METHODS FOR ACHIEVING EQUIVALENCE OF SAMPLES IN CROSS-NATIONAL SURVEYS: THE EUROPEAN SOCIAL SURVEY EXPERIENCE

Peter Lynn, Sabine Häder, Siegfried Gabler and Seppo Laaksonen

ISER Working Papers
Number 2004-09

Institute for Social and Economic Research

The Institute for Social and Economic Research (ISER) specialises in the production and analysis of longitudinal data. ISER incorporates the following centres:

- ESRC Research Centre on Micro-social Change. Established in 1989 to identify, explain, model and forecast social change in Britain at the individual and household level, the Centre specialises in research using longitudinal data.

- ESRC UK Longitudinal Studies Centre. This national resource centre was established in October 1999 to promote the use of longitudinal data and to develop a strategy for the future of large-scale longitudinal surveys. It is responsible for the British Household Panel Survey (BHPS) and for the ESRC's interest in the National Child Development Study and the 1970 British Cohort Study

- European Centre for Analysis in the Social Sciences. ECASS is an interdisciplinary research centre which hosts major research programmes and helps researchers from the EU gain access to longitudinal data and cross-national data sets from all over Europe.

The British Household Panel Survey is one of the main instruments for measuring social change in Britain. The BHPS comprises a nationally representative sample of around 5,500 households and over 10,000 individuals who are reinterviewed each year. The questionnaire includes a constant core of items accompanied by a variable component in order to provide for the collection of initial conditions data and to allow for the subsequent inclusion of emerging research and policy concerns.

Among the main projects in ISER's research programme are: the labour market and the division of domestic responsibilities; changes in families and households; modelling households' labour force behaviour; wealth, well-being and socio-economic structure; resource distribution in the household; and modelling techniques and survey methodology.

BHPS data provide the academic community, policymakers and private sector with a unique national resource and allow for comparative research with similar studies in Europe, the United States and Canada.

BHPS data are available from the Data Archive at the University of Essex
http://www.data-archive.ac.uk

Further information about the BHPS and other longitudinal surveys can be obtained by telephoning +44 (0) 1206 873543.

The support of both the Economic and Social Research Council (ESRC) and the University of Essex is gratefully acknowledged. The work reported in this paper is part of the scientific programme of the Institute for Social and Economic Research.

Acknowledgement: The authors of this paper constitute the expert panel on sampling of the European Social Survey (ESS). The paper arises from the work of the panel, which is funded by the European Union and the European Science Foundation. The second and third authors are members of the Statistics Department at the Zentrum für Umfragen, Methoden und Analysen (ZUMA), Mannheim, Germany; the fourth author is based at the University of Helsinki, Finland. The success of the European Social Survey is a result of the collaboration of many people in many organisations in many countries. We are grateful to all of them for their energy and commitment to the survey and particularly to the National Co-ordinators in each country, who have without exception being patient, understanding and polite in the face of our considerable demands. This Paper has been prepared for presentation in the session, “Sampling Methods,” at the RC33 Sixth International Conference on Social Science Methodology, Amsterdam, 16-20 August 2004

Readers wishing to cite this document are asked to use the following form of words:

> **Lynn, Peter, Häder, Sabine, Gabler, Siegfried and Laaksonen, Seppo (July 2004) ‘Methods for Achieving Equivalence of Samples in Cross-National Surveys: The European Social Survey Experience’, *Working Papers of the Institute for Social and Economic Research*, paper 2004-09. Colchester: University of Essex.**

For an on-line version of this working paper and others in the series, please visit the Institute’s website at: http://www.iser.essex.ac.uk/pubs/workpaps/

Institute for Social and Economic Research
University of Essex
Wivenhoe Park
Colchester
Essex
CO4 3SQ UK
Telephone: +44 (0) 1206 872957
Fax: +44 (0) 1206 873151
E-mail: iser@essex.ac.uk
Website: http://www.iser.essex.ac.uk

ABSTRACT

Most surveys carried out at national or sub-national level involve a single sample design and sampling frame. Where multiple frames are used, they are typically to access different sub-populations. For cross-national surveys, however, it is usually necessary to use a different design in each nation to sample from an analogous national population. Cross-national sampling frames are rare.

In this paper, we describe procedures used to obtain equivalence of sample designs in 22 nations in round 1 of the European Social Survey. We evaluate the implementation of the procedures and we summarise lessons for the design of cross-national surveys. We focus particularly on novel aspects of the procedures. These include specification of national sample sizes in terms of "effective sample size" and provision of guidelines on how to predict design effect components and how to use the predictions to determine the necessary sample size. We also discuss procedures for interaction between the various parties involved: the ESS central co-ordinating team, the ESS sampling panel, national co-ordinators and field work organisations.

Key words: Design effects, effective sample size, intra-cluster correlation, sample design, sampling frames

1. Introduction

Compared with surveys carried out within a single nation, cross-national surveys involve an extra layer of complexity in terms of both organisation and design. Special procedures are required in order to derive and implement appropriate standards for design and implementation (Lynn, 2003). This is particularly important in the case of sample design, as available sampling frames and other constraints tend to vary between nations. Additionally, field work is often organised at the national level (at least in the case of surveys in developed countries), resulting in the involvement of many persons and organisations in the survey process. It is important to find ways to ensure the comparability of the sample designs used in each nation, as substantive comparisons between nations, or between groups of nations, are often a key objective of the survey.

In this paper, we first discuss the objectives of sample design for cross-national surveys (section 2). We then 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 estimate design effects and to use these estimates in determining national sample sizes. Given the large number of persons and organisations typically involved in cross-national surveys, effective procedures for implementing the agreed principles and requirements are of paramount importance. The procedures used on the ESS are described in section 4, and some of the strengths and weaknesses are pointed out. In section 5, we evaluate the requirements regarding design effects due to both variable selection probabilities and clustering and make some suggestions for how such requirements might be better stated in future. Section 6 concludes.

2. Sample Design for Cross-National Surveys

To enable comparisons between nations, we suggest 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, 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, as this would represent an effective allocation of resources if a prime objective was to make between-nation comparisons.

To best meet these criteria, it is likely that details of the sample design will vary between nations. The goal is functional equivalence, not replication of parameters of the sample design. As Kish (1994, p. 173) wrote, "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 of the background and objectives of the survey can be found at www.europeansocialsurvey.org. The ESS is funded via the European Commission's 5th and 6th 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 as soon as possible: no-one involved in the survey has advance access and there are no restrictions on access. Data can be down-loaded from http://ess.nsd.uib.no.

The ESS consists of regular "rounds" of data collection, with each round involving an independent cross-sectional sample in each nation. The intention is that survey rounds will take place at 2-year intervals. The first round of field work took place in September-December 2002[1]. There is a core questionnaire (approximately 30 minutes) that will be administered in every round, along with modules of questions (in round 1, two modules of approximately 15 minutes each) that will change from round to round[2]. Nations are not asked to commit themselves to more than one round at a time, though of course continued participation is actively encouraged. Twenty two nations participated in round 1, namely the fifteen member states of the EU plus four accession states (Czech Republic, Hungary, Poland, Slovenia) and three non-EU members (Israel, Norway, Switzerland). All interviews are carried out face-to-face.

The ESS methodological functions have been organised into ten work-packages, overseen by a central co-ordinating team (CCT). One of the work-packages deals with sampling. This work-package is being carried out by a panel of four sampling experts which, at least for the first two rounds of the survey, consists of the four authors of this paper. The panel developed the requirements for participating nations – which will be described in the remainder of this section under five broad

[1] In a few nations fieldwork was not completed until 2003.

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[3] or legal status. It is worth noting in passing that this definition was subject to considerable discussion by a 21-country steering group (European Science Foundation, 1997) prior to agreement. Concerns raised related almost exclusively to cases in which implementation of the definition would represent a departure from normal practice in a particular country or in which practical difficulties were caused by the nature of the available sampling frames – not to any conceptual or substantive concerns.

Thus, 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:

---

[2] In round 1, mean interview length ranged from 51.8 minutes in both Spain and Italy to 69.7 minutes in Sweden.

[3] 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.

a) nations with reliable lists of *residents* that are available for social research such as the Danish Central Person Register that has approximately 99.9% 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 from "PTT-afgiftenpuntenbestand" in the Netherlands and from the "Postcode Address File" in the UK;

d) nations without reliable and/or available lists such as Portugal and Greece.

Drawing a sample is most complicated if no lists are available (group d). In this instance area sample designs (Särndal *et al*., 1992) were usually applied, in which the selection 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 (Häder and Gabler, 2003), about which some survey organisations were enthusiastic. 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 particularly 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. A new method was developed, resulting in known non-zero selection probabilities for all target population members living at addresses where at least one registered elector resided (Lynn and Pisati, 2004). Thus, under-coverage, while not zero, was restricted to persons at addresses with no registered electors. In Ireland, similar issues arose in the use of the electoral registers as a frame. 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. Additionally, all departures from the ideal were documented carefully and are available on the ESS website to all potential data users.

3.3 Response rates

Non-response is the next problem for achievement of the objective to represent the target population. A carefully drawn gross sample from a perfect frame can be worthless 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. This would be particularly challenging for some countries where response rates between 40 and 55 percent are common (Lyberg 2000). 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 (addresses, households or persons, depending on the sampling frame), 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. Definitions of outcome categories and response rate were also supplied.

While these targets and constraints were felt necessary, they were not expected to be sufficient to ensure high response rates, let alone similar (and small) non-response bias in each nation, which was the ultimate goal. The sampling panel and the CCT offered extensive advice on techniques for increasing response rates such as advance letters, toll-free telephone numbers for sample members, training of interviewers in response-maximisation techniques and doorstep interactions, calling patterns, incentives, etc.

3.4 Sample selection methods

We have already argued that strict probability sampling is a necessary pre-requisite for cross-national comparability. Without it, the second of the two fundamental criteria introduced in section 2 cannot be met. 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 tends to reduce the extent of interviewer efforts to gain a response at the original addresses/households (Elliot, 1993).

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). In these exceptional cases, provided the circumstance applied to only a very small proportion of the sampled PSUs, replacement with areas randomly selected from the same strata was allowed. (This happened in only one nation.)

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[4]. Explanation was provided as to what was meant by effective sample size and how it should be estimated. This involved predicting, under certain simplifying assumptions, the design effect due to variable 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 gross sample size required in order to produce the target number of completed interviews.

For most of the NCs and survey organisations, the concepts of effective sample size and design effect were completely new. The sampling panel invested considerable time and effort in explaining the concepts and helping NCs to estimate design effects. This often involved the NCs seeking out statistical information with which they had no familiarity, such as the national distribution of the number of persons aged 15 or over in a household. In return, we encountered almost universal enthusiasm to understand the concepts and to meet the requirements. Feedback obtained by the CCT from the NCs after round 1 suggested that the sample design process had been perceived as one of the most productive and useful aspects of taking part in the ESS and that several nations felt that both knowledge and methods had been improved in their country. Because this process was, to our

[4] 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. In practice, this applied to only 2 of the 22 nations participating in round 1: Slovenia and Luxembourg.

knowledge, novel and appeared to have a considerable impact, we will describe the requirements and how they were implemented in some detail.

Design effect due to variable selection probabilities (*DEFFp*)

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

$$D\hat{E}FF_P = \frac{m\sum_{i=1}^{I} m_i\left(w_i^2\right)}{\left(\sum_{i=1}^{I} m_i w_i\right)^2} \qquad (1)$$

where m_i and w_i denote respectively the number of interviews and the design weight associated with the i^{th} weighting class. This is in fact a simplification (Gabler *et al*., 1999) that assumes $Var\left(y_{ij}\right) = \sigma^2$, where $Var\left(y_{ij}\right)$ is variance of the target variable *y* over respondents *j* within weighting class *i*. This may be a reasonable assumption in many cases, but we would note that the extent of departure from this assumption will depend partly on the source of variation in design weights, which in turn will differ between nations.

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 (number of persons aged 15 or over). 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 used such an address-based design (e.g. Czech Republic, Greece, Ireland, Israel, Netherlands, Portugal, Spain, Switzerland, UK). There was considerable variation between these nations in $D\hat{E}FF_P$ due to differences in the household size distribution.

Another reason for variable selection probabilities is that minority groups are over-sampled for substantive reasons. Examples of this are Germany, where the East German population is over-sampled and Israel, where the Arab population is over-sampled. A third reason is that certain strata (typically, the largest cities) may be over-sampled in anticipation of lower response rates.

A fourth source of variation in selection probabilities occurs in countries where the PSUs are selected with probability proportional to a proxy size measure which does not correlate perfectly with the units sampled at the subsequent stage. Examples include Israel and Ireland, where the PSU size measures were numbers of persons and the selected units were households and addresses respectively.

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:

$$D\hat{E}FF_C = 1 + \left(\bar{b} - 1\right)\rho \qquad (2)$$

where $\bar{b}$ is the mean number of interviews per cluster and ρ is the intra-cluster correlation. From this, it is clear that the cluster sample size should be chosen as small as possible since the larger the average cluster size, the more interviews have to be conducted to reach the minimum effective sample size of 1,500. Participating nations were encouraged to seek estimates of ρ from other surveys in their country if possible, or alternatively to assume $\rho = 0.02$ (a value that was chosen as an approximate mean of estimates for attitude measures from surveys in a small number of countries).

Considerable variation in $D\hat{E}FF_C$ was observed, primarily because of the variation in proposed cluster sample size.

Combined design effect

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

$$D\hat{E}FF = D\hat{E}FF_P \times D\hat{E}FF_C \qquad (3)$$

This of course 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. These estimates of total design effect varied greatly between nations. At one extreme, there were 3 nations for whom both component design effects, and hence the overall design effect, was 1.00 (Denmark, Finland, Sweden). There were 3 nations with clustering but no variation in selection probabilities (Belgium, Hungary, Slovenia), 2 with variation in probabilities but no clustering (Luxembourg, Netherlands), and for the other 14 nations, both component design effects were predicted to be greater than one, in most cases substantially greater. The largest predicted design effects were 1.59 for Israel, 1.58 for France, 1.55 for UK and 1.52 for Germany. Variation in predicted design effects can be seen to be primarily influenced by two factors: the nature of available sampling frames (e.g. population registers that permit equal-probability sampling *versus* address-based methods) and the interaction between geographical spread of the population, field costs and available budget (which largely determine the extent of clustering, if any).

3.6 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 design weights and the use of appropriate methods for the analysis of data from a complex survey.

Meta-data to be supplied and made freely available included a detailed description of the sample design and sample selection process and a clear account of any ways in which the design fell short of the ideal, for example where the frame may have suffered some under-coverage.

4. Developing and Implementing Sample Designs

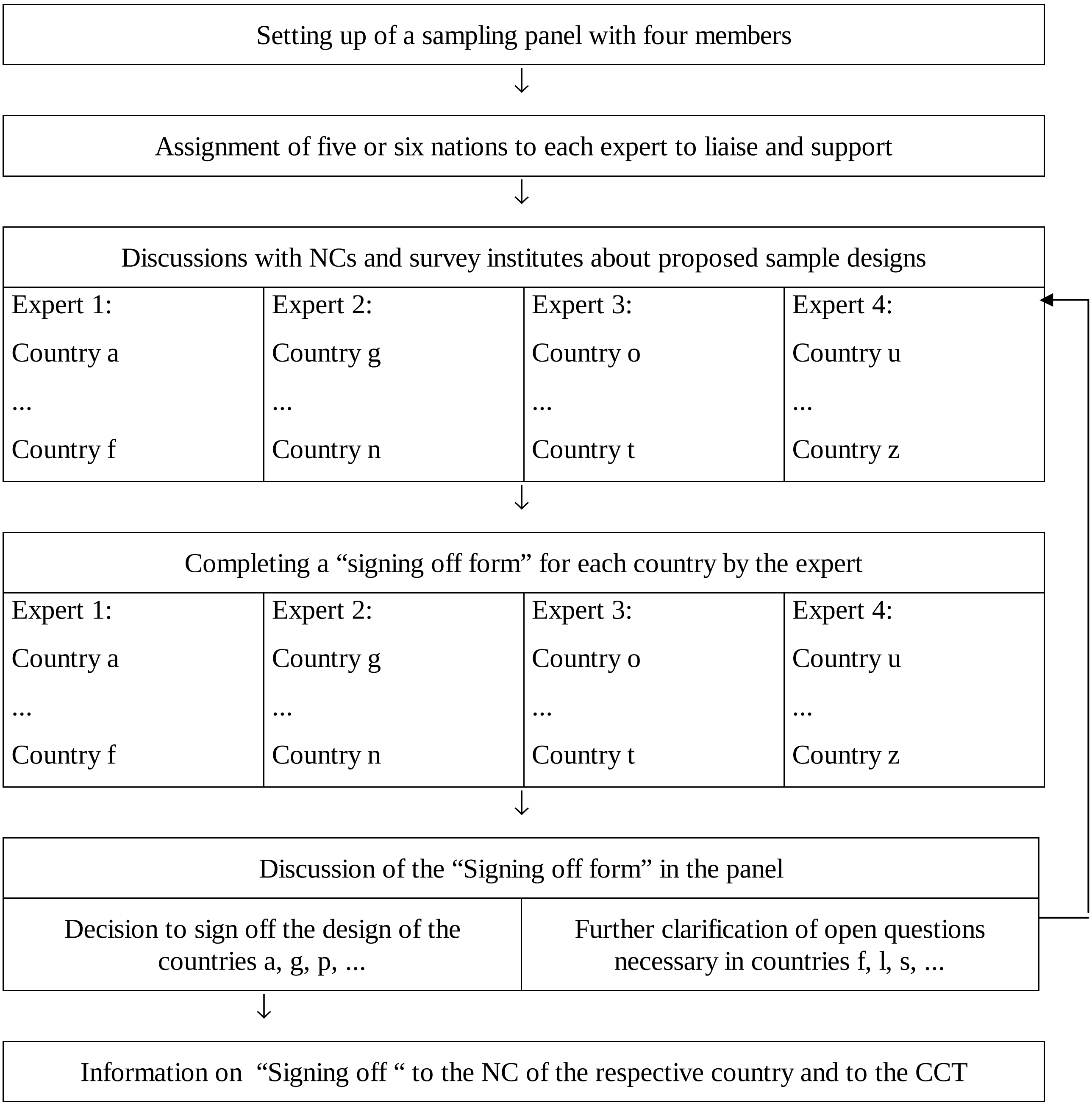
Providing clear written requirements for participating nations is necessary but does not guarantee that they will be met. In this section we describe the ESS process for developing acceptable sample designs in each nation. The process has two features which we think are particularly important. The first is that it is co-operative rather than authoritarian. The sampling panel saw its prime role as the provision of advice and assistance where needed. The role of monitoring the design and implementation was kept as implicit as possible. The second important feature is that interaction was intensive. Regular contact was made between relevant parties and this promoted the spirit of co-operation and enabled the building of rapport. To our knowledge, this process is unique amongst cross-national social surveys.

As already mentioned, the functions relating to sample design (development, agreement, documentation and evaluation) are the responsibility of a small panel. A deliberate decision was made to keep the panel small in order to minimise formality and engender the spirit of a team of co-researchers rather than a committee. At the first meeting of the panel, in December 2001, an allocation of round 1 participating nations to panel members was agreed, so that each member would liaise with and support 5 or 6 nations. Panel members contacted “their” National Co-ordinators (NCs) asking for information about the foreseen sampling design. Then, a process of co-operation began. In most cases, the sampling panel member worked closely with the NC; in some cases, the NC preferred instead that the panel member should work directly with the survey organisation while keeping the NC informed. Most of the communication with NCs and survey organisations was done by email, but special visits by panel members to participating nations were allowed where this was thought likely to be particularly beneficial (5 such visits were made). Additional face-to-face contact took advantage of other occasions such as conferences and the regular meetings of all NCs with the Central Co-ordinating Team (CCT).

In many countries completely new designs had to be developed to meet the ESS requirements. In other countries, it was only a matter of clarifying details. In particular, support in calculating the effective sample sizes was often necessary. The amount of time and effort committed by the sampling panel members therefore varied greatly over the nations.

Once all questions were clarified and the panel member was satisfied with the proposed design of a country, it was deemed ready for "signing off". The panel had developed a sample form, where details of the design of each country had to be completed in a systematic way. This task was done by the panel member for "his/her" countries, thus ensuring the use of standardised terms and ensuring that the design was clearly defined (otherwise the panel member was not able to describe it in the form). Then the panel member presented the form to the other panellists. If all of them agreed, the design was "signed off". Thus, the decision of signing off a design was always made by the whole team together. Otherwise, the discussion with the National Co-ordinator had to carry on. This happened in several cases where the panel did not initially agree with the design or requested additional information. Once signed off, the sample form was circulated to the CCT and eventually incorporated into the sampling panel's report on round 1. During the period of the development of designs for round 1 (December 2001 to October 2002), the panel met on three occasions – in London, Mannheim and Helsinki - to discuss general principles (for example whether random route methods could ever be acceptable and if so, in what circumstances), processes, and specific designs. Figure 1 summarises the process of “signing off” the sample designs.

Figure 1. The Process of “Signing Off” ESS Sample Designs



5. Evaluation of the ESS Procedures

5.1 Process

The process described in section 4 above was felt by all parties to have been a great success. Though the final designs do not strictly meet the stated requirements in every respect in all cases, no unanticipated major departures have yet been identified. Additionally, all known departures are documented and this information is in the public domain on the ESS website. In some countries, the implemented design represented a significant breakthrough in survey standards (see section 5.8 below). This would not have been possible without such an intensive process of co-operation.

5.2 Predictions of $DEFF_P$

The requirement to predict $DEFF_P$ was novel for all but a few of the participating nations. In several cases, this made the prediction subject to quite heroic assumptions because the necessary information was simply unavailable. Uncertainty about the distribution of selection probabilities was of three sorts:

a) Uncertainty about the distribution of household size (number of persons aged 15 or over);

b) Uncertainty about the relationship between a proxy size measure used at the PSU level and the actual size measure of relevance;

c) Uncertainty about the relationship between a proxy size measure used at the household/address level and the actual number of persons aged 15 or over.

There is of course a fourth source of variation in selection probabilities, namely planned over-sampling of domains. But in no cases was there uncertainty about the ratios of selection probabilities involved. Over-sampling of this kind was performed in four countries to combat expected response rate differences (Netherlands, Poland, Portugal, Spain) and in three countries for substantive reasons (Israel, UK, Germany). In all seven cases, the domains were defined by geographical areas.

In Table 1 we present the pre-fieldwork predictions of $DEFF_P$, as documented on the sample form ($D\hat{E}FF_P$) and the post-fieldwork estimate of $DEFF_P$ ($D\tilde{E}FF_P$) calculated using expression (1), for each nation subject to at least one of the three sources of uncertainty described above. There are 13 such nations; 6 nations used equal-probability designs and the remaining 3 (Germany, Norway and Poland) had complete control over the variation in selection probabilities (all used population registers as the frame), except in so far as non-response may have led to the distribution amongst the responding sample differing from the distribution amongst the selected sample – a source of variation that we shall ignore here. The first three columns of the table indicate which sources of uncertainty applied in which nation. By using expression (1) for both $D\hat{E}FF_P$ and $D\tilde{E}FF_P$, any difference in the calculated values is due solely to differences between the predicted and observed distribution of design weights.

We would note that type b) uncertainty is unlikely to have large impacts, as the proxy measures used in these 3 cases are likely to be highly correlated at the PSU level. For example, in Italy municipalities were selected with probability proportional to the number of residents aged 18 or over according to the municipal population registers ("elenco civile"), whereas for that component of the design to be self-weighting, the size measure should have been the number of addresses at which at least one person is a registered elector.

Table 1. Predicted and Estimated Values of $DEFF_P$

		Source of uncertainty				
		a)	b)	c)	$D\hat{E}FF_P$	$D\tilde{E}FF_P$
Austria	AT	X			1.25	1.25
Switzerland	CH	X			1.25	1.21
Czech Republic	CZ	X			1.16	1.25
Spain	ES	X			1.16	1.22
France	FR	X			1.30	
Greece	GR	X	X		1.18	1.22
Ireland	IE	X				1.04
Israel	IL	X	X		1.30	1.56
Italy	IT		X	X	1.01	1.16
Luxembourg	LU			X	1.40	1.26
Netherlands	NL	X			1.19	1.19
Portugal	PT	X				
United Kingdom	UK	X			1.23	1.22

Note: Missing values in the final two columns of the table are due to final data not having been provided and/or checked at the time of writing. These values will appear in the final version of this paper.

In most cases, then, the uncertainty concerned the distribution of number of persons aged 15 or over at a household/address. There were three nations where this was predicted from recent previous surveys that had used a similar selection method, though in all three cases a different age cut-off had been used (either 16 or 18): Switzerland, Netherlands and UK. In these three cases, differences between $D\hat{E}FF_P$ and $D\tilde{E}FF_P$ are small[5]. In the other 8 nations that used address-based sampling, there was considerable variability in the accuracy of $D\hat{E}FF_P$ as a prediction of $D\tilde{E}FF_P$. The predictions were too low, with the exceptions of Austria and Ireland.

5.3 Predictions of $\bar{b}$

The mean number of sample units per PSU in the selected sample was of course determined by the design. However, prediction of $\bar{b}$, the mean number of respondents per PSU, relied upon prediction of the response rate. In practice, $\bar{b}$ differed from the predicted value due to response rates that differed from expectations in several countries. In most cases, the achieved response rates were lower than the predictions, but in some they were higher. The greatest proportionate under-prediction was in Greece ($\hat{\bar{b}} = 4.8; \bar{b} = 5.9$), while the greatest

[5] The slightly larger difference for Switzerland may possibly be due to variability introduced by a low response rate: Switzerland had the lowest response rate of all nations in round 1 of the ESS: just 33%, compared to a mean response rate of 63% in the other 21 nations.

over-prediction was in Italy ($\hat{\bar{b}} = 18.0; \bar{b} = 11.0$), followed by Spain ($\hat{\bar{b}} = 6.0; \bar{b} = 5.0$), Czech Republic ($\hat{\bar{b}} = 3.2; \bar{b} = 2.7$) and United Kingdom ($\hat{\bar{b}} = 14.0; \bar{b} = 12.6$).

5.4 Predictions of combined DEFF

The pre-fieldwork predictions of the combined design effect (3) are compared with the predictions that would have been made had the realised values of $\{w_i\}$ and $\bar{b}$ been used in (1) and (2) respectively (Table 2). This is informative of the effect on the sample size calculation of inaccurate prediction of $\{w_i\}$ and $\bar{b}$. It is not informative of the relationship between the predicted values and values that may be estimated for specific survey estimates, as this depends also on the accuracy of the prediction of ρ and on sensitivity to the assumptions inherent in (1) and (2). The latter issue is dealt with elsewhere (Lynn and Gabler, 2004).

It can be seen that differences between the predicted and estimated values of DEFF are non-existent in some cases, but considerable in others. Three nations are trivial cases as the sample designs involved neither clustering nor variable selection probabilities (DK, FI, SE). In all other cases there was some uncertainty regarding the parameters of clustering, design weights, or both. In three nations (BE, HU, SI), the uncertainty only concerned $\bar{b}$, as an equal-probability sample was selected from the electoral registers. 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 PSUs (and associated reduction in the selected cluster sample size), subsequent to the prediction made on the sign-off form. In two nations (DE, PL), though not equal-probability designs, the weights were completely determined by the design so again the only uncertainty was the effect of response rate on $\bar{b}$. Response rate was lower than predicted in Germany, with a consequent over-prediction of DEFF and the opposite was true in Poland.

Table 2. Predicted and Estimated Values of the Combined DEFF

	$\hat{DEFF}$	$\tilde{DEFF}$		$\hat{DEFF}$	$\tilde{DEFF}$
AT	1.38	1.40	IE		1.26
BE	1.10	1.10	IL	1.59	1.92
CH	1.47	1.41	IT	1.53	1.50
CZ	1.28	1.36	LU	1.40	1.26
DE	1.52	1.49	NL	1.19	1.19
DK	1.00	1.00	NO	1.50	1.95
ES	1.39	1.42	PL	1.12	1.14
FI	1.00	1.00	PT		1.98
FR	1.58		SE	1.00	1.00
GR	1.36	1.45	SI	1.40	1.46
HU	1.19	1.14	UK	1.55	1.50

There were two nations (LU, NL) where the only uncertainty concerned $\{w_i\}$ as the sample was not clustered. In the Netherlands, the design weights depended only on the distribution of household size and this was well known from previous surveys. In Luxembourg, there were two sources of variation in the weights: a deliberate over-sampling of certain domains and an unavoidable many-to-many correspondence of frame units to eligible persons within households. The latter could not be well predicted as the relevant information had not been recorded on previous surveys that had used this frame. The prediction of the impact on the design effect turned out to have been pessimistic.

In the remaining nations, there was some uncertainty about both components of DEFF. In two (CZ, ES) $D\hat{E}FF_P < D\tilde{E}FF_P$, but $D\hat{E}FF_C > D\tilde{E}FF_C$ (see section 5.3 above), so the impact of the latter reduced the impact of the former. In Israel too $D\hat{E}FF_P < D\tilde{E}FF_P$, but there was no counter-balancing over-prediction of $DEFF_C$, so overall there was rather a dramatic under-prediction of DEFF with the consequence that the Israeli sample fails to achieve the ESS effective sample size requirement, even though it had been designed to achieve the requirement and the number of completed interviews exceeded the predicted number. In the other nations (AT, CH, GR, IT, UK) the predictions proved reasonably accurate.

Finally, we should note that in three nations (BE, PL, UK) a dual-design was used, involving one stratum where an unclustered sample was selected (the largest cities/municipalities in the case of Belgium and Poland; Northern Ireland in the case of the United Kingdom) and a second stratum where a clustered sample was selected. The ESS had not provided guidelines on how to predict the combined (national) DEFF in such situations and consequently a different method was used in each case. We believe that a weighted combination of two separately predicted design effects should be used in such cases, but the weights should depend on between-strata differences in the overall sampling fraction and in anticipated coverage/response rates (Gabler *et al*., 2004).

5.5 Predictions of ρ

Only a small number of nations used a value other than $\hat{\rho} = 0.02$ to calculate $D\hat{E}FF_C$. In these few cases, the values of $\hat{\rho}$ ranged from 0.03 to 0.05. These values were chosen either because previous "similar" surveys had produced estimates of this general magnitude or because the proposed cluster units were particularly small geographical areas. The ESS round 1 data now provide the opportunity for estimates of ρ to be compared across countries. This will be done in order to inform the design of future rounds but is not treated here.

5.6 Predictions of response and eligibility rates

As mentioned already in sections 5.3 and 5.4, predictions of response and eligibility rates were rather inaccurate in some nations. This affected $\hat{\bar{b}}$ and hence $D\hat{E}FF_C$. In no case was the eligibility rate over-estimated: the tendency was to be over-cautious in cases where the properties of the frame were not well known in relation to the ESS target population. But response rate was over-estimated in several cases. This may have been the result of the target response rate set (section 3.3). Participating nations who were sure that they would not meet the target may have been unwilling to admit the anticipated extent of their under-achievement. This is likely to change on future rounds of the survey, now that response rates for all nations for round 1 have been published and openly discussed in meetings of the NCs. In bilateral discussions, nations that have previously failed to achieve 70% response will be encouraged to aim for modest and realistic improvement.

5.7 Completeness of documentation

As introduced in section 3.6, three types of documentation were required:

1. Micro-level indicators of selection probabilities, PSU membership and stratum membership;

2. A detailed description of the sample design and selection process;

3. A description of any ways in which the design fall shorts of the strict ESS requirements, such as under-coverage of the sampling frame.

The type 1 documentation was supplied accurately and in a form consistent with the specification by most nations. However, in a minority of cases the data supplied initially were found by the sampling panel to be in error. This was either due to a misunderstanding of how the data items should be defined (this applied particularly to conditional selection probabilities) or a misconception regarding the sample design (the person providing the data was not necessarily the same person with whom the sampling panel had discussed the design). To identify these errors required careful analysis by the sampling panel (akin to "edit checks" typically run on survey data) and to correct them required extended liaison with NCs and survey organisations. Regarding selection probabilities, the specification asked only for the conditional probabilities at each stage, on the assumption that it would reduce the burden on data providers if the sampling panel were to calculate the overall probabilities and hence the design weights. However, for some designs the overall probabilities, or the product of those arising from two stages, would have been easier to provide and would also have provided a check on the components. With hindsight, asking the data providers to provide also the overall probability is likely to have averted some of the errors and to have made the process of error detection easier. This will be introduced on future rounds of the ESS.

The type 2 documentation was in most cases a natural by-product of the extended dialogue between the sampling panel member and the NC/survey organisation. The panel member was given the responsibility to produce a final document which was typically a synthesis of information provided and amended over a series of emails and other contacts. This worked well, as it was possible to ensure consistency of content and style while also ensuring that the documentation was produced by someone who had a detailed knowledge of the design. In a small number of cases, elements of the design changed after the documentation had been produced (which was at the "sign-off" stage). For future rounds, it may be desirable to produce two versions of the document – a preliminary (signed-off) version and a final (post-implementation) version. The former could be less detailed than the latter.

Type 3 documentation proved trivial for some aspects of sample design but a considerable challenge for others. An example of the former is sample size. There were two nations in which the selected sample size was clearly insufficient to provide the required minimum number of interviews. This happened due solely to national budgetary constraints. An example of the latter is sampling frame coverage. It was rare that a survey organisation or NC had detailed knowledge of the coverage properties of the frame. Rather, they were able to provide only general statements regarding the circumstances that would have led to omission from the frame – but not estimates of the proportion or characteristics of population units omitted. In some cases, sampling panel members were fruitfully able to suggest where such estimates may be obtained, but typically this was not possible and only the general statements could be provided in the documentation. In practice, the documentation is therefore only of limited use in assessment of likely survey error.

5.8 Examples of improvement in national procedures

In some countries, the implemented design represented a significant breakthrough in survey standards. For example, in Czech Republic a strict probability general population sample was implemented by a private sector survey company for the first time; in Greece an area-based probability sample was used for the first time, and with almost complete coverage (only the smallest islands were excluded), which is unusual; in Italy a new method of including non-electors in a sample drawn from the electoral registers was developed (Lynn and Pisati, 2004); and in France, a probability design was used for the first time in the experience of the survey organisation and researchers involved. In several other nations, more modest advances were made.

In Switzerland, the design incorporated an experiment designed to inform possible improvements in procedures in future years. The ESS specification required a face-to-face initial approach by interviewers to all sample persons/addresses. (Once initial contact had been made, subsequent contacts by telephone were permitted, for example to arrange appointments.) However, common practice in Switzerland is to make the first approach by telephone, for cost-efficiency reasons. As a compromise, a random proportion of the sample addresses were allocated to a “telephone” treatment, while the remainder were approached face-to-face. The intention was that the data resulting from this experiment could be used to assess the extent to which the ESS assumption, that refusal rates would be higher with an initial contact by telephone, held. Preliminary analysis appears in Joye and Bergman (2003). This is an example of the ESS extending methodological knowledge.

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. In particular, the process for developing and agreeing the 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).

The evaluation presented here of the estimation of design parameters 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. For example, uncertainty in the distribution of household size was found to be common for address-based samples (section 5.2). For repeating surveys, such as the ESS, the sample distribution observed at round t might provide a good estimate of the expected distribution at round t+1. For new surveys, analysis of data from other sources (e.g. other surveys or census microdata) might be warranted. Design effect estimates were found to be quite sensitive to predictions of $\bar{b}$, particular when $\bar{b}$ was expected to be small (sections 5.3 and 5.6). In this case, inaccuracy in the prediction of eligibility and response rates can produce relatively large shifts in ($\bar{b}$-1). It can be suggested that conservative estimates of $\bar{b}$ should be used when $\hat{\bar{b}}$ is small. The importance of guidance for estimation of design effects with dual designs (section 5.4) has also been highlighted.

The ESS has provided advances in survey practice in a number of specific nations (section 5.8). Additionally, we believe that the procedures for sample design that are the subject of this article represent a useful advance in the methodology of cross-national surveys. The documentation of the sample design (sections 3.6 and 5.7) provides researchers with considerable data for addressing methodological research questions as well as the opportunity to produce variance estimates that appropriately take into account the sample design. This is important for any survey, but particularly difficult to achieve on a cross-national survey. The ESS provides a model for how it can be achieved successfully.

Finally, much of the information presented in this article can, and will, be treated as quality indicators that will feed into a process of continuous quality improvement for the ESS. The sample design procedures have already been amended in the light of the round 1 experiences and will continue to be reviewed.

References

Elliot, D. (1993). The use of substitution in sampling. *Survey Methodology Bulletin* 33, 8-11.

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. (2004, forthcoming). Combining design effects for dual-design samples.

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.

Joye, D. and Bergman, M. (2003). ESS in Switzerland and modes of recruitment: some remarks. Paper presented to the methodology committee of the European Social Survey, London, 14 November.

Kish, L. (1987).Weighting in $deft^2$. *The Survey Statistician*. June 1987.

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

Lyberg, L. (2000): Review of IALS – a commentary on the technical report. In: *Measuring Adult Literacy. The International Adult Literacy Survey in the European Context*, London: Office for National Statistics.

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. (2004). Approximations to b^* in the estimation of design effects due to clustering. *Working Papers of the Institute for Social and Economic Research*, paper 2004-07. Colchester: University of Essex. Available from http://www.iser.essex.ac.uk/pubs/workpaps/pdf/2004-07.pdf.

Lynn, P. and Pisati, M. (2004, forthcoming). Improving the quality of sample design for social surveys in Italy.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model-assisted Survey Sampling*. New York: Springer-verlag.

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