Sampling and Estimation of Multi-National Surveys with Examples from the European Social Survey

Seppo Laaksonen
University of Helsinki and Statistics Finland
Seppo.Laaksonen@Helsinki.Fi [1]

Key words: Design effects, effective sample size, intra-cluster correlation, sampling frames, cross-country analysis of happiness

1. Introduction

This paper first discusses the objectives of sample design for cross-national surveys (section 2). Then we describe the principles and requirements for sample design that were developed for the European Social Survey (ESS) in order to meet these objectives (section 3). In particular, these include a requirement to predict design effects and to use these predictions in determining national sample sizes. The procedures used on the ESS are described in section 4, and some of the strengths and weaknesses are pointed out. In section 5, some cross-country analysis have been presented, just as examples how to do it and to motivate to exploit these data files that are freely available for everyone. Section 6 concludes and presents some ideas for improving the survey.

2. Sample Design for Cross-National Surveys

To enable comparisons between nations, the ESS sampling group suggests that national sample designs for cross-national surveys must meet two fundamental criteria:

- The study population must be equivalent in each nation. In practice, this will usually mean that the same population definition is applied in each nation and that no or minimal under-coverage can be permitted;
- Sample-based estimates must have known and appropriate precision in each nation. In practice, "known" precision means that a strict probability sample design must be used, and those aspects of sample design that affect precision (selection probabilities, stratum membership, primary sampling unit (psu) membership) must be available on the microdata to permit estimation of standard errors; "appropriate" precision may mean, a) meeting some minimum precision requirement in order for the estimates to be useful and, b) aiming for similar precision in each nation.

To best meet these criteria, it is likely that details of the sample design will vary between nations (Le and Verma, 1997). The goal is functional equivalence, not replication of parameters of the sample

design. As Kish (1994, 173) writes, "Sample designs may be chosen flexibly and there is no need for similarity of sample designs. Flexibility of choice is particularly advisable for multinational comparisons, because the sampling resources differ greatly between countries. All this flexibility assumes probability selection methods: known probabilities of selection for all population elements." Therefore, an optimal sample design for a cross-national survey should consist of the best probability sample design possible in each nation, where "best" can be interpreted as an optimum trade-off between cost and precision. The choice of a specific national design depends on the available frames, experiences, other constraints such as those that may be imposed by the national legal infrastructure and, of course, costs of sample selection and data collection (Häder and Gabler, 2003). If adequate estimators are chosen, the resulting estimates can be compared using appropriate statistical tests.

3. Requirements of Sample Design for the European Social Survey

3.1 The European Social Survey

The ESS is an academically-driven social survey designed to chart and explain the attitudes, beliefs and behaviour patterns of Europe´s diverse populations. In parallel with its substantive aims, it aims also to provide a model of best practice in methodology and to contribute towards improvement in methodological standards (further details: www.europeansocialsurvey.org). The ESS is funded via the European Commission's Framework Programmes, with supplementary funds from the European Science Foundation. In each participating nation, the cost of data collection and the appointment of a national co-ordinator (NC) is funded by the national research council or equivalent body. An important principal of the survey is that the data are made freely available: no-one involved in the survey has advance access and there are no restrictions on access. Data can be downloaded from http://ess.nsd.uib.no.

There is a core questionnaire that is administered in every round, along with modules of questions that will change from round to round. Nations are not asked to commit themselves to more than one round at a time, though of course continued participation is encouraged. All interviews are carried out face-to-face. However, after the interviewing a respondent is asked also to fill-in a self-completed supplementary questionnaire (the big part of this questionnaire includes Schwartz' life values) that will be submitted by mail to the country survey organisation (This was not done in Luxembourg and Italy). It should be noted that there is some second phase unit nonresponse since all first-phase respondents have not answered these supplementary questions.

The ESS consists of regular "rounds" of data collection, with each round involving an independent cross-sectional sample in each nation (it is a repeated survey, not a panel). The first

[1] The points relating to the sampling guidelines and conclusions have been made together with the other ESS sampling experts, that is, Sabine Häder (Zuma, Mannheim), Siegfried Gabler (Zuma. Mannheim) and Peter Lynn (Univ. of Essex).

round of field work took place in September-December 2002 (in a few nations fieldwork was not completed until 2003.). Consequently, the interviews for the second round were performed two years later. The third round is now in August 2006 approaching. Table 1 shows which countries have been participated in this survey. Until now, the 30 countries have contributed at least for one round. The 31th country in the list is Turkey for which the sampling design was accepted for round 2 but the data are still missing.

Table 1. Participation of countries in the ESS 2002-2007

Country	*Round 1*	*Round 2*	*Round 3*	*Country*	*Round 1*	*Round 2*	*Round 3*
Austria	Yes	Yes	Yes	Italy	Yes	No	?
Belgium (Flemish)	Yes	Yes	Yes	Luxembourg	Yes	Yes	No
Belgium (Francophone)	Yes	Yes	Yes	Netherlands	Yes	Yes	Yes
Bulgaria	No	No	Yes	Norway	Yes	Yes	Yes
Cyprus	No	No	Yes	Poland	Yes	Yes	Yes
Czech Republic	Yes	Yes	?	Portugal	Yes	Yes	Yes
Denmark	Yes	Yes	Yes	Romania	No	No	Yes
Estonia	No	Yes	Yes	Russia	No	No	Yes
Finland	Yes	Yes	Yes	Slovak Republic	Yes	No	Yes
France	Yes	No	Yes	Slovenia	Yes	Yes	Yes
Germany	Yes	Yes	Yes	Spain	Yes	Yes	Yes
Greece	Yes	Yes	Yes	Sweden	Yes	Yes	yes
Hungary	Yes	Yes	Yes	Switzerland	Yes	Yes	Yes
Iceland	No	Yes	?	Turkey	No	?	?
Ireland	Yes	Yes	Yes	Ukraine	No	Yes	?
Israel	Yes	No	No	United Kingdom	Yes	Yes	Yes

The ESS sampling design group developed the requirements for participating nations – which will be described in the remainder of this section under five broad headings – and then co-operated with participating nations in developing acceptable sample designs.

3.2 Population definition and coverage

The target population for each participating nation is defined as all persons 15 years or older resident in private households within the borders of the nation, regardless of nationality, citizenship, language or legal status. (In countries in which any minority language is spoken as a first language by 5 % or more of the population, the questionnaire must be translated into that language.) It is worth noting in passing that this definition was subject to considerable discussion by a 21-country steering group prior to agreement.

The requirement for sample design was that every person with the defined characteristics should have a non-zero chance of selection. In practice, the quality of available frames – e.g. coverage, updating and access - differs between nations, so careful evaluation of frames was necessary to assess the likely extent of under-coverage and ensure that any coverage bias was likely to be minimal.

Among others, we found the following kinds of frames:

a) nations with reliable lists of *residents* that are available for social research such as the Danish Central Person Register that has approximately 99 % coverage of persons resident in Denmark;
b) nations with reliable lists of *households* that are available for social research such as the "SIPO" database in the Czech Republic, that is estimated to cover 98% of households;
c) nations with reliable lists of *addresses* that are available for social research such as the postal delivery points in the Netherlands and in the UK;
d) nations without reliable and/or available lists such as Portugal, Russia and Greece.

Drawing a sample is most complicated if no lists are available (group d). In this instance area sample designs were usually applied, in which the selection of a probability sample of small geographical areas (e.g. Census enumeration areas within municipalities) preceded a complete field enumeration of households or dwellings within the sampled areas, from which a sample was selected. In nations where this approach was used (e.g. Greece), the sampling panel insisted that the selection stage should be separated from the enumeration and carried out by office staff or supervisors who had not been present for the enumeration. An alternative to area sampling in this situation is the application of random route sampling about which some survey organisations were enthusiastic. The basic idea of random route sampling is that within each sampled psu one address is selected by a random method to serve as a starting point and the interviewer then follows rules that specify the route he or she should take from there, sampling systematically using a pre-specified interval (Häder and Gabler, 2003). The question here, however, is the extent to which random routes can be judged to be "strictly random". That depends on both the rules for the random walk and the control of the interviewers by the fieldwork organisation in order to minimise interviewer influence on selection. A rigorous version of random route sampling was permitted in one country (Austria).

Even in countries where reliable lists exist, some problems had to be solved. For example, in Italy there is an electoral register available. But it contains, of course, only persons 18 years or older (and

only those who are eligible to vote). Therefore, it had to be used as a frame of addresses. This had not been attempted before and there were practical problems to be overcome, not least the fact that persons at the same address do not necessarily appear together on the list, making it difficult to ascertain the selection probabilities of addresses. Thus, under-coverage, while not zero, was restricted to persons at addresses with no registered electors. In countries with population registers, people with illegal status will be excluded because they are not registered. The practical task for the sampling panel was to ensure that levels of under-coverage were kept to an absolute minimum by considering all possible frames and evaluating the properties of each with respect to the ESS population definition.

3.3 Response rates

Non-response is the next problem for achievement of the objective to represent the target population. A carefully drawn sample from a perfect frame can be devalued if non-response leads to systematic bias. Therefore, it is essential to plan and implement adequate field work strategies to minimise non-contacts and refusals. For the ESS a target response rate of 70% was fixed although it was known that this would be challenging for the countries where response rates between 40 and 55 percent are common. Nevertheless, it was felt that a realistic but challenging target should encourage maximum efforts. Additionally, the ESS required that non-contacts should not exceed 3% of eligible sample units, that at least four personal visits must be made to a sample unit before non-contact was accepted as an outcome, and that the field period must last at least 30 days.

As expected, the target response rate was hard to achieve. Table 2 illustrates this from round 2 that also shows the net sample sizes by country. However, some success from round 1 was happened in Czech Republic and Switzerland, in particular.

Table 2. Response rates and realised interviews from round 2 based on the data from April 2006.

	Number of realised interviews	Rate of ineligibles (%)	Response rate (%)	Non-contact rate (%)	Refusal rate (%)
Austria	2256	1.7	62.5	7.8	28.6
Belgium	1778	4.9	61.5	7.1	22.7
Czech Republic	3026	1.3	55.5	n/a	n/a
Denmark	1487	6.4	65.1	5.6	23.9
Estonia	1989	12.7	79.5	5.1	11.4
Finland	2022	1.5	70.8	2.8	21.2

France	1806	7.1	44.2	12.1	39.5
Germany	2870	7.2	52.7	6.2	27.4
Greece	2406	0.1	78.8	3.7	16.4
Hungary	1498	13.5	70.3	6.0	16.0
Iceland	579	5.9	51.3	4.6	39.1
Ireland	2286	8.1	62.5	9.5	22.3
Luxembourg	1635	10.2	52.1	7.7	40.2
Netherlands	1881	3.0	64.5	2.7	28.0
Norway	1760	3.4	66.2	2.1	25.5
Poland	1716	3.8	74.4	2.3	18.2
Portugal	2052	6.4	70.9	2.8	20.0
Slovenia	1442	6.7	70.2	10.2	15.3
Spain	1663	7.8	56.1	13.6	18.6
Sweden	1948	2.3	66.5	4.3	22.6
Switzerland	2141	6.5	47.1	2.9	39.7
United Kingdom	1897	7.9	51.1	8.0	34.0

3.4 Sample selection methods

We have already argued that strict probability sampling is a necessary pre-requisite for cross-national comparability. However, partly as a measure to overcome the fear of non-response bias, many survey organisations habitually implement substitution of non-cooperative or not reachable primary sampling units, households or target persons by others. There are many varieties of substitution (Vehovar, 2003; Lynn, 2004), but none of them meet the requirement for probability sampling. Another important disadvantage of substitution in the field is that it can reduce the effort made by interviewers to gain a response at the original addresses/households.

For the ESS, substitution of non-responding households or individuals (whether 'refusals' or 'non-contacts') was not permitted in any circumstances. However, in exceptional circumstances substitution was permitted at the first stage of sampling. Administrative considerations may mean that addresses cannot be obtained for specific sampled areas (e.g. a particular municipality may refuse to grant access to the list, or be unable to co-operate within the available time).

3.5 Effective sample size

The ESS requirement was for a minimum estimated effective sample size of 1,500 completed interviews and a minimum of 2,000 actual interviews. (An exception was made for nations with a total population of less than 2 million persons, recognising that resources for funding surveys are considerably constrained in such nations. For such nations, the minimum requirement was an effective sample size of 800 and an actual sample size of 1,000.) Explanation was provided as to what was meant by *effective sample* size and how it should be predicted. This involved predicting, under certain simplifying assumptions, the design effect due to unequal selection probabilities (*DEFFp*) and the design effect due to clustering ($DEFF_C$). Additionally, realistic estimates of response rate and eligibility rate were required in order to calculate the sample size to select in order to produce the target number of completed interviews.

A reasonable approach to sample size determination is to predict the determinants of design effects within reasonable bounds. The aspects of the survey that make this possible are, 1) relatively low – and relatively stable over time - expected correlation between survey variables and psu's; 2) relatively small variation in selection probabilities; 3) prior estimates in several countries for similar variables on surveys with similar designs. Additionally, a repeating survey like ESS offers the opportunity to revise predictions at each round based on estimates from previous rounds.

3.5.1 Design effect due to unequal selection probabilities (DEFFp)

The ESS guidelines suggested that *DEFFp* should be predicted as follows:

$$\tilde{DEFF}_P = \frac{m\sum_{i=1}^{I} m_i \left(w_i^2\right)}{\left(\sum_{i=1}^{I} m_i w_i\right)^2} \qquad (1)$$

where m_i and w_i denote respectively the number of interviews and the design weight associated with the i^{th} weighting class. (This can be expressed equivalently as $1+cv_w^2$, where cv_w is the coefficient of variation of the weights)

In some nations, it is necessary to select the sample in stages, with the penultimate stage being addresses or households. In this case, each person's selection probability depends on the respective household size. The guidelines illustrated estimation of (1) with a hypothetical example of an address-based design of this sort, where the weighting classes were defined by the possible values of number of persons aged 15 or over resident at an address. Several nations use such an address-based design (e.g. Czech Republic, Greece, Ireland, Israel, Netherlands, Portugal, Russia, Spain, Switzerland, UK).

Another reason for unequal selection probabilities is that minority groups are over-sampled for substantive reasons. A third reason is that certain strata (typically, the largest cities) may be over-

sampled in anticipation of lower response rates, though in principle this should not affect variance of estimates as it will lead to equal inclusion probabilities if the response rate predictions turn out to be accurate.

A fourth source of variation in selection probabilities occurs in countries where the psu's are selected with probability proportional to a proxy size measure which does not correlate perfectly with the units sampled at the subsequent stage.

3.5.2 Design effect due to clustering ($DEFF_C$)

The cluster sample size and the intra-class correlation also influence the design effect. Following Kish (1987), the ESS guidelines suggested that $DEFF_C$ should be predicted as follows:

$$\tilde{DEFF}_C = 1 + \left(\bar{b} - 1\right)\rho \qquad (2)$$

where $\bar{b}$ is the mean number of interviews per cluster and ρ is the intra-cluster correlation. Expression (2) implies that, were cost not a consideration, the cluster sample size should be chosen as small as possible. The larger the average cluster size, the more interviews have to be conducted to reach the minimum effective sample size. The challenge, therefore, is to find the combination of $\bar{b}$ and *n* that delivers the desired effective sample size for the lowest cost. Participating nations were encouraged to seek estimates of ρ from other surveys in their country if possible, or alternatively to assume $\rho = 0.02$. In practice, ρ will take different values for different statistics and can also vary between subgroups for any particular statistic, but the ESS sample design requirements were stated only in terms of the total sample and only in terms of a "typical" ρ. Considerable variation in $\tilde{DEFF}_C$ was observed, primarily because of the variation in proposed cluster sample size.

3.5.3 Combined design effect

The ESS guidelines suggested that the total design effect should be predicted as:

$$\tilde{DEFF} = \tilde{DEFF}_P \times \tilde{DEFF}_C \qquad (3)$$

This ignores any design effect due to stratification of the sampling frame, but as this is generally modest in magnitude and beneficial in direction (i.e. less than one), ignoring this effect was felt to both simplify the calculation and build in a little conservatism to the required sampled size. Expression (3) also assumes no association between the weights and the clusters – see Lynn and Gabler (2005). Predictions of total design effect vary greatly between nations.

3.6 Summary of sampling procedure

When taking into account all the effects the gross sample size can be anticipated as Table 3 illustrates.

Table 3. Illustrative example of all factors related to anticipate an ideal gross sample size.

Operation	**Size calculation**
1. Target effective sample size - *neff* (size that can be received with *srs* without missingness).	**1500**
2. Anticipated missingness due to nonresponse (on average , may vary by strata, e.g.)	30% eli 1500/.7 = 2143
3. Anticipated missingness due to overcoverage (on average)	5% eli 2143/.95 = 2256
4. Anticipated cluster effect so that the final cluster size has been anticipated too * and intra-cluster correlation based on earlier experience on similar surveys	$DEFF_c$ = 1+(5.3- 1)*.025 = 1.108 2256*1.108= 2499
5. Anticipated design effect due to unequal inclusion probabilities used in the design*	$DEFF_p$ = 1.25 2499*1.25 = 3125
6. Anticipated risk in fieldwork and then we have the gross sample size (here net sample size = 3150*.7*.95 = 2095)	**3150**

* *should be consistent with figures in points 2 and 3*

3.7. Documentation

Comprehensive and clear documentation of all relevant methodological aspects of the survey was demanded. At the level of sampling units, this meant that indicators of sampling stratum, primary sampling unit and the selection probability at each stage of sampling should be included on a micro-level data file that carried the same identifiers as the questionnaire and other data files. A detailed file specification was provided. Supply of these data would allow the application of weights and the use of appropriate methods for the analysis of data from a complex survey.

A problem of the current procedure is that nonresponse has not been well taken into account. So, any adjusted weights due to this reason are not available on the web. It would first require to include more auxiliary data for the sampling file, both of respondents and non-respondents. This is very realistic for many countries but not for all. It is obviously the main reason that the central co-ordinating team (CCT) of the ESS has not required these operations even although I have proposed it.

4. Evaluation of the ESS Procedures

4.1 Predictions of DEFF

As already mentioned the ESS sampling system has not yet taken completely into account nonresponse. However, this present system as illustrated in Table 3 uses numbers of respondents when calculating the basic weights that are called design weights (variable DWEIGHT in the freely available web data file that are scaled so that the average for each country is equal to one).

The basic weights vary in all other countries except in those which have applied simple random sampling (three countries in round 1 and seven in round 2). This thus assumes that nonresponse is non-informative that does not hold, of course. On the other hand, the total DEFF is equal to one in these countries. In the case of the other countries, we have tried to analyse the DEFF's in order to improve the sampling procedure for the subsequent rounds. Since the DEFF's are variable-dependent we created nine variables from the round 1 file so that the different characteristics of the questionnaire were taken into account. Most variables were constructed from several initial variables, being thus like indicators. Table 4 gives some results on the DEFF's.

Table 4: Estimation of design effects for countries participating in both rounds

Country	Median ρ	max b^*	*DEFFc*	*DEFFp*	*DEFF*
AT	0.11	6.49	1.61	1.24	2.01
BE	0.04	6.56	1.22	1	1.22
CH	0.03	8.83	1.27	1.21	1.54
CZ	0.15	2.94	1.28	1.25	1.61
DE	0.06	18.85	2.03	1.11	2.26
ES	0.15	4.96	1.60	1.22	1.95
FR	0.05	7.42	1.34	1.23	1.65
UK	0.03	12.06	1.40	1.22	1.69
HU	0.05	8.68	1.36	1	1.36
NL	-	-	1	1.19	1.19
NO	0.01	30.03	1.41	1.43	2.03
PL	0.05	10.07	1.32	1.02	1.35
PT	0.14	5.07	1.57	1.83	2.88
SI	0.03	10.76	1.33	1	1.33

We see for example that the median intra-class correlation was in most countries higher than the initial minimum recommendation = 0.02 (this was given based on the UK experience and used due to the fact that many countries has no idea how high this correlation could be). Table 4 also shows that the average cluster size varies quite much. It was in most countries set to be initially quite constant but because the response rates may be varied substantially between clusters (psu's) some variability was present.

In most cases, the achieved response rates were lower than the predictions, but in some they were higher. The greatest proportionate under-prediction (round 1) was in Greece ($\tilde{\bar{b}} = 4.8; \bar{b} = 5.9$), while the greatest over-prediction was in Italy ($\tilde{\bar{b}} = 18.0; \bar{b} = 11.0$), followed by France ($\tilde{\bar{b}} = 12.0; \bar{b} = 8.9$). Where the response rate was less than predicted, this was not necessarily due to a failure to meet the ESS minimum requirements regarding contact efforts or indeed due to lack of efforts generally (Philippens and Billiet 2006).

Differences between the predicted and estimated values of *DEFF* are non-existent in some cases, but considerable in others. There was some uncertainty regarding the parameters of clustering, design weights, or both. In five nations in round 1, the uncertainty only concerned $\bar{b}$. These were three nations (BE, HU, SI) with an equal-probability sample selected from population registers and two (DE, PL) where the weights were completely determined by the sample design. The prediction turned out accurate in Belgium. In Slovenia, $\bar{b}$ was under-estimated as both the eligibility rate and response rate turned out higher than predicted. These two rates were also both under-estimated in Hungary, but this was more than compensated for by an increase in the number of psu's (and associated reduction in the selected cluster sample size), subsequent to the prediction made on the sign-off form.

Due to too optimistic anticipated DEFFs and response rates many countries failed to achieve the minimum requirement for the effective sample size in both rounds. This thus means that the confidence intervals are not to be very close to each other, and a user has to be careful with cross-country comparisons. Some countries made an improvement in round 2 either increasing the number of psu's, or increasing gross sample size or fighting better against nonresponse. It is unclear whether this tendency will continue in round 3 since more and more countries have met budgetary problems. Recently, it was also discussed whether it is necessary to lower the ESS sampling requirements, in the terms of effective sample size.

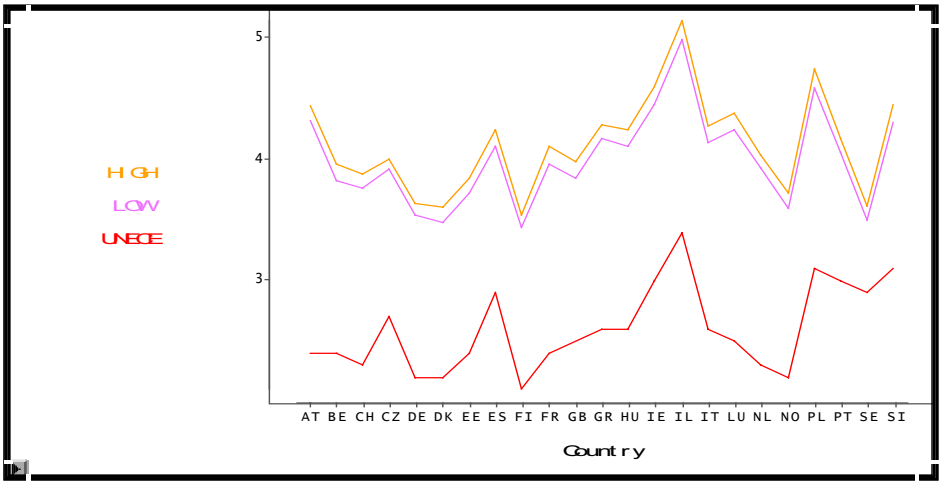
4.2 Need for nonresponse adjustments

Nonresponse has not taken into account enough well although its effect has been analysed in various reports. A problem is that there are no much individual-level data collected on nonrespondents in any country. It is possible to evaluate this problem indirectly, using outside-aggregate data, for example. I

performed one exercise which exploits the data from the United Nations web on one hand and the ESS micro data, on the other; see Figure 1.

Although it is not guaranteed that the UNECE data are complete, it is however rather correct. We see clearly that small households are much less represented in the ESS samples. It is not easy to see, why? Sampling procedure may be one reason but I think that the main reason is that single households have not contacted well and hence they have responded worsely than members of larger households. This is usual in most surveys, why not in the ESS. Such a bias could be reduced in field work to some extent but more using nonresponse adjustments by collecting more auxiliary data of non-respondents, and then constructing the adjusted weights (e.g. using the methodology presented in Laaksonen & Chambers 2006).

Figure 1. Average household size by country based on the UNECE data from early 2000 and from the ESS micro data from 2004-2005. HIGH = higher 95% confidence interval, LOW = lower 95% confidence interval.



5. Analysis of ESS data – and happiness example

The ESS micro files are easy to download and then to analyse. The web also includes rather good meta data and some para data derived from interviewing. Nevertheless, since the variables are coded so that all collected information is included in the files, a user has to be careful especially with such codes that give information about missingness which can be of different kinds. Hence a user always needs to make some refinements because starting real analysis.

The files also include the two types of weights, the ones based on the sampling design of each country and the others indicating the size of the target population of each country. A user thus has to select the first weights always and sometimes both weights. Nonresponse adjusted

weights are thus missing. When using the available weights in the analysis, the point estimates will be correct but it is not reasonable in the case of interval estimates. Some idea of the impact on interval estimates can be got from Table 4 but naturally it is best if a user can estimate the corresponding estimates him/herself since the DEFF's depend on each particular case. Currently, a big problem is that the information of psu's is not available on the web files. Obviously, this will be the case later. If you are using these files, ask for the data of psu's.

To illustrate a bit which types of analysis from the ESS files can be provided, I present some results. The variable of interest is called HAPPY and measured as follows:

Taking all things together, how happy would you say you are? Please use this card.

Extremely unhappy										Extremely happy	(Don't know)
00	01	02	03	04	05	06	07	08	09	10	88

I exclude the individuals with values '88' from the analysis and also those with other missing values (coded afterwards as '99'). The results in Figure 2 are direct averages for each country, however so that the reference country, Ukraine, has been set equal to 0. The people in the other countries are happier than Ukrainians, on average.

The standard errors are not in Figure 2 but these are rather even over countries without taking into account cluster effects. The standard errors are around 5.4-6.4 percent in most countries except Iceland (8.8 %). The cluster effect will increase this number to some extent. Since I had the German cluster data, I calculated the DEFF that was = 1.76.

I also looked the changes in happiness from round 1 to round 2. In most countries the change was not significant but it is interesting that in the following countries the average happiness was increased significantly: Hungary, Poland and Slovenia. All these joined the EU after round 1. Is it the explanation? On the other hand, some decrease was happened in Portugal, in particular, and also in Spain. I cannot explain these results although we have found that the quality of the Portuguese data seems to be weakened from round 1 to round 2.

Much further analysis can be performed. I briefly present some results in order to explain happiness/unhappiness. These are based on ordinary multivariate regression analysis using country weights. A number of explanatory variables were included in the model. I do here present only some results:

- Women are happier than men, and young people happier than older. On the other hand, the happiness of women decreases quite linearly by age but men are least happy at middle ages.

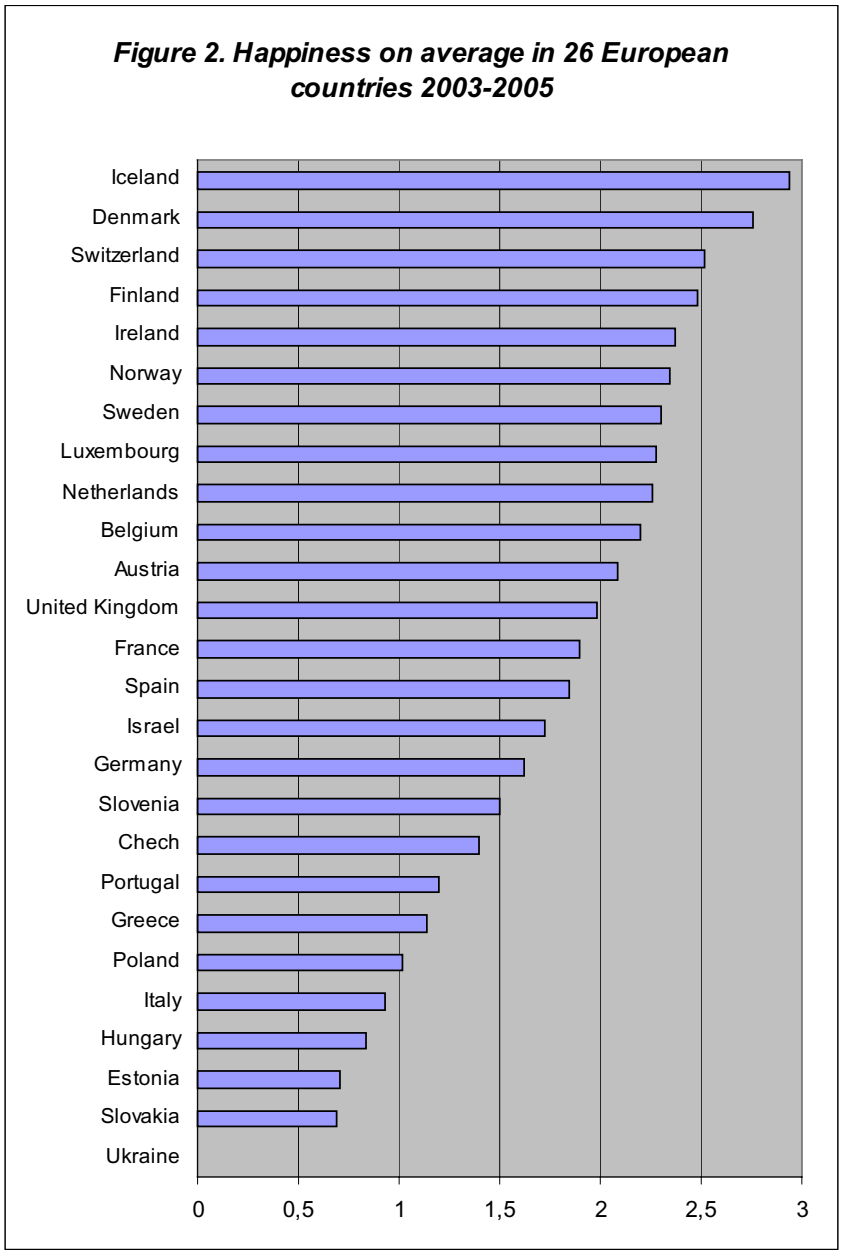


Figure 2. Happiness on average in 26 European countries 2003-2005

- Married are happier than never married who are happier than divorced and separated.
- Big household increases happiness to some extent.
- Trust on police and legal issues in the country are good for people's happiness. The same is concerned trust on administration including health organisations.
- People who feel to be discriminated by gender, race etc. are less happy.
- Very poor people seem to be least happy but there is not much difference if the income level is over some minimum.
- Bad health naturally decreases happiness.
- Active people are slightly happier than inactive.

- Foreigners are slightly less happy than native people.

The model includes also the country as one explanatory variable. Naturally, the differences between countries were reduced essentially after this modeling. The happiness order also changed to some extent. However, Iceland was still in the top, and Ireland the second before Denmark, but now Italy was the last, Slovakia the second last. Can a reader explain these?

6. Conclusion

The aims of the ESS, in terms of sample design standards and procedures for implementation of those standards, were ambitious. Though not realised in every detail, the ESS can be considered a great success. This is evidenced also so that the ESS (with subtitle "Innovations in comparative measurement") was one of the five winners of the 2005 EU Descartes Laureates (see http://www.sardinien.com/astronomie/pdf/pr02122005_annex_winners_dp_research2005_en.pdf). In particular, the process for developing and finalising sample designs can be considered successful both at a subjective level and in objective terms (guidelines used to estimate design parameters proved useful and estimates generally accurate; documentation is relatively complete).

In my opinion, the quality of the ESS is one of the best ones in the world if the demanding multinational surveys are concerned. This does not mean that the quality cannot be essentially improved. The evaluation of the estimation of design parameters presented here has provided several pointers to how such estimation might be improved on future cross-national surveys. The nature of uncertainty in the estimates has been described and the directions of errors documented.

In general, the ESS has provided advances in survey practice in a number of nations. Additionally, the procedures for sample design represent a useful advance in the methodology of cross-national surveys.

Oversampling has been used in some countries and also so that the anticipated differences in nonresponse/overcoverage between regions have been taken into account. But this could be exploited much more, also in *srs*-countries where it is well-known that response rates vary much by region and other domains. So, pre-stratification would be my recommendation for these countries too, and consequently leading to varying weights if the anticipation is not complete. Furthermore, I recommend to insert the new adjusted sampling weights into the ESS archive data files in addition to the current design weights (DWEIGHT).

In the first stage these weights should be required for the *srs* countries that have always the weights equal to one in the current integrated archive file. This is not even difficult since these countries have already created such weights, based on post-stratification or other calibration. Later, we should require all countries to add information for nonresponse analysis and adjustments. For example, all countries are able to add to a sampling file some variables of nonrespondents.

References

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

Gabler, S., Häder, S., Lynn, P. (2005). Design effects for multiple design samples, ISER Working Paper No. 2005–12, Colchester: University of Essex. http://www.iser.essex.ac.uk/pubs/workpaps/pdf/2005-12.pdf

Häder, S. and Gabler, S. (2003). Sampling and estimation. In *Cross Cultural Survey Methods*, eds. J. Harkness, F. van de Vijver, P. Mohler, New York: John Wiley and Sons.

Kish, L. (1992). Weighting for unequal P_i. *Journal of Official Statistics*, 8:2, 183-200.

Kish, L. (1994). Multipopulation survey designs: five types with seven shared aspects. International Statistical Review, 62, 167-186.

Kish, L. (1995). Methods for design effects. *Journal of Official Statistics*, 11:1, 55-77.

Laaksonen, S. & Chambers, R. (2006). Survey estimation under informative non-response with follow-up. *Journal of Official Statistics, 81-95*.

Le, T. and Verma, V. (1997). An analysis of sampling designs and sampling errors of the Demographic and Health Surveys, DHS Analytic Report No.3, Calverton, Maryland: Macro International Ltd. http://www.measuredhs.com/pubs/details.cfm?ID=4

Lynn, P. (2003). Developing quality standards for cross-national survey research: five approaches. *International Journal of Social Research Methodology*, 6:4, 323-336.

Lynn, P. (2004). The use of substitution in surveys. *The Survey Statistician*, 49, 14-16.

Lynn, P. and Gabler, S. (2005). Approximations to b^* in the prediction of design effects due to clustering. *Survey Methodology*, 31:1, 101-104.

Lynn, P., Gabler, S. & Häder, S. & Laaksonen, S. (2006). Methods for achieving equivalence of samples in gross-national surveys. *Journal of Official Statistics*. (in print)

Philippens, M. and Billiet, J. (2006). Nonresponse in cross-national surveys. Results of the European Social Survey. *ESS Working Paper.*

Vehovar, V. (2003). Field substitutions redefined. *The Survey Statistician*, 48, 35-37.