The Stata Journal (2019)
19, Number 1, pp. 112–128 DOI: 10.1177/1536867X19830909

piaactools: A program for data analysis with PIAAC data

Maciej Jakubowski
Faculty of Economic Sciences
University of Warsaw
and
Evidence Institute
Warsaw, Poland
mjakubowski@uw.edu.pl

Artur Pokropek
Institute of Philosophy and Sociology
Polish Academy of Sciences
and
Educational Research Institute (IBE)
Warsaw, Poland
artur.pokropek@gmail.com

Abstract. The OECD Programme for the International Assessment of Adult Competencies (PIAAC) is currently the only international survey of adult skills. It provides rich data on skills, work and life situations, earnings, and attitudes. To ensure representativeness and high reliability, the study is based on a complex survey design and advanced statistical methods. To obtain correct results from publicly available microdata, one must use special methods that are often too advanced for less experienced researchers. In this article, we present `piaactools`—a package of three commands that facilitate analysis with PIAAC data. The command `piaacdes` calculates basic statistics, `piaactab` computes frequencies of adults at each proficiency level, and `piaacreg` allows for the use of several regression models with PIAAC data. Output is saved as HTML files that can be opened in most spreadsheets and as Stata matrices that can be further processed in Stata. We also explain how to use these commands and provide examples that can be easily modified for use with different models and variables.

Keywords: st0551, piaactools, piaacdes, piaactab, piaacreg, PIAAC, OECD, regression, plausible values, replicate weights, adult survey, skills, proficiency, education

1 Introduction

The OECD Programme for the International Assessment of Adult Competencies (PIAAC) is a household survey that aims at measuring literacy, numeracy, and problem-solving skills. The assessment was launched in 2008, with the first round finalized in 2013 covering 24 countries and economies. Round 2 was finalized in 2016, bringing in 9 new countries for the dataset. Round 3 is scheduled to be finished in 2019. The study uses a complex survey design and advanced statistical methods to improve the representativeness and reliability of the results. As the only up-to-date source of internationally comparable information about adult skills, PIAAC is highly popular with researchers across the world.

However, to analyze microdata from the PIAAC study, one must use relatively complex methods. In Stata, one cannot obtain correct results from the PIAAC data without some programming, which is a difficult barrier to overcome for some researchers. In many cases, researchers use wrong approaches and later have difficulties when publish-

ing their results. To facilitate analysis with PIAAC data, we developed the package `piaactools`, which contains three commands: `piaacdes`, `piaacreg`, and `piaactab`. These commands allow for the analysis of plausible values available in PIAAC datasets and use the jackknife replication method to calculate standard errors (SEs).

The three commands are straightforward to use even for Stata beginners and guarantee that users will obtain correct point estimates and SEs. The results obtained with these three commands are virtually identical to those that researchers can obtain using the IDB analyzer provided by the PIAAC consortium for SPSS users. The command `piaacdes` calculates basic statistics like mean, median, percentiles, standard deviation, etc.; `piaacreg` allows for the use of several regression models; and `piaactab` computes frequencies of students at each proficiency level. Output is saved as HTML tables that can be easily opened and edited in Excel or any spreadsheet program or Internet browser. Estimation results are also saved in matrices for those users who would like to use them inside Stata for further analysis.

2 Statistical background

2.1 Complex sampling

Each country participating in PIAAC chose its own sampling design. The only requirement was to produce "a probability-based sample, representative of the target population of the country" (Mohadjer, Krenzke, and Van de Kerckhove 2013). One-, two-, three-, and four-stage stratified designs were used (for details, see table 1). Each country was free to define sampling units and stratification design. Poland, for example, used a two-stage sampling design with towns or villages as primary sampling units and individuals as secondary sampling units. It also used a simple rural-versus-urban stratification. Italy used a three-stage design with municipalities as primary sampling units, households as secondary sampling units, and persons as final sampling units, while stratification was based on geographic regions of equal size (Mohadjer, Krenzke, and Van de Kerckhove 2013).

Table 1. Sample designs for PIAAC country (rounds 1 and 2)

Sample design	Countries
One-stage sample designs	Round 1: Flanders (Belgium), Denmark, Estonia, Finland, Netherlands, Norway, Sweden
Two-stage stratified probability proportionate to size designs	Round 1: Cyprus, France, Germany, Japan, Poland, Slovakia, Spain; Round 2: Israel, Slovenia
Three-stage stratified probability proportionate to size designs	Round 1: Canada, Ireland, Italy, Korea, Northern Ireland; Round 2: Greece, Jakarta (Indonesia), Lithuania, Turkey
Four-stage stratified probability proportionate to size designs	Round 1: Australia, Czech Republic, Russian Federation, England, the United States; Round 2: Chile, New Zealand

NOTE: For details, see Mohadjer, Krenzke, and Van de Kerckhove (2013).

Sampling weights were computed to reflect different selection probabilities on the basis of the varying sampling designs to adjust for nonresponse and for "any known differences between the selected sample and the total target population" (Mohadjer, Krenzke, and Van de Kerckhove 2013).

The complex sampling design used in PIAAC has two major consequences. First, all point estimates must be computed using sampling weights. Second, one must use replication methods to obtain correct SEs. Although there are analytical methods that are used for SE computations in household surveys similar to PIAAC, replication methods have important advantages. They provide unbiased and efficient estimates and are relatively easy to implement for many applications (Efron 1982; Levy and Lemeshow 1999). For PIAAC, replication methods have the important benefit of enabling the unification of estimation procedures for samples from dozens of countries participating in the survey.

In short, replication methods involve selecting a set of replicated subsamples instead of a single sample. Subsamples are drawn using an identical or similar sample selection mechanism, and desirable estimates are computed. Because each estimate in each subsample follows an identical sampling process, it is assumed that sampling variance can be estimated from the variability of subsample estimates (for details, see Efron [1982]).

For PIAAC, two variants of the so-called jackknife procedure were proposed: delete-a-group jackknife (JK1) and the paired jackknife (JK2). The first variant is used in sample designs where sampling units are not stratified, while the second variant is used in stratified samples. In the JK1, in each replication one random group is sampled using the same sample design as in the whole sample. In JK2, in each replication one group is removed from a stratum, and weights are adjusted so the remaining groups retain the stratum's proportion in the total sample (Lee and Forthofer 2005). Table 2 provides information about the jackknife variants applied to each participant of the PIAAC survey.

Table 2. Variants of jackknife estimation used by different PIAAC participants (rounds 1 and 2)

Jackknife variant	PIAAC participants
JK1	Round 1: Australia, Austria, Canada, Denmark, Germany; Round 2: New Zealand, Singapore
JK2	Round 1: Cyprus, Czech Republic, England and Northern Ireland (UK), Estonia, Finland, Flanders (Belgium), France, Ireland, Italy, Japan, Korea, Netherlands, Norway, Poland, Russian Federation, Slovak Republic, Spain, Sweden, United States; Round 2: Chile, Greece, Israel, Jakarta (Indonesia), Lithuania, Slovenia, Turkey

International datasets provided by the PIAAC consortium contain both probability and replication weights. Estimation of SEs that does not involve measures of cognitive proficiency (for cognitive components, see next section) is based on computations using a standard replication approach. Computations are based on the variability of estimates in subsequent replicates obtained using replication weights,

$$\mathrm{SE}_\theta = \sqrt{f \sum_{r=1}^{R} \left(\widehat{\theta}_r - \widehat{\theta}_0\right)^2} \tag{1}$$

where

- R is the number of replicates;
- $\widehat{\theta}_r$ represents any statistic of interest (percent, mean, variance, regression coefficient, etc.) not involving plausible values for replicate $r = (1, \ldots, R)$;
- $\widehat{\theta}_0$ represents the statistic of interest (not involving plausible values) estimated using the whole sample and the final sample weight; and
- constant f depends on a jackknife variant used in each country for the delete-a-group variant $f = (R-1)/R$ and for the JK2 $f = 1$.

In `piaactools`, all variance estimates not involving cognitive components are estimated using the procedure described by (1). Determination of the jackknife variant is automatically done based on the country identification code. Optionally, users can also force `piaactools` to use a different variant, although this is not recommended.

2.2 Calculations using plausible values (measures of cognitive proficiency)

The PIAAC study uses advanced measurement and estimation methods to provide countries with reliable measures of adult proficiency. In PIAAC, cognitive tests use matrix sampling design with different sets of items, multistage adaptive testing, and different assessment modes. Similar designs are used in all modern surveys measuring the proficiency of students or adults, and their key feature is that respondents answer different sets of items. The main benefit of this approach is that it covers more cognitive material (more test questions). The main drawback is that it makes the analysis of survey outcomes more complex.

Similarly to other surveys of cognitive proficiency, PIAAC uses complex measurement design with item response theory models combined with plausible values methods to obtain unbiased estimates of respondents' proficiency (Yamamoto, Khorramdel, and von Davier 2013).

In short, plausible values represent random draws from an empirically derived distribution of latent variables that is conditional on the observed values of the scale (measurement model) and the covariates (latent regression),

$$P\left(\theta_j|\mathbf{y}_j,\mathbf{x}_j,\boldsymbol{\beta},\boldsymbol{\Gamma},\boldsymbol{\Sigma}\right) \propto P\left(\mathbf{x}_j|\theta_j,\boldsymbol{\beta}\right) P\left(\theta_j|\mathbf{y}_j,\boldsymbol{\Gamma},\boldsymbol{\Sigma}\right)$$

where $\boldsymbol{\theta}$ is the latent variable, $\boldsymbol{\beta}$ is the matrix of response ($\mathbf{x}_j$) parameters, $\boldsymbol{\Gamma}$ is a matrix of latent regression coefficients for the covariates ($\mathbf{y}_j$), and $\boldsymbol{\Sigma}$ is a common variance matrix for residuals.

Usually, plausible values are drawn from posterior distribution using the Markov Chain Monte Carlo methods (see a detailed description in Fox and Glas [2001, 2003]) in the Bayesian estimation framework. The procedure generates a set of plausible values, and each respondent receives several indicators (plausible values) of proficiency. Each plausible value is used once in each analysis. For instance, for 10 generated plausible values, 10 models are fit; in each model, one plausible value is used, and the final estimates are obtained using Rubin's rule (Little and Rubin 2002)—results from all analyses are simply averaged.

Because plausible values are random draws, the random component provides a way to model the uncertainty associated with the estimate that is related to the measurement error (Wu 2005). Moreover, plausible values have approximately the same conditional distribution as the latent trait being measured. The latent regression part (that is, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}$ matrices) adjusts plausible values to have conditional distribution on covariates (accounted for in the generating process), the same as the latent distribution conditioned

on covariates. Thus, the results obtained from the plausible values are assumed to be close to results that would be obtained using latent regression modeling (for more details, see Mislevy et al. [1992]).

2.3 Combining complex sampling with plausible values

Analysis of PIAAC cognitive outcomes is not straightforward because it requires additional computations to reflect complex sampling design and measurement of cognitive proficiency with plausible values. Estimates variance must reflect both the uncertainty related to sampling and the uncertainty related to the measurement of cognitive proficiency. SE for a survey statistic involving plausible values can be expressed as

$$\text{SE}_{\theta_P} = \sqrt{(\text{Sampling error})^2 + (\text{Measurement error})^2}$$

Implementation of (1) for data with plausible values is

$$\text{SE}_{\theta_P} = \sqrt{\left[\sum_{p=1}^{P}\left\{f\sum_{r=1}^{R}\left(\widehat{\theta}_{r,p} - \overline{\theta}_{0,P}\right)^2\right\}\frac{1}{P}\right] + \left\{\left(1+\frac{1}{P}\right)\frac{\sum_{p=1}^{P}\left(\widehat{\theta}_{0,p} - \overline{\theta}_{0,P}\right)^2}{P-1}\right\}} \quad (2)$$

where

- $\overline{\theta}_{0,P} = (\sum_{p=1}^{P}\theta_{0,p})/P$;
- P is the number of plausible values, $p = (1, \ldots, P)$;
- $\widehat{\theta}_{r,p}$ represents the statistic estimate for replicate r and the pth plausible value;
- $\widehat{\theta}_{0,p}$ represents the statistic estimate using the final sample weight for the pth plausible value; and
- $\overline{\theta}_{0,P}$ represents the unweighted average of the statistic for each plausible value using the whole sample and the final weight.

PIAAC datasets contain 10 plausible values and 80 replicate weights. Thus, implementing the above formula requires 810 computations for each statistic of interest. The procedure starts with 10 calculations for each plausible value with the final weight. Then, the procedure calculates statistics with 80 replicate weights and repeats this for each plausible value. These calculations might be time consuming, especially with complex statistical models. Thus, researchers often use simpler methods to obtain initial results and run full models for the final calculations. A shortcut approach to running the initial analysis could use only one of the plausible values and final weights to obtain point estimates with analytically calculated SEs. However, the final results must be obtained using the formula described above and in some cases might be different from the initial results because of a large sampling or measurement error.

3 Analyzing PIAAC data with Stata survey commands

Analysis without plausible values is possible using the Stata built-in commands for survey data. Below are examples of how to use `svyset` in Stata to set up analysis with PIAAC survey data.

For estimation with jackknife replicate weights, Stata by default sets the constant $f = (R-1)/R$ [see (1)] that is appropriate for the JK1 method. Thus, for this method, a standard setup can be used:

```
. svyset [pw=spfwt0], jkrweight(spfwt1-spfwt80) vce(jackknife) mse
```

For the JK2 method, this constant can be specified using the option `multiplier(1)`.

```
. svyset [pw=spfwt0], jkrweight(spfwt1-spfwt80, multiplier(1)) vce(jackknife) mse
```

After setting up data with `svyset`, we can use many Stata estimation commands with PIAAC data by using the `svy:` prefix. Below are examples calculating means and regression with PIAAC noncognitive variables.

```
. svy: mean readytolearn
. svy: regress readytolearn gender_r
```

However, in most cases users will want to analyze adult proficiency with PIAAC data and thus use plausible values in their analyses. While users can analyze plausible values in Stata using the built-in commands, it is usually too demanding for most users. One option is to manipulate the datasets to use them with Stata's multiple imputation commands. However, the current multiple imputation estimation command in Stata does not allow using jackknife replication weights, so users cannot obtain unbiased SEs this way. Another option is for users to write their own code that repeats calculations over plausible values and replicate weights using the formulas specified above. The `piaactools` package facilitates analysis with PIAAC data for the most commonly used commands. Users need not worry about using replicate weights or plausible values correctly because all calculations are done in the background. The package components, their options, and examples are discussed below.

4 Descriptive statistics with PIAAC data: piaacdes

4.1 Syntax

`piaacdes` [*varlist*] [*if*] [*in*], `save(`*filename*[, `replace`]`)`
[`countryid(`*varname*`)` `vemethodn(`*varname*`)` `weight(`*varname*`)` `rep(`*varlist*`)`
`stats(`*string*`)` `centile(`*string*`)` `pv(`*string*`)` `over(`*varname*`)` `round(`*int*`)`]

4.2 Description

`piaacdes` calculates basic statistics with PIAAC data for the given variables and plausible values listed in the `pv()` option. The `pv()` option takes the prefix of the plausible values variable without ending numbers (for example, `pvlit`, `pvnum`, `pvpsl`). Similarly, the prefix of plausible values can be specified in the `over()` option. In this case, the statistics will be calculated over proficiency levels. `piaacdes` saves results as an HTML file that can be further edited, for example, in a spreadsheet application.

4.3 Options

`save(`*filename*[`, replace`]`)` specifies the name of the output file. `save()` is required.

In addition, users can specify several options that can be used to replace original information available in PIAAC datasets and facilitate analysis with plausible values or different regression models.

`countryid(`*varname*`)` provides the name of a variable containing a list of countries for which you want to obtain results. The list can be numeric or string, with any possible values. Missing categories will be omitted from calculations, and averages across all countries will be calculated. By default, `CNTRYID` or the `cntryid` variable will be used to identify countries, and the OECD average will be calculated.

`vemethodn(`*varname*`)` provides the name of the numeric variable specifying the jackknife variance estimation method for each country. The variable can contain only values of 1 or 2. By default, `VEMETHODN` or the `vemethodn` variable will be used to identify the jackknife method. By default, `piaacreg` will look for `VEMETHODN` or the `vemethodn` variable to identify the jackknife method.

`weight(`*varname*`)` and `rep(`*varlist*`)` give the main weight and a list of jackknife replication weights. You do not have to specify these options if your dataset contains original weights `spfwt0`–`spfwt80` or `SPFWT0`–`SPFWT80`.

`stats(`*string*`)` gives a list of statistics to be calculated. You can list any statistic calculated by `summarize`. For example, `stats(mean sd)` will calculate mean and standard deviation. If `stats()` and `centile()` are not specified, means will be calculated.

`centile(`*string*`)` gives percentiles to be calculated. For example, `centile(5 50 75)` will result in calculating the 5th, median, and 75th percentiles.

`pv(`*string*`)` gives a list of plausible values prefixes. For example, use `pv(pvlit)` to calculate statistics for plausible value in literacy.

`over(`*varname*`)` specifies a categorical variable or plausible value for which you want to obtain statistics for each category. The variable must be numerical with a sequence of integers denoting each category. For proficiency levels, specify the prefix of plausible values without ending numbers (for example, `pvlit`, `pvnum`, `pvpsl`).

`round(`*int*`)` specifies how many decimal places you want to see in results tables. The default is `round(2)`.

4.4 Stored results

`piaacdes` also returns results in Stata matrices. Users can type `return list` after executing the command to see what is available for further analysis or reporting, as in the example below.

4.5 Examples

Example 1. Computation of statistics without involving plausible values

Suppose you want to estimate results for just three countries: Germany, Poland, and the United States. You can recode the `cntryid` variable from the original PIAAC dataset to a new variable denoting these countries by typing

```
. generate countryname="USA" if cntryid==840
. replace countryname="Poland" if cntryid==616
. replace countryname="Germany" if cntryid==276
```

Then, you can use `piaacdes` to compute means, standard deviations, and the 33rd and 66th centiles for the PIAAC index of readiness to learn (`readiness`) for these countries. Results are saved in the working directory in the file `example1.html`. Note the use of the `countryid()` option.

```
. piaacdes readytolearn, save(example1) countryid(countryname) stats(mean sd)
> centile(33 66)
```

Results are also returned in matrices that can be further processed directly in Stata. Type `return list` after running `piaacdes` to see the list of returned result matrices. For example, to see the point estimates, you can type

```
. matrix list r(b_readytolearn)
```

which will display the coefficients matrix in Stata.

```
r(b_readytolearn)[3,4]
            mean        sd          33          66
 Germany    1.9086076   .82491105   1.5515525   2.1524218
 Poland     1.9903249   .96758928   1.5553051   2.3102554
 USA        2.4425124   1.0242794   1.9577689   2.7420436
```

You can use this matrix to further process `piaacdes` results. Similarly, to see the matrix with SEs, type

```
. matrix list r(se_readytolearn)
```

Example 2. Computation of statistics with plausible values

You can compute the means and specific percentiles for the PIAAC numeracy proficiency scale (`pvnum`) across countries. Results are saved in the working directory in the file `example2.html`. Numbers are reported with five decimal points.

```
. piaacdes, save(example2) pv(pvnum) stats(mean) centile(5 10 25 50 75 90 95)
> round(5)
```

Example 3. Computation of statistics by categories

You can compute the means and standard deviations for the PIAAC index of readiness to learn (readiness). Results are reported separately for each country and by gender. Numbers are reported with two decimal points.

```
. piaacdes readytolearn, save(example3) stats(mean sd) round(2) over(gender_r)
```

Results are saved in the HTML file and in separate return matrices for each `over()` category. For example, to see the matrix with the SEs for the second `over()` category, type

```
. matrix list r(se_readytolearn_over2)
```

Example 4. Computation of statistics over proficiency levels

You can compute the means for the PIAAC index of readiness to learn (`readiness`) over the levels of problem-solving proficiency. The OECD's proficiency thresholds are used to obtain results using the plausible values in problem solving (`pvpsl*`).

```
. piaacdes readytolearn, save(example4) countryid(cntryid) stats(mean) round(2)
> over(pvpsl)
```

Results are saved in the HTML file and in the separate return matrices corresponding to different proficiency levels defined over plausible values in problem solving. For example, to see the matrix with the point estimates for the fourth `over()` category, type

```
. matrix list r(b_readytolearn_over4)
```

5 Regression analysis with PIAAC data: piaacreg

5.1 Syntax

`piaacreg` [*varlist*] [*if*] [*in*] , `save(`*filename*[, `replace`]`)`
[`countryid(`*varname*`)` `vemethodn(`*varname*`)` `weight(`*varname*`)` `rep(`*varlist*`)`
`over(`*varname*`)` `round(`*int*`)` `pvdep(`*pv_prefix*`)` `pvindep1(`*pv_prefix*`)`
`pvindep2(`*pv_prefix*`)` `pvindep3(`*pv_prefix*`)` `fast` `cons` `cmd(`*command*`)`
`cmdops(`*options*`)` `or` `r2(`*string*`)`]

5.2 Description

`piaacreg` facilitates regression analysis with PIAAC data. The first variable listed after the `piaacreg` command is the dependent variable unless you specify `pvdep()`. If your dependent variable is a vector of plausible values, you should specify the `pvdep()` option, providing the prefix of the plausible values variable without ending numbers (for example, `pvlit`, `pvnum`, `pvpsl`). The remaining variables listed after `piaacreg` are treated as independent variables. Options `pvindep1()`, `pvindep2()`, and `pvindep3()` allow for the use of plausible variables as independent variables. Similarly, the prefix of plausible values can be specified in the `over()` option. In this case, the regressions will be run over so-called proficiency levels, for example, ordinal scales that are based on continuous plausible values. The `piaacreg` command saves results as an HTML file that can be further edited, for example, in a spreadsheet application. The results are also saved in matrices for further analysis within Stata.

5.3 Options

`save(`*filename*[, `replace`]`)` specifies the name of the output file. `save()` is required.

Users can also replace original variables identifying countries, variance computation method, and weights using the options `countryid()`, `vemethodn()`, `weight()`, and `rep()` similarly as for `piaacdes`. Options `over()` and `round()` can be also used as for `piaacdes`. In addition, one can specify certain `piaacreg` options:

`pvdep(`*pv_prefix*`)` specifies plausible values as dependent variables. One can specify a plausible values prefix without ending numbers. For example, `pv(pvlit)` asks for plausible values `pvlit1`–`pvlit10` to be used as dependent variables.

`pvindep1(`*pv_prefix*`)`, `pvindep2(`*pv_prefix*`)`, and `pvindep3(`*pv_prefix*`)` specify plausible values to be used as independent variables. One can specify a plausible values prefix without ending numbers. For example, `pvindep1(pvlit)` asks for plausible values `pvlit1`–`pvlit10` to be used as independent variables. If additional plausible values must be used, one could specify the option `pvindep2()` or `pvindep3()`.

`fast` specifies speeding up calculations at the cost of not fully valid estimates of SEs. Point estimates are correct, while SEs are obtained analytically and usually differ from those obtained with the jackknife method.

`cons` specifies saving estimates for the regression constant.

`cmd(`*command*`)` specifies running a regression model different from `regress`.

`cmdops(`*options*`)` passes options to the regression command.

`or`, along with `cmd("logit")` or `cmd("logistic")`, allows one to obtain odds ratios instead of coefficients. In this case, a standard Stata approach is taken, and the SEs are derived using the delta rule.

`r2(`*string*`)` specifies to report any scalar returned in `e()`. For example, `r2(r2_a)` reports adjusted R-squared.

5.4 Stored results

`piaacreg` also returns results in Stata matrices. Type `return list` after executing the command.

5.5 Examples

Example 1. Regression without plausible values

You can compute regression without plausible values by typing

```
. piaacreg readytolearn gender_r, save(example1) round(5)
```

Results are saved in the working directory in the file `example1.html`. Numbers are reported with five decimal points. Results are also returned in matrices that can be further processed directly in Stata. Type `return list` after running `piaacreg` to see the list of returned result matrices. For example, to see the point estimates, type

```
. matrix list r(b)
```

To see the SEs, type

```
. matrix list r(se)
```

To see the R-squared matrix, type

```
. matrix list r(r2)
```

Example 2. Regression with plausible values as a dependent variable

Plausible values in numeracy are declared as the dependent variable by using the `pvdep()` option.

```
. piaacreg readytolearn gender_r, save(example2) pvdep(pvnum) round(5)
```

Results are saved in the HTML file and in the return matrices that can be accessed as in example 1.

Example 3. Regression with plausible values as an independent variable

Plausible values in numeracy are declared as one of the independent variables by using the `pvindep1()` option.

```
. piaacreg readytolearn gender_r, save(example3) pvindep1(pvnum) round(5) cons
```

One can include the second or the third set of plausible values as independent variables by using the `pvindep2()` or `pvindep3()` option. The regression constant is reported in the output tables because the `cons` option was specified.

Example 4. Logistic regression with plausible values as an independent variable

Other regression models can be run by using the `cmd("")` option. This option specifies the regression command. For example, to run logistic regression, type `cmd("logit")`.

```
. recode computerexperience (1=1) (2=0), gen(compexp)
. piaacreg compexp readytolearn gender_r, save(example4) pvindep1(pvnum)
> cmd("logit")
```

6 Tables with PIAAC data: piaactab

6.1 Syntax

`piaactab` *varname* [*if*] [*in*], `save(`*filename*[`, replace`]`)`
[`countryid(`*varname*`) vemethodn(`*varname*`) weight(`*varname*`) rep(`*varlist*`)`
`over(`*varname*`) round(`*int*`) twoway fast missing`]

6.2 Description

The `piaactab` command calculates cell percentages with PIAAC data. Percentages for proficiency levels will be calculated if the prefix of the plausible values variable is provided without ending numbers (for example, `pvlit`, `pvnum`, `pvpsl`). Similarly, the prefix of plausible values can be specified in the `over()` option. In this case, the percentages will be calculated across proficiency levels. Users can also use their own categorical

variables based on plausible values. Names of these variables should start with a `pv` prefix, and the `varname` should include them with an underscore instead of 1 to 10 numbers denoting each plausible variable. For example, `piaactab pv_level, ...` will calculate statistics for variables for `pv1level` to `pv10level` that are based on plausible values. Similarly, one's own variables based on plausible values can be specified in the `over()` option. For example, `over(pv_level)` will calculate statistics over categories of `pv1level-pv10level` variables based on plausible values. `piaactab` saves results as an HTML file that can be, for example, further edited in a spreadsheet application. `piaactab` also returns results in Stata matrices. Type `return list` after executing the command.

6.3 Options

`save(`*filename*`[ , replace ])` specifies the name of the output file. `save()` is required.

Users can also replace original variables identifying countries, the variance computation method, and weights using the options `countryid()`, `vemethodn()`, `weight()`, and `rep()`, similarly to `piaacdes`. Options `over()` and `round()` can also be used as for `piaacdes`. Additional options with `piaactab` are the following:

`twoway` calculates cell percentages in a two-way table (over both the main and the `over()` variable).

`fast` speeds up calculations by not reporting SEs.

`missing` reports statistics for missing observations as a separate category.

6.4 Stored results

Results are saved as HTML files and returned in matrices that can be further processed directly in Stata.

6.5 Examples

Example 1. Calculation of the table without involving plausible values

You can compute the percentage of males and females by country with the appropriate SEs by typing

```
. piaactab gender_r, save(example1) countryid(cntryid) round(5)
```

Results are saved in the working directory in a file named `example1.html`. Numbers are saved with five decimal points. Results are also returned in matrices that can be further processed directly in Stata. For example, to see the point estimates, type

```
. matrix list r(b)
```

To see the SEs, type

```
. matrix list r(se)
```

Example 2. Calculation of the table involving plausible values

You can compute the percentage of respondents at each proficiency level in literacy with the appropriate SEs. Calculations are based on the OECD's thresholds using plausible values `pvlit*`.

```
. piaactab pvlit, save(example2) countryid(cntryid) round(5)
```

Results are saved in the HTML file and in the return matrices. See example 1.

Example 3. Calculations by subgroups using the over() option

You can compute calculations by subgroups using the `over()` option by typing

```
. piaactab computerexperience, save(example3a) over(gender_r) countryid(cntryid)
> round(5)
. piaactab pvlit, save(example3b) over(gender_r) countryid(cntryid) round(5)
. piaactab gender_r, save(example3c) over(pvlit) countryid(cntryid) round(5)
```

Results are saved in the HTML file. Results are also saved in the separate return matrices for each `over()` category. For example, to see the matrix with the SEs for the second `over()` category, type

```
. matrix list r(se_over2)
```

7 Other community-contributed commands that work with PIAAC data

Two other community-contributed commands allow for the analysis of PIAAC data. The `ssc install pv` command will install the `pv` and `pvtest` commands written by Macdonald (2008). Both commands are quite flexible and allow for the use of standard Stata commands to estimate different models with plausible values and replicate weights. An example below shows how `pv` can be used with PIAAC data:

```
. pv, pv(pvlit*) jrr jk(1) weight(spfwt0) rw(spfwt1 - spfwt80): oprobit
> monthlyincpr @pv [aw = @w]
```

In addition, `pvtest` allows for the testing of linear hypotheses after estimation with plausible values. The `repest` is a similar command written by Avvisati and Keslair (2014) from the OECD. The example below demonstrates that `repest` can be used even with Stata commands that use more complex syntax:

```
. repest PIAAC, estimate(stata: sureg (lnwage pvlit@ yrsqual) (lnwage pvnum@
> yrsqual)) by(cnt)
```

These and additional examples are provided in the help files for these commands.

8 References

Avvisati, F., and F. Keslair. 2014. repest: Stata module to run estimations with weighted replicate samples and plausible values. Statistical Software Components S457918, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s457918.html.

Efron, B. 1982. *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Fox, J.-P., and C. A. W. Glas. 2001. Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika* 66: 271–288.

———. 2003. Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika* 68: 169–191.

Lee, E. S., and R. N. Forthofer. 2005. *Analyzing Complex Survey Data*. 2nd ed. Thousand Oaks, CA: Sage.

Levy, P. S., and S. Lemeshow. 1999. *Sampling of Populations: Methods and Applications*. 3rd ed. New York: Wiley.

Little, R. J. A., and D. B. Rubin. 2002. *Statistical Analysis with Missing Data*. 2nd ed. Hoboken, NJ: Wiley.

Macdonald, K. 2008. pv: Stata module to perform estimation with plausible values. Statistical Software Components S456951, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s456951.html.

Mislevy, R. J., A. E. Beaton, B. Kaplan, and K. M. Sheehan. 1992. Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement* 29: 133–161.

Mohadjer, L., T. Krenzke, and W. Van de Kerckhove. 2013. Sampling design. In *Technical Report of the Survey of Adult Skills (PIAAC)*, chap. 14. Paris, France: OECD.

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

Yamamoto, K., L. Khorramdel, and M. von Davier. 2013. Scaling PIAAC cognitive data. In *Technical Report of the Survey of Adult Skills (PIAAC)*, chap. 17. Paris, France: OECD.

About the authors

Maciej Jakubowski works at the Faculty of Economic Sciences, University of Warsaw, and at the Evidence Institute, Poland. His main areas of research are economics of education and impact evaluation methods.

Artur Pokropek is an assistant professor at the Institute of Philosophy and Sociology, Polish Academy of Sciences, and a research scientist at the Educational Research Institute (IBE), Warsaw, Poland. His main research interests include statistics, psychometrics, survey methodology, and social inequalities.