Chapter 3
Plausible Values: Principles of Item Response Theory and Multiple Imputations

Lale Khorramdel, Matthias von Davier, Eugenio Gonzalez, and Kentaro Yamamoto

Abstract This chapter introduces the principles of item response theory (IRT) and the latent regression model, also called population or conditioning model, which is central for generating plausible values (multiple imputations) in PIAAC. Moreover, it is illustrated how plausible values can reduce bias in secondary analyses compared to the use of customary point estimates of latent variables by taking explanatory variables into account. An overview of standard techniques for utilizing plausible values (PVs) in the analyses of large-scale assessment data will be provided, and it will be discussed how to calculate the different variance components for statistics based on PVs, which play an important role in the interpretation of subgroup and country differences.

The Programme for the International Assessment of Adult Competencies (PIAAC) provides a rich international database that can be used by policymakers, stakeholders, and educational researchers for examining differences in educational systems and outcomes across countries, groups of test-takers within countries, and over time for the measurement of trend. The PIAAC database includes measures of cognitive domains, such as literacy, numeracy, and problem solving in technology-rich environments (PS-TRE), as well as background information and non-cognitive measures obtained from a background questionnaire (BQ). For each cognitive domain and background variable, test-takers' raw responses are available in addition to proficiency estimates in the form of plausible values (PVs) for the cognitive domains and item response theory (IRT)-based estimates for some of the non-cognitive measures. For the computer-based assessment, two types of process data

L. Khorramdel (✉) · M. von Davier
National Board of Medical Examiners, Princeton, NJ, USA
e-mail: lale.khorramdelameri@gmail.com

E. Gonzalez · K. Yamamoto
Educational Testing Service, Princeton, NJ, USA

D. B. Maehler, B. Rammstedt (eds.), *Large-Scale Cognitive Assessment*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-47515-4_3

are included in the database as well—the number of actions (e.g. number of mouse clicks when interacting with an item on the computer) and the total response time—as well as the time to first action for each item. As we will see later in this chapter, utilising a latent regression model is necessary to reduce bias in the estimation of means and variances of the subgroups of interest. The source of this bias is the fact that, while the domains measured are broad, we have a limited amount of assessment time during which we can assess the respondent's skills, and therefore we need to resort to statistical techniques that will borrow information to correct for the unreliability of measurements.

In order to facilitate broad domain coverage while limiting individual testing time, which is aimed at reducing test-takers' burden, the PIAAC data are based on a variant of matrix sampling where different groups of respondents answered different sets of items (see Chap. 2 in this volume). Therefore, it is not appropriate to directly compare the group performance using conventional statistics such as the total score. This would only be feasible if one made very strong assumptions—for instance, that the different test forms are perfectly parallel and that there is hardly any measurement error. Since this is almost never the case, conventional scoring methods show several limitations, such as ignoring the variability and dissimilarities of proficiencies of subgroups. These limitations can be overcome in part by using IRT scaling where respondents as well as items can be characterised on a common scale, even if not all respondents take identical sets of items (e.g. in adaptive testing). This makes it possible to describe performance distributions in a population or subpopulation and to estimate the relationships between proficiencies and background variables.

As stated above, to improve the statistical properties of the group-level proficiency estimates, PIAAC uses PVs, which are multiple imputations. These imputations are drawn from a posterior distribution that is the result of combining information from the cognitive assessment and the BQ. To compute PVs, a latent regression model, also called population or conditioning model, is estimated that combines an IRT model with an explanatory model regressing proficiency on background data. In this model, which is tailored for use in large-scale assessments, IRT item parameter estimates are fixed to values from previous item calibrations, and the background variables are used as predictors.

The remainder of this chapter is organised as follows: First, we describe the IRT model and the scaling of item parameters. This is followed by a description of the latent regression model used for generating the PVs in PIAAC. It will be illustrated how the use of PVs can reduce bias (by accounting for measurement error) in secondary analyses and lead to more accurate results. It will also be described how PVs can be used appropriately in statistical analyses to avoid errors and biases when analysing the PIAAC data. Moreover, we will give an outlook on how the predictive power of the population model can be improved by including information from process data obtained from computer-based assessments.

3.1 IRT Scaling

3.1.1 IRT Models and Calibration of Item Parameters

The proficiency values θ for the PIAAC cognitive domains literacy, numeracy, and PS-TRE cannot be directly observed, as each respondent provides only a small number of answers, and respondents will only answer a subset of the domains. Hence, we do not have a very accurate picture on the individual level, but we have a large number of responses on the level of the 5000, or so, respondents per country (see the national sample requirements based on the PIAAC test design in Chap. 2 in this volume). Even if a person takes a long test, a case can be made that we never directly observe variables such as reading ability, general intelligence, or neuroticism, but that we rather observe only behavioural indicators that we believe are related to underlying individual differences.

In addition, tasks such as literacy items differ with respect to how well they measure aspects of literacy and in terms of how difficult they are on average. IRT is a model that takes into account interindividual differences as well as differences between items, and can be used to derive estimates that represent proficiencies on the one hand, and parameters representing features of the tasks, such as item difficulty, as well as discrimination, which can be described as the ability of an item to differentiate between high- and low-proficient respondents.

Latent variable or IRT models can disentangle differences between items from differences between test-takers and therefore have a number of advantages when it comes to statistical analyses of data from assessments such as PIAAC. Interested readers are referred to van der Linden and Hambleton (2016) for an overview of IRT, and to Rutkowski et al. (2014) for a handbook that describes in great detail the methods used in PIAAC, but also in student assessments such as the Programme for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), and the Progress in International Reading Literacy Study (PIRLS). IRT is used to estimate the proficiency values as well as the item parameters in PIAAC using the two-parameter logistic model (2PLM; Birnbaum 1968) for items with two response categories and the generalised partial credit model (GPCM; Muraki 1992) for items with more than two response categories.

The *2PLM* is a mathematical model for the probability that an individual will respond correctly to a particular item depending only on the following parameters: the individual's ability or proficiency (the person parameter) and the difficulty and discrimination of the particular item (the item parameters). This probability is given as a function of the person parameter and the two item parameters and can be written as follows:

$$P\left(X = x|\theta, \beta_i, \alpha_i\right) = \frac{\exp\left(D\alpha_i\left(\theta - \beta_i\right)\right)}{1 + \exp\left(D\alpha_i\left(\theta - \beta_i\right)\right)} \tag{3.1}$$

with $X \in \{0, 1\}$ and $X = 1$ indicating a correct response to a binary coded item. The θ, β_i are real-valued parameters, commonly referred to as ability and difficulty parameters, respectively, and α_i is the discrimination or slope parameter (similar to a factor loading). $D > 0$ is a positive constant of arbitrary size, often either 1.0 or 1.7, depending on the parameterisation used in the software implementation; in PIAAC, D took on the value of 1.7. Note that for $\alpha_i > 0$ (a commonly made, but not necessary, assumption in IRT), this is a monotone increasing function with respect to θ; that is, the conditional probability of a correct response increases as θ increases.

For polytomous items, the *GPCM* is used. This is a generalisation of the 2PLM for responses to items with two or more ordered response categories and reduces to the 2PLM when applied to dichotomous responses. For an item i with $m_i + 1$ ordered categories, $x \in \{0, \ldots, m_i\}$, the GPCM can be written as

$$P\left(\mathrm{X} = x|\theta, \boldsymbol{\beta}_i, \alpha_i\right) = \frac{\exp\left\{\sum_{r=1}^{x} D\alpha_i \ (\theta - \beta_{ir})\right\}}{\sum_{u=0}^{m_i} \exp\left\{\sum_{r=1}^{u} D\alpha_i \ (\theta - \beta_{ir})\right\}} \tag{3.2}$$

where $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{im})$ are the category threshold parameters. For only two categories, there is only a single threshold parameter that is equivalent to the item difficulty in the 2PLM.

A central assumption of the 2PLM and the GPCM, and most IRT models, is conditional independence (sometimes referred to as local independence). Under this assumption, item response probabilities depend only on θ and the specified item parameters. There is no dependence on any demographic characteristics of the respondents, on responses to any other items presented on the test, or on the survey administration conditions. Moreover, the 2PLM assumes unidimensionality—that is, a single latent variable (θ) accounts for the performance on the full 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:

$$P\left(\boldsymbol{x}|\theta, \boldsymbol{\beta}, \boldsymbol{\alpha}\right) = \prod_{i=1}^{n} P\left(\mathrm{X} = x_i|\theta, \boldsymbol{\beta}_i, \alpha_i\right) \tag{3.3}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n)$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$.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 maximised with respect to the item parameters. To do this, it is assumed that respondents provide their answers independently of one another and that the student's proficiencies are sampled from a distribution, $f(\theta)$. The (marginal) likelihood function for i.i.d. respondents $j = 1, \ldots, J$ and locally independent responses $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{nj})$ can be written as

$$P\left(\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\alpha}\right) = \prod_{j=1}^{J} \int \left(\prod_{i=1}^{n} P\left(X = x_{ij}|\theta, \boldsymbol{\beta}_i, \alpha_i\right) \right) f\left(\theta\right) d\theta. \tag{3.4}$$

Typically, the marginal log likelihood function, $L = \log P(\boldsymbol{X}|\,\boldsymbol{\beta},\boldsymbol{\alpha})$, is maximised using customary approaches such as the EM algorithm (Dempster et al. 1997). The item parameter estimates obtained by maximising this function are used as fixed constants in the subsequent estimation of the latent regression model. This is a convenient choice that enables using fixed parameter linking across groups, as the item parameters are typically found by maximising the likelihood for a sample of respondents drawn from all countries. While PISA used only 500 students per country up until the 2012 cycle, PIAAC, as well as previous international adult assessments, such as the International Adult Literacy Survey (IALS) and the Adult Literacy and Lifeskills Survey (ALL), and PISA since 2015, use all available data in this item calibration, and the resulting item parameters represent the evidence on item difficulties and item discrimination parameters aggregated across all participating countries.

To ensure that the IRT model provides adequate fit to the observed data, different types of model checks are applied. One of these checks is the evaluation of the fit of the estimated item parameters to the observed empirical data. To assess differences in item fit across countries, or relative to previously calibrated parameters, the country-specific mean deviation (MD) and the root mean square deviation (RMSD) were computed for each item in each group of interest (i.e. the different country and language groups in PIAAC). For simplicity, the MD and RMSD are presented here for dichotomous variables only:

$$\mathrm{MD} = \int \left(P_0\left(\theta\right) - P_e\left(\theta\right)\right) f\left(\theta\right) d\theta \tag{3.5}$$

$$\mathrm{RMSD} = \sqrt{\int \left(P_0\left(\theta\right) - P_e\left(\theta\right)\right)^2 f\left(\theta\right) d\theta} \tag{3.6}$$

$P_0(\theta) - P_e(\theta))$ describes the deviation of the pseudo-counts-based ('observed') item characteristic curve from its model-based expected counterpart for a given ability level θ, and $f(\theta)$ is the density of ability distribution at this ability level. More details can be found in Yamamoto et al. (2013). *MD* and *RMSD* both quantify the magnitude and direction of deviations in the observed data from the estimated item characteristic curves. The *MD* is more sensitive to deviations of observed item difficulties than the *RMSD*. The *RMSD* is more sensitive to the deviations of both the item difficulties and discriminations (Yamamoto et al. 2013). In PIAAC, *MD* values between -0.1 and 0.1 and *RMSD* values smaller than 0.1 indicated acceptable item fit.

3.1.2 Treatment of Missing Values

Because of the matrix sampling and the multistage testing (MST) design in PIAAC, the treatment of different types of missing values in the IRT scaling had to be considered.

1. *Missing by design:* Items that were not presented to each respondent due to the matrix sampling design (structural missing data) do not contribute information to respondents' cognitive skills and were excluded from the likelihood function of the IRT model.
2. *Not reached items:* Missing responses at the end of an item block or cluster (see Chap. 2 in this volume) 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. Hence, these missing responses were also excluded from the likelihood function of the IRT model.
3. *Omitted responses:* Any missing response to an item that was administered to a particular respondent and that was followed by a valid response (whether correct or incorrect) was defined as an omitted response. Omitted responses in the paper-based assessment (PBA) were treated as incorrect responses and added information to the estimation. In the case of the computer-based assessment (CBA), where response times and the number of actions per item were available, nonresponses due to little or no interaction were treated differently from nonresponses after some interaction with the item took place. More specifically:

   (a) If a respondent spent less than five seconds on an item (a threshold defined in the literature on response latencies; see Setzer and Allspach 2007; Wise and DeMars 2005; Wise and Kong 2005) and showed only 0–2 actions, the nonresponse was considered not attempted and therefore excluded from estimation (similar to missing by design and not reached items).
   (b) In all other cases, omitted responses were treated as incorrect and included in the estimation. More precisely, if a respondent spent less than five seconds on an item but showed more than 0–2 actions, or if a respondent spent more than five seconds on an item (independent of the number of actions), these not observed responses were treated as incorrect responses.

Nonresponse in cases of refusal to participate or an inability to provide a written response due to a physical disability was considered as not related to the cognitive proficiencies and was therefore not included in the estimation.

3.1.3 Scaling, Linking, and Measurement Invariance

The IRT scaling in PIAAC had to provide a valid, reliable, and comparable scale for each cognitive domain to allow for meaningful group comparisons and stable trend measures. More precisely, the scaling needed to achieve the following goals:

- Linking across different sets of items and delivery modes (paper- and computer-based assessments) to provide a common scale for each cognitive domain for the *international comparison* of the average proficiencies of countries within the PIAAC cycle.
- Linking PIAAC to previous educational adult surveys (IALS and ALL) to provide a common scale for the *measurement of trends*.
- Examining and establishing the extent to which comparability or invariance of the item parameters across countries, languages, and surveys can be assumed. Only if the majority of item parameters are common (i.e. have the same characteristics) across different groups can it be assumed that the same construct is measured and groups can be compared with regard to that construct.
- Examining and establishing stable item parameters and sufficient model–data fit to achieve sufficient *reliability of the measures* to allow for accurate group comparisons. This can only be achieved by treating differential item functioning (DIF) and other sources of systematic error (such as translation deviations or technical issues) through the estimation of group-specific or unique item parameters or the exclusion of particular items.

3.1.3.1 Scaling and Linking Through Common Item Parameters

To create a common scale across countries, languages, and administration modes (paper- and computer-based modes) within one assessment cycle and across surveys over time, common sets of items must be used and linked together in the test design. More precisely, certain items were administered in both the paper-based and the computer-based branch in PIAAC (note that this pertains to literacy and numeracy items, as problem solving was available only for the CBA) as well as in different booklets/modules. Moreover, 60 items of the literacy and numeracy items administered in PIAAC came from IALS and ALL (note that numeracy was first introduced in ALL).

The initial IRT scaling was based on a large joint dataset including the data from prior large-scale adult skill surveys (IALS and ALL) and the data from PIAAC Round 1 (22 countries). A mixed 2PLM and GPCM IRT model was applied in the form of a multiple group model for a concurrent calibration of the PIAAC (and IALS and ALL) items across countries. More precisely, the IRT scaling accounted for country-by-language-by-cycle groups and estimated common (or international) item parameters across all groups. The same item difficulty and slope parameters were assumed for all groups in a first step using equality constraints in the IRT modelling.

By retaining as many common, international item parameters as possible, a high level of comparability of the IRT scales was maintained across countries, administration modes, and surveys. However, the appropriateness of the fit of these common item parameters to the empirical data had to be examined for each country and language in a subsequent step of the scaling as described in the next section.

3.1.3.2 Balancing Measurement Invariance and Model Fit Through Common and Unique Item Parameters

To ensure validity and accuracy of the measurement, the fit of the estimated common item parameters to the empirical data was examined through item fit statistics (RMSD and MD) as described above. Item-by-country interactions in the form of misfitting item parameters were examined and either treated by assigning unique (or country- and language-specific) item parameters—by relaxing the equality constraints in the scaling model—or excluded from the scaling, depending on the source of misfit (see procedures outlined in Glas and Jehangir 2014; Glas and Verhelst 1995; Oliveri and von Davier 2011, 2014; Yamamoto, 1997).

If the misfit was due to errors in the administration that were unable to be fixed, such as translation errors, items were excluded from the scaling in the affected groups. In case of group-level differential item functioning (DIF), unique item parameters were estimated for a particular country and language or a group of countries that showed DIF in the same direction. In the latter case, the unique item parameter was different from the international one, but common for the group of countries that showed similar patterns of DIF (those item parameters could be referred to as partially common). This approach was favoured over dropping the group-specific item responses for these items from the analysis in order to retain information from these responses. While the items with group-specific DIF treated with unique item parameters no longer contribute to the international set of comparable item parameters, they continue to contribute to the reduction of measurement uncertainty for the specific country and language group(s).

For countries participating in PIAAC Rounds 2 and 3 (i.e. at different time points but using the same instruments), the common item parameters obtained from the joint calibration of PIAAC Round 1, IALS, and ALL data were fixed, and their fit was evaluated as described above. Through this approach, the different countries participating in PIAAC at different time points were linked through a common scale for each domain, and their results were made comparable.

While establishing a high level of comparability (in terms of a high percentage of invariant parameters across countries) of the PIAAC scale was one of the main goals of PIAAC, achieving good model–data fit for sufficient measurement accuracy for each of the participating countries and language groups was important as well. An increasing number of unique item parameters will increase the model–data fit but decrease the measurement invariance across the relevant comparison groups. Hence, a balance between these two goals had to be achieved. In PIAAC, the majority of items received international item parameters common to all or almost all countries, while unique item parameters had to be estimated for a subset of items providing a comparable and reliable scale for group-level comparisons (more details can be found in Yamamoto et al. 2013, Chap. 17).

3.1.3.3 Software

The software used for the IRT scaling, *mdltm* (von Davier 2005), provides marginal maximum likelihood estimates (MML) obtained using customary expectation–maximisation methods (EM), with optional acceleration. Furthermore, it implements an algorithm that monitored DIF measures and that automatically generated a suggested list of group-specific item treatments for the estimation of unique parameters for an individual country-by-language group or multiple country-by-language groups that showed the same level and direction of DIF. The international and national calibrations were conducted simultaneously for all countries—that is, all estimated item parameters (common and unique) are on a common scale. During the item calibration, sample weights standardised to represent each country equally were used.

3.2 Latent Regression Model

In the latent regression model, the posterior distribution of the proficiency variable (θ) is assumed to depend on the cognitive item responses (X) as well as on a number of predictors (Y) obtained from the BQ (such as gender, education, occupation, employment status, etc.). Both the item parameters from the IRT scaling stage and the estimates from the latent regression analysis are needed to generate plausible values.

3.2.1 The Latent Regression Model

The regression uses the BQ variables to predict the proficiency variable θ. It is assumed that

$$\boldsymbol{\theta} \sim N(\mathbf{y}\Gamma, \Sigma) \tag{3.7}$$

The latent regression parameters Γ and Σ are estimated conditional on the previously determined item parameter estimates. 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 written as follows:

$$\theta_i = \gamma_{s0} + \gamma_{s1} y_1 + \cdots + \gamma_{sL} y_L + \varepsilon_s \tag{3.8}$$

where ε_s is an error term.

The residual variance–covariance matrix is given by the following equation:

$$\Sigma = \Theta\Theta^t - \Gamma\left(YY^t\right)\Gamma t \tag{3.9}$$

The conditional distribution from which plausible values for each respondent j are drawn can be written as follows:

$$P\left(\theta_j|\boldsymbol{x}_j, \boldsymbol{y}_j, \Gamma, \Sigma\right) \tag{3.10}$$

Using standard rules of probability, this posterior probability of proficiency can be represented as follows:

$$\begin{aligned} P\left(\theta_j|\boldsymbol{x}_j, \boldsymbol{y}_j, \Gamma, \Sigma\right) &\propto P\left(\boldsymbol{x}_j|\theta_j, \boldsymbol{y}_j, \Gamma, \Sigma\right) P\left(\theta_j|\boldsymbol{y}_j, \Gamma, \Sigma\right) \\ &= P\left(\boldsymbol{x}_j|\theta_j\right) P\left(\theta_j|\boldsymbol{y}_j, \Gamma, \Sigma\right) \end{aligned} \tag{3.11}$$

where θ_j is a vector of scale values (these values correspond to the performance on each of the three cognitive domains literacy, numeracy, and PS-TRE), $P(\boldsymbol{x}_j|\theta_j)$ 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 BQ responses and item parameters Γ and Σ. As described above, the item parameters are assumed to be fixed constant in the estimation.

An expectation–maximisation (EM) algorithm is used for estimating Γ and Σ; the basic method for the single scale case is described in Mislevy (1985). The EM algorithm requires the computation of the mean and variance of the posterior distribution in the equation above.

3.2.2 Generating Plausible Values

After the estimation of the regression parameters (Γ and Σ) is complete, plausible values are randomly drawn in a three-step process from the joint distribution of the values of Γ for all sampled respondents:

1. First, a value of Γ is drawn from a normal approximation to $P(\Gamma, \Sigma|\boldsymbol{x}_j, \boldsymbol{y}_j)$ that fixes Σ at the value $\hat{\Sigma}$ (Thomas 1993).
2. Second, conditional on the generated value of Γ (and the fixed value of $\Sigma = \hat{\Sigma}$), the mean $m_j{}^p$, and variance $\Sigma_j{}^p$ of the posterior distribution of θ are computed using the same methods applied in the EM algorithm.
3. In the third step, the θ are drawn independently from a multivariate normal distribution with mean $m_j{}^p$ and variance $\Sigma_j{}^p$.

These three steps were repeated ten times, producing ten independent PVs of θ for each sampled respondent in each administered cognitive domain. Each set of PVs is equally well designed to estimate population parameters; however, multiple PVs

are required to appropriately represent the uncertainty in the domain measures (von Davier et al. 2009).

Because the presence of extensive background information related to respondents' cognitive skills is necessary to implement any method for the imputation of proficiency scores, cases where respondents did not answer a sufficient number of background questions (< 5 BQ items) were considered as incomplete cases and not used in the latent regression model. These cases did not receive plausible values.

Respondents who provided sufficient background information but did not respond to a minimum of five items per domain (<2% of cases in PIAAC) were not included in a first run of the latent regression to obtain unbiased regression parameters (Γ and Σ). In a second run of 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. This procedure aimed at reducing the uncertainty of the measurement.

3.2.3 Overview of the Analytic Steps in the Latent Regression Model

The latent regression modelling in PIAAC involves multiple steps. Some involve a comprehensive analysis across all participating countries to establish international scales of literacy proficiency variables, ensuring internationally comparable results, and some involve utilising country-specific models in order to reduce bias and support country-level analyses of explanatory variables:

1. *IRT scaling:* Estimation of IRT-based common and unique item parameters (slopes and difficulties) for dichotomous and polytomous items using the 2PLM and GPCM as described in the section above.
2. *Contrast coding* of the BQ items, by contrasting each level as well as a code for missing (omitted) and routed (skipped by design) responses for each variable, creating a very large number of contrast-coded variables.
3. *Principal component analyses* of the contrast-coded variables to reduce the number of variables needed in the model and to remove collinearity. Principal components were extracted, explaining 80% of the variance represented by the background questions to avoid overparameterisation. The use of principal components also served to incorporate information from examinees with missing responses to one or more background variables. Note that the principal component analysis was conducted separately for each country based on international variables (collected by every participating country) as well as national background variables (country-specific variables in addition to the international variables).
4. *Latent regression analysis* with IRT item parameter estimates (X) treated as fixed values and the principal components of the BQ variables as predictors (Y) for estimating the latent regression parameters Γ (regression coefficients)

and $\sum$ (residual variance–covariance matrix). Note that latent regression models are estimated separately for each country to take into account the differences in associations between the background variables and the cognitive skills. The regression model for each country consisted of two steps:

(a) First, the model was estimated on a dataset that excluded cases with fewer than five responses to cognitive items to estimate the regression parameters (Γ and $\sum$).
(b) Second, the model was applied to the full dataset, including cases with fewer than five responses to cognitive items but with the regression parameters (Γ and $\sum$) fixed to the values obtained in the first step.

This ensured that the population model was calculated based on cases that included a reasonable amount of information in the domain of interest, avoiding the potential bias from poorly measured cases, while at the same time being able to then calculate scores for all respondents, regardless of the amount of cognitive information collected.

5. *Plausible values* (PVs) are randomly drawn from the resulting posterior distribution for all sampled respondents in a three-step process described below. A total of ten plausible values are independently drawn for each respondent per cognitive domain. Note that paper-based respondents have PVs only for the literacy and numeracy domains that were administered to them (i.e. paper-based respondent did not receive any PS-TRE items and hence did not receive PVs for PS-TRE). Also note that respondents with an insufficient amount of background information (i.e. less than five BQ items) did not receive PVs. The PVs that were made available in the public use file (PUF) can be used in secondary analyses of the PIAAC data.

3.2.3.1 Software

The software DGROUP (Rogers et al. 2006) was used to estimate the latent regression model and generate plausible values. In PIAAC, a multidimensional variant of the latent regression model was used that is based on Laplace approximation (Thomas 1993).

3.3 Analyses with Plausible Values

As outlined above, PVs are based on a latent regression model that was specifically designed to estimate population characteristics. They should never be used to draw inferences at the individual level, as they are not a substitute for test scores for individuals. When the underlying population model is correctly specified, PVs will provide consistent estimates of population characteristics, even though they are not generally unbiased estimates of the proficiencies of individuals (von Davier et al.

2009). Moreover, if PVs are correctly used in statistical analyses, the accuracy of derived test statistics enables fair and meaningful group-level inferences. In the following, we explain how PVs are used properly.

First, it is important to remember that the proficiency values θ for the cognitive domains cannot be directly observed and that latent variable (IRT) models had to be used to make inferences about these latent variables. Hence, we follow the approach taken by Rubin (1987) and treat the latent variable θ as missing information. Any statistic $t(\theta,y)$, for example, a scale or composite subpopulation sample mean, is approximated by its expectation given the observed data (x,y):

$$t^* (\overline{x}, \overline{y}) = E\left[t\left(\overline{\theta}, \overline{y}\right)|\overline{x}, \overline{y}\right] = \int t\left(\overline{\theta}, \overline{y}\right) p\left(\overline{\theta}|\overline{x}, \overline{y}\right) d\theta \tag{3.12}$$

It is possible to approximate t^* using PVs instead of the unobserved θ values. For any respondent, the value of θ used in the computation of t is replaced by a PV.

Second, Rubin (1987) argued that this process should be repeated several times so that the uncertainty associated with the imputation can be quantified. For example, the average of multiple estimates of t, each computed from a different set of PVs, 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. This sampling variance is another important component of the total error variance of any statistic calculated in surveys.

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 and Rust 1992) was used to estimate the sampling variability as well as the imputation variance associated with the plausible values. Variance estimates are crucial in the comparison of proficiencies across groups. In surveys such as PIAAC, several variance components are integrated into the estimate of variances, for example, the variance of the mean of literacy in a country.

The correct use of PVs to compute any statistics for an arbitrary function T and the computation of the different variance components are described in the following:

1. Calculate the statistic of interest using the first PV (i.e. the vector of the first PV across respondents). Call this T_1.
2. Calculate the sampling variance of T_1. Call this $SVar(T_1)$.
3. Repeat steps 1 and 2 for each of the remaining PVs obtaining T_2 through T_{10}, and $SVar(T_2)$ through $SVar(T_{10})$, thus obtaining T_u and $SVar_u$ for u = 1, …,10.
4. The statistic of interest, or T, would be the average of T_1 to T_{10}:

$$\mathrm{T} = \frac{\sum_{u=1}^{10} \mathrm{T_u}}{10} \tag{3.13}$$

5. The sampling variance of T is the average of $SVar(T_1)$ to $SVar(T_{10})$:

$$SVar(T) = \frac{\sum_{u=1}^{10} SVar_u}{10} \tag{3.14}$$

This sampling variance reflects uncertainty due to sampling from the population (i.e. the selection of a subset of respondents from the total population). This is potentially the largest contributor to the uncertainty of the estimated statistic.

6. The imputation variance is $Var(T_1 \text{ to } T_{10}) * (11/10)$:

$$Var(T) = \frac{\sum_{u=1}^{10} (T_u - T)^2}{10 - 1} \left(\frac{11}{10}\right) \tag{3.15}$$

This imputation variance is related to the lack of precision of the measurement instrument and reflects uncertainty because the respondents' proficiencies θ are only indirectly observed through x and y. This variance component is captured (approximately) by the variability of the PVs.

7. The overall error variance of T is sampling variance + imputation variance. An example of partitioning the error variance in the two error components (i.e. sampling and measurement error) is provided in the PIAAC Technical Report (Yamamoto et al. 2013, Chap. 17). The standard errors, or the square root of the overall error variance of the statistic T, can be used to evaluate the magnitude of the statistic. This error variance plays an important role in interpreting subpopulation results and in comparing the performances of two or more subpopulations or countries.

3.3.1 Software Tools

Different software tools based on STATA, R, SPSS, or SAS are available for utilising PVs in analysis using the procedures described above. They will be introduced and illustrated on practical examples in other chapters in this volume.

3.4 Why Plausible Values Should Be Used for Secondary Data Analyses

Plausible values (PVs) are multiple imputed proficiency values obtained from a latent regression or population model. PVs are used to obtain more accurate estimates of group-level proficiency than would be obtained through an aggregation of point estimates (Mislevy 1991; Mislevy and Sheehan 1987; Thomas 2002; von Davier et al. 2006, 2009). The aim is to reduce uncertainty and measurement error for quantities used in the analyses of large-scale surveys aiming at valid group-level comparisons rather than optimal point estimates for individual test-takers. In

contrast to tests that are concerned with the measurement of skills of individuals (e.g. for the purposes of diagnosis or selection and placement), PIAAC aims to provide group-level test scores to describe populations and subpopulations. Usually, the amount of measurement error can be reduced by increasing the number of items for each individual. However, PIAAC uses matrix sampling as well as MST for the test design, resulting in the test-taker responding to a subset of items only. The reasons for this design are described in more detail in Chap. 2 of this volume. Thus, the survey solicits relatively few responses from each respondent while maintaining a wide range of representation of the constructs when responses are aggregated. In other words, the PIAAC test design facilitates the estimation of population characteristics more efficiently, while the individual measurement accuracy is reduced.

The IRT scaling in PIAAC solves the problem of the comparability of groups responding to different set of items by placing both the items and the proficiencies on the same scale. Point estimates of the proficiencies obtained from the IRT scaling could lead to seriously biased estimates of population characteristics due to the uncertainty in the measurement (Wingersky et al. 1987). Therefore, PIAAC provides PVs obtained from the latent regression model, thereby ensuring that the group-level effects are properly controlled for in the regression, thus eliminating this bias in group-level comparisons while reducing measurement error.

3.4.1 An Example Using Plausible Values and Background Data

We will use a simulated dataset to exemplify the limitations encountered when aggregating individual 'scores' for reporting group-level results and the advantages of using an approach as described in this chapter where IRT is implemented in combination with population modelling to obtain PVs. We will also illustrate some of the risks incurred when not using the PVs properly.

The advantage of using a simulated dataset is that we know the exact values (the 'truth') on which we based our simulation, and therefore we can test whether our proposed methods give us the right results.

For our example, we generated data from nine different hypothetical proficiency groups, each responding to different sets and combinations of a total of 56 items. We chose 56 items, as this is the number of items in the PIAAC numeracy domain. The 56 items were grouped into seven blocks or subsets of eight items each. Each item is included in one, and only one, of the subsets. We chose the seven subsets with eight items each, as this would allow us to experiment with the amount of items that each individual would be asked to respond to, similar to the design implemented in PIAAC, even if not exactly the same.

Table 3.1 above shows descriptive statistics for the item discrimination and difficulty of the simulated item pool. The statistics are presented overall and block

Table 3.1 Descriptive statistics of item parameters used in the simulation

	Discrimination			Difficulty		
Block	Average	Minimum	Maximum	Average	Minimum	Maximum
A	1.19	0.57	1.50	−0.13	−1.72	1.51
B	0.94	0.50	1.47	−0.63	−1.72	0.22
C	1.09	0.76	1.39	0.22	−1.51	1.94
D	0.90	0.55	1.38	0.05	−1.71	1.45
E	1.00	0.68	1.44	0.12	−1.98	1.72
F	0.70	0.56	0.91	−0.69	−1.79	1.86
G	1.05	0.53	1.43	0.56	−0.68	1.74
Overall	0.98	0.50	1.50	−0.07	−1.98	1.94

Table 3.2 Descriptive statistics of the simulated samples

Group	Mean	Standard deviation	Number of blocks							
			0	1	2	3	4	5	6	7
1	1.02	0.76	760	2036	2103	1977	2049	2095	1946	2034
2	0.75	0.76	724	2080	2042	2067	1942	2023	2065	2057
3	0.50	0.75	745	2022	2015	2058	2029	2031	2029	2071
4	0.26	0.75	737	2036	2055	2024	2030	2035	2085	1998
5	0.01	0.76	716	2122	2026	2055	1957	2032	2028	2064
6	−0.26	0.76	797	2069	1987	2077	1970	1930	2148	2022
7	−0.51	0.75	678	2041	2053	2030	2038	2016	2052	2092
8	−0.76	0.76	752	1988	2052	2080	2037	2007	2011	2073
9	−1.01	0.75	725	2035	2042	2019	2052	2097	2007	2023

by block. While these are not exactly the item parameters of the numeracy item pool, they resemble them closely enough for the purposes of this simulation.

The nine simulated proficiency groups ranged in average 'true' ability between −1.01 and 1.02, each with a standard deviation of 0.75–0.76. They go from a high average proficiency group (Group 1) to a low average proficiency group (Group 9), with Groups 4, 5, and 6 being of about average proficiency.

In total, we generated 15,000 respondents for each one of these proficiency groups, and each of these respondents was simulated to respond to all items, or a subset of 6, 5, 4, 3, 2, or 1 block of eight items each. To further test the strength of the statistical model described in this chapter, we deleted the responses for about 5% of the cases in the simulated sample. This was done to test what would happen if we used these models to estimate the ability of groups of respondents who did not respond to any of the items in the assessment, and all we knew was their group membership.

Table 3.2 above shows descriptive statistics (mean and standard deviation) for each of the subgroups and the number of cases that responded to a particular number of blocks from the simulated assessments.

We then calculated item parameters using the combined simulated sample of 135,000 cases. The items were calibrated using Parscale Version 4.1 (Muraki and

Bock 1997), and these item parameters were used to assign scores to each of the respondents using the following methods:

(a) Expected a posteriori (EAP)
(b) Maximum likelihood estimates (MLE)
(c) Warm's maximum likelihood estimates (WML)
(d) Plausible values taking into account group membership (PV1)
(e) Taking the average of ten plausible values (PVA)

Please note that PVA scores are not (!) recommended, and they are shown in this simulation to illustrate their deficiency as a group-level score. The EAP, MLE, and WML scores were computed using Parscale Version 4.1. The PVs were computed using Dgroup (Rogers et al. 2006). The syntax for Dgroup was generated using the windows interface DESI (Gladkova et al. 2006). Notice also that for the purpose of this example, we will use only the first plausible value, although the proper way to work with these is to compute the statistics with each of these and report the average of these statistics, and the variance associated with them, as is explained later in this chapter.

The results of the simulation by proficiency group are presented in Table 3.3. In particular, notice in the panel where means are presented. While we are able to reproduce relatively well the group means using the MLE, WML, PV1, and PVA scores, the mean of the EAP scores show a consistent regression towards the overall mean. Notice also in the panel where the standard deviations are shown for the different groups that the PV1 consistently reproduces the standard deviation of the generating scores, whereas the EAP and PVA consistently underestimates them, and the MLE and WML consistently overestimated them.

The results from the simulation by number of blocks taken (each block consisting of eight items) are presented in Table 3.4. Notice in the means panel that we are not able to estimate the means using the EAP, MLE, or WML scores for those who did not take any items. However, the average overall score is reproduced with the PV1 and consequently the PVA scores. Then, looking at standard deviation panel, we see that the EAP and PVA underestimate the standard deviation as we use fewer

Table 3.3 Summary statistics of estimated means and standard deviations by proficiency group

	Means						Standard deviation					
Group	Theta	EAP	MLE	WML	PV1	PVA	Theta	EAP	MLE	WML	PV1	PVA
1	1.02	0.89	1.00	1.01	1.02	1.02	0.76	0.72	0.85	0.85	0.75	0.66
2	0.75	0.65	0.73	0.73	0.74	0.74	0.76	0.73	0.85	0.85	0.76	0.67
3	0.50	0.44	0.49	0.49	0.50	0.50	0.75	0.74	0.85	0.85	0.76	0.67
4	0.26	0.24	0.26	0.26	0.26	0.26	0.75	0.74	0.83	0.83	0.76	0.67
5	0.01	0.01	0.01	0.00	0.00	0.00	0.76	0.75	0.85	0.84	0.76	0.68
6	−0.26	−0.22	−0.24	−0.24	−0.25	−0.25	0.76	0.74	0.84	0.84	0.75	0.67
7	−0.51	−0.45	−0.50	−0.51	−0.51	−0.51	0.75	0.74	0.84	0.84	0.75	0.67
8	−0.76	−0.68	−0.74	−0.76	−0.77	−0.76	0.76	0.75	0.85	0.85	0.76	0.67
9	−1.01	−0.89	−0.98	−1.00	−1.02	−1.01	0.75	0.73	0.83	0.83	0.75	0.67

Table 3.4 Summary statistics of estimated means and standard deviations by number of blocks

	Means						Standard deviation					
Number of blocks	Theta	EAP	MLE	WML	PV1	PVA	Theta	EAP	MLE	WML	PV1	PVA
0	0.00				−0.01	0.00	0.99				1.00	0.69
1	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.83	1.05	1.13	1.00	0.87
2	−0.01	0.00	0.00	−0.01	0.00	0.00	1.01	0.91	1.08	1.09	1.01	0.93
3	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.93	1.06	1.05	0.99	0.94
4	0.00	0.00	0.01	0.00	0.00	0.00	1.01	0.97	1.07	1.06	1.01	0.97
5	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	1.01	0.97	1.06	1.05	1.01	0.97
6	0.00	0.01	0.01	0.01	0.01	0.01	0.99	0.96	1.04	1.02	0.99	0.96
7	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.97	1.04	1.02	0.99	0.97

items, and even if all 56 items are used, the standard deviation is underestimated. On the other hand, the MLE and WML scores consistently overestimate the standard deviation. The only score type that estimates the means and standard deviations consistently, regardless of the number of items used in the estimation, is the PV1 score.

As can be seen from the tables presented above, we are able to reliably reproduce the mean and standard deviation for groups of different abilities, regardless of the proficiency level with respect to the average item difficulty, and also regardless of the number of items that are administered, to the extreme of being able to estimate the mean and standard deviation of the proficiency even in the case when no items are administered, and all we know is the group membership of the respondent.

3.5 Summary and Outlook

PIAAC uses a latent regression model to estimate plausible values (PVs) by incorporating item responses and background data. These can be used by researchers, policymakers, and stakeholders to conduct research in the area of adult competencies (including literacy, numeracy, and problem solving in technology-rich environments) and their relation to economy and society. The latent regression model uses item parameters of test items obtained from IRT scaling as fixed values and background variables obtained using a principal component analysis of contrast-coded background questionnaire items as predictors.

PVs are multiple imputations that are randomly drawn from the posterior proficiency distribution resulting from this modelling approach and are designed to facilitate comparisons at the group level to describe population and group-level characteristics. They should never be used to draw inferences at the individual level. PIAAC provides ten plausible values for each cognitive domain for all respondents with sufficient background information (i.e. responses to five or more BQ items). PVs provide less biased and more accurate measures than point estimates

can for group-level comparisons and allow consistent estimates of population characteristics. If used correctly in statistical analyses as described above, they provide fair and meaningful results and subgroup comparisons and allow variance estimation accounting for measurement and sampling error.

In the first cycle of PIAAC, the latent regression model is based on item parameters and background variables only. However, the modelling approach can be improved in future cycles by including process or logfile data, such as response times and the number and sequence of actions (mouse clicks and interactions of the respondent with the test item), which are available in the computer-based assessment branch (e.g. Shin et al. 2018). Especially, since future PIAAC cycles will likely move the current paper-based assessment branch to a tablet administration mode (at least for the majority of test-takers), process data will be available for even more respondents. Moreover, more simulation-based tasks might be developed to better assess life-relevant skills and new aspects of the PIAAC framework (such as adaptive problem solving in the second cycle of PIAAC). Including additional process data information into the latent regression model may further decrease the bias related to measurement error and increase the accuracy of PVs (von Davier et al. 2019), especially at the extreme ends of the proficiency scale and for lower-performing countries and subgroups (Shin et al. 2018). However, the option of including additional variables in the already extensive latent regression model is challenged by the problem of overparameterisation and requires careful considerations and additional research before being considered for operational procedures (von Davier et al. 2019).

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: Addison-Wesley.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1997). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*, 1–38.

Gladkova, L., Moran, R., & Blew, T. (2006). *Direct Estimation Software Interactive (DESI) – Manual*. Princeton: Educational Testing Service.

Glas, C. A. W., & Jehangir, K. (2014). Modeling country specific differential item functioning. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment*. Boca Raton: CRC Press.

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 69–95). New York: Springer.

Johnson, E. G. (1989). Considerations and techniques for the analysis of NAEP data. *Journal of Educational Statistics, 14*(4), 303–334. https://doi.org/10.3102/10769986014004303

Johnson, E. G., & Rust, K. F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics, 17*, 175–190. https://doi.org/10.2307/1165168

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal 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. https://doi.org/10.1007/BF02294457

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: Educational Testing Service.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–177. https://doi.org/10.1177/014662169201600206

Muraki, E., & Bock, R. D. (1997). *PARSCALE: IRT item analysis and test scoring for rating-scale data [Computer software]*. Chicago: Scientific Software.

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling, 53*(3), 315–333. Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011_20110927/04_Oliveri.pdf

Oliveri, M. E., & von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing, 14*(1), 1–21. https://doi.org/10.1080/15305058.2013.825265

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

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

Rutkowski, L., von Davier, M., & Rutkowski, D. (Eds.). (2014). *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. Boca Raton: CRC Press.

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. http://www.psyc.jmu.edu/assessment/research/pdfs/SetzerAllspach_NERA07.pdf

Shin, H. J., Khorramdel, L., von Davier, M., Robin, F., Yamamoto, K. (2018). *Incorporating response time into population modeling for large-scale assessments*. Paper presented at the conference of the International Test Commission (ITC), Montreal, 2.–5. July 2018.

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics, 2*, 309–322. https://doi.org/10.1080/10618600.1993.10474614

Thomas, N. (2002). The role of secondary covariates when estimating latent trait population distributions. *Psychometrika, 67*(1), 33–48. https://doi.org/10.1007/BF02294708

Van der Linden, W. J., & Hambleton, R. K. (2016). *Handbook of modern item response theory* (2nd ed.). New York: Springer.

Von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report No. RR-05-16). Princeton: 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., Khorramdel, L., He, Q., Shin, H., & Chen, H. (2019). Developments in psychometric population models for technology-based large-scale assessments: An overview of challenges and opportunities. *Journal of Educational and Behavioral Statistics, 44*(6), 671–705. https://doi.org/10.3102/1076998619881789

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 & S. Sinharay (Eds.), *Handbook of statistics (Vol. 26): Psychometrics* (pp. 1039–1056). 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: 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. https://doi.org/10.1207/s15326977ea1001_1

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. https://doi.org/10.1207/s15324818ame1802_2

Yamamoto, K. (1997). A chapter: Scaling and scale linking. In *International Adult Literacy Survey technical report*. Ottawa: Statistics Canada.

Yamamoto, K., Khorramdel, L., & von Davier, M. (2013). Scaling PIAAC cognitive data. In OECD (Ed.), *Technical report of the survey of adult skills (PIAAC)* (pp. 406–438). Paris: OECD Publishing.