User's Guide for a SAS Macro enabling the computation of design based standard errors in the Survey of Adult Skills (PIAAC)

Version 1.2
July, 2014

Author: Vanessa Denis (Vanessa.DENIS@oecd.org)

Contents

INTRODUCTION

The Survey of Adult Skills (PIAAC) is a study administered in multiple countries based on complex survey samples involving both household survey methodologies and direct psychometric assessment methodologies. Both sets of methodologies are complex and require advanced methods and special procedures to obtain accurate statistical results. In the Survey of Adult Skills these complex methodologies are uniquely combined creating the need for even more specialised procedures that are not easily available in standard commercialized softwares such as SAS, SPSS or STATA. Standard functions in these commercial softwares (e.g. PROC MEANS, PROC FREQ in SAS) may be used on data from the Survey of Adult Skills to produce unbiased point estimates (i.e., means, percentiles, proportions, regression parameters…), as long as proper weighting procedures are followed, but they cannot be used to produce unbiased standard error estimates. Thus it is necessary to write specialized macros that can be read in each of these softwares in order to implement procedures that generate design-based standard errors for each point estimate (i.e., standard errors of means, percentiles, proportions, regression parameters…). Unbiased standard error estimates are desirable for assessing the quality of point estimates accurately and also for producing valid inferential statistics such as t-values and p-values.

Accordingly, a SAS macro was developed by the OECD to obtain unbiased point estimates and associated standard errors for commonly used descriptive statistics (e.g. means, percentiles, proportions) and also statistics based on linear and logistic regression. This macro builds on existing commands available in SAS and adapts them to specific variance formulas associated with the Jackknifing method (for details see Efron, 1982; Levy and Lemeshow, 1999). The macro is straightforward to use, even for novice users of SAS. The results obtained from the SAS macro are identical to those that researchers can obtain using the IDB analyser, which was designed for use by SPSS users, and was made available by the consortium who implemented the Survey of Adult Skills. The SAS macro and IDB analyser, however, do not necessarily produce all the same statistics. For example, the IDB analyser can produce unbiased standard errors associated with correlation analysis. Important features of this SAS macro are that it produces estimates for logistic regression analysis and outputs can be saved as SAS tables or exported directly into EXCEL.

BACKGROUND ON VARIANCE FORMULAS

As mentioned before, complex sampling designs were used in the Survey of Adult Skills. For details see The Survey of Adult Skills: Reader's Companion (OECD, 2013) and Technical Report of the Survey of Adult Skills (PIAAC, 2013) (www.oecd.org/site/piaac/). This has two consequences for statistical estimation based on the survey data. First, all point estimates must be computed using sampling weights. Second, it is necessary to use special procedures for standard error computations. While different analytical procedures are available for computing standard errors, the replication approach was chosen since it is an efficient and versatile approach accommodating many different types of sampling (see Efron, 1982; Levy and Lemeshow, 1999). In particular, the *jackknife* replicate procedure was chosen to be used in the Survey of Adult Skills.

The specific variance formula that is relevant depends on the type of sampling (i.e., with or without stratification) and also whether there are psychometric scores (i.e., with or without *plausible values*) involved in the statistical estimation.

In the estimation of standard errors not involving psychometric scores (i.e. not involving *plausible values*), only the error associated with the sampling of persons (i.e., *sampling error*) is taken into account.

Formula (1) is based on a sum of squares principle which summarizes the variability of estimates in subsequent subsets of samples called replicates:

$$SE_\theta = \sqrt{f\sum_{r=1}^{R}\left(\hat{\theta}_r - \hat{\theta}_0\right)^2} \quad (1)$$

Where:

R is the number of replicates;

$\hat{\theta}_r$ represents any statistic of interest (e.g. mean) (not involving *plausible values*) for replicate r=(1,...,R);

$\hat{\theta}_0$ represents the statistic of interest (not involving *plausible values*) estimated using the whole sample and final sample weights.

f is a constant, which depends on the sampling procedures used in each country. In the Survey of Adult Skills, two types of sampling were used: non-stratified and stratified. In each situation, the constant has different values, respectively: $f = \frac{R-1}{R}$ or $f = 1$.

In the estimation of standard errors involving psychometric scores (i.e. involving *plausible values*), both the error associated with the sampling of persons (i.e. *sampling error*) and the sampling of psychometric items administered to each respondent (i.e. *measurement error*) are taken into account (for details see Wu, 2005). The formula can be expressed as:

$$SE_{\theta_p} = \sqrt{(\text{Sampling error})^2 + (\text{Measurement error})^2} \quad (2)$$

Implementation of equation (2) in the Survey of Adult Skills is as follows:

$$SE_{\theta_p} = \sqrt{\left[\sum_{p=1}^{P}\left(f\sum_{r=1}^{R}\left(\hat{\theta}_{r,p} - \bar{\theta}_{0,P}\right)^2\right)\frac{1}{P}\right] + \left[\left(1+\frac{1}{P}\right)\frac{\sum_{p=1}^{P}\left(\hat{\theta}_{0,p} - \bar{\theta}_{0,P}\right)^2}{P-1}\right]} \quad (3)$$

Where:

$\bar{\theta}_{0,P} = \frac{\sum_{p=1}^{P}\theta_{0,p}}{P}$;

P is the number of plausible values, *p*=(1,...,*P*);

$\hat{\theta}_{r,p}$ represents the statistical estimate for replicate r and the pth *plausible value*;

$\hat{\theta}_{0,p}$ represents the statistical estimate using the final sample weight for the pth *plausible value*;

$\bar{\theta}_{0,P}$ represents the unweighted average of the statistic for each *plausible value* using whole sample and the final weights;

In the Survey of Adult Skills, there are 10 *plausible values* and 80 *replicate weights* for each observation. In practice, the above formula "jackknifes" or summarizes variability in 800 additional estimates (80 *replicates* for each of 10 *plausible values*).

KNOWN FAULTS AND OTHER PRACTICAL NOTES

- Results must be produced by country
- The variable VEMETHOD denoting whether it is the JK1 or JK2 formula that is applicable to different countries must be in the dataset
- The variable VENREPS denoting the number of replicates weigths computed in each country must be in the dataset
- If there are too few observations, the macro may crash when results are crossed by two or more variables
- The macro doesn't function properly with versions before SAS 9.2

USING THE MACRO

1. GENERAL STRUCTURE

The macro code is available in the SAS file PIAAC_Tool.sas.
The structure for running the macro function is :

```
%PIAAC_TOOL(      method= ,
                  table= ,
                  wgt=SPFWT0 ,
                  rwgt=SPFWT ,
                  nrep=80 ,
                  nb_pv=1 ,
                  dvar= ,
                  dvarpv=0 ,
                  ivar= ,
                  ivarpv= ,
                  byvar= ,
                  byvarpv= ,
                  vemethod=VEMETHOD,
                  clvar= ,
                  clvarpv= ,
                  refclvarpv= ,
                  event=1,
                  percentile= ,
                  R2= ,
                  Probit=0,
                  Predict=0,
                  Table_pred=pred,
                  where_cond= ,
                  ficout= ,
                  path_out ,
                  excel_sheet=sas_output ,
                  dbms= );
```

2. PARAMETERS

- **Method** (mandatory)

4 types allowed :

- MEAN: computes means, percentiles
- FREQ: computes proportions
- REG: computes linear regressions (parameters, R-square, predictions and residuals)
- LOGISTIC: computes logit or probit regressions (parameters, odds ratios, R-square)

- **Table** (mandatory)

Name of the SAS database.
Warning: the table used should be in the work library, do not include the librairy in the name.
Example: Table=piaac (and note Table=work.piaac)

- **Wgt**

Name of the population weight variable used in the computation. By default wgt=SPFTW0
Example: wgt=SPFWT0

- **Rwgt**

Name of the replicate weights, only the root of the variable. By default rwgt=SPFWT.
Example: Rwgt=SPFWT

- **Nrep**

Number of sets of replicate weights. By default nrep=80.
Example: Nrep=80

- **Nb_PV**

Number of sets of plausible values. By default put to 1 (if no plausible values).
Example: Nb_pv=10

- **Dvar** (mandatory)

Name of the dependant variable or only the root if based on plausible values

- **Dvarpv**

Flag indicating if the dependant variable is based on plausible values. Put dvarpv=1 in this case. By default dvarpv=0. Mandatory if Dvar is based on plausible values.

- **Ivar**

List of the independant variables which are not based on plausible values.

- **Ivarpv**

List of the independant variables which are based on plausible values. Indicate only the root of the variables.

- **Byvar** (mandatory)

List of the subpopulation variables which are not based on plausible values.
N.B: at least a country variable should be included here.

- **Byvarpv**

List of the subpopulation variables which are based on plausible values.

- **Vemethod**

Name of the variable specifying the jackknife variance estimation method for each country. The variable can only contain values "JK1" or "JK2". By default, VEMETHOD variable will be used to identify jackknife method.

- **Where_cond**

WHERE condition in SAS language enabling to define the scope of the analysis, to exclude population or selected modalities of a variable.
Example : where_cond= age between 16 and 65.
Warning : no parentheses allowed in the code

- **Ficout**

Name of the Excel file that will contain the output results
*Examples: Table_1**.xls***
*Table_1.**xlsx***
N.B: include the Excel extension. The macro doesn't work with the extension .xlsm

- **Path_out**

Path where the Excel output file will be stored, without quote or double quotes.
Example : Z:\FIR\SAS Outputs\Chp2
N.B: "/" should be used at the end.

- **Excel_sheet**

Name of the sheet in the Excel output file. By default the sheet is named OUTPUT_SAS.
Example : MEAN_LIT

- **Dbms**

Specifies the type of external data source the EXPORT procedure creates. To export to a DBMS table, specify DBMS= using a supported database identifier:

Data source Identifier	**Output Data Source**	**File Extension**
Excel*	Excel 97, 2000, 2002, 2003 or 2007 spreadsheet	.xls .xlsb .xlxs
Excel4	Excel 4.0 spreadsheet	.xls
Excel5	Excel 5.0 or 7.0 spreadsheet	.xls
EXCELCS	Excel spreadsheet connecting remotely through PC files Server	.xls .xlsb

Examples : Dbms=excel2007
Dbms=EXCELCS

- **Clvar**

For logistic regression only, indicates the list of the variables (not based on plausible values) which should be considered as class variables and not continuous variables. You could define here the reference class in SAS language format. By default, the reference group will be the highest modality.
Example : clvar=AGEG10LFS (Ref="3") GENDER BORNLANG(ref="1") EDCAT3(ref="2")

- **Clvarpv**

For logistic regression only, indicates the list of the variables (based on plausible values) which should be considered as class variables and not continuous variables.The definition of the reference class is not allowed here (see the refclvarpv parameter). Indicate only the variable root.
Example : clvarpv=LITLEVEL

- **Refclvarpr**

For logistic regression and if the Clvarpv is used in the macro, you can define the reference class of the variable based on plausible value .
Example : refclvarpv=(Ref="1")

- **Event**

For logistic regression only, you can define the probability modeled. By default the event modeled is the modality 1.
Example: event=0

- **Percentile**

For mean computation only (method=MEAN), you can obtain percentiles using this parameter. Indicate the list of the percentiles you want.
Example: percentile= 5 25 65 78 96
N.B: this percentile option requires at least the SAS version 9.2

- **R2**

For linear regression, you can ask the computation of the R-Square in the output file. The results will be available by default in the same Excel file in a new sheet called "_R2" if an export is asked. To change the name of the sheet you can use the optional parameter out_R2. The results can be found in the sas table "R2"
Example: R2=1

- **Out_R2**

For linear regression, when the export and the R2 option are asked, you can choose the name of the sheet where will be stored the results.
Example: Out_R2=rsquared

- **Probit**

For logistic regression, you can ask to modelise the probit regression (by default the logit is used)
Example: Probit=1

- **Predict**

For linear regression, you can ask the computation of the prediction and the residuals. The results will be available in a new SAS table (by default the name of the table will be *pred*)

Example: Predict=1

- **Table_pred**

For linear regression and if you have asked for the option *PREDICT*, you can choose the name of the table in which the prediction and residuals will be saved.
Example: Table_pred=myTable

3. OUTPUT RESULTS

The results of the macro are stored in the table **work.resul**.

They are shown in the output window if the export in Excel is not asked (by default).

You can ask the export of the results in Excel (if the ficout parameter is used).
In this case the 3 parameters should be used:
- ficout
- path_out
- dbms

The Excel_sheet parameter is optional (by defaut the sheet name is SAS_OUTPUT)

4. EXAMPLES OF SYNTAX FOR DIFFERENT FUNCTIONS

Method MEAN

Mean with dependant variable based on plausible values

```
%PIAAC_TOOL(method=mean ,
            table=travail ,
            nb_pv=10 ,
            dvar=PVLIT ,
            dvarpv=1 ,
            byvar= cntry_out ageg10lfs,
            where_cond=age>=16 and age<=65 ,
            ficout=3.1(L).xls ,
            path_out=Z:\FIR\SAS Outputs\Chp3\,
            dbms=EXCELCS);
```

Computation of mean of literacy score (PVLIT) across countries by age group (AGEG10lfs). Results will be saved in file 3.1(L).xls (in Z:\FIR\SAS Outputs\Chp3\).

Mean with dependant variable not based on plausible values

```
%PIAAC_TOOL(method=mean ,
            table=travail ,
            nb_pv=10 ,
            dvar=AGE_R ,
            dvarpv=0 ,
            byvar= cntry_out,
            where_cond=age>=16 and age<=65);
```

Computation of mean of age (AGE_R) across countries.

Percentiles computation

```
%PIAAC_TOOL(method=mean ,
            percentile= 5 10 25 50 75 90 95,
            table=travail ,
            nb_pv=10 ,
            dvar=PVNUM ,
            dvarpv=1 ,
            byvar= cntry_out,
            where_cond=age>=16 and age<=65 ,
            ficout=2.3.A.xls,
            path_out=Z:\FIR\SAS Outputs\Chp2\,
            dbms=EXCELCS);
```

Computation of mean and chosen centiles of numeracy score (PVNUM) across countries. Results will be saved in file 2.3.A.xls (in Z:\FIR\SAS Outputs\Chp2\).

Method FREQ

Percentage with dependant variable based on plausible values

```
%PIAAC_TOOL(method=freq ,
            table=travail ,
            nb_pv=10 ,
            dvar=LITLEVEL ,
            dvarpv=1 ,
            byvar= cntry_out ageg10lfs,
            where_cond=age>=16 and age<=65);
```

Computation of percentage of adults at each proficiency level for plausible values in literacy (LITLEVEL) at each age group across country.

Percentage with subpopulation variable based on plausible values

```
%PIAAC_TOOL(method=freq ,
            table=travail ,
            nb_pv=10 ,
            dvar=LFSINO, dvarpv=0 ,
            byvar= cntry_out ISCOSKIL4 ,
            byvarpv= &LITLEVEL ,
            where_cond=age>=16 and age<=65);
```

Computation of percentage of labour force participants (LFSINO) by proficiency level for plausible values in literacy (LITLEVEL) across country and by occupation status (ISCOSKILL4).

Method REG

OLS regression with dependant variable based on plausible values and option R-Square

```
%PIAAC_TOOL(method=reg ,
            table=travail ,
```

```
            nb_pv=10 ,
            dvar=PVLIT ,
            dvarpv=1 ,
            ivar=AGE1624 AGE2534 AGE4554 AGE5565 WOMEN FOREIGNBORN FOREIGNLANG
            ED1 ED3 PARED1 PARED3 SKILL1 SKILL3 SKILL4,
            byvar= cntry_out ,
            where_cond=AGEG10LFS>=0 and GENDER>=0 and EDCAT3>=0 and PARED>=0
            and ISCOSKIL4>=0 ,
            ficout=B.3.2 (L).xls,
            R2=1,
            path_out=Z:\FIR\SAS Outputs\Chp3\,
            dbms=EXCELCS);
```

Literacy proficiency, adjusted for socio-demographic characteristics. Regression with plausible values as a dependent variable: PVLIT=β_0 + βX+ e, by countries.

OLS regression with dependant variable and independent variable based on plausible values

```
%PIAAC_TOOL(method=reg ,
            table=travail ,
            nb_pv=10 ,
            dvar=PVLIT ,
            dvarpv=1 ,
            ivarpv=PVNUM,
            byvar= cntry_out ,
            where_cond= age>=16 and age<=65);
```

Computation of the regression coefficients with plausible values as a dependent variable and independant variable: PVLIT=β_0 + βPVNUM+ e, by countries.

Method LOGISTIC

Logistic regression with independant variable based on plausible values

```
%PIAAC_TOOL(method=logistic ,
            table=travail ,
            dvar=UNEMPO ,
            event=1,
            ivar=  GENDER BORNLANG,
            ivarpv=PSLLEVEL ,
            clvar= GENDER  BORNLANG(Ref="1"),
            clvarpv=PSTLEVEL ,
            refclvarpv=(Ref="1") ,
            nb_pv=10 ,
            byvar= cntry_out age3c ,
            where_cond= age>=16 and age<=65 and UNEMPO>=0 and LFSINO=1,
```

```
ficout=6.1.A(P).xls ,
path_out=Z:\FIR\SAS Outputs\Chp6\,
dbms=EXCELCS);
```

Computation of the regression coefficients and odd ratios: logit(UNEMPO)=β_0 + β_1GENDER + β_2BORNLANG+ β_3PSTLEVEL + e.

REFERENCES

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans.* Philadelphia Pennsylvania: SIAM

OECD (2013), *The Survey of Adult Skills: Reader's Companion*

OECD (2013), *Technical Report of the Survey of Adult Skills*

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

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*(2), 114-128.