the third most difficult task in the PIAAC numeracy assessment. It presents respondents with an advertisement claiming it is possible for an investor to double an amount invested in seven years, based on a 10 percent fixed interest rate each year. Adults are asked if it is possible to double $1,000 invested at this rate after seven years and have to support their answer with their calculations. A range of responses was accepted as correct as long as a reasonable justification was provided, with relevant computations. Respondents were free to perform the calculation any way they wanted, but they could use a "financial hint," which accompanied the advertisement and presented a formula for estimating the worth of an investment after a specified number of years. Those who used the formula had to enter information stated in the text into variables in the formula (principal, interest rate and time period) and then perform the needed computations and compare the result to the expected amount if $1,000 is doubled. All respondents could use a handheld calculator provided as part of the assessment.

This task proved difficult because it involved percents, and the computation, with or without the formula, required the integration of several steps and several types of operations. Performing the computations without the formula required understanding of compound interest procedures. This task required adults to use a range of reasoning strategies, including algebraic reasoning and informal or invented procedures. It also required the use of formal mathematical information and deeper understanding of nonroutine computational procedures, all of which may not be familiar or accessible to many adults.

Numeracy Level 5 **376 to 500**

Tasks in this level require the respondent to understand complex representations and abstract and formal mathematical and statistical ideas, possibly embedded in complex texts. Respondents may have to integrate multiple types of mathematical information where considerable translation or interpretation is required; draw inferences; develop or work with mathematical arguments or models; and justify, evaluate and critically reflect upon solutions or choices.

21.3.3 Problem solving in technology-rich environments

The PSTRE domain is organized around three core dimensions: the cognitive strategies and processes a person uses to solve a problem, the tasks or problem statements that trigger and condition problem solving, and the technologies through which the problem solving is conducted. Variations within and across all of those dimensions were expected to contribute to the overall difficulty of the problems presented in the PIAAC assessment. For example, a problem is likely to be more complex if it is ill-defined as opposed to explicitly stated, if it requires complex problem solving strategies such as defining goals and resolving impasses, and/or if it requires the use of multiple technology environments (e.g., respondents must utilize both emails and spreadsheets).

In order to explain how proficiency can be affected by the three dimensions of PSTRE, the problem-solving proficiency scale was divided into three levels as shown below. In this section, we describe the essential features of tasks at each of these three levels.

Table 21.1: Technology, task and cognitive characteristics of problems at each of three main levels of proficiency

Level	Technology features	Task features	Cognitive processes
Level 1	• Generic applications • Little or no navigation required • Relevant information is directly available • Use of facilitating tools not required	• Few steps • Single operators	• Reach a given goal • Apply explicit criteria • Minimal monitoring demands • Simple relevance match • Categorical reasoning • No integrate or transformation
Level 2	• Both generic and novel applications (e.g., Web-based services) • Some navigation required to acquire information or perform actions • Use of tools facilitates operations	• Multiple steps • Multiple operators	• Goal may need to be defined • Apply explicit criteria • Generally higher monitoring demands • Generally involves resolving impasses • Some evaluation of relevance • Some integrate or transformation • Inferential reasoning
Level 3	• Generic and novel applications • Some navigation required to acquire information or perform actions • Use of tools required to efficiently solve the problem	• Multiple steps • Multiple operators	• Goal may need to be defined • Establish and apply criteria • Generally high monitoring • High inferential reasoning and integration • Evaluate relevance and reliability • Generally involves resolving impasses

The proficiency levels of PSTRE are defined as follows:

PSTRE Below Level 1 **0 to240**

Tasks are based on well-defined problems involving the use of only one function within a generic interface to meet one explicit criterion without any categorical, inferential reasoning or transforming of information. Few steps are required and no subgoal has to be generated.

Though the current set of tasks included very simple problems, none of those fell within the Below Level 1 category. The simplest item on the assessment had an RP67 of 268. The expert group did, however, consider the characteristics of tasks that might fall at this level. Based on the PSTRE framework (OECD, 2012), such problems would have the following characteristics. They would be well-defined problems involving the use of only one function on a generic interface to meet one explicit criterion without any categorical, inferential reasoning or transforming of information. Few steps would be required and no subgoal would have to be generated. PSTRE problems at this level would still differ from simple ICT literacy in that the goal would extend beyond the mere use of ICT functions and commands. Thus, respondents would still need to implement a set of actions aimed at solving the problem through the use of technology.

It should be noted that more than a quarter of the PIAAC participants were excluded from the PSTRE survey because they reported no prior experience using computers, they were not willing to take the survey on a computer, or they were not able to demonstrate the basic ICT skills required to complete the assessment such as clicking, highlighting and simple typing. This proportion is likely to decrease in future surveys, as more and more people become familiar with using computers and other digital devices such as smartphones and tablets. It is likely that future assessment would include a larger percentage of the total population, most of which would likely display modest levels of proficiency. Therefore, in future assessments it will become increasingly important to include easier tasks to better describe in more detail the lower end of the proficiency scale

PSTRE Level 1 **241 to 290**

At this level, tasks typically require the use of widely available and familiar technology applications, such as email software or a Web browser. There is little or no navigation required to access the information or commands required to solve the problem. The problem may be solved regardless of one's awareness and use of specific tools and functions (e.g., a sort function). The task involves few steps and a minimal number of operators. At a cognitive level, the person can readily infer the goal from the task statement; problem resolution requires one to apply explicit criteria; there are few monitoring demands (e.g., the person does not have to check whether he or she has used the adequate procedure or made progress toward the solution). Identifying contents and operators can be done through simple match; only simple forms of reasoning, for example, assigning items to categories are required. There is no need to contrast or integrate information.

Party Invitations (U01A)
Difficulty: 286

This task presents a problem where respondents are asked to organize a set of email responses they had received in response to a party invitation. The necessary folders are present in the email environment; respondents need to sort a set of emails into those existing folders. The email interface is presented with five emails in an inbox and the respondent is asked to organize the responses to keep track of who can and cannot attend the party. In terms of the three PSTRE dimensions, the item requires the respondent to categorize a small number of messages in an email application in existing folders according to a single criterion. This is typical of a Level 1 item because the goal is explicitly stated in operational terms, the task is performed in a single environment, and it can be solved in a relatively small number of steps using a restricted range of operators. Thus, the task does not require the user to learn a novel environment, nor does it necessitate a significant amount of monitoring across a large number of actions.

PSTRE Level 2 **291 to 340**

At this level, tasks typically require the use of both generic and more specific technology applications. For instance, the person may have to make use of a novel online form. Some navigation across pages and applications is required to solve the problem. The use of tools (e.g., a sort function) can facilitate the resolution of the problem. The task may involve multiple steps and operators. In terms of cognitive processing, the problem goal may have to be defined by the person, though the criteria to be met are explicit. There are higher monitoring demands. Some unexpected outcomes or impasses may appear. The task may require evaluating the relevance of a set of items to discard distractors. Some integration and inferential reasoning may be needed.

Club Membership (U19B)
Difficulty: 296

This task consists of responding to an information request and demands locating information in a spreadsheet. Respondents must identify an undefined number of members of a biking club who meet the provided eligibility requirements to serve as club president. The information can most efficiently be located within the long spreadsheet by using a sort function. The respondent is presented with two environments: a word processor page containing information about the two conditions required for club presidents, and a database with 200 entries where the relevant information can be found. In terms of the three PSTRE dimensions, the item requires the respondent to organize large amounts of information in a multiple column spreadsheet using multiple explicit criteria and locate and mark relevant entries. This is typical of Level 2 because the task requires switching between two different applications and involves multiple steps and operators. It also requires some amount of monitoring. Making use of the available tools (e.g., the sort function) greatly facilitates the identification of the relevant entries.

PSTRE Level 3 **341 to 500**

At this level, tasks typically require the use of both generic and more specific technology applications. Some navigation across pages and applications is required to solve the problem. The use of tools (e.g., a sort function) is required to make progress toward the solution. The task may involve multiple steps and operators. In terms of cognitive processing, the problem goal may have to be defined by the person, and the criteria to be met may or may not be explicit. There are typically high monitoring demands. Unexpected outcomes and impasses are likely to occur. The task may require evaluating the relevance and the reliability of information in order to discard distractors. Integration and inferential reasoning may be needed to a large extent.

Meeting Rooms (U02)
Difficulty: 346

This task requires respondents to check a number of email requests regarding reservations for a meeting room on a particular date and schedule those reservations based on multiple constraints (including the number of rooms available and reservations already made). Impasses due to conflicting constraints have to be resolved by initiating a new subgoal, that is, issuing a standard message to decline one of the requests. Two environments are present: an email interface with a number of emails containing the requests for meeting dates and times, and a novel Web application that allows respondents to assign rooms to meetings at certain times. Upon discovering that one of the requests cannot be accommodated, the respondent has to use a specific command on the website in order to issue a standard message declining the request. In terms of the three PSTRE dimensions, the item requires the respondent to use information from a novel Web application and several email messages, establish and apply criteria to solve a scheduling problem where an impasse must be resolved, and communicate the outcome. This is typical of Level 3 as the task involves multiple applications, a large number of steps, a built-in impasse, and requires the respondent to discover and use ad hoc commands in a novel environment. The respondent has to set up and monitor the application of a plan in order to minimize the number of conflicts. Furthermore, the respondent has to transfer information from one application (email) to another (room reservation).

21.4 Final remarks

This chapter focused on described proficiency scales, an important reporting tool that enhances the understanding of what has been measured in large-scale surveys such as PIAAC and allows policymakers and other stakeholders to better interpret survey results. Each of the PIAAC expert groups reviewed the Main Study data and analyzed the characteristics of tasks that fell along the scale for each domain, defining proficiency levels and describing the cognitive skills and knowledge required at each level.

References

Hauser, R. M., Edler, C. F. Jr., Koenig, J. A., & Elliott, S. W. (Eds.) (2005). *Measuring literacy: performance levels for adults*. Retrieved from http://www.nap.edu/catalog.php?record_id=11267

Kirsch I., de Jong, J., Lafontaine, D., McQueen, J., Mendelovits, J., & Monseur, C. (2002). *Reading for change – performance and engagement across countries.* Retrieved from Organisation for Economic Co-operation and Development website: http://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/33690904.pdf

National Center for Education Statistics. (2011). *NAEP technical documentation*. Retrieved from NCES website: http://nces.ed.gov/nationsreportcard/tdw/analysis/describing_itemmapping.asp

Organisation for Economic Co-operation and Development. (2000). *Literacy in the information age – final report of the International Adult Literacy Survey*. Retrieved from OECD website: http://www.oecd.org/edu/skills-beyond-school/41529765.pdf

Organisation for Economic Co-operation and Development. (2009). *PISA 2006 technical report.* Retrieved from OECD website: http://www.oecd.org/pisa/pisaproducts/pisa2006/42025182.pdf

Organisation for Economic Co-operation and Development. (2011). *Proficiency levels in PIAAC,* (Report No. COM/DELSA/EDU/PIAAC(2011)14). Paris, France: Author.

Organisation for Economic Co-operation and Development. (2012). *Literacy, numeracy and problem solving in technology-rich environments – framework for the OECD Survey of Adult Skills*. Paris, France: Author.

Trends in International Mathematics and Science Study (TIMSS). (2007). TIMSS 2007 Technical report. Retrieved from TIMSS website: http://timss.bc.edu/timss2007/PDF/T07_TR_Chapter13.pdf

Chapter 22: Generating Results for PIAAC

Alfred Rogers and John Barone, ETS

22.1 Data processing and analysis

The ETS data analysis systems are set up to process the PIAAC data in both SPSS file format and "flat" file ASCII text format. It was therefore imperative for both sets of data files across all countries to be perfectly synchronized with respect to the currency and content of the constituent data fields.

SPSS data files are completely self-documented, containing variable labels, data value labels and missing value definitions in addition to the data. However, many of the scaling and analytic tools used by ETS required the input data to be represented in "flat" file ASCII text rectangular format, where each data field is in the same position on every record in the file. ETS developed a procedure that extracts the data from an SPSS file into an ASCII text file and also extracts the metadata (labels, formats, missing value definitions, etc.) into a proprietary XML data dictionary file. Any program or procedure that uses the ASCII data file must first process the XML dictionary file to map the contents of the data file onto the set of variables to be analyzed or processed.

22.2 Receipt processing

When the data files were received from the IEA-hosted secure FTP site, they were unzipped and placed in a date-tagged folder before transfer to the operational folder.

Many of the data variables in the survey component were long text responses that could not be reduced to numeric codes and needed to be retained in the database for future interpretation. These responses were usually encoded in the native language of each country and could contain extended ASCII codes (Unicode) to represent certain characters. When placed in an ASCII file, these codes corrupt the rectangular structure of the data file and cause errors in processing the data. Because these responses have no analytic utility, they were identified and stripped from the SPSS data files before transfer to the operational folder.

There were also a number of variables that needed to be created or derived from existing variables which ETS uses to identify or track the data through the analytic processes. Because these variables have no intrinsic value outside of these processes, they were not provided to IEA for the master database but were only generated and retained in ETS operational data files. An SPSS macro was implemented to create and add these variables to the SPSS data files as they were transferred to the operational folder.

After the SPSS files were transferred to the operational folder, the last step in the process was to produce the ASCII extract data file and its accompanying XML data dictionary file.

22.3 Updating/adding data

The results of the several analytic processes at ETS produced new variables (or new data for existing variables) that required merging into the operational data files for internal quality and consistency checking before addition to the master database at IEA. These various data sources included, but were not limited to, the following activities (which are described elsewhere in the documentation):

- the production of scale scores for the literacy, numeracy and problem solving in technology-rich environments (PSTRE) components
- the development of indices for the skill use categories
- the derivation and imputation of income variables as specified by ROA
- the creation of variables to be used for trend analyses with the IALS and ALL surveys

Some of these data came in ASCII files that first needed to be converted to SPSS files before merging, some were already in SPSS file format, and some were represented in SPSS macro code that had to be applied to the operational SPSS files to be created and saved as separate files.

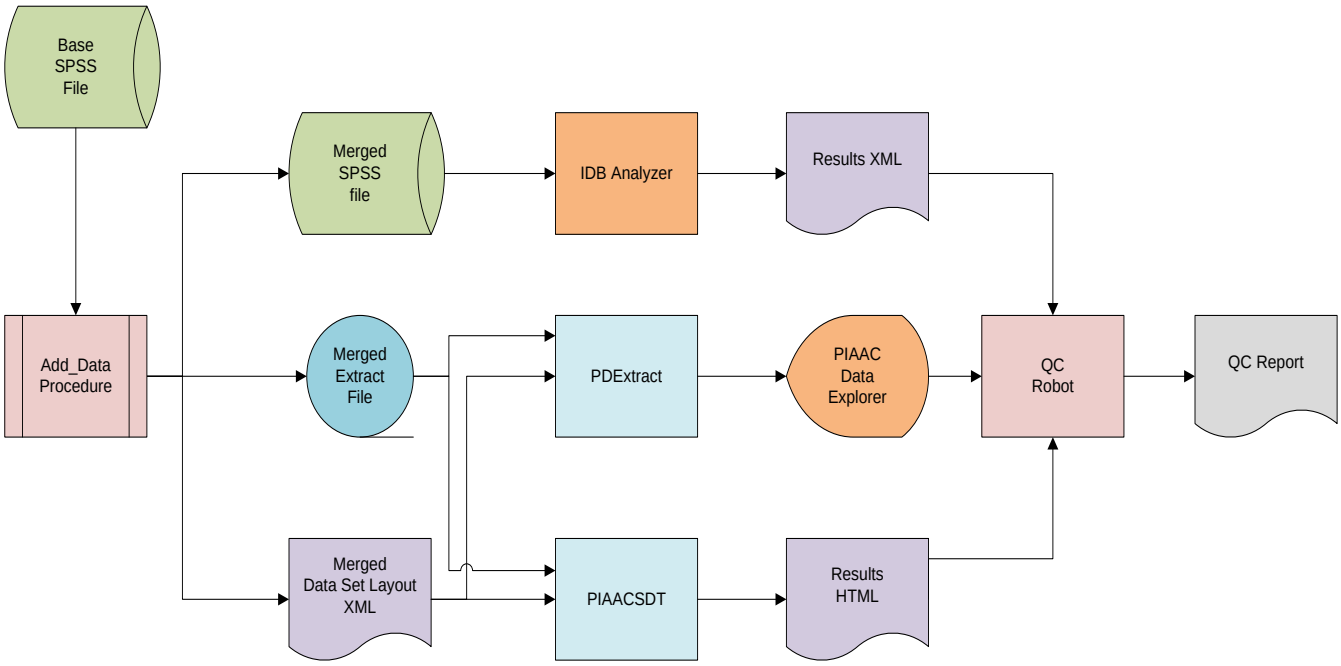
To efficiently, consistently and accurately perform these merging operations using a variety of input data sources, ETS developed a Python-based procedure that would iteratively process the data files for each country. Each application of the procedure only required as input a parameter file that specified the operations to be performed and the folder and file names for the input and output files.

The critical outputs for each application of this procedure were the SPSS file containing the new or updated variables, an SPSS file containing the merger of the operational file and the new or updated variables and the data and dictionary extracts of the merged SPSS file. Once these files were checked and approved by ETS, the SPSS file containing new or updated variables was sent to IEA for addition to or updating of the master database and the merge files became the new operational files.

22.4 Population and quality check of the PIAAC Data Explorer

The process to populate the PIAAC Data Explorer database and confirm the results it produces is summarized in Figure 22.1 below. For the purpose of explanation, consider that this process was applied separately to the data from each country.

Figure 22.1: PIAAC database population and quality control



The Base SPSS File contained the data as received from IEA/DPC and as forwarded to the appropriate country for its analysis and reporting.

The Add_Data procedure performed two functions. The first was conditional on whether a country provided supplemental data that was collected or derived and merged these data with the Base file. The second function created two files from the enhanced Base file: an ASCII text rectangular file containing the data values extracted from the Base file and an XML file containing information about the extracted data variables (location, format, labels). This Data Set Layout (DSL) XML is structured in a proprietary ETS schema.

The PDExtract program used the information from an input parameter file to process the data from the Extract file and metadata from the DSL file to produce a series of text files suitable for loading into the appropriate tables in the PIAAC Data Explorer (PDX) database. The program also produced a SQL script that is customized for performing the loading of these tables and contains a procedure for forming the data tables used by the PDX.

The PIAACSDT program also used the information from an input parameter file as well as a list of data variable names to calculate and produce summary data tables (SDT) – one analysis for each scale score. Each table in the analysis was a one-way tabulation of various statistics for each category of a given variable. The statistics pertained to a scale score and include percentage, average score and percentages within the benchmark levels. Each statistic was accompanied by the standard error estimate, degrees of freedom, number of cases on which the statistic is based and number of strata on which the standard error was based. All of these results

were stored in an HTML document in full precision. This document may be viewed with any of the popular Internet browsers when accompanied by the appropriate Cascading Style Sheet (CSS) document, which ETS provides. The document may also be parsed or translated to produce Excel workbooks and report quality tables, among others.

In the QC Robot procedure, the results HTML document from the PIAACSDT program were used to generate analysis requests for the PDX, one for each variable, and the results returned from the PDX were compared with those in the HTML document. The results of these comparisons are posted to the QC Report document where differences above specified criteria are flagged and subsequently examined.

The only statistics that can be reported in the PDX which cannot be calculated by the PIAACSDT program are the percentiles. Because the calculation of the percentiles within the PDX uses more resources than the other statistics, only a subset of critical variables was selected for quality-assurance analysis. The IEA IDB Analyzer reads data from the Base SPSS file, uses SPSS macros to calculate the desired percentile statistics, and writes the results to an XML file. The QC Robot procedure processed this XML file in the same way as the HTML file from the PIAACSDT program and added the comparison results to the QC Report file.

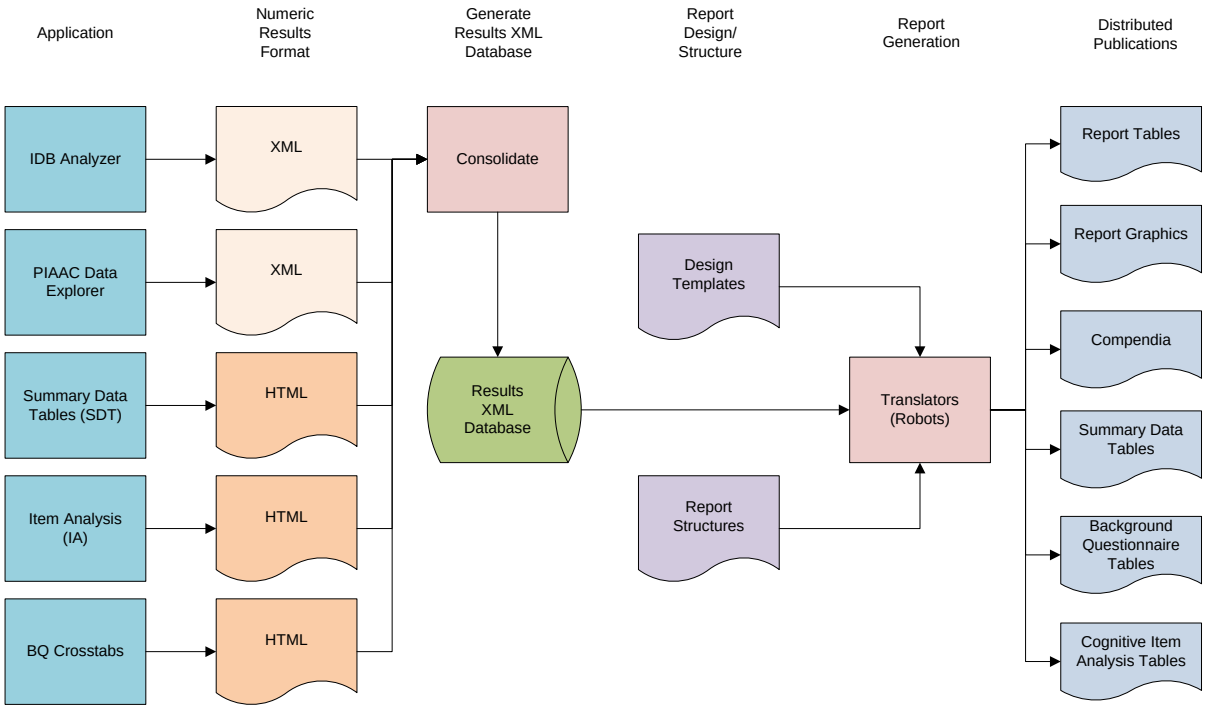
Prior to the first execution of the procedure described above, the IEA IDB Analyzer and the PIAACSDT programs were extensively calibrated with each other to ensure that the Merged SPSS and Merged Extract files were isomorphic and produced identical results for the statistics common to both programs.

22.5 Dynamic reporting system

The PIAAC dynamic report translation and publication system streamlined report and form generation by separating the data extraction and statistical computation process from the report design layouts and generation of publication formats. The generation process is shown in Figure 22.2. In the first stage, PIAAC-based Data Explorer, Data Analyzer, and procedural language applications performed data extraction and statistical computation across the entire PIAAC database and provided data files containing the numeric results for BQ and cognitive items in tagged language (XML, HTML) formats. In this stage, numeric computations were done once, and the numeric results were available to the second stage for efficient repurposing for quality control and report generation processing.

In the second stage, the publication system accepted report design and layout templates that were created with common desktop applications, and rendered XML structures based on those templates. Using these well-formed XML reporting structures, the system then applied XML-XSLT style sheet language or PYTHON-based scripts to transform the numeric results into viewable and publication-ready formats (PDF, RFT, Excel, HTML, and so on) for distribution. By increasing flexibility for rapid report generation customization through XML translation processing and the availability of a common numeric results archive, this two-stage phase approach to reporting dramatically reduced technical resource requirements and delivery times, enabling PIAAC to accommodate iterative data cleaning cycles while maintaining fixed publication delivery timelines.

Figure 22.2: PIAAC report translation and publication system



PIAAC BQ crosstabulations, summary data tables, item analyses, tables and graphical displays for possible inclusion in international and national reports, and compendia were generated by the Consortium using the PIAAC dynamic report translation and publication system. Following are descriptions and examples of each of these tables. All but the compendia are secure and not available for public view.

22.6 Summary data tables (SDT)

Via a secure FTP site, the Consortium delivered sets of files individually to each country containing summary data tables (SDT) that provided descriptive statistics for every categorical background variable in the respective country's PIAAC data file. For each country, the SDT included both international and idiosyncratic national background variables. The SDT were used by the Consortium and the countries for quality control and validation purposes: plausibility of 1) distributions of background characteristics and 2) performance results for groups, especially in the extent to which they agree with expectations or external/historical information.

For each variable, these tables contain weighted summary statistics, including variable identification, sample size, number of valid cases, weighted percentages of individuals corresponding to each valid response option, weighted percentages of individuals for whom none of the valid response options were selected, and within each categorical cell, the average score on one of the three PIAAC scale score domains. Standard errors were also included where applicable. An individual set of tables was provided for each scale score domain – literacy, numeracy, and for those countries that administered it, PSTRE. The SDT were provided in two formats – HTML and Excel.

The HTML files are suitable for viewing in a browser application, using the accompanying CSS that was provided. Two HTML files were provided for each of the three scales – literacy (LIT), numeracy (NUM), and problem solving (PSL) – separately by the set of international variables (INT) and the set of national adaptations and extension variables (NAT). The "INT" SDT files include the original BQ variables, the OECD-derived variables, and the quintile categorical variables derived from the skill use indices. The "NAT" SDT files include the original idiosyncratic national BQ variables. An additional analysis was performed for the reading component (RC) scores by selected BQ variables. When viewed in a browser, each file has a link at the top of the file to a Table of Contents at the bottom; after clicking on the link, a user can scroll to the left of the display to see links to each of the variables processed in the analysis.

Two types of Excel files were provided for each HTML file. These correspond to two modes of presenting the results:

1. A "Data" worksheet. Each row of the data sheet contains a statistic from the tables presented in the HTML file across all values within each variable and across all variables in the file. Each statistic is accompanied by its standard error estimate, estimated degrees of freedom, estimated population count (weighted N), and number of cases from the data file. The organization of this sheet allows for post-processing of the results by secondary analysis procedures.

2. "Report" worksheets. This consisted of one worksheet for each variable in the analysis, using the variable name as the name of the tab, and a sheet named Table of Contents that contains hyperlinks to the individual worksheets. Each variable-named worksheet contains the analysis results in tabular form, mimicking the tables in the HTML file display.

22.7 BQ crosstabulations

The BQ crosstabulations were produced for internal Consortium quality control and data validation during the initial stages of PIAAC data processing and cleaning. Their contents are similar to the SDT contents that were subsequently provided to the individual countries.

22.8 Item analysis tables, weighted and unweighted

Similar to the summary data tables, the item analysis tables contain summary information about the response types given by the respondents to the cognitive items. They contain, for each country, the percent of individuals choosing each option for multiple-choice items or the percent of individuals receiving each score in the scoring guide for the constructed-response items. They also contain the international average percentages for each response category. The item analysis tables were used by the Consortium and the countries for quality control, verifying data structure accuracy, and validation purposes. A brief description of the details of the calculation of item statistics for the PIAAC data follows.

PIAAC introduced many features for the first time to the large-scale population surveys of cognitive skills. Two main unique features that impact item analysis are: 1) the use of two modes of assessment – CBA and PBA, and 2) adaptive testing on computer.

Both features interact with the background characteristics and skills of the respondents who received particular sets of items. Even a simple statistic such as the proportion correct across two groups of respondents may not be directly comparable if, for example, it involves comparisons between groups taking two different modes of assessment, or groups following different adaptive-testing paths due to variation in skills. In general, younger and more educated respondents tended to receive CBA rather than PBA items based on their ICT skills. However, statistics for the items in a set administered to a group of respondents are comparable within a country. For example, item statistics of PBA items can be compared to each other. But because CBA items are clustered in smaller sets for multistage adaptive testing, direct comparison of item statistics among CBA items is limited.

All respondents with nonzero weights were included in the item analysis. Item analysis of cognitive data involves calculation of a set of statistics to describe the data in terms of quantity and quality before we apply any measurement model. Two sets of statistics were calculated on the unweighted data to represent the number of cases and structures of data using the uniform weight of 1, and also on the final weights to calculate similar statistics to describe the data in comparison to the reference of choice, such as international means.

Unweighted item analysis results are particularly useful to verify the accuracy of the data structure. Seven worksheets are provided in each Excel item analysis file for unweighted and weighted items: literacy core items, numeracy core items, literacy and numeracy PBA items, literacy CBA items, numeracy CBA items, PSTRE CBA items and reading component items.

Each worksheet has eight columns unless there are polytomous items in a set. Each row represents a unique item, identified in the first column entry with the item ID used throughout all phases of PIAAC. The second and third columns are the number of respondents for "not administered" and "not reached." Due to the matrix sampling design, in addition to the two modes of administration, each respondent was given only a fraction of the items in the item pool. By design, these missing responses were termed "not administered." In some cases, respondents were given tasks they did not attempt or reach during the time period allotted for the survey. Consecutively missing responses at the end of a block were termed "not reached." Both "not administered" and "not reached" respondents are excluded in calculating percent correct.

In some cases, responses were missing because respondents chose not to perform a task. Any missing responses that were followed by a valid response (whether correct or incorrect) were termed as "omitted" responses for PBA items. This means a missing response on the last item in a PBA booklet was not treated as omitted. For the adaptively administered CBA items, the position of an item is not nearly as informative as the duration of time each respondent spent on it, as well as the type of input that the respondent provided using the keyboard or mouse. Clearly, the absence of keyboard or mouse responses from a respondent who skips items without having the chance to examine them is not a good indication of his or her skills. A heuristic decision was made that the absence of response when less than five seconds was spent on an item was treated as "not administered" even though it might have been followed by a valid response later on. Omitted responses were treated as wrong. The total consists of the sum of omitted, correct and incorrect responses. Percent correct is calculated as the number of respondents with the correct response divided by the total number of respondents who attempted the item.

Because statistically equivalent samples received either the literacy or numeracy PBA booklet, item statistics are comparable within a country, that is, an item with percent correct of 0.4 was more difficult than another item with percent correct of 0.65 for the PBA population. The comparability of PBA item statistics is limited across countries due to the population characteristics of the PBA respondents of each country, which is primarily driven by the ICT skills of respondents instead of good representation of the national population.

Using only the final weight to calculate item statistics means they are not comparable across countries due to the differential proportions of respondents who took a particular adaptive path. In particular, the total number correct would be greatly biased based on the distribution of paths. In order to increase comparability across countries, path weights were standardized using the international average of path proportions in addition to the final sample weights. The final weights (prior to the application of path proportions) were standardized to 5,000 for each country.

22.9 Compendia

Using the public-use files (PUF) as the source data, the compendia are sets of tables that provide categorical percentages for both cognitive and background items. The compendia are essentially redacted versions of the summary data tables. The purpose of the compendia is to support PUF users so they can gain knowledge of the contents of the PUF and use the compendia results to be sure that they are performing PUF analyses correctly. The item statistics reported in the compendia differ from the item analysis tables in two ways: 1) for confidentiality reasons, some countries have altered data or removed respondent records from their PUF files; and 2) the compendia do not use the routing methods employed in the item analysis. As a result, comparing compendia item statistics across countries for reporting purposes is not appropriate. The compendia reside on the OECD PUF Web site.

22.10 Report tables

The report tables are publication-ready tables that were provided by the Consortium to support the OECD international report. These tables were derived using the ETS Dynamic Reporting System. The data source is the PIAAC Data Explorer database. The PIAAC Data Explorer analysis and reporting engines generated the required reporting statistics.

Chapter 23: International Database and Data Analysis Tools

Ralph Carstens and Tim Daniel, IEA Data Processing and Research Center;
Eugenio Gonzalez, ETS

23.1 Overview

Designing, collecting, validating and analyzing PIAAC data was a very complex, highly demanding and collaborative process involving all Consortium partners, a broad range of external experts, all participating countries, and the OECD Secretariat. Naturally, this in turn led to a data product that reflects the design complexities. To support and promote secondary analyses, the OECD is making a public-use version of the international database and this technical report available to interested analysts and users in the scientific community as well as the general public. The international public-use version of the PIAAC database is made available in two different ways: i) as a database underlying a Web-based data analysis software, the PIAAC Data Explorer (PDX), and ii) a set of public-use files (PUF) which comprise person-level microdata from those countries that gave permission to release their national data.

This chapter is intended to provide a basic introduction to the PIAAC public-use database and the software tools capable of replicating the descriptive and inferential analysis presented in the initial publication "Skills Outlook 2013: First Results firm the Survey of Adult Skills (PIAAC)" (OECD, 2013). First, the chapter will discuss the contents of the public-use data both at the record as well as the variable level, the approach to identifying missing data under a complex, multi trajectory design, and the available database formats. Then, the chapter will describe general analytical considerations followed by the types of analysis supported by the two software tools provided by the Consortium: the International Data Explorer and the IDB Analyzer.

This chapter, however, does not intend to cover and illustrate the full range of possible analytical techniques appropriate for PIAAC and therefore does not describe, for example, advanced modeling of data such as structural equation modeling (SEM). Nonetheless, analysts wishing to use the public-use microdata to undertake advanced analysis not covered by the provided software or those wishing to use alternative statistical software packages will find sufficient technical information on the structure of the database, the included measures, and the variance estimation approaches to successfully configure such software and statistical models.

23.2 Files in the database

As described in Chapter 13 on data management, a large number of raw response data files and documentation were processed to form a series of files that jointly made up the national master databases for PIAAC, that is, all variables collected or derived as part of PIAAC. These national databases consisted of one main flat file holding respondent/household level information, a set of files holding information relating to the study of scoring reliability within and across countries, an audit log file holding interview process and timing data, and, for each respondent, a set of cognitive log files

native to the CBA platform used in PIAAC. Of these files, only the main flat file is of key analytical interest and thus forms the basis of the public-use database described in this chapter. Other parts of the national master database, such as the cognitive log data, did not have a high analytical priority and in light of time and budgetary constraints are not part of the public-use data described here. However, the OECD may make derivatives of these files available to the public in the future.

At the time of processing, analysis, weighting, validation and reporting, all data for a particular PIAAC participant were kept separate from that of other participants. This partitioning per participant also holds for the PUFs and allows for a more flexible, staggered release of files to public users. This is especially useful given that a number of additional countries are currently implementing a second "round" of the first cycle of PIAAC. It is expected that these participants will be added to the public-use database in due time and be available through both the Data Explorer and in the form of a public-use microdata file. Further, certain PIAAC participants may require confidentiality agreements to be signed before public users may receive and use the data.[1] This and related information will be communicated by the OECD via the PIAAC website.

For the naming of physical files, lists of available samples and assigning value labels within the variables identifying countries and subnational entities, operational identifiers based on the ISO 3166/UN M49 standard were used. Table 23-1 provides details. Physical data files are named using the alpha-3 code of the national entity. Within databases, the variable CNTRYID holds the numerical codes and labels of the national entity to which the data belong. The variable CNTRID_E holds the numerical codes and labels of the subnational entity.

With the exception of three participants, it was a national entity that participated in the assessment; therefore, the codes and labels for CNTRYID and CNTRYID_E are identical. In the case of Belgium, only the Flemish part participated. In the case of Canada, the English- and French- speaking parts are identified as subnational entities. In the case of the United Kingdom, the database includes the data from two subnational entities: England and Northern Ireland. Keeping this information in two separate variables allows for analysis at the level of the national as well as subnational entities (as domains) as appropriate. The initial reporting in "Skills Outlook 2013: First Results firm the Survey of Adult Skills (PIAAC)" (OECD, 2013) was done at the level of national entities. Combined data for "England (UK)" and "Northern Ireland (UK)" was reported as "England/N. Ireland (UK)" in the international reporting. Data for Belgium (Flemish part only) was reported as "Flanders (Belgium)."

[1] At the time of writing, this applied to Australia.

Table 23-1: Operational participant codes and names used in PIAAC

National entity name	National entity numeric code	National entity alpha-3 code	Subnational entity name	Sub-national numeric code	Sub-national alpha-3 code
Australia	36	AUS	n/a	n/a	n/a
Austria	40	AUT	n/a	n/a	n/a
Belgium	56	BEL	Flanders (Belgium)	956	BFL
Canada	124	CAN	Canada (English)	1241	CEN
			Canada (French)	1242	CFR
Cyprus[2]	196	CYP	n/a	n/a	n/a
Czech Republic	203	CZE	n/a	n/a	n/a
Denmark	208	DNK	n/a	n/a	n/a
Estonia	233	EST	n/a	n/a	n/a
Finland	246	FIN	n/a	n/a	n/a
France	250	FRA	n/a	n/a	n/a
Germany	276	DEU	n/a	n/a	n/a
Ireland	372	IRL	n/a	n/a	n/a
Italy	380	ITA	n/a	n/a	n/a
Japan	392	JPN	n/a	n/a	n/a
Korea	410	KOR	n/a	n/a	n/a
Netherlands	528	NLD	n/a	n/a	n/a
Norway	578	NOR	n/a	n/a	n/a
Poland	616	POL	n/a	n/a	n/a
Russian Federation[3]	643	RUS	n/a	n/a	n/a
Slovak Republic	703	SVK	n/a	n/a	n/a
Spain	724	ESP	n/a	n/a	n/a
Sweden	752	SWE	n/a	n/a	n/a
United Kingdom	826	GBR	England (UK)	926	ENG
			Northern Ireland (UK)	928	NIR
United States	840	USA	n/a	n/a	n/a

23.3 Records in the database

This section describes the records included in the database. PIAAC used a highly complex assessment design that resulted in a number of possible trajectories through the interview process. It is therefore important for users to understand this design in order to make appropriate use of the database.

23.3.1 Records included in the database

As a general principle, each national master database and, by extension, each national public-use database includes the exact same records that were considered to be suitable for analysis. One

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

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

exception to this rule in discussed below. More specifically, each record in the database generally corresponds to a responding sampled person. Each record in the database also conforms to the international target population definition, that is, adults between the ages of 16 and 65. All records in the database were adjudicated, weighted and used in the computation of response rates.

While the vast majority of records in the database are "true completes," that is, sampled respondents that followed the intended interview workflow until the end (regardless of administration mode and flow), there are noteworthy exceptions. The two main groups of respondents that are included in the database with weights and replicate but with only very little or partial information are i) literacy-related nonresponse cases, that is, respondents who were unable to take the assessment or discontinued it for one of three reasons,[4] and ii) certain types of break-offs, i.e., respondents who decided to discontinue the assessment after it commenced.

The inclusion of these two types of records directly relates to the PIAAC Technical Standard 4.3.3 (OECD, 2011b) that defines a "completed case." A completed case is one that minimally has:

a. responses to key background questions (age, gender, highest level of education and employment status) and a completed Core instrument (i.e., the interviewer asked the respondent all Core questions), or
b. responses to age and gender for literacy-related nonrespondents to the BQ and the Core instrument.

The original plan was to assign imputed scores at the lowest level of proficiency for these cases. However, this was not warranted from a psychometric point of view and the additional information reviewed. As a consequence, these types of records in the database are likely incomplete and not fully usable and, in the case of literacy-related nonresponse, will not have plausible values. In the analytical tools described below in this chapter, these cases will be reported as "not classified" in certain types of analysis.

23.3.2 Records excluded from the database

As part of the data collection, validation, and weighting, certain cases in the original master databases were excluded from the analysis and the public-use databases. The types of cases dropped from the databases include, but are not limited to: i) out of scope respondents, ii) households with no sampled persons, iii) noninterviews, meaning sampled persons who were not interviewed due to refusal or other reasons, iv) a small number of suspected falsified cases detected as part of the validation and quality control, v) respondents with less than the minimally required BQ items (age, gender, highest level of education and employment status) or age and gender in the case of literacy-related nonresponse, and vi) cases with certain anomalies or unclear origin. These cases were flagged accordingly and no weights were computed.

In relation to this, two notes should be made:

a. Two countries targeted respondents not part of the international target population definition (adults from 16 to 65). In the case of Denmark, this related to an oversample of Programme for International Student Assessment (PISA) students. In the case of Australia, an oversample targeted individuals at the age of 15 and between 66 and 75 years. Both groups of cases were excluded from the respective PUFs.

[4] The three types of literacy-related nonresponse are: i) language problems (disposition code 7), ii) reading and writing difficulty (code 8), and iii) learning/mental disability (code 9).

b. In the case of Canada, disclosure risk assessment demanded the reweighting of a small number of cases from a particular domain of respondents in order to comply with Statistics Canada's minimum weight reporting standards. As a consequence, some cases were excluded from the public-use microdata file for Canada and its corresponding weights were loaded onto other cases in the domain. This means the full set of cases used for the international reporting and the revised set of cases included in the PUF for Canada are not identical. Therefore, it will be impossible to replicate reported estimates precisely using the PUF. However, these small weight adjustments should have no practical relevance and should not affect the agreement of estimates published by the OECD, those produced by the Data Explorer, and those made on the basis of the PUF.

23.4 Variables in the database

The PIAAC design is a highly complex one that integrates sophisticated sampling and weighting approaches, a multitrajectory assessment, rich BQ, CBA and PBA modes, innovative item formats, related process information, and a range of derived measures, indicators and indices. In total, each national database includes 1,712 common variables. There were a total of 1,206 country specific variables.

With that said, it is obvious that such a rich database contains variables of varying analytical utility and priority. For example, a large number of variables only include process-related information or temporary information that is necessary, for example, during the computation of weights. This section therefore describes the types of variables included in each participant's public-use database. It also describes those excluded because they carry no analytical utility for international comparisons or address identified and/or assumed disclosure risks.

23.4.1 Variables included

The public-use database underlying the PDX and the PUF contains different sets of variables. The PUF includes a comprehensive set of 1,328 variables. Of these, only 575 are included in the Data Explorer database, implying that certain sets are not informative for analysis in the PDX yet are included in the PUF for secondary analysis. The majority of variables included only in the PUF relate to the individual cognitive item scores and process information. Table 23-2 provides a breakdown of variables by type, name or naming convention, and whether the respective group is available in the PDX or the PUF.

Table 23-2: Variable groups and their description, count, naming convention, and inclusion in public-use database

Variable group	Description	Count	Names or naming convention[5]	Inclusion
Identifiers	National entity, subnational entity and respondent identifier	3	CNTRYID, CNTRYID_E, SEQID	DX and PUF
Resolved demographics	Resolved age and gender	2	AGE_R, GENDER_R	DX and PUF
Derived disposition codes	Summary disposition codes derived from detailed disposition codes	3	DISP_CIBQ, DISP_MAIN, DISP_MAINWRC	DX and PUF
BQ	Originally collected BQ responses (after mapping from national data where applicable)	249	{A-J}_{Q/D}*{a-m}*, e.g., B_Q01a	DX and PUF
BQ – Coded responses	Coded values for respondents' language, education, occupation, industry, country, and region	13	LNG_*, ISCED_HF, ISCO08_*, ISIC4_*, CNT_*, REG_TL2	DX and PUF
BQ – Derived background information	Background information derived from original or coded BQ items	30	AGE10LFS, AGEG5LFS, BIRTHRGN, BORNLANG, CTRYQUAL, CTRYRGN, FIRLGRGN, FORBILANG, FORBORNLANG, HOMLANG, HOMLGRGN, IMGEN, IMPAR, IMYRCAT, IMYRS, ISCO*, ISCOSKIL4, ISIC*, NATBILANG, NATIVELANG, NOPAIDWORKEVER, PAIDWORK12, PAIDWORK5, SECLGRGN,	DX and PUF
BQ – Derived education information	Education information derived from original or coded BQ items	26	AETPOP, EDCAT*, EDWORK, FAET*, FE12, FNFAET*, FNFE12JR, LEAVEDU LEAVER1624, NEET, NFE*, PARED, YRSQUAL, YRSGET, VET	DX and PUF
BQ – Derived earnings information	Earnings variables (continuous, continuous purchasing power parity (PPP) corrected, deciles) for BQ earnings items	17	EARN*, MONTHLYINCPR, YEARLYINCPR	DX and PUF
BQ – Derived skill use information / scale scores	Scales scores (standardized and categorized weighted likelihood estimation) for skill use items in BQ	26	LEARNATWORK*, READYTOLEARN*, ICTHOME*, ICTWORK*, INFLUENCE*, NUMHOME*, NUMWORK*, PLANNING*, READHOME*, READWORK*, TASKDISC*, WRITHOME*, WRITWORK*	DX and PUF
BQ – Derived trend information	Recoded versions of BQ responses to facilitate trend analysis with IALS/ALL data	44	As for original BQ variables yet with suffix "_T" or "T1"	DX and PUF

[5] {Brackets} indicate the possible characters used in variable names. Asterisks (*) indicate name stems.

Variable group	Description	Count	Names or naming convention[5]	Inclusion
BQ – Derived coarsened information	Coarsened versions of BQ responses (collapsed, categorized or top-coded)	29	As for original BQ variables yet with suffix "_C"	DX and PUF
BQ – Derived cognitive routing	Variables derived from BQ at the time of collection to determine adaptive routing	3	COMPUTEREXPERIENCE, NATIVESPEAKER, EDLEVEL3	PUF only
Cognitive scores, pass flags, random numbers	Core scores, pass status, and random module allocation recorded at the time of collection	13	CBA_CORE_STAGE*_SCORE, CORESTAGE*_PASS, RANDOM_CBA_*, CBA_START, PPC_SCORE, RANDOM_PP	PUF only
Cognitive routing – Derived	Variables derived from the actual routing describing the module allocation	9	PAPER, CBAMOD*, PBROUTE	DX and PUF
Observation module	Interviewer's descriptions of the assessment session	13	ZZ*	PUF only
Cognitive item responses and process information	Cognitive item information: actual response (R), scored response (S), total time (T), time to first action (F), number of actions (A)	720	{C/D/E/M/N/P/U}*{A/F/R/S/T}, e.g., C301C05S	PUF only
Numeracy, literacy and problem-solving scale score status	Status flags indicating availability of scale scores for the respective domain	3	LITSTATUS, NUMSTATUS, PSLSTATUS	DX and PUF
Numeracy, literacy and problem-solving scale scores	Scale scores (plausible values) for each of three domains	30	PVLIT1 to PVLIT10, PVNUM1 to PVNUM10, PVPSL1 to PVPSL10	DX and PUF
Reading components scores	Total correct scores (point estimates) for reading components	3	PRC_PV_SCR, PRC_SP_SCR, PRC_PC_SCR	DX and PUF
Reading components timers	Timing values for reading component parts	5	PRC_PV_Q1, PRC_SP_Q1, PRC_PF_Q1, PRC_PF_Q2, PRC_PF_Q3	DX and PUF
Variance estimation	Variables controlling variance estimation stratification, method, and number of replicates	6	VEMETHOD, VEMETHODN, VEFAYFAC, VENREPS, VARSTRAT, VARUNIT	DX and PUF
Full weight and replicates	Complex sample estimation weights	81	SPFWT0, SPFWT1 to SPFWT80	DX and PUF
Total		1,328		

23.4.2 Variables excluded, suppressed or coarsened for some or all countries

The public-use databases only include a subset of the information available in the master databases. The public-use database does not include any data collected using national adaptations and extensions. It only includes data that were collected or derived across all countries. Further, a sizable number of variables were excluded in consultation with the OECD Secretariat and the BPC because they i) have no or little analytical utility, ii) were intended for internal or interim purposes only, iii) relate to secure item material, or iv) include personally identifiable data, or at least data that may increase the risk of unintended or indirect disclosure.

The groups of variables excluded from the public-use database are:

a. direct, indirect, and operational identifiers for respondents, interviewers, scorers, key operators, and paper materials
b. interim sampling, disposition, data availability, demographic, and weighting information
c. certain BQ or process variables that are available in coded or derived form (for example, country and language), especially detailed write-ins
d. all national adaptations and extensions in the BQ
e. interviewer's scoring of paper-based core items
f. detailed response information for secure problem-solving items
g. original scale score values (theta) before standardization to an international metric

National data is not of key interest in an international large-scale assessment and comparison. However, national data might be available by directly contacting the concerned PIAAC participant.

A particularly important issue is to preserve the confidentiality of individual respondents in the release of the public-use aggregate (PDX) and microdata (PUF) in order to prevent unintended or indirect disclosure. The risk of such disclosure is greatest in cases where the combined characteristics of a respondent in a sample lead to a unique individual in the population. The higher the sampling fraction, the more likely a unique record in the sample will also be unique in the population. As agreed by the BPC, countries were given the possibility to either coarsen or suppress their data prior to submission to the Consortium and the OECD and/or afterward during the production of the public-use database. PIAAC participants were asked to suppress information only when deemed absolutely necessary to meet national legislative requirements.

The database underlying the PDX and PUF was subject to around 700 instances of suppression (participant x variable) at the cell or column level. The majority of these instances relates, but are not limited, to:

a. detailed age
b. detailed language, country of birth, or region information
c. detailed education information (BQ section B)
d. detailed occupation (ISCO) and industry (ISIC) information
e. detailed, original, or derived earnings variables (BQ section D)
f. variance strata and unit information

Suppressed data are represented in the database by means of missing codes. As with national data, more detailed data might be available directly from the concerned participant.

Database users should note that the most complete set of information was available to the Consortium for analysis and the OECD for reporting and archiving. The PDX is based on a reduced database, that is, it includes fewer variables and less information as a function of suppressions. Finally, the PUF is the most restricted database in PIAAC.

In almost all cases where more than one participant requested the suppression of a particular variable for the PDX or PUF, a coarsened version of this variable (suffix "_C") was created that includes the level of detail deemed suitable for public release by the concerned countries (see group "BQ – Derived coarsened information" in Table 23-2 above). Analysts are therefore recommended to use such a

coarsened variable if the aim of the analysis is to include the most complete set of countries, albeit with a reduced level of detail.

As a result (and similar to other data collections), public users of the databases in the PDX or PUF may be unable to fully replicate particular tables, figures, and other exhibits in the international reporting because such reporting was based on the most complete set of confidential information, which is not available to the general public.

23.5 Representing valid and missing data

As in all survey projects, missing data is a natural phenomenon. PIAAC is no exception, and despite the intention to collect complete or almost complete information, there are related gaps in the database. In principle, missing data in a survey may occur when there are no or almost no observed data as well as no administrative data for a respondent (unit nonresponse) or when some variables for a respondent are unknown or cannot be known (item nonresponse). Missing data can further be distinguished semantically in two broad groups: i) data that cannot exist due to the way a survey is designed and ii) data that were supposed to be observed but were not.

To understand the missing data pattern in PIAAC, users are reminded of the complex assessment design. Missing data in PIAAC can occur for a number of reasons. The main ones are:

a. Data are missing by design (that is, it is known a priori they will not collected) for some or all respondents because of the way the assessment is designed.
    i. Respondents with literacy-related dispositions (see above) were not administered the interview.
    ii. A small number of PIAAC participants did not participate in one or both of the international options: i) problem solving in technology-rich environments and ii) reading components.
    iii. Certain sections in the BQ were intentionally presented to subpopulations (domains) only with reference to responses given to prior questions (“valid skip”).
    iv. Respondents were by default administered the CBA or, as a result of their lack of computer familiarity, inability or refusal to take the exercise on the computer and/or performance on core modules, a full or reduced PBA was administered.
    v. Respondents following the paper-based path were not administered problem-solving items and therefore have no plausible values for problem solving.
    vi. Domain item clusters (CBA and PBA) were assigned based on random allocation and previous proficiency information collected (in the case of CBA).
b. Data are missing as a result of the response process.
    vii. Respondents may have broken off the interview after it was started as a function of, for example, time, motivation, fatigue, or sensitive questions being asked.
    viii. Respondents may have explicitly refused (“refused”) to respond to questions in the BQ or they may not have known the answer to a question with sufficient certainty (“don’t know”).
c. Data in a few instances are missing due to logistics, processing, or analysis.

ix. Data were captured yet paper booklets and/or CBA result files were lost during transfer.

x. Erroneous routing in national versions of the BQ collected fewer data items for particular respondents than intended.

xi. Certain data items (variables and/or a subset of values) were not provided or suppressed due to regulations relating to confidentiality of information.

xii. Respondents with literacy-related dispositions (see above) were usually not assigned domain scores.

xiii. A small number of values were obvious outliers, otherwise useless, or erroneously coded in the original national databases.

It should be noted that no imputation was intended for missing item responses except for i) the imputation of earnings from precise and/or broad categories, and ii) the multiple imputation of proficiency scale scores for the literacy, numeracy and problem-solving domains.

Table 23-3 below provides an overview of the main missing values and their semantic, scope and representation in SAS and SPSS PUFs. The representation of missing values differs in these two statistical packages. In SAS, the standard missing code (.) and special missing values (.A thru .Z) were used. In SPSS, a "dynamic" code that depends on the length of the numeric variable was used. Variables of length 1 use missing values 6, 7, 8 and 9; those of length 2 use missing values 96, 97, 98 and 99; variables of length 3 use 996, 997, 998, 999; and so on unless missing values conflicted with payload values, in which case the variable lengths were increased.

The PIAAC public-use databases also include a small number of coded variables that are defined as strings because the respective coding schemes are defined as string. For example, occupational codes may appear as using a numerical scheme but need to be stored as strings because codes include leading zeros that would be lost if converted to a number. The use of string variables and, therefore, string missing values relates to: i) ISCO codes for occupation, ii) ISIC codes for industry, iii) region codes, and iv) language codes. In these cases, number-based strings such as "9999" were used to represent missing data.

Table 23-3: Generally used missing values in the public-use database (DX and PUF)

Semantic	Scope	Label	SAS	SPSS
Valid skip	BQ and any variables derived from it; reading components	"Valid skip"	Numeric: .V String: "996," "9996"	Numeric: 6, 96 … String: "996," "9996"
Don't know	BQ and variable derived from it	"Don't know"	Numeric: .D String: "997," "9997"	Numeric: 7, 97 … String: "997," "9997"
Refused	BQ and variable derived from it	"Refused"	Numeric: .R String: "998," "9998"	Numeric: 9, 98 … String: "998," "9998"
Not stated/inferred, invalid, not codeable, omitted, not provided, or suppressed	Almost all variables	"Not stated or inferred" (general) "Not reached/Not attempted" (cognitive items)	Numeric: .N String: "999," "9999," "99999"	Numeric: 9, 99 … String: "999," "9999," "99999"
Not administered / not applicable (missing by design)	Cognitive items	n/a	Numeric: (.)	Numeric: (.)

In addition to the general missing scheme described above, which applies to the largest set of variables, the specifications of some derived variables included missing schemes specific to a particular variable or, in some cases, a small set of variables. These missing values are fully documented in the SPSS files and SAS format scripts. Given that the number space for missing values (or letters in case of SAS special missing values) is limited, some of the per-variable missing schemes may use the same missing code, yet the semantic of these codes may vary from one variable to the next. Database users are strongly encouraged to review the coding of missing values in derived BQ variables carefully, using the information provided as part of the SAS/SPSS files and earlier in this report prior to analysis.

23.6 Public-use file (PUF) formats

While the database underlying the Data Explorer is not directly accessible to users, the PUFs are. They are being made available in two standard formats – SPSS and SAS – allowing for data to be loaded and used in these and many other standard packages.

SPSS data files are standard, Windows-based .sav files and encoded in Unicode (UTF-8). SPSS data files include full dictionary information from the applicable metadata maintained in the codebooks: i) variable types and formats, ii) variable labels, iii) value labels (including any missing value labels), iv) missing value definitions, and v) variable measurement levels.

SAS formatted files are standard, compressed .sas7bdat data files for Windows environments and encoded in Unicode (UTF-8). Variable types, widths, decimals, and labels are assigned to all variables according to the labels defined in the metadata. SAS does not provide for a way to permanently stored value labels on the file. Therefore, each PUF file in SAS format is accompanied by an equivalently named .sas file that includes syntax to assign formats (value labels). The SAS format syntax files include the relevant LIBNAME (in), PROC FORMATS, DATA and FORMATS statements. These syntax files can be executed against each individual SAS export file in order to display value labels in analytical procedures such as PROC UNIVARIATE, PROC FREQ, and so on.

23.7 Data analysis and software tools

23.7.1 General considerations for data analysis using PIAAC data

For analysts familiar with population estimation using other large-scale educational survey databases such as those produced by, for example, the OECD PISA program or IEA studies, the analysis of PIAAC data will present relatively few difficulties after becoming familiar with the conceptual foundation and the methodological, operational, and analytical details of the study, especially the BQ framework (OECD, 2011a) and the BQ itself (OECD, 2010). For those unaccustomed to working with complex survey sample data, the technical report as a whole, this chapter in particular, and the analytical tools provided by the Consortium should contain sufficient technical information and references to support statistically correct analysis.

The three main analytical requirements that any analysis of PIAAC data needs to account for are i) the use of sampling weights, ii) the complex multistage cluster sample design that was implemented to balance the research goals and cost-efficient operations, and iii) the use of multiply imputed proficiency estimates, the so-called "plausible values." The key challenge for analyzing PIAAC data, especially when one or more of the proficiency scales are involved, lies at the intersection of the uncertainty in estimating population characteristics due to sampling and the uncertainty introduced by the use of multiple imputations. In addition, another key challenge for PIAAC – in contrast to other international studies – is that there was not a common variance estimation procedure across all participating countries. Chapters 14 and 15 include details of the sampling, weighting and variance estimation techniques intended for PIAAC, the approach adopted by each country, and the mathematical combination of sampling and imputation variance. Chapter 17 includes details on the IRT and latent regression models used in deriving plausible values.

Standard analytical packages for the social sciences and educational research do not readily recognize or support handling the complex sample and assessment design. This gap is filled by the two software tools made available by the Consortium to assist database users to access and analyze PIAAC data and produce basic outputs: i) the PIAAC Data Explorer (PDX) and ii) the IEA's International Database Analyzer (IDB Analyzer). Each of these two software tools addresses a slightly different set of needs. While the PDX is a web based application that allows relatively easy and publication-ready access to basic estimates of means, totals and proportions, the IDB Analyzer used in conjunction with the PUFs provides unit record access to the public-use database and the opportunity to conduct analysis offline, derive additional variables, and produce various estimates for further use and reporting. The PDX and the IDB Analyzer are described in turn in the remainder of this chapter.

A variety of statistical software packages are available as alternatives for the analysis of complex sample data with support for the jackknife and/or BRR replication methods implemented in PIAAC. Still, all of these packages would require participant-by-participant runs or custom scripting to configure the variance estimation used by each participant. Further, these packages may or may not support the simultaneous integration of sampling and imputation variance as required when using PIAAC data. WesVar (2008; Westat Inc., 2007, 2008) software for complex sample analysis is available free of charge from Westat. Commercial packages that include support for the weighting and replication methods used in PIAAC, among others, are SAS 9.4, SUDAAN 11 (2013), and Stata 13 (2013).[6] A (slightly outdated) feature comparison of these packages is available in Heeringa, West and Berglund (2010, Appendix A, pp. 399+).

[6] Only current versions are mentioned. Previous versions may have support for complex sample analysis.

More generally, a detailed description of contemporary sampling and weighting approaches as well as analytical approaches and techniques for complex survey data analysis can be found in Heeringa, West, and Berglund (2010). Additional analytical advice and background useful for PIAAC and the previous international adult assessments may also be found in the user guides for IALS (Statistics Canada, 2001) and ALL (Statistics Canada, 2002) studies.

23.7.2. ETS PIAAC Data Explorer (PDX)

The PDX is a web based application developed by ETS that allows the user to query the PIAAC International Database via a web browser. The PDX can be used to compute a diverse range of statistics including, but not limited to, means, standard deviations, percentages by subgroup, percentages by levels, and percentiles. All statistics are computed taking into account the sampling and assessment design. In addition, the PDX has the capability of conducting significance testing between statistics from different groups, and displaying the results in graphical form. Results from the PDX can be directly exported and saved is MS Word, MS Excel and HTML formats. The PDX is accessible from any computer connected to the internet from the following address: http://piaacdataexplorer.oecd.org/ide/idepiaac.

23.7.3 IEA IDB Analyzer

The IEA International Database Analyzer (IDB Analyzer, 2013) is an application developed by the IEA Data Processing and Research Center (IEA DPC) in Hamburg, Germany. The IDB Analyzer can be used to combine and analyze data from IEA's large-scale assessments such as Trends in International Mathematics and Science Study (TIMSS), TIMSS Advanced, Progress in International Reading Literacy Study (PIRLS), the Second Information Technology in Education Study (SITES), the Teacher Education and Development Study in Mathematics (TEDS-M), the Civic Education Study (CivEd), and the International Civic and Citizenship Education Survey (ICCS) as well as analyze data from the OECD's Teaching and Learning International Survey (TALIS), PISA and PIAAC.

The IDB Analyzer creates SPSS syntax that can be used to perform analysis with these international databases. In other words, it requires SPSS (Version 15 or above) to be installed on the user's system. The syntax generated and the referenced macros take into account information from the sampling design in the computation of sampling variance. In addition, it handles plausible values. The resulting code can be used to calculate estimates of achievement and their corresponding standard errors, combining sampling and imputation variance. The code generated by the IDB Analyzer enables the user to compute descriptive statistics and conduct statistical hypothesis testing among groups in the population without having to write any programming code.

The IDB Analyzer is licensed free of cost by the IEA and is for use only in accordance with the terms of the licensing agreement. While users can use the software for free, they do not have any ownership of, copyright or other intellectual property rights to the software itself or its components, including the SPSS macros. Users are only licensed to use the SPSS enclosed macros in combination with the IDB Analyzer unless explicitly authorized by the IEA in writing.

The Analyzer is available from the following permanent URL: http://www.iea.nl/data.html. The software license expires at the end of each calendar year, when users will again have to download and reinstall the most current version of the software. Features will be added on a continuous basis to support additional surveys and databases or include additional types of analysis, options or outputs. Technical support for the IDB Analyzer can be obtained by contacting the IEA Data Processing and Research Center's Software Unit at software@iea-dpc.de.

The IDB Analyzer is fully self-documenting, and each version comes with a comprehensive help manual as part of the installation (International Association for the Evaluation of Educational Achievement [IEA], 2013). Users of the PUFs (PUF) are referred to this more detailed documentation with respect to the use and interpretation of the Analyzer's features, options, and outputs.

The IDB Analyzer consists of two modules – the Merge Module and the Analysis Module – which are integrated in one common application window.

23.7.3.1 The Merge Module

The Merge Module is used to combine data files from different study participants, and when necessary, merge data files from different sources like student BQs and achievement files, or student background files with teacher- or school-level files. The Merge Module is only available to use with IEA databases and others in which the data are published separate by participant, currently TALIS and PIAAC. In the case of PIAAC, each participant corresponds to a single, flat data file and hence only requires vertical merging, that is, that of one or more participants.

The Merge Module also allows the user to easily select individual or groups of variables to create a smaller and more manageable dataset. When running the Merge Module, the IDB Analyzer creates SPSS code that merges and combines files specified by the user, keeping only the selected variables yet automatically adding all mandatory variables for correct variance estimation.

Merged data files created using the Merge Module can be processed either with the Analysis Module (see below) of the IDB Analyzer or by any other analysis software that accepts SPSS files as input.

23.7.3.2 The Analysis Module

The Analysis Module of the IDB Analyzer provides procedures for the computation of means, percentages, standard deviations, correlations, and regression coefficients for any variable of interest overall for a participant, and for specific subgroups within a participant. It also computes percentages of respondents in the population that are within, at, or above benchmarks of performance or within user-defined cut points in the proficiency distribution, percentiles based on the achievement scale, or any other continuous variable.

The Analysis Module can be used to analyze data files from the above mentioned studies, regardless of whether they have been preprocessed with the IDB Analyzer Merge Module. The Analysis Module can create code for several analysis procedures. Like the Merge Module, the Analysis Module creates SPSS code that computes the statistics specified by the user.

The following analyses can be performed with the Analysis Module:

a. Percentages and means: Computes percentages, means and standard deviations for selected variables by subgroups defined by the user. The percentage of missing responses is included in the output.
b. Percentages only: Computes percentages by subgroups defined by the user.
c. Regression: Computes regression coefficients for selected variables predicting a dependent variable by subgroups defined by the user. New in this version of the IDB Analyzer is the capability of including plausible values as dependent or independent variables in the regression equation. When more than one set of plausible values is specified in the analysis,

the analysis is carried out using the first of the plausible values, then the second, and so on. In the end, the results are summarized across the plausible values.

d. Benchmarks: Computes the percentage of students meeting a set of user-specified performance or achievement benchmark by subgroups defined by the user. It computes these percentages in two modes: cumulative (percentage of students at or above given points in the distribution) or discrete (percentage of students within given points of the distribution). It can also compute the mean of an analysis variable for those at a particular achievement level. As an additional feature, the IDB Analyzer allows the user to compute percentages of people at each of the proficiency levels including, or excluding those that did not participate in the assessment.

e. Correlations: Computes correlation for selected variables by subgroups defined by the grouping variable(s). New in this version of the IDB Analyzer is the capability of computing the correlation between plausible values. When more than one set of plausible values is specified in the analysis, the analysis is carried out using the first of the plausible values, then the second, and so on. In the end, the results are summarized across the plausible values.

f. Percentiles: Computes the score points that separate a given proportion of the distribution of scores by subgroups defined by the grouping variable(s).

Prior to every analysis, the Analyzer calculates unweighted and weighted descriptive statistics for the analysis variables (means, standard deviations, minimum and maximum), and frequencies by analysis subgroups. In addition, except when computing percentiles, the estimate of the population size for each of the subgroups processed (sum of the sampling weights) and the corresponding standard errors are computed. Bar or line charts are drawn by default when computing percentages, percentages and means, and when calculating the percentages of the population within benchmarks with or without an analysis variable.

When calculating these statistics, the IDB Analyzer has the capability of using any continuous or categorical variable in the database, or makes use of scores in the form of plausible values. When using plausible values, the IDB Analyzer generates code that takes into account the multiple imputation methodology in the calculation of the variance for statistics as it applies to the corresponding study.

All procedures offered within the analysis module of the IDB Analyzer make use of appropriate sampling weights and standard errors of the statistics that are computed according to the variance estimation procedure required by the design as it applies to the corresponding study. In the case of PIAAC, this functionality extends to the level of participants, as the variance estimate method (VEMETHOD) and number of replicate weights (VENREPS) is encoded in the respective PUF.

For a complete list of features, options, and output fields and parameters, users are referred to the help manual that is part of every installation.

References

Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). *Applied survey data analysis*. Boca Raton, FL: Chapman & Hall/CRC. See also http://www.isr.umich.edu/src/smp/asda/

IDB Analyzer (Version 3.1) [Computer software], (2013). Retrieved from the International Association for the Evaluation of Educational Achievement website: http://www.iea.nl/data.html

International Association for the Evaluation of Educational Achievement. (2013). *Help manual for the IDB Analyzer* (Version 3.1). Hamburg, Germany: IEA Data Processing and Research Center.

Organisation for Economic Co-operation and Development. (2010). *PIAAC background questionnaire* (MS Version 2.1). Retrieved from the OECD website: http://www.oecd.org/edu/48442549.pdf

Organisation for Economic Co-operation and Development. (2011a). *PIAAC conceptual framework of the background questionnaire. Main survey*. Retrieved from the OECD website: http://www.oecd.org/site/piaac/PIAAC(2011_11)MS_BQ_ConceptualFramework_1%20Dec%202011.pdf

Organisation for Economic Co-operation and Development. (2011b). *PIAAC technical standards and guidelines. December 2011*. Paris: Author.

Organisation for Economic Co-operation and Development. (2013). *Skills Outlook 2013: First Results firm the Survey of Adult Skills (PIAAC).* Paris, France: OECD.

Stata 13 [Computer software], (2013). College Station, TX: StataCorp LP.

Statistics Canada. (2001). *International Adult Literacy Survey. Microdata user's guide*. Ottawa, Canada: Statistics Canada.

Statistics Canada. (2002). *The Adult Literacy and Life Skills Survey, 2003. Public use microdata file user's manual*. Ottawa, Canada: Statistics Canada.

SUDAAN 11 [Computer software], (2013). Research Triangle Park, NC: RTI International.

Westat Inc. (2007). *WesVar 4.3 user's guide*. Rockville, MD: Author.

Westat Inc. (2008). *Addendum to the WesVar user's guide. New features in WesVar 5.1*. Rockville, MD: Author.

WesVar (Version 5.1), Replication-Based Variance Estimation for Analysis of Complex Survey Data [Computer software], (2008). Rockville, MD: Westat Inc.