ESS Round 8 Sample Design Data File: User Guide

Peter Lynn
INSTITUTE FOR SOCIAL AND ECONOMIC RESEARCH, UNIVERSITY OF ESSEX

07 February 2019

v2

Contents

1. **Introduction**

This User Guide documents the contents of the ESS Round 8 integrated Sample Design Data File (SDDF) and provides guidance on use of the data. For the first time, the sample design data are being released in a form that is suitable for complex standard error estimation. This means that the sample design indicators that were provided by each country have been recoded following the principles set out in Kaminska and Lynn (2017), such that there is no duplication of stratum or PSU values between countries and such that every survey respondent has a valid value of both these variables. This should make it easy for data users to estimate standard errors in an appropriate way. How this can be done is illustrated in section 3 of this Guide.

2. Variables

Aside from indicators of the edition of the data release and the ESS Round to which the data pertain (each of which are invariant within the file), the file contains six variables for each responding sample member. The variables are described in this section.

2.1 CNTRY

This is a string variable containing a two-letter country code. The codes are shown in table 1 below. This variable must be used in combination with IDNO (see section 2.2 below) when merging SDDF data to the integrated file of questionnaire data, as identical values of IDNO may occur in multiple countries. (See section 3.1 for guidance on merging files.)

2.2 IDNO

This is a numeric variable containing the respondent's individual identification number, which is unique within a country. It is used in merging files (see section 3.1).

2.3 PSU

This numeric variable indicates the primary sampling unit within which the respondent was selected. Respondents from the same primary sampling unit will have the same value of PSU. In many countries, the first stage in the sample design was to select a sample of small geographical areas, while at subsequent stages a number of individuals were selected in each of these small areas. Thus, these areas constitute the PSUs and the sample is clustered within a set of PSUs. In other countries, no sample clustering was used. Instead, addresses or persons were selected independently within each of a number of sampling strata, which collectively encompassed the whole country. In these countries, the address or person constitutes the PSU and each respondent will have a unique value of PSU.

2.4 DOMAIN

Some countries use a different sample design in each of two or more parts of the country. These parts are referred to as sampling domains (Gabler et al, 2006). A typical two-domain design might consist of a single-stage unclustered sample in urban areas, but a multi-stage clustered design in other areas, in order to provide cost-efficient fieldwork.

The numeric variable DOMAIN indicates the sampling domain to which the respondent belongs. At Round 8, nine countries used a two-domain design (and no countries used more than two domains). For the 14 countries with a single-domain design, DOMAIN is set to missing.

2.5 STRATUM

All except one of the 23 countries participating in ESS8 used some form of stratified sampling at the first stage of selection. The numeric variable STRATUM indicates the sampling stratum to which the PSU and therefore the respondent belongs. In the Netherlands a simple random sample was selected, so the entire country is treated as a single stratum. In countries that employed systematic sampling with implicit stratification at the first stage, the values of STRATUM reflect groups of PSUs in the order that they were selected, with boundaries based on the underlying variables that defined the implicit stratification. This provides a reasonable approximation in terms of explicit strata to the true implicit stratification (Lynn, 2018).

2.6 PROB

The numeric variable PROB is the unscaled sample selection probability of the respondent. This probability forms the basis of the ESS design weight, DWEIGHT, which can be found in the integrated questionnaire data file. However, PROB can not be recreated from DWEIGHT for all cases, as the production of DWEIGHT involves some truncation and rescaling (see Lynn & Anghelescu, 2018).

Table 1. Value ranges for variables STRATUM and PSU

Countries		Observations (net sample size)	STRATUM	PSU
Austria	AT	2010	1 to 97	1 to 640
Belgium	BE	1766	98 to 108	641 to 912
Switzerland	CH	1525	109 to 115	913 to 2437
Czech Republic	CZ	2269	116 to 171	2438 to 3002
Germany	DE	2852	172 to 191	3003 to 3183
Estonia	EE	2019	192 to 201	3184 to 5202
Spain	ES	1958	202 to 258	5203 to 5636
Finland	FI	1925	259 to 270	5637 to 7561
France	FR	2070	271 to 365	7562 to 8063
United Kingdom	GB	1959	366 to 401	8064 to 8313
Hungary	HU	1614	402 to 571	8314 to 9318
Ireland	IE	2757	572 to 632	9319 to 9798
Israel	IL	2557	633 to 645	9799 to 10048
Iceland	IS	880	646 to 649	10049 to 10753
Italy	IT	2626	650 to 677	10754 to 11212
Lithuania	LT	2122	678 to 704	11213 to 12783
Netherlands	NL	1681	705	12784 to 14464
Norway	NO	1545	706 to 747	14465 to 16009
Poland	PL	1694	748 to 909	16010 to 16903
Portugal	PT	1270	911 to 925	16904 to 17504
Russia	RU	2430	926 to 933	17505 to 17688
Sweden	SE	1551	934 to 941	17689 to 19239
Slovenia	SI	1307	942 to 988	19240 to 19539

3. Using the File

3.1 Merging with other files

There is little of use that can be done with this file alone. We anticipate that the file will mainly be used for incorporating sample design indicators (PSU and STRATUM) into substantive analyses in order to obtain design-unbiased estimates of standard errors and/or to specify the multilevel structure of the data. To do this, the first step is to merge the two files together, based on the combination of CNTRY and IDNO. Merging is easy to do in most statistical software. For example, in Stata you could use the syntax in box 1 (where ESS8e02.dta is the integrated questionnaire data file). The result (a successful match of all records) is also shown.

Box 1: File merging in Stata

```
use ESS8SDDFe1_2.dta, clear
merge 1:1 cntry idno using ESS8e02.dta, force

(note: variable edition was str3 in the using data, but will be float
now)

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                            44,387  (_merge==3)
    -----------------------------------------
```

3.2 Estimating standard errors

It is good practice to estimate standard errors in an appropriate way that takes into account the sample design. There are several analytical methods for doing this, and most statistical analysis software supports at least some of these methods (West et al, 2018). In order to do this, the data file must contain indicators – in an

appropriate form – of sampling stratum, primary sampling unit, and (adjusted) design weight. Once you have merged the SDDF and questionnaire data as described in section 3.1 above, your file will contain all of these variables, namely STRATUM, PSU and PSPWGHT. In fact, for most analyses we recommend that the population size adjustment, PWEIGHT, should also be used, so we begin by creating the analysis weight, which we shall call ANWEIGHT. This is simply the product of PSPWGHT and PWEIGHT. The first line of syntax in box 2 shows how this is done in Stata.

The next step is to specify the sample design. How this is done in Stata is shown in the second line in box 2. Note that the output, also included in box 2, confirms the parameters of the design and the estimation method that will be used subsequently. As we have not specified a method, Stata will use its default of Taylor Linearization. Other options are available.

Box 2: Specifying the sample design in Stata

```
ge anweight=pspwght*pweight
svyset psu [pweight=anweight], strata(stratum) single(cen)

      pweight: anweight
          VCE: linearized
  Single unit: centered
     Strata 1: stratum
         SU 1: psu
        FPC 1: <zero>
```

Now that the design has been specified, it is a simple matter to produce estimates that take the design into account. The syntax in box 3 shows how this can be done using the "svy:" prefix in Stata. The example is an estimate of a mean of an eleven-point scale (ppltrst), but the svy: prefix can be used with a wide range of estimation commands, including OLS regression, logistic regression and many others. The same is true of equivalent procedures in other major statistical analysis packages.

Box 3: Obtaining a design-adjusted estimate

```
svy: mean ppltrst if ppltrst<11


Survey: Mean estimation

Number of strata =     987        Number of obs   =     44,272
Number of PSUs   =  19,473        Population size = 51,961.607
                                  Design df       =     18,486

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
     ppltrst |   4.863305   .0310671       4.80241     4.9242
--------------------------------------------------------------
Note: Strata with single sampling unit centered at overall
      mean.
```

3.3 Illustration of the effect of sample design on estimation

It is instructive to compare the estimate obtained in box 3 with the estimate that would have been obtained if we had not taken the design into account. In box 4 we make two estimates that do not take the stratum and psu into account – with and without weighting. Comparing these with each other, and with the correctly specified estimate in box 3, we observe:

- The unweighted estimate (5.27) is substantially different from the weighted one (4.86). The confidence intervals for these two estimates are far from overlapping. This demonstrates the importance of applying the weights: the difference between the two estimates can be assumed to be indicative of bias that is corrected by the weights;
- The standard error of the estimate is greatly under-estimated if the weights are ignored (0.0112 rather than 0.0225). This will lead to biased significance tests (null hypothesis will be incorrectly rejected too often) and biased model fitting (over-fitting);

- Even if the weights are used, the standard error is under-estimated if the sample design (STRATUM and PSU) is not taken into account (0.0225 rather than 0.0311). This is not a universal truth, as the effects of stratification and clustering tend to work in opposite directions, but it is a very common finding that the increase in variance due to clustering outweighs the reduction in variance due to stratification.
- Taking STRATUM and PSU into account does not affect the point estimate of the mean, only the variance (standard error). This is always the case.

Box 4: Obtaining naïve estimates

```
mean ppltrst if ppltrst<11

Mean estimation                   Number of obs   =     44,272

------------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
     ppltrst |   5.268928   .0112418      5.246894    5.290963
------------------------------------------------------------------

mean ppltrst if ppltrst<11 [pw=anweight]

Mean estimation                   Number of obs   =     44,272

------------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
     ppltrst |   4.863305   .0224732      4.819257    4.907353
------------------------------------------------------------------
```

As the ESS sample design differs between countries, the effect of sample design on estimation can also be expected to differ between countries. This is illustrated by the analysis in box 5. Here we estimate the mean of ppltrst separately for four countries, two of which have single-stage designs (CH and EE) and two of which have multi-stage clustered designs (AT and BE). For each country, estimates of the

mean are obtained in three different ways: first, ignoring weights and sample design; second, using weights but ignoring sample design; and third, using weights and taking into account sample design. We observe the following:

- For both CH and EE, both weighting and sample design have very little effect on the estimated standard errors;
- For AT, both weighting and sample design have a substantial effect on the estimated standard errors;
- For BE, weighting has a small effect and sample design has a large effect;
- For both CH and EE, the direction of the effects is that weighting slightly increases the standard errors and sample design slightly reduces them. The latter result is because these countries have a (slightly) beneficial effect of stratification and no effect of clustering as the design is single-stage;
- For both AT and BE the direction of the effect of sample design is to increase the standard errors. This indicates that the negative effect of sample clustering far outweighs the positive effect of sample stratification. For both these countries, the standard errors would be under-estimated if the design is not taken into account (by a factor of around 0.913 for AT and 0.827 for BE).

Box 5: Comparing naïve and complex estimates across countries

```
mean ppltrst if ppltrst<11 & (cntry2<4|cntry2==6), over(cntry2)

Mean estimation                   Number of obs   =      7,316

------------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
ppltrst      |
          AT |    5.36753   .0515579      5.266462    5.468598
          BE |   5.246319   .0513778      5.145604    5.347035
          CH |    5.98622   .0542012      5.879971     6.09247
          EE |    5.71556   .0463617      5.624678    5.806442
------------------------------------------------------------------


mean ppltrst if ppltrst<11 & (cntry2<4|cntry2==6) [pw=anweight],
over(cntry2)

Mean estimation                   Number of obs   =      7,316

------------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
ppltrst      |
          AT |   5.450543   .0613481      5.330283    5.570803
          BE |   5.167452   .0535956      5.062389    5.272514
          CH |   5.966669   .0550286      5.858797    6.074541
          EE |   5.733973   .0467657      5.642299    5.825647
------------------------------------------------------------------


svy: mean ppltrst if ppltrst<11 & (cntry2<4|cntry2==6), over(cntry2)

Survey: Mean estimation

Number of strata =     125        Number of obs   =      7,316
Number of PSUs   =   4,483        Population size = 2,501.4198
                                  Design df       =      4,358

------------------------------------------------------------------
             |             Linearized
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
ppltrst      |
          AT |   5.450543   .0671638      5.318868    5.582218
          BE |   5.167452   .0647914      5.040428    5.294476
          CH |   5.966669   .0539696      5.860861    6.072477
          EE |   5.733973   .0464553      5.642897    5.825049
------------------------------------------------------------------
```

References

Gabler, S., Häder, S. & Lynn, P. (2006) 'Design effects for multiple design samples'. *Survey Methodology* 32 (1), 115-120. https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X20060019256

Kaminska, O. & Lynn, P. (2017) 'Survey-based cross-country comparisons where countries vary in sample design: issues and solutions'. *Journal of Official Statistics* 33(1), 123-136. https://doi.org/10.1515/jos-2017-0007.

Lynn, P. (2018) The advantage and disadvantage of implicitly stratified sampling. *Methods, Data, Analyses*. Published online 10 October 2018. https://doi.org/10.12758/mda.2018.02.

Lynn, P. & Anghelescu, G. (2018) European Social Survey Round 8 Weighting Strategy. Available at https://www.europeansocialsurvey.org/docs/round8/methods/ESS8_weighting_strategy.pdf

West, B.T., Sakshaug, J.W. & Aurelien, G.A.S. (2018) 'Accounting for complex sampling in survey estimation: a review of current software tools'. *Journal of Official Statistics* 34(3), 721-752. https://doi.org/10.2478/jos-2018-0034