Design of PIAAC survey of adult skills and its implications to data analysis

Introduction

The Programme for the International Assessment of Adult Competencies (PIAAC) is an OECD survey measuring key information-processing skills in literacy, numeracy and problem solving among the adult population aged 16–65. The first cycle of PIAAC was conducted in three rounds (2011–2012, 2014–2015 and 2017) in 37 countries. The first round of the second cycle was conducted between 2022 and 2023 in 31 countries and its data will be published in December 2024 in conjunction with the first international report.

PIAAC data are made openly available for research purposes. They provide unique opportunities to analyse adults' cognitive skills needed in the labour market and society. Moreover, PIAAC data make it possible to analyse the connection between cognitive skills and a range of variables, such as educational attainment, participation in formal and non-formal education, labour market status, income, health, and the use of skills at work and outside work.

Valid statistical analyses of PIAAC data require methodology which respects relevant characteristics of the survey. First, the sampling design used in collecting the data must be considered, as the appropriate approach to estimation of variances and standard errors depends on it. Second, in PIAAC as well as in many other international assessments, the cognitive skills are assessed using an approach based on so-called plausible values (Wu 2005). A valid analysis of plausible values data requires specific approach, which has its roots in the methodology of multiple imputation and analysing multiple-imputed data (Rubin 1987).

This document aims to serve as an introduction to the key issues mentioned below:

1. What kind of survey designs have been adopted in PIAAC, what kind of methodology is available for analyses which successfully take the design into account, and what are the consequences if the design is not respected in the analyses.
2. What are the plausible values, why they have become the standard approach in the international large-scale assessments of skills, and how valid analyses of plausible values data are conducted.

In addition, references to available technical tools for valid analysis of PIAAC data are given.

The purpose of this document is not to present the detailed theory behind the above issues. Rather, it aims to introduce the principles, on which the methodological solutions supported in PIAAC are based. Neither this document aims to serve as a user manual for performing valid analyses of PIAAC data with specific software. This is not advisable, because the choice of available software changes continuously: new tools and new versions of old tools come out, while some tools may disappear, or at least they will not be supported any more. The exact and up-to-date information on how to perform the desired analyses with chosen software, must be fetched from the manuals and handbooks of that software.

Table of content

Importance of recognizing survey design 3
Stratification and clustering in PIAAC 4
Survey weights 5
Model-based approach to analysis of PIAAC data 6
Design-based approach to analysis of PIAAC data 7
Jackknife variance estimation in PIAAC 8
Plausible values in measuring adult skills 10
Summary of software for analyzing PIAAC data 13
References 15

Importance of recognizing survey design

International large-scale assessment studies (including PIAAC, PISA, PIRLS, TIMSS, etc.) follow strictly planned complex sampling designs in their data collection. The survey designs almost invariably involve stratification of the target population, and, in many cases, clustering. Furthermore, the inclusion probability of an individual in the sample typically varies across the population in these studies. This results from stratification and specific sampling mechanisms. A consequence of the unequal inclusion probabilities is that the data as such can be distorted: Some subgroups in the population can be overrepresented at the cost of other subgroups, thus introducing bias in the data. Similarly, bias in the data is often caused by non-response, which commonly is not random, but associated with characteristics of the individuals in the sample. Thus, there can be systematic differences in realized response rates across subgroups of individuals, which further manifests in distortions in the collected data, with respect to the target population.

The problems in terms of representativeness, resulting from both sampling and non-response, can be corrected by employing survey weights in data analysis. The weights are derived according to the sampling design, and auxiliary information from statistical registers is further employed to reduce bias due to non-response. By appropriate survey weights it can be ensured that the results based on sample data are representative for the population and the possible biases in the results are small.

Recognizing the features of the sampling design is essential when performing statistical analyses on assessment data. In particular, the standard errors of the statistics of interest can be severely inaccurate if the characteristics of the design and realized response pattern are ignored. Additionally, the estimates of population parameters can be biased, which in turn may lead to inappropriate statistical inference. To obtain accurate standard errors, it is particularly important to account for clustering because individuals within a cluster tend to be more homogeneous than individuals coming from different clusters. The common assumption of independence of observations, required for valid statistical inference in standard statistical analyses, no longer holds. If the within-cluster homogeneity is not considered in the analysis, standard errors will be underestimated.

A valid analysis of data collected by a complex sampling design calls for specific methodology, which accounts for stratification, clustering, and the characteristics of the eventual group of respondents, and correspondingly, software into which this methodology is implemented. In addition, weights should be used to correct distortions caused by unequal inclusion probabilities and non-response. The key issue is to obtain an appropriate estimate of the sampling variance (and standard error as its square root) of the statistic of interest, to enable valid statistical inference concerning the study findings.

There are two possible approaches to analyse complex sample data: Design-based and model-based. In the context of international large-scale assessments, such as PIAAC, the design-based approach is generally supported, and references to appropriate computing tools are provided. The advantage of design-based inference is that it can match completely with all features of the survey design (stratification, clustering, unequal inclusion probabilities). In particular, the standard errors obtained by design-based inference are practically unbiased. The disadvantage of the design-based approach is that the range of options for statistical analyses is rather limited, containing typically just descriptive statistics, linear and logistic regression, and analysis of

cross-tabulations. This contrasts the model-based approach in which practically all kinds of statistical analyses are available: Linear and generalized linear models (e.g., linear and logistic regression), multilevel and mixed models, multivariate statistics (e.g., structural equation models), survival analysis, and so on. In addition, every well-known statistical software contains a wide range of procedures for model-based analysis, making the tools for model-based inference easily available. The disadvantage of the model-based approach, however, is that it cannot always be applied successfully to complex survey data. In particular, the obtained standard errors can be biased. Thus, the model-based approach is not generally recommended to be used with large-scale assessment data like PIAAC, where survey design is almost always complex. However, it is sometimes useful to perform model-based analysis anyway. One such situation occurs when there is no desired analysis available under the design-based framework. Because of its ease of use (no need for special software), the model-based approach may also be convenient for conducting preliminary analyses, to get an approximate idea of possible findings based on the data – for example, to identify the variables that seem important in explaining variation in specific outcome variables. Both approaches will be discussed in more detail below.

This introduction draws mainly from the Technical Report of PIAAC Cycle 1 (Third Edition, OECD 2019), which is a comprehensive reference to all topics discussed in this article. It is freely available on OECD's PIAAC website: PIAAC_Technical_Report_2019.pdf (oecd.org)

Stratification and clustering in PIAAC

Due to different national circumstances, sampling designs in PIAAC vary across countries. Consequently, sampling variances cannot be estimated in the same way for all countries. However, practically all countries in PIAAC Cycle 1 employed stratification in their design. Stratification improves the national representativeness of the sample by ensuring the inclusion of all the relevant subpopulations (specified by the strata). In stratified sampling, a random sample from each stratum is selected separately, and then these samples are combined to obtain the actual national sample. The stratification criteria may vary across countries. The variables used most often in stratification in PIAAC Cycle 1 were age group and gender. Geographical or administrative region, educational level and immigration status were also used in many countries as stratifying variables. The PIAAC Technical Report contains explicit information on how the stratification was implemented in each participating country.

In general, stratification reduces the random variation concerning the composition of the sample data as it guarantees that the data contain individuals from every stratum. This further increases the efficiency of statistical analysis of the data. Because the population and sample sizes are seldom equal across the strata, the stratification may cause systematic differences in the inclusion probabilities of the sampling units. The resulting bias is corrected in data analysis by using survey weights derived under the stratified design. The reduced random variation in the sample composition is best utilized in data analysis by employing approaches that fully take the stratification into account in the computation of standard errors. Replication methods, in particular the jackknife estimator JK2 supported in PIAAC and introduced later in this article, are examples of this.

Several countries in the first cycle of PIAAC employed stratified multi-stage clustered sampling in their data collection, with geographically determined areas (e.g., provinces, municipalities, or localities) typically serving as clusters. Such design is particularly advisable in countries that are geographically large. In the first stage of multi-stage sampling, the primary sampling unit is

clusters, and a number of these are sampled from the population of clusters. In a two-stage design, a random sample of persons is then drawn from these clusters. Some participating countries employed three-stage or even four-stage designs, where the sampling was conducted in three or four levels of hierarchy. An example of a country with a three-stage design is Italy, where municipalities were sampled in the first stage, households from the municipalities in the second stage, and persons from the households in the third stage. An example of a country with a four-stage design is Czech Republic, where districts were sampled in the first stage, streets from the districts in the second stage, dwelling units from the streets in the third stage, and finally persons from the dwelling units in the fourth stage.

In fact, only eight countries used one-stage stratified sampling in the first cycle of PIAAC. That is, there was no clustering present in their sampling design and, consequently, in the data. The only sampling unit in this design is persons. The countries Denmark, Estonia, Finland, Norway, and Sweden, included in the Nordic-Baltic PIAAC network, all used this type of sampling design.

As noted above, from the viewpoint of data analysis it matters whether the data are clustered or not. The independence of observations is a basic assumption in several traditional approaches to statistical data analysis. In a clustered sample, however, the observations within a cluster cannot be assumed independent because the individuals that belong to the same cluster (e.g., school, workplace, region, neighbourhood) are exposed to similar circumstances and, consequently, tend to be more homogeneous. Therefore, their values of the outcome variables tend to be positively correlated (this correlation is commonly called intra-cluster correlation or ICC). Basic statistical methods fail to recognize this correlation or homogeneity, which causes the sampling variances and, consequently, standard errors to be underestimated, and the resulting statistical inference becomes overly liberal. To put it differently, statistical significance of the findings is overestimated.

In case of clustered data, valid statistical methods account for clustering either by adding the cluster variable(s) in the model used in data analysis (model-based approach) or by utilizing the clustering hierarchy of the sampling design in computing standard errors of the statistics (design-based approach).

Survey weights

In complex sampling designs, the probability of being included in the sample is not necessarily equal for all individuals in the population of interest. That is, certain individuals may be sampled with higher probability than others, depending on the design. For instance, individuals within a certain stratum can be selected with higher probability than individuals in other strata to obtain a sufficiently large data set on the specific subgroup. This is typically when the stratum in question is small, but represents an important minority (e.g., immigrants) in the population. Individuals with higher selection probabilities will then be overrepresented in the sample (while others will be underrepresented, respectively). An additional distortion may be caused by the fact that the extent to which the selected individuals participate in the survey differs. It is consistently observed in survey studies that individuals with certain characteristics tend to have lower response rates than others (e.g., young men as opposed to older women). Accordingly, subgroups with large non-response rate will be underrepresented in the data. These distortions can be corrected by applying survey weights in data analysis, together with auxiliary information from statistical registers. In general, derivation of survey weights is complicated. They are based on inverses of inclusion probabilities generated by the sampling design, which are then adjusted

to correct for non-response and calibrated to make certain pre-defined totals of the weighted data match with those of the target population. The survey weights take sample design and response pattern into account, through using information about both respondents and non-respondents from statistical registers.

In international assessment studies like PIAAC, the weights are ready-computed and provided in the public-use data sets, to be used in analyses. In the PIAAC data sets, the variable containing weights is named SPFWT0. The weights make the collected data represent the target population accurately. In analysing PIAAC data, the weights should always be applied, because the estimated statistics will be biased with high probability without weighting.

Model-based approach to analysis of PIAAC data

As stated above, there are two general approaches to analyze survey data: the model-based approach and the design-based approach. Most of the traditional statistical methods presented in textbooks and taught in statistics courses belong to the model-based category. In the model-based approach, it is assumed that the observed data is a random sample from an imaginary infinite population, so-called superpopulation, which is characterised by some probability model (e.g., normal distribution) with certain mean and variance. The real population, from which the actual sample is drawn, is considered a realization of the superpopulation. Under this approach, the sampling variances and standard errors needed in the inference are derived by fitting a postulated model, with an appropriate distributional assumption, to the observed data. The model parameters and standard errors are estimated by methods derived under this assumption. The maximum likelihood estimation with its variants is probably the most widely used example of such method. Typical examples of model-based approaches to data analysis are the t test, linear regression and analysis of variance, logistic regression, or structural equation modelling. All well-known statistical packages like SPSS, SAS, STATA, MPLUS and R contain procedures for performing model-based analyses.

Under the model-based approach, there are some possibilities to account for the features of sampling design, but they can be limited. Survey weights can be incorporated in the analysis easily, but they should be rescaled to match with the number of observations in the realized data set before using them in procedures for model-based analyses. Regarding PIAAC, the weight variable SPFWT0 adds up to the size N of the target population, not to the size n of the realized set of responses. A consequence of this is that the model-based analysis procedures with weights SPFWT0 erroneously assume that there are N observations available instead of n. This leads to overly liberal inference, i.e., significance tests invariably give p values that are practically zero. This happens with procedures designed for model-based analysis in general, regardless of software. The solution is to replace the original weight SPFWT0 with a rescaled weight, which adds up to n (the mean of these rescaled weights equals 1). In PIAAC, the rescaling is done by the transformation RESCALED_WEIGHT = n/N*SPFWT0.

Clustering is handled by applying multilevel models or mixed regression-type models, where random effects of the cluster are added to the model. The simplest and often sufficient way to do this is to incorporate a random intercept for each cluster in the model. The stratification variables could also be added to the model as fixed controlling variables, but this may not be necessary, as some features of stratification can also be handled implicitly by the weights.

Design-based approach to analysis of PIAAC data

As mentioned earlier, the design-based approach is the primary approach supported in PIAAC. It strictly accounts for all relevant features of the national sampling designs in PIAAC, that is, stratification, clustering, and unequal inclusion probabilities.

Survey weights should always be applied in analyses of PIAAC data as they correct for bias in parameter estimation resulting from both unequal selection probabilities in the design and non-response. Using weights is similar in both model-based and design-based approaches, except that rescaling of weights is not needed when using procedures tailored for design-based estimation. Several well-known statistical packages contain such procedures. In SPSS and SAS they are known as procedures for complex samples and procedures for survey analysis, respectively. R and STATA have special modules for the design-based approach as well.

When weights are applied, both model-based and design-based approaches produce unbiased point estimates of statistics like means, proportions, standard deviations, and regression coefficients. The crucial difference between the approaches is the estimation of sampling variance and, consequently, the standard error of those point estimates. As the design-based estimation matches completely with the properties of survey design, it is the recommended approach to be used with large-scale assessment data whenever possible.

In general, the design-based approach assumes that the random variation associated with the sample is solely determined by the sampling mechanism applied to the target population of known number of individuals (e.g., adult population aged 16–65 living in a country). The target population is not characterised by any statistical model or distributional assumptions (e.g., normal distribution with certain parameters). The underlying idea is that if one repeatedly drew random samples from a fixed population using the same survey design every time (possibly containing stratification, clustering, etc.), there would in any case be variation between the statistics computed from these samples. The design-based variance estimate tries to measure this variation. The challenge is that only the actual sample is available, because repeated sampling is not possible in practice.

Because no distributional assumptions are given, there are usually no precise closed-form computation formulas for sampling variance under complex survey designs. Instead, approximate methods must be used. A widely used method is based on Taylor series linearization. It is implemented in several popular statistical packages like SPSS (procedures for complex samples), SAS (survey analysis procedures), R and STATA. Another option is to use replication methods, which try to imitate the underlying idea of repeated sampling using the actual sample data only. The common idea of these methods is to create a series of subsamples or replicate data sets from the actual sample according to a specific scheme, which is consistent with the survey design of the actual sample. As each replicate data set respects the original design, the features of this design are incorporated in the estimation. The desired statistics are then computed from each replicate data, and the variation between these statistics is used in estimating sampling variance. Such methods include bootstrapping, balanced repeated

replication (BRR), and jackknife (JK). Of these, jackknife was the method supported in PIAAC Cycle 1[1].

The jackknife estimation can be performed with several computing tools. We have already mentioned SAS survey analysis procedures, the MPLUS software, and certain modules in STATA and R environments. Additionally, the PIAAC consortium has provided a SAS macro called PIAAC_Tool for design-based analysis. The International Association for the Evaluation of Educational Achievement (IEA) has developed a free software tool called *IDB Analyzer* (https://www.iea.nl/data-tools/tools) which can be used to combine and analyze data from various large-scale assessments, including PIAAC, under the design-based approach. It is an external collection of macros build to communicate with SPSS, SAS, and R software, thus requiring that at least one of these packages is available.

Jackknife variance estimation in PIAAC

In this section, the focus will be on the jackknife variance estimation approach supported by PIAAC. This approach is fully consistent with the sampling designs adopted in the participating countries, including stratified designs and cluster designs.

The idea of jackknife is to imitate repeated sampling by removing a subset of the observations in the realized full sample (i.e., the complete set of observed responses) by turns, according to a given scheme, to produce a specified number, say g, of slightly varying subsamples from the original realized full sample. Except for the removed subset, these g subsamples contain exactly the same observations. A different subset is removed in each subsample. The remaining observations are then reweighted so that in each subsample the sum of new weights – so called replicate weights –equals the sum of weights in the full sample.

In each PIAAC country, the scheme of removing subsets was specified to match with the survey design implemented in that country. Each subsample can be considered a realized random sample of the population. For instance, if the full sample was drawn by a stratified design from the target population, all subsamples would also be stratified samples of that target population. In this sense the jackknife approach is truly design-based. The subsample sizes are slightly smaller than the size of the full sample, but these (usually small) differences are eliminated from the estimation by reweighting.

The desired statistic is then computed from the full sample and from each subsample. As a result, there are 1+g estimates of the desired statistic. The original weights are used in computations with the full sample, while the replicate weights are used with the subsamples. The sampling variance of the statistic is estimated through the sum of squared deviations between the full sample estimate and the subsample estimates. The standard error of the statistic is naturally the square root of the estimated sampling variance.

In PIAAC Cycle 1, the number of replicates was g=80 in most countries. This holds for the Nordic countries Denmark, Finland, Norway, and Sweden, as well as the Baltic countries Estonia and Lithuania (Lithuania participated in the second round of Cycle 1, while the others participated in

[1] According to advance information, Fay-modified BRR will be the method supported in Cycle 2. This document will be updated accordingly after the Cycle 2 data have been published.

the first round)[2]. To create the subsamples, the full sample of each country was divided into 80 subsets. In the first subsample, the observations within the first subset were removed (i.e., their weights were set to zero) and the observations in the other 79 subsets were reweighted so that the sum of their (replicate) weights equals the corresponding sum in the full sample. In the second subsample, the observations within the second subset were removed and the rest of the observations were reweighted. This procedure is repeated for each subset to obtain 80 desired subsamples.

In practice, the jackknife resampling is organized in PIAAC so that the consortium, or alternatively, the participating countries themselves, have produced the 80 vectors of replicate weights for each country in advance, and provided them as variables in the public-use data sets to be used in data analysis. In each country, the replicate weights are based on the national survey design. The names of replicate weight variables in the data are SPFWT1–SPFWT80. (Recall that the full sample weight is in variable SPFWT0.) As the vectors of replicate weights are readily available in the data, there is no need to create explicit subsamples at the actual data analysis stage. Instead, the jackknife estimation is performed by conducting the desired data analysis first with the weight SPFWT0 to obtain the full sample estimate. Then the same analysis is replicated 80 times using the weights SPFWT1–SPFWT80. The standard error of the full sample estimate is finally computed from the observed variation across the 80 replicate estimates.

In PIAAC Cycle 1 there were two variations of jackknife in use: Delete-one jackknife or random-groups jackknife (JK1) and paired-jackknife (JK2). The version used depended on the sampling design of the country in question. Of the Nordic and Baltic countries in PIAAC Cycle 1, only Denmark adopted JK1 while all other countries adopted JK2.

JK1 is slightly simpler than JK2 as it does not explicitly utilize the information of strata. In JK1, the full sample is randomly divided into 80 subsets of almost equal size. These subsets are called variance units. The first replicate sample (or to be precise, the replicate weight SPFWT1 in the data) is formed by recoding the original weights SPFWT0 of the observations in the first variance unit to zero and adjusting the weights of the other observations to add up to the total of original weights. This procedure is repeated for all variance units, to obtain 80 vectors SPFWT1–SPFWT80 of replicate weights.

In JK2, the full sample is first divided into 80 subsets called variance strata. The variance strata are nested within the strata employed in the national sampling so that it explicitly respects the actual stratification used in the sampling design. If possible, the variance strata are of almost equal size. Then, each variance stratum is randomly divided into two halves, which are now called variance units. Within a variance stratum, the weights of observations in one variance unit are set to zero, while the weights of the other variance unit are doubled. Like in JK1, these weights as well as the weights in the other variance strata are finally adjusted to add up to the total of original weights. In turn, this procedure is repeated for all 80 variance strata to obtain 80 vectors of replicate weights.

Recall that in all national PIAAC data sets the full sample weight variable is named SPFWT0 and the replicate weight variables are named SPFWT1–SPFWT80. In addition, the data sets contain

[2] The four Nordics countries Denmark, Finland, Norway, and Sweden, as well as the three Baltic countries Estonia, Latvia, and Lithuania are participating in PIAAC Cycle 2. Latvia did not participate in Cycle 1.

variables VARSTRAT and VARUNIT, which represent the variance strata and variance units used in jackknife. If a country has used JK1, the values of VARSTRAT are missing. The constant variable VEMETHOD indicates which version of jackknife has been used in the country, and its possible values are JK1 and JK2. There is also a variable called VEMETHODN in the data, which is a numeric version of VEMETHOD and it has the values 1 (=JK1) or 2 (=JK2). The constant variable VENREPS indicates the number of replicates used in jackknife. In most countries, its value is 80. At least some of these variable names must be given to the computing tools when running the jackknife estimation. However, some tools tailored for PIAAC, such as the IDB Analyzer, read these variable names automatically from the data so that the user does not need to provide them explicitly.

Plausible values in measuring adult skills

A methodological speciality of international large-scale assessments, including PIAAC, is that individuals' proficiency, e.g., in literacy, numeracy and problem solving, is measured with so-called plausible values. These are fairly complicated quantities, which cannot be interpreted simply as scores measuring an individual's success in a skills test. Rather, they are predictions of the individual's skills that utilize information on both his/her test success and background characteristics.

There are good reasons for using plausible values instead of raw test scores. To start with, it is important to recognize that the proficiency of an individual is a latent variable that cannot be observed directly. Instead, it is measured as performance to a collection of test items delivered to the individual. However, it is difficult to measure the proficiency of an individual comprehensively in limited response time and with a limited number of test items. In addition, in a single test the respondent might underachieve or overachieve with respect to his/her true ability. This implies that there is a lot of uncertainty in the proficiency measurement. This uncertainty is captured well with the plausible values approach.

Second, it has been proven that the plausible values approach (unlike many simpler approaches) gives approximately unbiased estimates of national variance. Instead of assessing individuals, the goal of international assessments like PIAAC is to map the national proficiency distribution, both totally and in subgroups of the population, as comprehensively as possible. In particular, the goal is to estimate the national variation in proficiency as reliably as possible. As skills such as literacy and numeracy are multifaceted areas, obtaining a thorough picture of the national distribution of skills requires a large collection of questions with various themes and varying difficulty. However, from a practical perspective it is not advisable if the test requires too much time from a respondent. To restrict the time needed for answering the questions, the assessment in PIAAC (and in other international assessments too) is organized so that each individual receives only a subset of all possible tasks, and different individuals receive different subsets. When the test results of individuals are then put together, a comprehensive picture of the national proficiency distribution (rather than individual skills) is obtained. This practice is called matrix design or item rotation. In it, the test items are divided into testlets or blocks, which contain both easier and more difficult items, and these cover various themes. In PIAAC Cycle 1, each respondent received four blocks of tasks. Each block contained 3–6 tasks. The assessment design of PIAAC was constructed so that the duration of test was 60 minutes, on average.

Because different respondents receive different tasks, it is possible that some receive more difficult tasks than others. In determining the final proficiency, an item analysis is conducted for

all tasks. In particular, the difficulties of the tasks are estimated from the international data and considered when estimating proficiency. Success in more difficult tasks corresponds with higher estimated proficiency than success in easier tasks. With help of item analysis, the results from different block combinations can be transformed into a comparable scale. Through the international scaling, the results from different countries also become comparable.

Due to the matrix design, the skills of each individual are assessed only with a relatively small collection of tasks, which leaves much uncertainty about the true level of skills of the individual. The idea of plausible values is to improve the estimation by using information on the background variables to strengthen the inference. This can be understood as a way to impute missing information by external knowledge. Because the sampled individuals in PIAAC answer a background questionnaire before moving to the cognitive assessment, the data contain rich information on the respondents' background. The correlations of the background variables with the assessed proficiency can be estimated from the data. This is called latent regression, as proficiency is actually latent and observed through a small number of test items only. The correlations help to improve the estimation of the national skills distribution. For instance, it could be observed from the data that young respondents, or individuals with higher educational background, tend to have good success in the relatively short cognitive test. It could then be further assumed that on average, their latent skills would also appear high if the cognitive test would have been more comprehensive. As mentioned above, the plausible values are actually predictions of the individual's skills, conditioned on both his/her test success in the received items and background characteristics. However, at the individual level they are biased, because using background variables moves the estimate derived only from the individual's cognitive test result towards the average estimate of individuals with similar background. A plausible value is an estimate of the expected level of proficiency for individuals who have succeeded in the cognitive test in a certain way and have certain background characteristics. As the data also give information on the national distribution of background variables, plausible values give a better picture on the national proficiency distribution despite individual-level bias. It has been proven that through plausible values it is possible to estimate the variation of skills reliably at the group level, and particularly at the national level (Wu 2005).

The computing of plausible values employs Bayesian statistics. First, a national probability distribution of proficiency is defined. Typically, this distribution is assumed normal, meaning that the average levels of proficiency are the most common in the population, while the extremely high and low proficiency levels are rare. Additionally, it is assumed that certain kinds of individuals tend to have higher/lower proficiency than others, and such differences are associated with various background characteristics. This is the basis of latent regression. The associations are not specified in advance, but they are estimated from the observed data simultaneously with the plausible values. After the data have been collected, i.e., the cognitive test results and the answers to the background questionnaire are available for estimation, they are used in adjusting the prior distribution to a posterior distribution. This is the probability distribution of the latent proficiency after the test results and background variables have been observed. The posterior distribution is derived separately for each individual and indicates the likely level of skills in a quantitative form for individuals with certain test success and background characteristics.

The fact that an individual's estimated proficiency is expressed in the form of probability distribution, instead of a single point estimate, helps to adequately capture the uncertainty related to assessing one's skills through a cognitive test limited to 60 minutes.

The individual posterior distributions cannot be expressed in a closed form, but numeric realizations from them can be generated with Monte Carlo simulation. These numeric realizations are called plausible values, and they serve as outcome variables, through which the national distributions of proficiency and its associations with background variables can be analysed empirically.

In PIAAC, ten plausible values are generated for each individual in the response set from each assessed domain (e.g., literacy, numeracy and problem solving). In other words, ten plausible values are randomly drawn from each individual's posterior distribution. Each plausible value is a potential estimate of individual's proficiency in that domain. Through the variation of these ten plausible values the uncertainty of the assessment becomes visible and measurable. The plausible values are generated by the consortium and readily available in the public-use PIAAC data sets. In the PIAAC Cycle 1 data sets, the variables PVLIT1–PVLIT10 contain the plausible values for literacy, the variables PVNUM1–PVNUM10 contain the plausible values for numeracy and the variables PVPSL1–PVPSL10 contain the plausible values for problem-solving in technology-rich environments[3].

A valid statistical analysis of plausible values requires a specific approach, which has its roots in Rubin's methodology of multiple imputation of missing values (Rubin 1987). This methodology is further discussed below.

The idea behind the analysis of plausible values is that each plausible value is regarded as an imputed value of the missing latent proficiency. The ten plausible values are seen as ten different imputations of this unobserved proficiency. Following the principles of analysis of multiple imputed data, the same analysis is performed in PIAAC separately with each plausible value as the target variable. This gives ten analysis results (e.g., means, proportions, regression coefficients, and so on), one for each plausible value. The ten results are then averaged to obtain the final estimation result. The sampling variance of the final estimate is obtained by averaging the ten sampling variances obtained from the ten analyses and adding a variance component measuring the variation between the ten analysis results (i.e., estimates of the desired statistics) to this. The latter term is commonly called imputation error variance, as it quantifies the extra variation resulting from using ten different plausible values in the same analysis, instead of just one observed variable. The standard errors of the final estimates are the square roots of these sampling variances. The exact computation formulas are given in PIAAC Technical Report, Chapter 15.2.3 (OECD 2019).

Using all ten plausible values is essential for obtaining valid standard errors and valid inference. If only one plausible value is used, standard errors will be underestimated, because the imputation error variance is ignored. Consequently, significance tests may be too liberal; that is, statistically significant effects are found too easily. However, sometimes it is convenient to conduct some approximate analyses with a single plausible value, for instance, when computing tools for handling multiple plausible values are not available, or when a quick preliminary analysis is desired (the procedures for multiple imputation analysis can be complicated and slow to run). Although the results are strictly incorrect, they can be close to the correct ones obtained from all ten plausible values and therefore they can serve as guidance for further steps in the data analysis, e.g., when deciding which variables should be included in the modelling. The final

[3] In PIAAC Cycle 2, problem-solving in technology-rich environments was replaced with adaptive problem solving.

data analysis, however, should be carried out with all ten plausible values, whenever possible, to ensure that the obtained results and related conclusions are valid.

The computing tools which are tailored for analysis of large-scale assessment data, normally contain facilities also for handling plausible values. A summary of these is given below.

Summary of software for analyzing PIAAC data

As for computation, valid analyses of assessment data like PIAAC usually require tools that are specifically developed for such data sets. These tools must be able to handle all the features of complex survey design as well as plausible values in the analyses. Information on appropriate tools is given on the OECD PIAAC website PIAAC 1st Cycle Database | OECD. Examples of such tools are the IDB Analyzer and the SAS macro PIAAC_Tool[4] provided by the PIAAC consortium. The Repest module of STATA software is also worth mentioning. It is an external STATA module that needs to be installed separately within a STATA session. In the R environment, there is a corresponding external module for analysing data from international assessment studies, including PIAAC, called intsvy. The macros, modules, and packages, which are built for jackknife variance estimation and plausible values in assessment studies, consist essentially of loops that run the desired analysis over the replicate samples and plausible values and then combine the obtained analysis results into one final result.

The MPLUS software, which is mainly designed for structural equation modelling (i.e., a model-based analysis), contains several options for the analysis of assessment data like PIAAC. The principle is that plausible values are handled with the option TYPE=IMPUTATION of DATA command, and the design features with the option TYPE=COMPLEX of ANALYSIS command as well as the STRATIFICATION, CLUSTER and WEIGHT options of VARIABLE command. Replication methods for variance estimation, including jackknife methods JK1 and JK2, are also available in MPLUS by the REPSE option of ANALYSIS command. More detailed information is provided in the MPLUS user manual.

It is important to recall that the basic procedures in packages such as SPSS, SAS, STATA or even R will not do as such as they cannot handle complex survey data properly and do not have facilities for multiple imputation, which is needed when dealing with plausible values. However, these packages contain advanced procedures or external modules for complex samples and survey analysis under the design-based approach. These can be used if the data analyst knows how to set the options (e.g., for jackknife replication) of the procedures required for correct or approximately correct analysis. In some situations, this holds also for procedures of the model-based approach (particularly mixed models or multilevel models), which are found practically in any package. The plausible values analysis can be performed with these procedures by repeating the same analysis for each plausible value separately and averaging the 10 obtained results, following the rules of analysis of multiple-imputed data.

In any case, it is vital for users to familiarize themself with the manuals and handbooks of the package they are using. All widely used packages have comprehensive documentation on the

[4] According to advance information, in PIAAC Cycle 2 the replicate weights will be calculated with Fay-modified BRR method (FAY), instead of jackknife methods (JK1 and JK2), at least in some countries. The current version of PIAAC_Tool supports JK1 and JK2 only. The other tools mentioned in this section can also handle the FAY method.

available procedures, their facilities and limitations, required command syntax, and illustrative examples of how to use the procedures.

Finally, it can be mentioned that Educational Testing Service (ETS) has also provided an interactive interface to explore the international PIAAC data. This application is called PIAAC Data Explorer or International Data Explorer, and it is available on OECD's PIAAC website PIAAC - Select Criteria (oecd.org). The application can be used in producing tables of basic descriptive statistics like percentiles or means and their standard errors on PIAAC variables as well as simple significance tests. It cannot be considered a software with which advanced statistical analyses could be performed. However, it respects the recommended methodology in that survey weights are employed, and standard errors are calculated under the design-based framework.

References

OECD (2019). Technical Report of the Survey of Adult Skills (PIAAC). Third Edition. PIAAC_Technical_Report_2019.pdf (oecd.org)

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

Wu, M. (2005). The role of plausible values in large-scale surveys. Studies In Educational Evaluation, 31, 114–128.