Estimation of Design Effects for ESS Round II

Documentation

Matthias Ganninger

Tel: +49 (0)621-1246-189
E-Mail: ganninger@zuma-mannheim.de

November 30, 2006

1 Introduction

For the calculation of the effective sample size, $n_{\text{eff}} = {}^{n_{\text{net}}}/_{\text{deff}}$, as well as for the determination of the *required net sample size* in round III, $n_{\text{net-req(III)}} = n_0 \times \text{deff}$, the design effect is calculated based on ESS round II data. The *minimum required effective sample size*, n_0, is 800 for countries with less than 2 Mio. inhabitants aged 15 and over and 1 500 for all other countries.

In the above formulas, n_{net} denotes the *net sample size* and deff is a combination of two separate design effects: firstly, the design effect due to unequal selection probabilities, deff_p, and, secondly, the design effect due to clustering, deff_c. These two quantities, multiplied with each other, then, give the overall design effect according to Gabler et al. (1999):

$$\text{deff} = \text{deff}_p \times \text{deff}_c.$$

This document reports in detail how these quantities are calculated.

The remaining part of this section gives an overview of the sampling schemes in the participating countries. In section 2, a description of the preliminary work, foregoing the actual calculation, is given. Then, in section 3, introduces the calculation of the design effect due to clustering (deff_c) and due to unequal selection probabilities (deff_p). The last section closes with some remarks on necessary adjustments of the ESS data to fit the models. The design effects based on ESS round II data are compared to those from round I in section 4.

Table 1 on the following page gives an overview of the countries participating in ESS round II and their respective sampling schemes (Häder and Lynn 2006).

2 Data Requirements and Data Preparation

For the calculation of deff_c and deff_p, ESS data has to be prepared in a way that makes computation convenient. Two data files underlay the calculation of deff_c and deff_p. On the one hand, for the calculation of deff_p, only the sample design datafile (SDDF) is needed. It includes the inclusion probabilities for every respondent in the sample. Additionally, for the calculation of the design effect due to clustering, access to some selected study variables is necessary. These study variables are taken from the ESS round II data file.

2.1 Sample Design Datafile

The data in the sample design datafile accounts for the distinctive features of sampling within a country. The SDDF contains information about an element's inclusion probability, the primary sampling unit (PSU) it belongs to, and other features related to sampling and fieldwork. The SDDF is generated under the supervision of a country's National Co-ordinator who, in

Country	Design	Units
Austria	strat, clus, 3 stages	H
Belgium	Cities: srs	P
	Rest: strat, clus, 2 stages	
Czech Republic	strat, clus, 4 stages	A
Denmark	srs	P
Estonia	systrs	P
Finland	systrs	P
France	strat, clus, 3 stages	A
Germany	strat, clus, 2 stages	P
Greece	strat, clus, 3 stages	A
Hungary	Cities: srs	P
	Rest: strat, clus, 2 stages	
Iceland	srs	P
Ireland	strat, clus, 3 stages	A
Israel	strat, clus, 3 stages	H
Italy	strat, clus, 4 stages	A
Luxembourg	stratrs	P
Netherlands	stratrs	A
Norway	srs	P
Poland	Cities: srs	P
	Rest: strat, clus, 2 stages	
Portugal	strat, clus, 3 stages	A
Slovakia	srs	P
Slovenia	strat, clus, 2 stages	P
Spain	strat, clus, 2 stages	P
Sweden	srs	P
Switzerland	strat, clus, 3 stages	A
United Kingdom	GB: strat, clus, 3 stages	A
	NI: srs	
Ukraine	strat, clus, 4 stages	A

Table 1: ESS II sampling schemes of participating countries.
Details refer to designs used at round II, except in the cases of Israel and Italy (who did not participate in round II), where they refer to round 1. strat: stratified; clus: clustered; srs: simple random sample; systrs: systematic random sample; stratrs: stratified (unclustered) random sample; stages: number of stages of selection; pts: number of sample points; H: household; P: person; A: address;

turn, is instructed by the responsible sampling expert. A detailed description how SDDF variables are being generated is given in a separate user's guide (see appendix A).

The most important information given in the SDDF are an element's inclusion probabilities on every sampling stage. The maximum number of sampling stages in ESS round II is four. Inclusion probabilities at each stage are stored in the variables PROB1 to PROB4. These values indicate an element's probability of being selected into the sample at the corresponding sampling stage.

For example, if the last stage in a three-stage clustered sample are persons (aged 15 and over) in households, a specific person in a four-person household has a 1/4 chance of selection (if a household member is selected at random, for example via the next/last-birthday-method or a kish-grid); thus, PROB3=0.25.

2.2 Study Variables

In order to calculate deff$_c$, 26 variables, which can be reasonably assumed to be *typical* ESS (core) variables[1], are selected from a country's data file, prepared for calculation, and stored in a separate dataset. These variables are referred to as *study variables*.

The set of $O = 26$ ($o = 1, \ldots, O$) variables is grouped to form $Z = 9$ ($z = 1, \ldots, Z$) aggregated study variables. Table 2 shows the original and the newly generated variables.

New	Original	ESS II Variable no.	z(o)	o
DISCRIM	dscrrce	124	1	1
	dscrrlg	126	1	2
	dscrage	129	1	3
	dscrgnd	130	1	4
SATISFAC	stfeco	92	2	5
	stfgov	93	2	6
	stfdem	94	2	7
	stfedu	95	2	8
PPLOK	ppltrst	14	3	9
	pplfair	15	3	10
	pplhlp	16	3	11
POLACT	contplt	46	4	12
	wrkprty	47	4	13
	wrkorg	48	4	14
	badge	49	4	15
	sgnptit	50	4	16
	pbldmn	51	4	17
	bctprd	52	4	18
TRUSTLEG	trstlgl	21	5	19
	trstplc	22	5	20
TRUSTGOV	trstprl	20	6	21
	trstplt	23	6	22
	trstep	25	6	23
NETUSE	netuse	13	7	24
LRSCALE	lrscale	90	8	25
STFLIFE	stflife	91	9	26

Table 2: ESS II study variables considered relevant for the calculation of deff$_c$.

The grouping is done by summing up the non-missing values of the ith

[1]Note that deff$_c$ will take on different values for different variables. However, there exists no criterion, which variables to select in order to achieve a *good* estimation of the design effect of the whole study. Thus, the selection of specific ESS variables is rather subjective and leaves room for discussion.

individual of those original variables, o, that belong to the same group, z. Then, this sum is divided by the number of non-missing entries on the original variables of a given group, $\widetilde{n}_{iz(o)=z}$. Thus, the aggregated study variable for individual i is the average of the original study variables, y_{io}, that make up the group of the aggregated study variable, namely

$$y_{iz} = \frac{1}{\widetilde{n}_{iz(o)=z}} \sum_{\substack{z(o)=z \\ y_{io} \in V}} y_{io}$$

where V denotes the set of non-missing values on the original study variables, y_{io}.

Example:

Assume a person has the following values on the SATISFAC, $z(o) = 2, o = \{5, 6, 7, 8\}$, variables:

IDNO	...	stfeco	stfgov	stfdem	stfedu	...
0073	...	6	88	7	8	...

Note that the respondent did not answer the question how satisfied he is with the government nowadays (stfgov). Accordingly, in the dataset, the missing value is indicated by the value 88. Thus, $\widetilde{n}_{0073\,2} = 3$.
Then, the respondent's mean value on the original variables is the sum of the valid values, 6+7+8, divided by the number of valid values, 3, giving an average satisfaction score (SATISFAC) of $y_{0073\,2} = 7$. The remaining variables are generated accordingly.

3 Calculation of deff$_p$ and deff$_c$

3.1 Design effect due to unequal selection probabilities - deff$_p$

In the following, let $n = \text{n}_{\text{net}}$. Based on the inclusion probabilities given in the SDDF, a *design weight* is calculated for each $i = 1, \ldots, n$ individual in the sample. The design weight is defined as the inverse of the product of the single inclusion probabilities as

$$w_i = \frac{1}{\prod_{k=1}^{K} \text{PROB}_{ik}}$$

where K depends on the number of sampling stages in a given country. These weights are then standardized on their mean value. Thus, the standardized individual design weights, w_i^*, are defined as

$$w_i^* = \frac{w_i}{\sum_{i=1}^{n} w_i} \times n = \frac{w_i}{\bar{w}}.$$

It is common in survey sampling to truncate design weights at a specific boundary. In ESS II, this boundary was set to 4.0. An algorithm is run on the w_i^* that truncates weights greater than 4.0 and, simultaneously, re-standardizes the weights. The design effect due to unequal selection probabilities, then, is defined as

$$\text{deff}_p = \frac{\sum_{i=1}^{n} w_i^{*2}}{\left(\sum_{i=1}^{n} w_i^{*}\right)^2} \times n. \tag{1}$$

Example:
Assume a country with three stage cluster sampling. Let the first stage correspond to the selection of municipalities and the second stage to the selection of households within these municipalities. The final stage corresponds to the selection of one person within each household. The country under consideration has a total population aged 15 and over of $N = 48\,665\,774$. In this country, there are 9 827 municipalities of which 187 are to be selected into the ESS sample by probability proportional to their size (pps). Let us assume the net sample size is $n = 1\,844$ in this country.
Then, the inclusion probability of a specific municipality is given by $187 \times \frac{N_g}{N}$ where N_g is the number of persons aged 15 and over in the gth municipality. It is easily seen that every individual living in this specific municipality receives the same PROB1 inclusion probability.
We are interested in the inclusion probability of the 24th ($g = 24$) municipality which is of size $N_{24} = 26\,985$. Then, the inclusion probability on the first stage, PROB1, for every individual in the sample belonging to the 24th municipality is $187 \times \frac{26\,985}{48\,665\,774} = 0.10369084$.
For the calculation of PROB2, we have to know the total number of households in the municipality under consideration, H_g and, additionally, the number of households which are to be selected in the ESS sample, h_g. Assume that H_{24} is 10 053 and $h_{24} = 12$. Then, the household inclusion probability in the 24th municipality is PROB2 $= \frac{12}{10\,053} = 0.00119367353$ for all individuals in the 24th municipality.
Finally, the inclusion probability on the third stage, PROB3, is the inverse of the total number of persons aged 15 and over in the jth household of the gth municipality, m_{gj}. Assume there are three persons living in the 4th selected household in the 24th municipality. Then, using a random selection scheme (e.g. Kish-grid, last/next-birthday-method, etc.), PROB3 $= \frac{1}{3} = 0.\bar{3}$.

Now, a respondent's design weight is easily calculated as the inverse of the product of his individual inclusion probabilities, PROB1, PROB2, and PROB3. According to the above case, assume the $i = 845$th respondent lives in a three person household in the 24th municipality. Then, his/her design weight, is

$$w_{845} = \frac{1}{0.10369084 \times 0.00119367353 \times 0.\bar{3}} = 24\,237.9175.$$

Assume we had calculated the weight for every individual as well as the mean weight, $\bar{w}$. Then, the 845th respondent's standardized design weight is

$$w^*_{845} = 24\,237.9175/\bar{w} = 24\,237.9175/22\,954.25 = 1.05592287.$$

Finally, deff_p is calculated in the following manner: first, each design weight of the $n = 1\,844$ respondents in the sample is squared. Then, these squared weights are summed up and divided by the squared sum of the 1 844 weights multiplied by the sample size.

For our example, assume that $\sum_{i=1}^{n} w_i^{*2} = 2\,099.236$ and $\left(\sum_{i=1}^{n} w_i^*\right)^2 = 3\,316\,743$. Then the design effect due to unequal selection probabilities is

$$\text{deff}_p = \frac{2\,099.236}{3\,316\,743} \times 1\,844 = 1.1671.$$

3.2 Design effect due to clustering - deff$_c$

The design effect due to clustering is defined as the ratio of the variance of a given clustered sample's estimate to the estimate's variance under SRS. Following Kish (1965, p. 258), the true design effect is defined as

$$\text{DEFF}_c = \frac{\text{Var}(\theta)_{clu}}{\text{Var}(\theta)_{srs}}, \tag{2}$$

where θ denotes an estimate under consideration, $\text{Var}(\cdot)_{clu}$ is the variance of an estimate under cluster sampling, $\text{Var}(\cdot)_{srs}$ under simple random sampling[2]. This ratio always refers to a specific variable. However, to make a statement about the overall design effect due to clustering in a given country, deff_c is calculated for a set of nine ESS core variables[3].

In the ESS sampling context, formula (2), which yields a *design based* design effect, seems inappropriate since for the variance estimator of the clus-

[2]In the following, the terms *cluster* and *PSU* are used interchangeably.

[3]See table 2 on page 3.

ter design, PSUs of equal size are assumed as well as only one sampling stage. Thus, a *model-based* design effect is used for the calculation (Gabler et al. 1999). According to this formula, the model-based design estimator for the effect due to clustering is defined as

$$\mathrm{deff}_c{}' = 1 + (\bar{b} - 1)\rho \qquad (3)$$

where $\bar{b}$ denotes the average cluster size and ρ is a *measure of homogeneity* – often ρ is referred to as the *intra-class correlation coefficient* (icc). The intra-class correlation coefficient measures the homogeneity (with respect to the variable of interest) of the individuals within a cluster. Since individuals that belong to the same cluster tend to be more similar to each other than to individuals of another cluster, ρ tends to take on small positive values and, clearly, cannot exceed unity. Typical values of ρ range from 0.02 to 0.10 and rarely exceed 0.2.[4]

Example:
For illustrative reasons assume a cluster sample where the number of respondents in each cluster is constant, so b_m, the size of the mth cluster equals $\bar{b}$, the mean cluster size. Let us assume $b_m = \bar{b} = 9$. The net sample size in this sample is $\mathrm{n_{net}} = 1\,800$ and thus M, the total number of clusters, is 200. For simplicity, also assume that deff_p equals one.
Now, assume there is perfect homogeneity within the clusters which means that all respondents that belong to the same cluster show the same value on the study variable under consideration. Then, y_{ij}, the value on study variable y for person j in the ith cluster is constant for all j and given i thus $\bar{y}_i$, the mean of y in the ith cluster, equals y_{ij}.
In other words, there is no variation within a cluster. Thus, all variation arises from between-cluster differences, namely the differences of the 200 cluster means, $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_m, \ldots, \bar{y}_{200}$. In this case ρ equals one by definition and $\mathrm{deff}_c = 1 + (9 - 1) = 9$, which simply is the cluster size.
This means that the effective sample size, $\mathrm{n_{net}}/\mathrm{deff}$, is the net sample size decreased by the factor of 9, yielding the total number of clusters, $M = 200$. This is equivalent to the claim that only one person per cluster has to be interviewed to gain the same amount of information.

If the cluster size shows great variation, an alternative measure of size, b^*, is used (Lynn and Gabler 2005). It is given by

$$b^* = \frac{\sum_{m=1}^{M} \sum_{j=1}^{b_m} w_{mj}^{*2}}{\left(\sum_{j=1}^{M} w_j^*\right)^2}.$$

[4] From (3) it is obvious that, for fixed ρ, the design effect increases linearly with the mean cluster size. Thus, small cluster sizes are desirable.

The design effect due to clustering based on b^*, then, is defined as

$$\text{deff}_c{''} = 1 + (b^* - 1)\rho. \qquad (4)$$

Example:
Assume we had calculated ρ for the variable SATISFAC as 0.0535. Then, for the calculation of deff_c, we only need the mean cluster size, $\bar{b}$, which is readily calculated as the sample size, n, divided by the number of clusters, M. In our example, $\bar{b} = \frac{1\,844}{187} = 9.86$ and $b^* = 9.23$. Now, two estimates of DEFF_c, the first based on $\bar{b}$, the second on b^*, are given by the estimators

$$\text{deff}_c{'} = 1 + 8.86 \times 0.0535 = 1.47$$

and

$$\text{deff}_c{''} = 1 + 9.23 \times 0.0535 = 1.49.$$

Note that missing values are excluded in the calculation of $\text{deff}_c{'}$ and $\text{deff}_c{''}$. This leads to the effect that, especially in small clusters, the number of non-missing values may reduce to one or even zero units per cluster. Thus, the *effective cluster size* defined as the number of individuals that have non-missing values on the study variable. Since, for example, the estimate of the intra-cluster variance is undefined for clusters of size one, also the summation of the individual intra-cluster variances runs only over those clusters where the number of respondents with non-missing values is greater than one.

Another note concerns the special case where some clusters are selected via srs as for example in the Northern Ireland part of the UK sample, the clusters in Berlin in Germany, and the Warsaw part of the Polish sample. These PSUs are excluded for the calculation of $\text{deff}_c{'}$ and $\text{deff}_c{''}$.

4 Comparison of design effects in ESS round I and II

In table 8 the median values of ρ and the design effects (deff_c, deff_p, and deff) are given for ESS round I (columns 3-6) and II (columns 7-10). All countries which have at least participated in ESS round II are included. The differences between the round I and round II design effects are given in columns 11-14. In column 15 the net sample size in round II is displayed. The next column contains the effective sample size, n_{eff}, which is defined as $n_{\text{net}}/\text{deff}$. The last two columns contain information about changes in sampling from round I to round II (column 15) and the total number of PSUs in round II (column 17).

Example:

For a specific country, deff_c is calculated in the following way: fist, ρ and b^* are calculated for every study variable specified above. Then, the overall deff_c is calculated as $\text{deff}_c = 1 + (\max(b^*) - 1) \times \tilde{\rho}$, where $\tilde{\rho}$ is the median value of the ρs of the nine study variables.

The following table gives estimates of the ρs and b^*s for every study variable in the case of Spain:

	ρ	b^*
DISCRIM	0.0879	4.7339
SATISFAC	0.0726	4.6801
PPLOK	0.1400	4.7339
POLACT	0.0637	4.7310
TRUSTLEG	0.1623	4.6386
TRUSTGOV	0.1181	4.5839
NETUSE	0.1097	4.7103
LRSCALE	0.0992	3.5932
STFLIFE	0.1066	4.6836

From the above table it is easily seen that the median value of ρ is 0.1066 and the maximum of b^* is 4.7339. Thus, the Spanish deff_c is given by

$$1 + (4.7339 - 1) \times 0.1066 = 1.3980 \approx 1.40$$

as denoted in column eight of table 8.

The differences of the round II and round I design effects are given in columns 11-14 of table 8 above. These differences are calculated by subtracting the values of deff_c, deff_p, and deff of round I from those of round II. Thus, negative values indicate a decrease (desired; marked green) in the design effect, positive values, in turn, indicate an increase (not desired; indicated by orange for values between +0 and +.25; red for differences >.25).

As is easily seen, most values in columns 11-14 are green. This indicates that most countries were able to improve their sample design from round I to II. Even more, in all countries that increased the number of PSUs (Austria, Switzerland, Spain, United Kingdom, Greece, Poland, and Portugal), the design effect decreased. In Norway, where the sample design changed from a clustered sample to simple random sampling, the greatest improvement in deff among all countries is achieved.

In The Netherlands and in Slovenia, slight increases in deff can be observed. These increases, at least in the first and in the last case, go back to increases in deff_c, which can be explained by changes in the composition of clusters from round I to II that may occur by chance.

Table 8: ESS I & II design effects.
*=SDDF and MS missing; deff$_c$ was calculated using b^*; deff is the product of deff$_p$ and deff$_c$ (see (1) on page 1); Mean and Median are based on all participating countries in the respective round. All values rounded on two decimals.

		ESS I				ESS II				Difference						
ID	Country	ρ	deff$_c$	deff$_p$	deff	ρ	deff$_c$	deff$_p$	deff	ρ	deff$_c$	deff$_p$	deff	n_{net}	Change	PSUs
1	Austria	0.11	1.61	1.25	2.01	0.10	1.47	1.25	1.84	-0.01	-0.15	0.01	-0.17	2556	PSUs (250 to 360)	360
2	Belgium	0.04	1.22	1.00	1.22	0.04	1.19	1.00	1.19	0.00	-0.02	0.00	-0.02	1778	none	324
3	Switzerland	0.03	1.27	1.21	1.54	0.04	1.25	1.22	1.53	0.00	-0.02	0.01	-0.01	2141	PSUs (222 to 282)	282
4	Czech Republic	0.15	1.28	1.25	1.61	0.14	2.60	1.50	3.91	-0.01	1.32	0.25	2.31	3026	complete sampling	275
5	Germany	0.06	2.03	1.11	2.26	0.06	2.01	1.11	2.23	0.01	-0.03	0.00	-0.03	2870	none	163
6	Denmark		1.00	1.00	1.00		1.00	1.00	1.00		0.00	0.00	0.00	1487	none	1
7	Estonia						1.00	1.06	1.06					1989	not in ESS I	1
8	Spain	0.15	1.60	1.22	1.95	0.11	1.40	1.01	1.41	-0.04	-0.20	-0.21	-0.54	1663	PSUs (346 to 479)	479
9	Finland		1.00	1.00	1.00		1.00	1.00	1.00		0.00	0.00	0.00	2022	none	1
10	France	0.05	1.34	1.23	1.65	0.04	1.36	1.19	1.62	-0.02	0.02	-0.04	-0.03	1806	PSUs (169 to 200)	200
11	United Kingdom	0.04	1.39	1.22	1.69	0.04	1.34	1.26	1.69	0.00	-0.05	0.04	-0.01	1897	PSUs (163 to 168)	168
12	Greece	0.14	1.64	1.22	2.00	0.10	1.36	1.20	1.64	-0.04	-0.28	-0.01	-0.36	2406	PSUs (438 to 526)	526
13	Hungary	0.05	1.36	1.00	1.36	0.05	2.41	2.16	5.21	0.00	1.05	1.16	3.85	1498	complete sampling	54
14	Ireland	0.08	1.92	1.04	2.01	0.08	1.87	1.31	2.44	0.00	-0.05	0.26	0.43	2286	none	193
15	Iceland						1.00	1.00	1.00					581	not in ESS I	1
16	Luxembourg		1.00	1.26	1.26		1.00	1.15	1.15		0.00	-0.11	-0.11	1635	none	1
17	Netherlands		1.00	1.19	1.19		1.00	1.20	1.20		0.00	0.01	0.01	1881	none	1
18	Norway	0.01	1.60	1.00	1.61		1.00	1.00	1.00		-0.60	0.00	-0.61	1761	CLU to SRS	1
19	Poland	0.06	1.83	1.02	1.87	0.07	1.57	1.01	1.59	0.01	-0.26	-0.01	-0.28	1717	PSUs (158 to 313)	313
20	Portugal	0.14	1.57	1.83	2.88	0.16	1.88	1.39	2.61	0.02	0.31	-0.45	-0.28	2052	PSUs (150 to 323)	323
21	Sweden		1.00	1.00	1.00		1.00	1.00	1.00		0.00	0.00	0.00	1948	none	1
22	Slovenia	0.03	1.33	1.00	1.33	0.05	1.39	1.00	1.39	0.01	0.06	0.00	0.06	1325	none	151
23	Slovakia						1.00	1.00	1.00					1512	not in ESS I	1
24	Turkey*														not in ESS I	
25	Ukraine					0.16	1.95	1.73	3.38					2031	not in ESS I	213
Mean		0.08	1.53	1.15	1.62	0.08	1.67	1.20	1.83	0.00	0.05	0.05	0.21	1911		
Median		0.06	1.57	1.15	1.61	0.07	1.47	1.15	1.53	0.00	-0.01	0.00	-0.02	1889		

The sampling schemes in the Czech Republic and in Hungary changed completely from round I to round II. These changes caused a dramatic increase in both, deff_c and deff_p in both countries. Unfortunately, by now even intensive investigations shed no light on the exact circumstances that have lead to these dramatic increases in deff_c, deff_p, and deff in both countries. Clearly, ρ in the Czech Republic is very high in both waves. Thus, the increase in deff_cmust be attributed almost solely to an increase in b^*. This claim is supported by the increase in deff_p, since both quantities take into consideration the standardized weights.

In Hungary, increases of the design effects are even more dramatic. But also in this case, the increases are – presumably – mainly caused by an increase in the variance of inclusion probabilities and thus in the weights.

Turning to the effective sample size, n_{eff}, it has to be noted that only six countries (Estonia, Finland, The Netherlands, Norway, Sweden, and Slovakia) achieved the required effective sample size as specified for ESS round II. Belgium, Denmark and Greece only slightly missed it.

As a conclusion, taking into consideration the median values of deff_c, deff_p, and deff of the first and the second round, it is easily seen that the median values of deff_c clearly decreased, deff_p stayed the same, and deff also decreased. Calculating the mean value of deff_c, deff_p, and deff for only those countries that participated in both rounds and excluding the two outliers (The Czech Republic and Hungary), the average of deff_c decreased by -.10 from 1.57 in round I to 1.47 in round II, the mean of deff_p also decreased from 1.16 to 1.13 and the mean value of deff even decreased by -.11 from 1.64 to 1.53.

5 Appendix

A User's Guide to Sample Design Datafile Creation

European Social Survey Round 2
SAMPLE DESIGN DATA FILE

This is the description of the data file needed regarding the sample design.

The file should contain one record for each selected address, in the case of address-based samples, or for each selected individual, in the case of individual-based (population register) samples. (In the rare event of a design in which where it is possible to select more than one household at an address, extra records will be needed for any extra households thus selected.) Please make sure that you include the non-responding units as well as the responding ones.

Column 1: Name of the variable
Column 2: Short description of the variable - including logical relationship with other variables.
Column 3: Source of the variable.
S = initial office-based sample selection process
F = field-based sampling process (interviewers)
Column 4 : Reason for variable
W = For the construction of design weights
SE = For estimation of design effects and standard errors

The annex contains examples of how the variables should be interpreted for different types of sample design. It is suggested that you look closely at these examples. National co-ordinators may seek advice from their nominated member of the sampling expert panel if necessary.

Note on missing values: Use code 9 for items that are missing through interviewer error. This should apply only to certain items (F in column 3) in the case of address-based sampling. Use 'blank' for items that are not applicable (e.g. 3rd-stage probabilities, in the case of a 2-stage design).
We assume that SPSS or SAS system files will be delivered.

NAME of variable	MEANING OF VARIABLE (label) and codes	I	Info
IDNO	Address number *(same as in main datafile)* Code = number	X	SE
IDEXTRA	Extra code to distinguish between households/persons selected at the same address (*applies only to the Netherlands*) Code = 1, 2, … (within IDNO)	X	SE
PROB1	Selection probability at first stage of sampling. Must be numeric > 0, with 12 decimal places (or as many as necessary to provide at least 4 significant figures for all cases). For multi-stage samples, all records with the same value of PSU should have the same value of PROB1.	S	W
PROB2	Only to be used if there are at least 2 stages of selection. Otherwise, PROB2 should be blank. Selection probability at second stage of sampling (conditional upon first stage). Must be numeric > 0, with 12 decimal places (or as many as necessary to provide at least 4 significant figures for all cases). All records with the same combination of PSU/SSU should have the same value of PROB2.	S or F	W

PROB3	Only to be used if there are at least 3 stages of selection. Otherwise, PROB3 should be blank. Selection probability at third stage of sampling (conditional upon first two stages). Must be numeric > 0, with 12 decimal places (or as many as necessary to provide at least 4 significant figures for all cases).	F	W
PROB4	Only to be used if there are 4 stages of selection. Otherwise, PROB4 should be blank. Selection probability at fourth stage of sampling (conditional upon first 3 stages). Must be numeric > 0, with 12 decimal places (or as many as necessary to provide at least 4 significant figures for all cases).	F	W
PSU	To be used if there is more than 1 stage of selection. For 1-stage designs, PSU should be blank. Indicator of the first stage unit (primary sampling unit) to which the sample address/person belongs. Integer. Note that a key to this variable is NOT necessary. In other words, the PSUs can be completely anonymous.	S	SE
SAMPPOINT	To be used only for designs where PSUs and sample points are not co-terminous. For 1-stage designs, and designs where PSU and sampling point are co-terminous, PSU should be blank. Indicator of the sampling point to which the address/person belongs. Integer. As for PSU, a key is not necessary, so sampling points can be anonymous.	S	SE
SSU	Only to be used if there are at least 3 stages and more than one 3rd-stage unit is selected within each 2nd-stage unit (only Netherlands?). Otherwise, SSU should be blank. Indicator of the second stage unit (secondary sampling unit) to which the sample person belongs. Integral. Need only be unique within PSU.	S	SE
STRATEX1	Only to be used if explicit stratification was used to select the PSUs (independent selections from each stratum). Otherwise, STRATEX1 should be blank. Indicator of stage 1 explicit stratum to which the sample address/person's PSU belongs. All records with the same value of PSU should have the same value of STRATEX1.	S	SE
STRATIM1	Only to be used if implicit stratification was used to select the PSUs (systematic selections - possibly PPS - from an ordered list). Otherwise, STRATIM1 should be blank. Indicator of order on the list (order of selection) of the sample address/person's PSU. All records with the same value of PSU should have the same value of STRATIM1. All records with the same value of STRATIM1should have the same value of PSU.	S	SE
STRATIM2	Only to be used if there were at least 3 stages and if implicit stratification was used to select the SSUs (systematic selections from an ordered list). Otherwise, STRATIM2 should be blank. Indicator of order on the list (order of selection) of the sample address/person's SSU within the PSU. All records with the same combination of PSU/SSU (if any) should have the same value of STRATIM2.	S	SE

STRATVAL1	Only to be used if implicit stratification was used to select the PSUs. Otherwise, STRATVAL1 should be blank. The numeric value of the (first) variable used to stratify PSUs.	S	SE
STRATVAL2	Only to be used if implicit stratification was used to select the PSUs, based upon a combination of at least two variables. Otherwise, STRATVAL2 should be blank. The numeric value of the second variable used to stratify PSUs.	S	SE
STRATVAL3	Only to be used if implicit stratification was used to select the PSUs, based upon a combination of at least three variables. Otherwise, STRATVAL3 should be blank. The numeric value of the third variable used to stratify PSUs.	S	SE
OUTCOME	Summary of field outcome. Code = 1 for respondent (data in questionnaire data file), 2 for eligible non-respondent (refusal, non-contact, etc), 3 for ineligible	F	W

References

European Social Survey: 2003, Specification for participating countries, *Technical report*, European Social Survey. http://naticent02.uuhost.uk.uu.net/proj_spec/round_2/r2_spec_participating_countries.doc.

Gabler, S., Häder, S. and Lahiri, P.: 1999, A model based justification of kish's formula for design effects for weighting and clustering, *Survey Methodology* **25**(1), 105–106.

Häder, S. and Lynn, P.: 2006, How representative can a multi-nation survey be?, *in* R. jowell, C. Roberts, R. Fitzgerald and G. Eva (eds), *Measuring attitudes cross-nationally. Lessons from the European Social Survey*, Sage. forthcoming.

Kish, L.: 1965, *Survey Sampling*, John Wiley & Sons.

Lohr, S.: 1965, *Sampling: Design and Analysis*, Duxbury Press.

Lynn, P. and Gabler, S.: 2005, Some title, *Survey Methodology* **1**(2).

Särndal, C.-E., B., S. and Wretman, J.: 1965, *Model Assisted Survey Sampling*, Springer-Verlag.