Survey Research Methods (2007)
Vol.1 , No.2 , pp.55-67
ISSN 1864-3361
http://www.surveymethods.org

Surveys from inside: An assessment of unit nonresponse bias with internal criteria

Ulrich Kohler
Social Science Research Center, Berlin

The article uses the so called "internal criteria of representativeness" to assess the unit nonresponse bias in five European comparative survey projects. It then goes on investigating several ideas why unit nonresponse bias might vary between surveys and countries. It is proposed that unit nonresponse bias is either caused by country characteristics or survey methodology. The empirical evidence presented speaks more in favour of the latter than of the former. Among the survey characteristics the features that strengthen the leverage to control interviewers' behaviour have top priority.
Keywords: Unit nonresponse bias, survey quality, sampling, reachability, response rate.

Introduction

The growing accessibility of European comparative datasets is one of the major developments of survey research in recent time. Sociologists use comparative datasets to investigate their theories under varying societal conditions, and to study the extent to which societal conditions are related to certain phenomena. In addition, European social reporting uses comparative data to monitor social cohesion among the European member states.

However, European comparative datasets not only are of interest for the investigation of European societies. They also provide a methodological quasi-experiment that applies diverse fieldwork procedures under various nation specific conditions. Stated that way, it is an interesting question whether specific country conditions, survey regulations, or sampling methods correlate in a systematic way with the quality of the achieved survey. The aim of this article is to investigate plausible causes for one specific aspect of the survey quality, namely the bias due to unit nonresponse (Groves 2004:11).

Three steps are necessary for this undertaking. The first step is to select the data to be used. This article grew out of a larger project that was aimed at the analysis of societies in the European Union (Fahey et al. 2003; Alber et. al. 2004; Alber et al. 2007). Therefore, survey programs with an emphasise on countries of the EU were selected. These survey programs were: the Eurobarometer (EB) 62.1; the European Quality of Life Survey (EQLS) 2003; the European Social Survey (ESS) of the years 2002 and 2004; the European Value Study (EVS) 1999; and the International Social Survey Program (ISSP) 2002.[1] These survey programs will be described in more detail in the data section.

The second step is the crucial one: the measurement of the unit nonresponse bias. Unit nonresponse bias is just one of several sources for errors in sample surveys (cf. Biemer and Lyberg 2003; Groves 2004). Bias, in general, is a type of error were positive and negative errors do not cancel over many different implementations of a survey (Biemer and Lyberg 2003:47). Bias is due to unit nonresponse if it is caused by sampling units that cannot be located or refuse the request of the interviewer for the interview (Groves 2004:11). To investigate in the causes of unit nonresponse bias, its operationalisation therefore must not vary with any of other possible sources of survey bias, like for example coverage errors, item-non response, sampling bias, interviewer errors, etc. Because of this requirement, the so called "internal criteria of representativeness" (Sodeur 1997) are applied here. The idea of these internal criteria is to measure unit nonresponse bias only for a subgroup of the sample for which the true value of a statistic is known. For the topic of this article these internal criteria have a number of advantages, which will be fully described below. However it should be clear from the beginning that the conclusions of this article are based on a specific subsample. More specifically, high unit nonresponse bias measured in the way proposed here, only gives an indication when something has gone wrong. Absence of such unit nonresponse bias does not guarantee the absence of bias for the entire sample.

The third step is to analyse possible causes for the observed sample bias. These causes should arise from theoretical considerations that cannot be placed into the introduction; hence, they will be introduced in the section on causes and correlates of unit nonresponse bias. However, the general idea that is followed in this article is that unit nonresponse bias is related either to country characteristics or to survey characteristics on the one hand, or to the interaction between these two features on the other hand. The results of this ar-

Contact information: Social Science Research Center, Reichpietschufer 50, 10785 Berlin, Germany (kohler@wzb.eu)

[1] EB, EVS, and ISSP are available from the Central Archive for Empirical Social Research, University of Cologne. The study numbers (ZA-Nr.) are as follows: EB 62.1: s4230; Euromodule s4063; ISSP 2002: s3880; EVS 1999: s3811. The EQLS 2003 is available from http://www.esds.ac.uk/International/access/eurofound.asp and the ESS is available via the homepage of the European Social Survey on http://www.europeansocialsurvey.org.

ticle will suggest that survey characteristics are more important for unit nonresponse bias than country characteristics. Among the survey characteristics the features that strengthen the leverage to control interviewers' behaviour have top priority.

Data

In order to analyse country characteristics that affect unit nonresponse bias it is necessary to select samples from different countries. Moreover, to investigate the causes of unit nonresponse biases that stem from survey characteristics there should be some variation in the fieldwork procedures used. It also is necessary that the fieldwork procedures vary within each country, and that a specific fieldwork procedure was used in more than just one country. At the same time, all samples should be samples of the same underlying population, which in turn implies that the samples should be drawn in a similar time frame. All these features may be achieved by using available datasets of recent survey projects with an emphasis on the EU. Restricting on these survey projects guarantees a reasonable overlap of countries, and at the same time offers considerable variance of country and survey characteristics.

The following comparative survey programs were selected:

- Eurobarometer 62.1: This survey is part of a continuing cross-national research project that has implemented at least two surveys in all EU member states since the early 1970's. The EB 62.1 was carried out in late 2004. The sample consists of citizens age 15 and above. The sample sizes usually amount to approximately 1,000 persons, and to 500 in smaller countries (Luxembourg, Cyprus and Malta). In the EB 62.1, the actual sample sizes vary between 500 in Malta and 1,561 in Germany (see Table 1).

- European Quality of Life Survey 2003: the EQLS '03 was carried out on behalf of the European Foundation for the Improvement of Living and Working conditions in all 25 current EU member states, and in Bulgaria, Rumania, and Turkey. It is the first round of a just started survey program on living conditions in the EU. The samples cover each country's residential population 18 years of age and older. Similar to the Eurobarometer, the sample size is approximately 1,000 persons, and 600 in the smaller countries (Luxembourg, Cyprus, Malta, Slovenia, and Estonia). In 2003, sample sizes varied between 591 in Estonia and 1,071 in Slovakia.

- European Social Survey 2002: This is the first round of a biennial multi-country survey funded jointly by the European Commission, the European Science Foundation, and academic funding bodies in each participating country. In 2002, it covered 21 nations, 18 of which were current EU member states or EU candidates. The samples covered the residential population 15 years of age and older in each participating country. The ESS tries to achieve *effective* sample sizes of around 1,500 respondents (800 for small countries).[2] The actual sample sizes in 2002 varied between 1,207 in Italy and 2,919 in Germany. The German sample, however, was stratified into two separated samples for East and West Germany with 972 and 1,947 observations, respectively.

- European Social Survey 2004: Second round of the ESS. Generally, it shared the features of the first round, but increased the number of participating countries to 26 (21 EU members or EU candidates). At the time of this analysis, data from only 24 (19) participating countries was available. Sample sizes varied between 579 in Iceland and 3,036 in the Czech Republic.

- European Value Study 1999: The EVS '99 was the third round of a cross-national survey research program started in the late 1970s by the European Value Systems Study Group. It covered 32 countries, all of which were EU members and candidates except Cyprus. The target population was adult citizens 18 years of age and older, and the sample sizes varied between 1,000 and 2,000 in most countries. The lowest and highest sample sizes were achieved in Iceland and Russia, respectively.

- International Social Survey Programme 2002: Round 15 of a continuing program of cross-national surveys. Between the end of 2001 and February 2004, surveys were carried out in 33 countries (20 EU members or EU candidates). The target population of the samples were residents 18 years of age and older. Sample sizes varied between 1,000 in Latvia and 2,947 in United Kingdom, with the latter stemming from a stratified sample of Northern Ireland and Great Britain with 987 and 1,960 observations, respectively.

Interviews were conducted entirely face-to-face for all survey programs except for four countries of the ISSP 2002 – France, Finland, Denmark, and Sweden used mail surveys. These surveys have been excluded from all analyses presented in the section on nonresponse bias.

The sampling methods of the survey programs used here not only differed among the programs, but also within them. Sampling methods depend on the available sampling frames, which naturally lead to different sampling methods for different countries. Figure 1 illustrates important features of the sampling methods used in the six survey programs.[3] The first panel of the figure roughly describes the sampling method.

[2] In multistage probability samples, the *effective* sample size is generally much smaller than the number of observations. The decreasing factor depends on the size of sampling units and the similarity among persons within a sampling unit (Kish 1965:187–190; Schnell and Kreuter 2005). In effect, the ESS collects more observations if the sampling-method is a multistage probability sample than if it is a simple random sample.

[3] The sources for information about the survey-organisation are

Table 1: Target population, number of countries and number of observations by survey program [a b c]

Study	Target	Countries	EU+	Min Obs.	Max Obs.
EB 62.1	Citizens 15+	25	25	500	1561
EQLS 2003	Residents 18+	28	28	591	1071
ESS 2002	Residents 15+	22	19	1207	2919
ESS 2004	Residents 15+	24	20	579	3026
EVS 1999	Citizens 18+	32	28	968	2500
ISSP 2002	Residents 18+	33	20	1000	2947

[a]The Eurobarometer 62.1 samples the citizens of age 15 and above. Samples are drawn in 25 countries, all of them are EU+ countries. The sample sizes vary between 500 observations (in Malta) and 1561 observations (in Germany).
[b]All analyses performed for this article were fully programmed with Stata do-files, which can be downloaded from http://www.wzb.eu/~kohler/publications/repraes07/index.htm. The names of the do-files are mentioned below each figure or table
[c]Do-File: ansvydes.do

In particular, it shows whether "simple random sampling" (SRS), several variants of "multistage probability sampling", or "quota sampling" were used. The term "unspecified" refers to samples that applied multistage probability sampling without documentation of the technique used to draw the individuals within the primary sampling units (PSUs). Hence, these may be anything from a multistage probability sample that draws the sampling units from an individual register to a random route sample. It is not documented whether the collection of and contacts to addresses are conducted independently from another for the samples that are categorised as "random route". Only the ESS states explicitly that it has separated these steps.

To date, use of SRS has been infrequent. However, it is fairly common in Denmark, Finland, and Sweden where suitable sampling frames for SRS exist. SRS also has been applied in Malta for the European Value Study, and in Estonia and Slovakia for the ESS 2004. The sampling frame in Malta was a list of all registered voters, making it unsuitable for samples that include persons below the age of 18. Multistage probability sampling with random route is the most often used sampling method. Unavailable sampling frames may be one reason for the popularity of this technique. However, it is only in France, Cyprus, Lithuania, and Turkey where sampling methods based on registers have not yet been applied. In particular, the ESS makes substantial efforts to bypass random route by applying (or even producing) registers as sampling frames. In light of this, the figure reveals that random route generally is not the only available alternative for many countries. Quota sampling, although common in market research, generally is not used in national social surveys. However, it was used in several EVS participating countries and in two ISSP countries. Note that all observations from quota samples were excluded from the analyses presented in the section on nonresponse.

The second panel in Figure 1 contains information on substitution regulations. The figure reveals that substitutions were not allowed in ESS surveys or the EQLS, but always were allowed in the Eurobarometer. Substitution was forbidden in only some of the participating countries of the ISSP and the EVS. The substitution regulations also reflect the use of quota samples in the EVS. It is a basic property of quota sampling that the interviewers arbitrarily select persons with certain properties. This implies that they can substitute every non-cooperative person by another person with the same properties.

Finally, panel three in Figure 1 shows whether institutionalised back-checks were applied. Unfortunately, the information is not available for the Eurobarometer and the EQLS. The figure therefore only displays the back-checking regulations for the three other survey programs; they generally applied back-checks.

The *reported* response rates[4] of the six survey programs are summarised in Table 2. The table illustrates that three survey programs achieved very high response rates for at least one country. The EQLS reports response rates above 90 percent for Germany and Malta, the ISSP achieved a response rate of 99 percent for Spain, and the EVS has a similarly high value for Slovenia. These three survey programs also have very low response rates for at least one other country. The EQLS reports a response rate of 32 percent for Ireland, the EVS reports 15 percent for Spain, and the ISSP reports 20 percent for France (which used a postal survey). The average response rate was about 58 percent for three of the six survey programs, and somewhat higher for the two rounds of the ESS. The response rates of the Eurobarometer are not documented.

Internal Criteria for Representativeness

Consider gender heterogenous couples living together in two-person households. Among this subgroup, precisely 50 percent of the persons are women and 50 percent are men.

as follows. EQLS: Ahrend (2003); ESS 2002: European Social Survey (2004); ESS 2004: European Social Survey (2006); EVS: Information from the methodological questionnaires that are part of the Data delivery package; ISSP: Klein and Harkness (2004). Limited information for the Eurobarometer 62.1 is in the document ebs_215_en.pdf available at http://europa.eu.int/comm/public_opinion.

[4] Remember that although high response rates often are interpreted as a signal of good sample quality, they also might be indicative of insufficiently controlled surveys. A more detailed discussion of response rates will be given in the section on nonresponse.

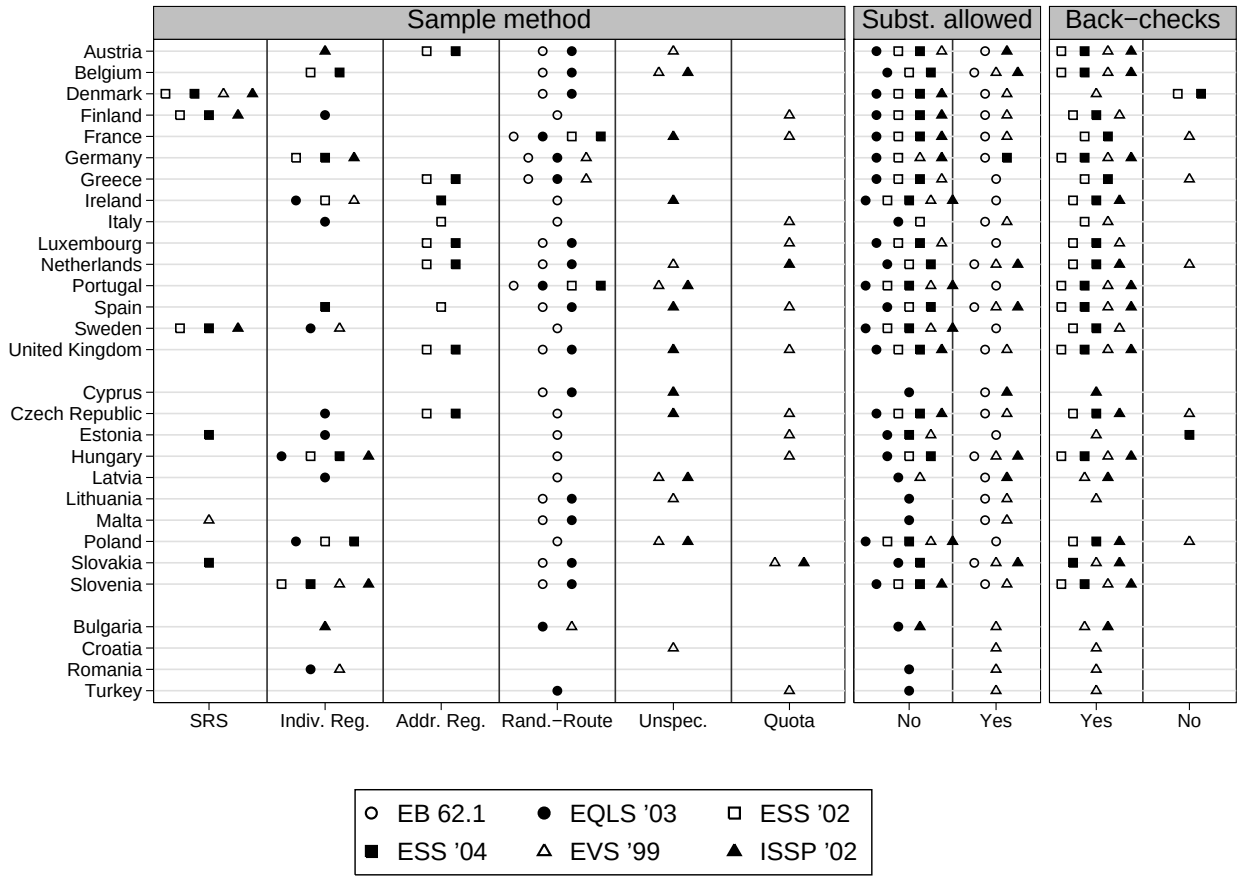

Figure 1. Components of the sampling process by country and survey program [a] [b]

[a]In Austria, all survey programs used multistage sampling. Thereby the ISSP used individual registers as sampling frames within the PSUs, the ESS 2002 and 2004 used registers of addresses; and the EQLS and the Eurobarometer applied a random route technique within the PSUs. For the EVS, the precise regulation of sampling within the PSU is not known.
[b]Do-File: ansample1.do

Also consider that a sample was drawn from a population, and that respondents living together in two-person households are included in that sample. If we were able to select persons from the realized sample belonging to this subpopulation, we could calculate the fraction of women in the sample. Deviations from the true value of 0.5 beyond some acceptable random fluctuation can then be regarded as "internal criteria for representativeness" (Sodeur 1997).

For a formal expression of the above idea, denote the known parameter of the subpopulation with p, and the observed value of that parameter in the sample with $\hat{p}$. Expressed this way, $|\hat{p} - p|$ should not go beyond the limits of random fluctuation. If the variance of $\hat{p}$ also is known, one may calculate

$$B_{\mathrm{UNR}} = \frac{\hat{p} - p}{\sqrt{\mathrm{Var}(\hat{p})}} \qquad (1)$$

with $\mathrm{Var}(\hat{p})$ being the variance of the statistic $\hat{p}$. This will provide a measure of how much $\hat{p} - p$ differs from the expectation of pure random fluctuation. The formula for B_{UNR} resembles the well known Z statistic used for inference procedures on the population mean. In direct analogy to the common practice for the Z-test, values of $|B_{\mathrm{UNR}}|$ above 1.96 might be regarded as high.

An empirical application of the concept of "internal criteria of representativeness" normally requires information about persons not being interviewed, which is often not available. Fortunately, a solution exists for gender-heterogenous couples in two person households. If we select only those respondents who live in two person households, who are married, and who live together with their spouse, it will be quite certain that they belong to the subpopulation of gender heterogenous couples in two-person households. As this approach can be applied for all survey programs, it will be used

Table 2: Reported response rates by survey program [a] [b]

	EB 62.1	EQLS 2003	ESS 2002	ESS 2002	EVS 1999	ISSP 2002
Minimum	n.a.	33	43	46	15	20
Average	n.a.	58	61	62	58	58
Maximum	n.a.	91	80	79	95	99
Missing (num. of countr.)	25	0	0	0	3	1

[a]For the EVS 1999, the average reported response rate was 58 percent. The lowest rate was 15 percent (in Spain), and the highest rate was 95 percent (Slovakia). Response rates were not reported for 3 of the 28 EU+ countries.
[b]Do-File: anresp.do

in the remainder.[5]

For an analysis of causes and correlates of the unit nonresponse bias the application of the above quantity is particular useful:

- As it is fixed by definition, the value of p is not affected by any sort of measurement error.
- It seems unlikely that sampling frames have gender related differences in the coverage of members of two-person households. The value of $\hat{p}$ is therefore unlikely to be effected by "frame coverage errors" (Biemer and Lyberg 2003:63). If the sampling frame is a list of households, or if the sample is drawn by random route, frame coverage errors cannot even affect the value of $\hat{p}$.
- If the sampling frame for a sample of individuals is a list of households, individuals of small households will be over-represented in the entire sample, and this might lead to sampling bias. Restricting on two-person households controls this source of sampling bias such that $\hat{p}$ stays unaffected. This also applies to the variance of $\hat{p}$ in the denominator of equation (1).
- Assuming that gender is commonly assessed by the interviewers (Wolf and Hoffmeyer-Zlotnik 2003:261), $\hat{p}$ is likely to be unaffected by item-nonresponse.
- Assuming that gender can be measured with high validity (Wolf and Hoffmeyer-Zlotnik 2003:261), $\hat{p}$ is likely to be unaffected by observational errors.

Of the sources for survey bias listed by Groves (2004:8–12), only one source of variation of B_{UNR} is left over, and that is unit nonresponse. Hence, if B_{UNR} differ between samples, the reason for it should be unit nonresponse. Therefore the term "unit nonresponse bias" will used to refer to B_{UNR} In the remainder. Unlike the more conventional measures for nonresponse bias listed by Biemer and Lyberg (2003:63) B_{UNR} does not rely on a questionable estimation of the mean of nonrespondents. It is obvious though that B_{UNR} can be computed only for a specific subgroup of the data. The conclusions of this article are therefore restricted to the subgroup of gender heterogenous couples, who are older than the average respondents, but do not differ systematically regarding their household income or their geographic location.[6] High unit nonresponse bias measured for this subgroup gives an indication when something has gone wrong, but absence of unit nonresponse bias for gender heterogenous couples does not guarantee the absence of unit nonresponse bias for the entire sample.

Figure 2 illustrates the quantities used to calculate the above defined measure for the unit nonresponse bias (B_{UNR}). The dots in the figure show the observed fraction of women amongst gender heterogenous couples ($\hat{p}$). The true fraction of $p = 0.5$ is indicated by a vertical line. One might argue that the more the dots deviate from the vertical line, the worse the sample is. An observed fraction of $\hat{p} = 0.6$ means that the fraction of women is 10 percentage points higher than it should be. Note that deviations of that size are not uncommon. Overall, the figure indicates a tendency towards overrepresentation of women for the Eurobarometer and the ISSP, but not as much for the other survey programs. Moreover, the distances between the observed and the true values seem to be larger for the EQLS and the Eurobarometer, somewhat smaller for the ISSP, and relatively small for the other survey programs. A more detailed discussion of these results appears in Kohler (2007).

However, as mentioned above, it is reasonable to accept a certain amount of random fluctuation in the differences. The amount of acceptable random fluctuation largely depends on the standard error of the observed fraction. As the true fraction of women is known to be $p = 0.5$, the variance may be calculated with $(p \times (1-p))/n = 0.25/n$, which can then be applied to calculate the 95% confidence interval with CI $= .5 \pm 1.96 \times \sqrt{.25/n}$.[7] One might consider values

[5] The true fraction of women in a subpopulation defined as above can deviate from 0.5 for two reasons:

1. Marriage among homosexual partners has become legal in some survey countries. It might be that homosexual marriage is gender specific, i.e. more frequent among men than among women (or vice versa).
2. The drop-out from the sampling population might be gender specific. For example, some survey programs restrict themselves to country's native citizens. It might be that for gender heterogenous couples in 2 person households men (or women) more often are excluded from this definition. In couples that are formed by a foreigner and a native citizen this will happen, if the foreigner is more often male (or female) .

Overall, the effect of both mechanisms is considered small, however.

[6] ansubgroup01.do

[7] Northern Ireland and Eastern Germany have been oversampled in the EVS, the ISSP, and the Eurobarometer. For these cases $\hat{p}$ was calculated with weights, and Var($\hat{p}$) was multiplied with the

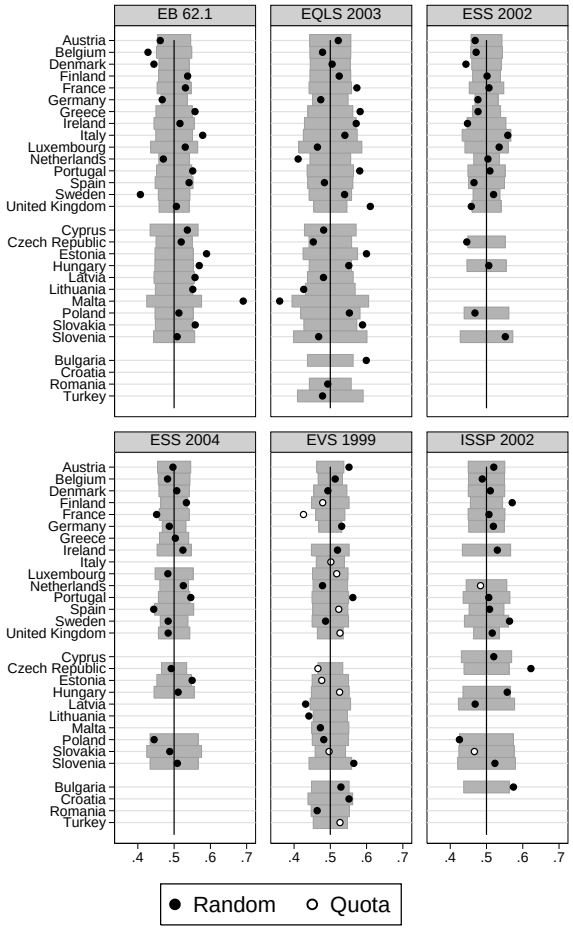

Figure 2. Fraction of women amongst gender heterogenous couples by survey program and country [a] [b]

[a]In the EB 62.1 the fraction of women is around 45 percent, which is too low, but still within the confidence bounds. The proportion of female in Austria also is too low for the ESS 2002, and too high for the EQLS, the EVS, and the ISSP. The EVS provides the only Austrian value that is outside the confidence bounds.
[b]Do-File: anpwomen.do

outside the 95% confidence interval as problematic.

Figure 2 uses shaded bars to display the boundaries of the confidence intervals around the true values, making it possible to identify samples with unit nonresponse biases that are critical (above 1.96). As it stands, 41 (29 %) of the 139 values displayed in the figure are too large in this respect. Given that only five percent of the values should be outside the confidence bounds, 29 percent is quite a large amount. At first glance, the fraction of critical unit nonresponse biases seem to be higher for the Eurobarometer and the EQLS than for the other four surveys, but this will be dealt with in more detail in the section on nonresponse. For the calculation of the unit nonresponse bias, both quantities - the difference between the dotted values and the vertical lines, as well as the size of the standard error - are taken into account. B_{UNR} is the difference between a plotted dot and the vertical line, divided by the standard error. Values outside the confidence bounds will have a unit nonresponse bias equal to or higher than 1.96. In what follows, the unit nonresponse bias will be described more thoroughly by showing the correlations with several country and survey characteristics.

Causes and correlates of the unit nonresponse bias

This section explores the relationship between unit nonresponse bias and some potential causal factors in order to discern *why* biases vary between survey programs and countries. Two possibilities will be considered. First, unit nonresponse bias may be a function of country properties. It is assumed that some countries may present a difficult environment for sample surveys, and that unit nonresponse biases should be high in these countries regardless of how hard the survey administration tries to avoid it. Second, unit nonresponse bias may be a function of survey methodology. Sound methodological practices should yield to low unit nonresponse bias no matter how difficult the environment for drawing the sample may be.

Before starting, a disclaimer on causality and the investigated attributes seems necessary. The empirical evidence provided here does not claim to be a strict causal analysis. In general, causality has to be established by experiments or sophisticated statistical models.[8] The data are not from a true experiment, and there is not enough independent information in the available datasets to apply more sophisticated statistical models. Thus, the following analysis essentially is descriptive. However, by investigating *different* implications of the same causal hypothesis this analysis may provide at least some piece of evidence for causality, and in this way may be considered more than simply descriptive.

The first check of the idea that unit nonresponse bias is related to country characteristics may be found in Figure 3. The figure displays the absolute values of B_{UNR} by country. Values far right indicate high unit nonresponse biases. As it stands, Figure 3 is more or less just a regrouping of Figure 2 with the exception that it is based only on face-to-face surveys and probability samples. The data from quota samples and from four postal surveys were excluded, and they were not used in any of the analyses presented in this section.

If some countries posed a difficult environment for conducting surveys of the general population, one might anticipate high unit nonresponse biases regardless of the survey program. Using $|B_{\text{UNR}}|$ as a measure of unit nonresponse bias, support for this idea would come if all the biases for a "difficult" country fell above the critical $|B_{\text{UNR}}|$ value of 1.96 (and, conversely, that $|B_{\text{UNR}}|$ would always be below 1.96 in countries that are "survey friendly"). Contrary to

respective design effects.

[8] Refer to King et al. (1994) and Winship and Morgan (1999) for an introductory overview on causal analysis with non-experimental data, and to Lieberson (1985) or Berk (2004) for some critical remarks.

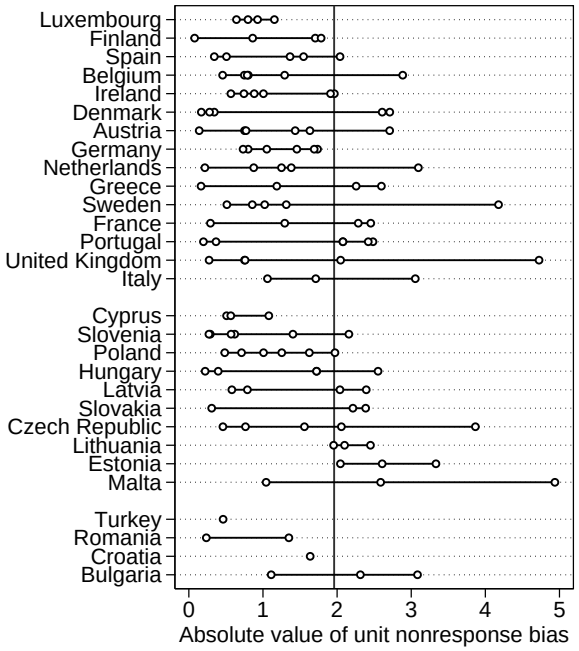

Figure 3. Absolute values of the unit nonresponse bias ($|B_{\text{UNR}}|$) by country [a] [b]

[a]Values far right indicate high unit nonresponse bias. Virtually any country that exhibits a problematic sample bias also shows an example of good survey quality.
[b]Do-file: anBctry.do

this expectation, the figure reveals that most countries evidence both small *and* large unit nonresponse biases. For virtually any country where at least one survey achieved a unit nonresponse bias above 1.96, at least one example of a low unit nonresponse bias also is found. In addition, the countries with at least one small unit nonresponse bias also show at least one problematic sample. Luxembourg, Finland, Germany, Cyprus, and Rumania are the only countries where some evidence suggests a "survey-friendly" environment; Lithuania and Estonia might have a difficult environment for survey research.

A somewhat less strict implication of the idea that country characteristics are responsible for unit nonresponse bias is that the variance of the biases within countries should be smaller than the variance between countries. However, we find no evidence to support this hypothesis either. By visual inspection, the variance within most countries seems to be quite large, while between country variance is at least not obviously inflated. Two numbers point out this visual impression more specifically: the average of the within country variances is around 1.03, while the variance of the country means is around 0.29.

It could be possible to observe a large within-country variance of the unit nonresponse bias even if countries differed in their survey-friendliness. This would be the case if some survey programs applied a fieldwork methodology that was able to cope with the specific environment, while others do not. Thus, if we could measure the survey-friendliness of countries more directly, we might be able to find a correlation between this measure and the unit nonresponse bias. Typical gender related reachability differences might be a useful first operationalization of survey-friendliness. In all surveys used here, interviewers needed to conduct the interview with only one person of a household, and this target person is either given by a preselected name or by a selection criteria to be applied after contacting the household (i.e. last birthday method, Kish selection grid, etc.). Now, let us assume for a while that interviewers do always follow the selection criteria correctly, i.e. they do not conduct interviews with persons that are not the target persons. It might than turn out that the target person in some household is essentially unreachable. In this case interviewers will be forced to skip the target person in question and to move on to the next sampling unit in another household. However, if the skipped target persons in gender heterogenous couples were predominately men, the sample should show an over-representation of women, and vice versa. Hence, one generally would expect an over-representation of that gender that is easier to reach.

As not employed persons tend to be at home more often than employed persons, employment status might be used as an indicator for reachability. Starting from this, the difference between the male and female employment rate might be an aggregated measure for the gender-related reachability structure in a country. The higher the employment rate of men compared with the employment rate of women, the more difficult it should be to reach the male part of the population and the stronger the over-representation of women should get. If men and women had similar employment rates, the gender related reachability difference should diminish, and in turn also the unit nonresponse bias among gender heterogenous couples.

Figure 4 contradicts this claim. It displays the raw values of unit nonresponse bias B_{UNR} by gender related reachability differences, calculated by subtracting the female from the male employment rate. Values at the top of the graph indicate strong over-representation of women; values at the bottom indicate over-representation of men. Values far right represent a prevalence of what is called the "male breadwinner model". The figure indicates that reachability differences measured that way do not lead to an over-representation of those that tend to be easier-to-reach (r =0.03, p =0.74).

Clearly, the aggregate difference between male and female employment rates for the entire population only can be a rough indicator for reachability differences within gender heterogenous couples. The EQLS, however, allows for a more direct measure of reachability differences within a household; it contains the employment status for each person within a household. Therefore, it is possible to investigate whether the unit nonresponse bias diminishes when both parts of the gender heterogenous couple are equally reachable. If selection within a household was caused by reachability, one would anticipate a selection bias towards the gender of the unemployed person of the household, and no selection bias for households where both members are either

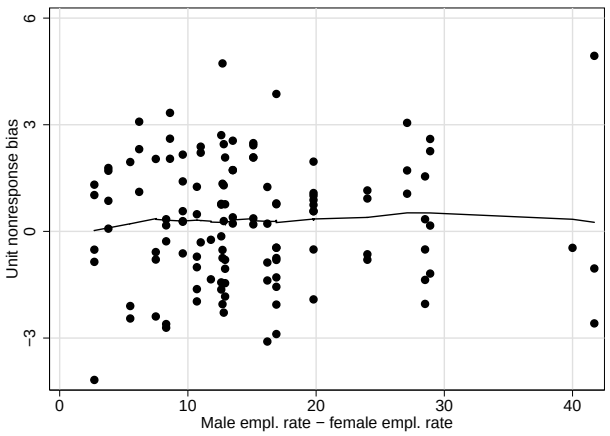

Figure 4. Unit nonresponse bias (B_{UNR}) by male employment rate minus female employment rate [a] [b]

[a]Values at the top of the graph indicate strong over-representation of women, and values far right represent higher prevalence of what is called the "male breadwinner model". The figure indicates that gender related reachability differences cannot explain unit nonresponse bias
[b]Do-File: anBreach.do

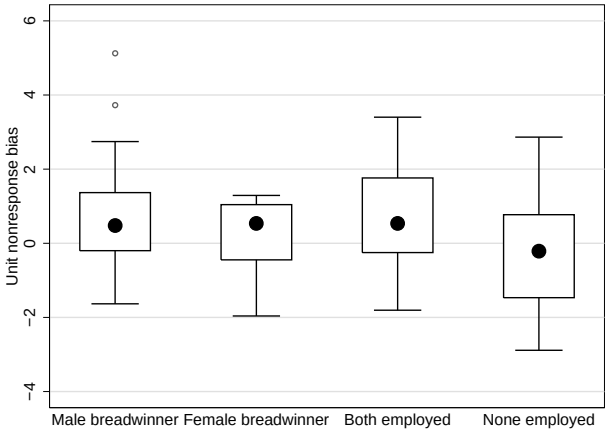

Figure 5. Box plots of the unit nonresponse bias (B_{UNR}) by within household reachability [a] [b]

[a]Boxes at the top of the figure indicate strong over-representation of women. The figure indicates that gender related reachability differences cannot explain unit nonresponse bias
[b]Do-File: anBrwithin.do

employed or unemployed.

Figure 5 shows box plots[9] of the raw values of the unit nonresponse bias (B_{UNR}) for four different household types. The first two types are one-earner households, which have been termed male or female breadwinner households, respectively. The two other types represent households where both parts are equally reachable. From the reachability hypothesis, one should expect an over-representation of women in the male breadwinner households, an over-representation of men in female breadwinner households, and no over-representation for the two others. The empirical results do not quite reflect these expectations. First of all, one has to state that the differences in the unit nonresponse biases between the household types are not very large. Moreover, although women are in fact over-represented in male breadwinner households, they also are over-represented in households where both respondents were employed, and even in female breadwinner households. Reachability differences also can not be responsible for the under-representation of women in households where both persons are not employed. Somewhat in favour of the reachability hypothesis, however, is the finding that women more often are part-time employed than man so that the over-representation of women in dual earner households might be an effect of these *unmeasured* reachability differences. As anticipated, the over-representation of women is also stronger for male breadwinner households than for female breadwinner households. Overall, the evidence that is supportive for the reachability hypothesis is weak, however.

One possible reason for the apparent inadequacy of the reachability hypothesis might be that the higher reachability of men in female breadwinner households is diminished by cultural norms that encourage survey participation in females more than males. If such norms existed, one also should find an over-representation of women in households where both persons either are employed or unemployed. Such a tendency is visible for the dual earner households, but not for the households where both parts of the couple are unemployed. There is even a tendency of male over-representation in the latter. Thus, the results do not fit well with the cultural norm hypothesis. Remember, however, that the results presented in Figure 5 only refer to data of the EQLS, which might be a special case in certain respects. There will be more on this later.

In the previous paragraph, a cultural explanation for unit nonresponse biases was suggested. Another cultural explanation might be widespread objections against survey research. Assuming that persons with such objections are likely to refuse participation in a population survey, the overall response rates may be used as an indicator for such objections. It is clear, however, that the response rates also are connected to the efforts a survey administration spent to reduce unit nonresponse. In this sense a correlation between response rates and unit nonresponse bias can also be viewed as a survey characteristic; there will be more on this later.

Conceptually, the overall response rate is the division of the number of realised interviews divided by the total number of eligible sampling units. High overall response rates leave less space for systematic drop out, and hence for unit nonresponse bias. Response rates are sometimes seen as a

[9] The filled circles of the box plots display the median. The upper and lower ends of the boxes are the upper and lower quartiles, and the vertical lines are used to indicate the spread and shape of the tails of the distribution. Little white circles indicate outliers. Box plots were invented by Tukey (1977) and described in some detail by Cleveland (1994:139–143), Schnell (1994:18–20), and many others.

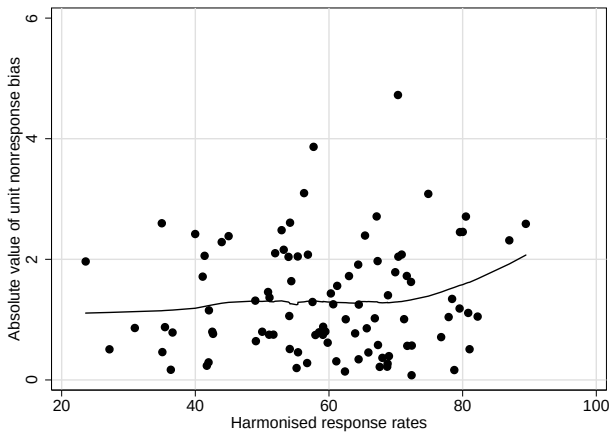

Figure 6. Absolute values of the unit nonresponse bias ($|B_{\text{UNR}}|$) by harmonised response rate [a] [b]

[a]The figure shows a slight positive correlation between the harmonised response rate and the deviation from randomness. This means that samples with high response rates tend to have a higher unit nonresponse bias. The results are taken as a manifestation of the problem that poorly controlled surveys tend to have high response rates.
[b]Do-File: anBhresp.do

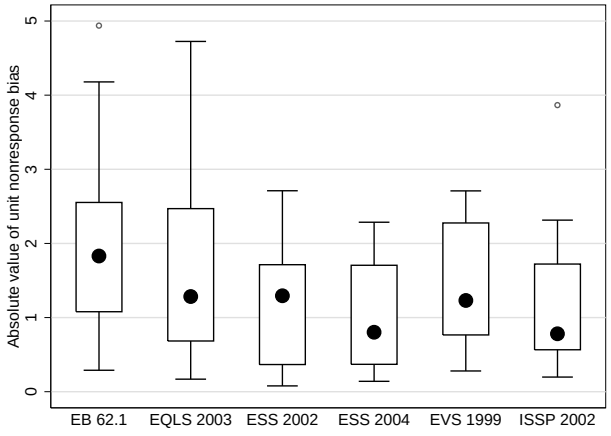

Figure 7. Box plots of the absolute values of the unit nonresponse bias ($|B_{\text{UNR}}|$) by survey program [a] [b]

[a]Boxes at the top of the figure indicate large unit nonresponse biases. The figure reveals that the bias is higher for the Eurobarometer 62.1 and the EQLS, and lowest for the two ESS surveys.
[b]Do-File: anBsurvey.do

measure of sample quality. However, although the general idea of the response rate is fairly ubiquitous, the actual formulas for calculating the response rates widely differ. Differences especially exist for *non eligible respondents* (Schnell 1997:19–27). For the following analysis, the response rates therefore were recalculated such that only non-residential and non-occupied addresses were treated as ineligible sampling units. Drop-outs that occured because of noncontacts, moving abroad, poor language skills, or illness were kept inside the gross sample. This recalculation was not possible for the Eurobarometer and some countries of the EVS, which were excluded from the analysis. Moreover, the harmonised response-rate of 99 percent for Spain in the ISSP has been regarded as meaningless and excluded as well.

Figure 6 shows the relationship between $|B_{\text{UNR}}|$ and the harmonised response rate. In marked contrast to the expectation that high response rates lead to low unit nonresponse bias, the figure reveals no (or even slight positive) relationship between the two measures (r =0.09, p =0.37).

The result that high response rates do not correlate with lower unit nonresponse bias is not very surprising, however. It merely reflects an argument that often has been made against using response rates as an indicator for sample quality (Schnell 1997:26;58): During fieldwork, a more or less natural defection of the interviewer from the sampling scheme is to substitute a not easy-to-reach sampling unit with a contact person at hand. Normally, interviewers will not report back their *defective* behaviour, with the result that the initial response rate (r) will raise by the factor of $\frac{n+1}{n}$. But even for the unlikely case that interviewers report back the substitution, the response rate will be inflated. In this case, both the number of realised interviews and the gross sample will be raised by one, so that the initial response rate will be raised by the factor $\frac{n+1}{n+r}$. The response rates will be inflated by the described interviewer behaviour because the they do not capture some of the unit nonresponse that has occurred; one might call this the "hidden nonresponse". Hidden nonresponse are likely to counteract the otherwise benignant affects of high response rates. High response rates only should be taken as an indicator for sample quality if a survey were well controlled.

Extensive fieldwork control might be a cause for a low unit nonresponse bias in itself. This brings up the second type of conjectures about the causes for unit nonresponse biases. To begin with, one might argue that the unit nonresponse bias largely depends on the sum of the decisions of the survey administration. These decisions include the overall sampling methods, the selection, briefing, supervision, payment, and control of the interviewers, the strategies for refusal conversions, the format of the advance information and much more. Clearly, taken together these decisions should affect the unit nonresponse bias. It is therefore reasonable to analyse the correlation between unit nonresponse bias and survey program.

Figure 7 shows the distribution of $|B_{\text{UNR}}|$ for each survey program with box plots.[10] Looking solely at the median, one gets the impression that the Eurobarometer is worse than other surveys, and that the ESS 2004 and the ISSP 2002 share a similar relatively good overall unit nonresponse bias. This impression has to be slightly modified, however, if one also looks at other features of the distributions. As indicated by the height of the boxes and whiskers, the variation of the

[10] See footnote 9.

bias is largest for the Eurobarometer, and the EQLS samples. Both show a strong skewness towards the right, with the result that more than 25 percent of the sample qualities are beyond the critical value of 2. The fraction of critical deviances are 44 and 39 percent in the Eurobarometer and the EQLS, respectively, while this fraction lies between 16 and 26 percent for the other surveys. The lowest fraction of deviances that are critical is observed for the ESS 2002 (16%). Using this metric, the ESS 2002 is the "best" of six. The second best survey in this respect is the ESS 2004 with critical deviances of 20 percent. All together, Figure 7 strengthens the notion that the sample qualities of the EB 62.1 and the EQLS are somewhat more problematic than of the other survey programs.

Several quantities habitually reported in the study descriptions of survey data may be used as more specific indicators for fieldwork control. First, there is the sampling method. Clearly, the sampling method is not an indicator for fieldwork control *per se*, but effective control over interviewer behaviour may differ systematically across different sampling methods. More specifically, the leverage of interviewers on the selection of respondents defines the efforts that are necessary to control their work (Schnell 1997:58). If the interviewers are dealing with respondents whose names and addresses are known beforehand, it will be easy to recontact these known persons and to ask them whether they were interviewed. If, on the other hand, the interviewer selects the address of a target household by applying a random route technique, the survey administration will need to reapply the random route of the interviewers in order to check whether the right respondent had been selected. It may be expected that simple random sampling and multistage probability sampling with individual registers produces better sample qualities than household samples based on applications of random route techniques.

Back-checking regulations are another indicator of the efforts spent on fieldwork control. Back-checking means that the survey administration re-contacts respondents in order to find out whether the respondent really exists and whether an interviewer has in fact conducted the interview with that respondent. With certain limits, back-checks can also be used to control whether the *right* respondent has been interviewed. In general, one should expect a lower unit nonresponse bias in surveys with institutionalised back-checks. Finally, there is the issue of the allowance of substitutions. During fieldwork, survey administrations have to deal with the problem of non-cooperative or unreachable research units. An often applied solution is to substitute such research units with newly drawn sampling units. It has been claimed that substitution leads to over-representation of cooperative and easy-to-reach respondents, and tends to decrease the extent of interviewer efforts to gain response from the original research units (Elliot 1993). One should therefore expect that the unit nonresponse bias increases when substitutions are allowed.

The conjectures about the effects of fieldwork control may be evaluated with Figure 8. The figure shows box plots[11] of $|B_{\text{UNR}}|$ by the three features of survey methodology just discussed. The first panel compares categories of the sampling method. It reveals that the unit nonresponse bias of multistage probability samples with random route tend to be higher than for the other techniques. Almost 50 percent of the samples that applied random route have a unit nonresponse bias beyond the critical value of 2, and there is nothing in the distribution of $|B_{\text{UNR}}|$ that is favourable for random route. Simple random samples (SRS) and the two other variants of multistage probability sampling have relatively low unit nonresponse biases.[12] Special attention is necessary regarding the results of samples with unspecified sampling methodology. As explained in the data section, these samples are multistage probability samples where the technique to select the respondents within the PSU is not fully documented. We now see that the median unit nonresponse bias of these samples is quite low, but that the distribution is heavily skewed to the right. Further investigation shows that the distribution is in fact bimodal. Slightly above 50 percent of the samples have values below 1, while almost all other values are beyond the critical value of 2. Given the other results shown here, this suggests that parts of the samples in this category are random route samples, while the others use registers to select the sampling units.

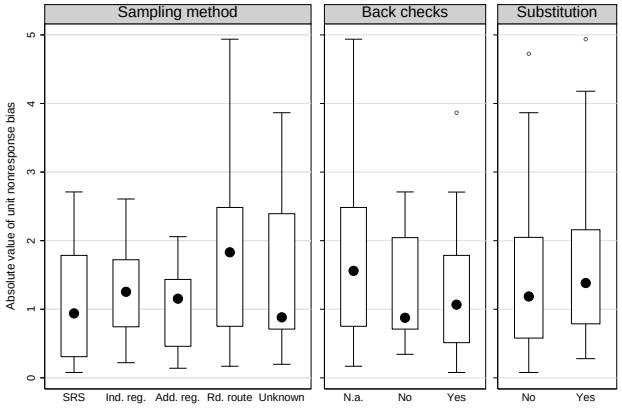

Figure 8. Box plots for absolute values of the unit nonresponse bias ($|B_{\text{UNR}}|$) by dimensions of the fieldwork [a] [b]

[a]Boxes at the top of the figure indicate large biases. The figure reveals that samples from random route samples tend to higher biases than from samples working with registers. Back-checks and restrictive substitution regulations reduce unit nonresponse bias.
[b]Do-File: anBmethod.do

The second panel of Figure 8 shows the distribution of the absolute values of B_{UNR} by back-checking regulations (the figure for the surveys without back-checking is based on 5 observations only, so that these results cannot be trusted).[13] The figure reveals that surveys with explicitely stated back-

[11] See footnote 9.

[12] Note that all multistage probability samples with selection from address registers are actually ESS-samples (cf. fig. 1).

[13] Figure 1 shows that back-checks were not applied in 8 samples. However, two of them were not used because they are quota samples, and the Greek EVS was not used because of data limitations that makes it impossible to isolate gender heterogenous couples.

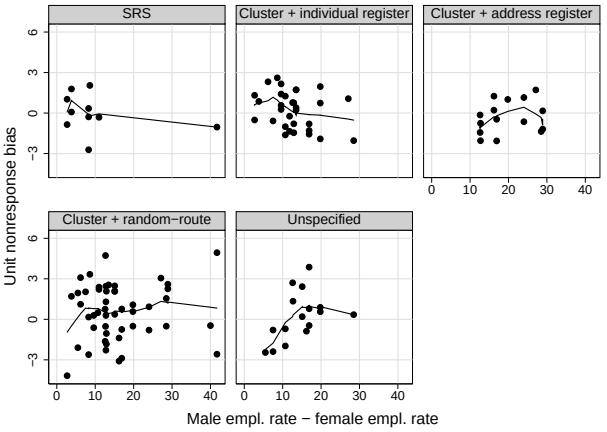

Figure 9. Unit nonresponse bias (B_{UNR}) by gender related reachability and sampling method [a] [b]

[a]Random route and unspecified sampling methods are probably less well suited to cope with gender related reachability differences.
[b]Do-File: anBreachmeth.do

checking regulations tend to have lower values of $|B_{\mathrm{UNR}}|$ than those that do not document any back-checking regulation. The latter are all EQLS and Eurobarometer samples.

Finally, the figure shows the effect of the allowance of substitutions on the unit nonresponse bias. As expected, the samples that allow substitutions tend to have a higher bias.

So far the analyses have implicitly assumed that country and survey characteristics affect the unit nonresponse bias independently. In practice, it is however likely that country characteristics that are problematic for conducting a survey can be dealt with by a suitable survey methodology. If this were true, country characteristics might be mediated or reinforced by survey methodology. In statistical terms, this suggests an interaction effect between country characteristics and survey methodology. In what follows, the interaction effects of the survey methodology with the aggregated reachability-differences and harmonised response rates will be described.

The sampling method is used as an indicator for the overall survey methodology. As argued before, the sampling method affects the leverage the interviewers have on the selection of respondents and thereby influence the efforts that are necessary to control interviewers' work. More specifically, it has been argued that random route is a problematic technique because interviewers' behaviours can not be controlled as well as in other sampling methods discussed here. Therefore, it was expected that the unit nonresponse bias might be higher for random route as for other sampling methods, and this has been confirmed above. It also was expected that country characteristics that create a difficult environment for population surveys would correlate more strongly with the unit nonresponse bias when random route is used.

Figure 9 shows the unit nonresponse bias by gender related reachability, separated for the five different sampling methods. The figure shows a very weak positive relationship for the random route samples (r =0.09, p =0.54) while there is an – again very weak – negative relationship for the samples that draw the target persons from individual registers (SRS: r =-0.30, p =0.39; Cluster + Individuals: r =-0.29, p =0.11). Even though none of the correlation coefficients between unit nonresponse bias and gender related reachability significantly differ from zero, and even though there are no significant differences between these correlations, the overall impression of the graphs in Figure 9 is in favour of the above expectation that random route samples are less suited to cope with gender related reachability differences than samples that do not rely on the interviewer to identify the sampling unit. At least the results presented here do not reject this claim.

Note that many of the EQLS samples are based on random route, so that the results presented here relate to the analysis of the effects of the within household reachability differences presented in Figure 5. In that analysis, a weak effect of within household reachability was found. A repetition of that analysis using only the random route samples strengthens the evidence in favour of the reachability hypothesis.

The strongest positive relationship between unit nonresponse bias and gender related reachability differences has been observed for the samples with unknown sampling methods (r =0.45, p =0.08), and for the samples with address frames (r =0.25, p =0.37). The latter sampling method also involves the interviewer for the selection of the sampled unit, and the former might include a substantive fraction of random route and/or household samples.

Another country characteristic that should interact with the sampling method is the overall response rate. It was argued above that the overall response rate is a somewhat problematic measure of survey quality. On the one hand, high response rates may be an indicator of little space for systematic drop-outs; on the other hand, they might be highly inflated because of "hidden nonresponse" (see page 63). In the case of hidden nonresponse, a high response rate is just an indicator for a less well controlled survey. One may expect that high response rates in well-controlled surveys would be indicative of high sample qualities, whereas high response rates in less well-controlled surveys would be indicative of low sample qualities. Hence, one would expect an interaction effect between the response rates and the survey methodology on the unit nonresponse bias.

To investigate the interaction between survey methodology and response rates, Figure 10 displays the relationship between the absolute values of the unit nonresponse bias and the harmonised response rates, separated by the five distinguished sampling methods. The unit nonresponse bias decreases with high response rates for the multistage probability samples with selection from address registers (r =-0.49, p =0.07). It generally stays the same for the multistage probability samples that used individual register (r =0.01, p =0.97), and it slightly increases for the random route samples (r =0.27, p =0.20) and those with unknown sampling

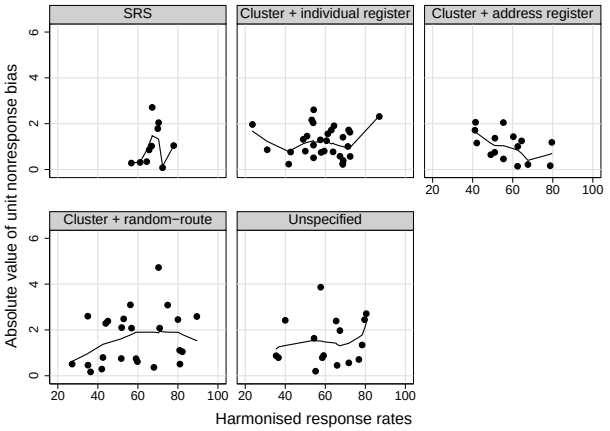

Figure 10: Absolute value of the unit nonresponse bias ($|B_{\text{UNR}}|$) by harmonised response rate and sampling method [a] [b]

[a] High response rates leads to large biases for random route and unspecified sampling, while there is no (or even negative) relationship for the other sampling methods. This can be taken as indication that the random route samples are less well controlled.
[b] Do-File: anBhresmeth.do

method (r =0.15, p =0.59).[14] In this sense, the figure supports the claim that response rates are indicative for high unit nonresponse biases only for the sampling methods that offer better opportunities for fieldwork control.

The results of Figure 10 also may be interpreted in a slightly different way. So far it has been assumed that random route samples are, in fact, less well controlled; one might want to argue against this assumption. Figure 10, however, seems to suggest that random route samples are less controlled. This is because, in an ideal world, high response rates *must* lead to a lower unit nonresponse bias. When they do not, the only possible explanation is that there is hidden nonresponse. Hence, if unit nonresponse bias increases with the response rate, there must be hidden nonresponse. And, if one observes more hidden nonresponse for certain sampling methods, one must conclude that this sampling method performs less well because it produces hidden nonresponse. In this sense, the surveys that have used random route (or unknown sampling methods) are suspicious for being less well controlled, or less well controllable.

It was stated at the beginning of this section that although the analyses were grounded in causal hypotheses, the results are largely descriptive. However, we also claimed that all the results together produce some causal evidence as well. The following summarising section tries to bring together these causal implications.

Discussion

This previous section has described the correlation between unit nonresponse bias and several plausible implications of the idea that unit nonresponse bias is either caused by country characteristics or survey methodology. The empirical evidence presented speaks more in favour of the latter than of the former. Almost *no* country had only either very high or very low unit nonresponse biases. Moreover, unit nonresponse bias and gender related reachability proved unrelated in these analyses (although there were some weak indications from the EQLS that reachability was positively related to survey participation). It is also *not* the case that countries with low response rates typically had larger biases. There are, however, several dimensions of survey methodology that show consistent and plausible correlations with the bias. The two European Social Surveys that have been shown to have the most rigid fieldwork procedures of the surveys included in this analysis (Kohler 2007) had the lowest unit nonresponse biases, while the Eurobarometer and the EQLS were somewhat higher. Sampling methods that offer more possibilities for extensive fieldwork control yielded lower unit nonresponse biases than the random route technique. Back-checking regulations and substitution allowance affected the unit nonresponse bias in the expected direction. Finally, it was shown that some sampling methods are able to cope with difficult country characteristics better than others. Again, random route sampling proved to be the most problematic in this respect.

Many of the correlates suggest that the unit nonresponse bias is a result of interviewers' behaviour. As it stands, all examined sampling methods are, in fact, probability samples. Hence, they all should lead to the expected fraction of women amongst gender heterogenous couples. If not, it either is because target respondents who have certain characteristics in common are not reachable or not willing to participate, or it is because the interviewer starts the interview with the wrong person. If it was the former, there should be no difference in unit nonresponse biases between different sampling methodologies. Men might be more difficult to reach than women but *why* should they be more difficult to reach if one applies random route sampling? Defective interviewer behaviour as a cause for unit nonresponse biases, on the other side, fits well to all the correlates investigated here. Suppose that men are more difficult to reach in a specific country. In that case, it would be harder for interviewers to get interviews from men than women, and they might consider conducting the interview with a substitute–but only if they did not fear negative consequences from their survey organisation. Thus, one would expect a higher unit nonresponse bias in surveys were the interviewers are less well controlled. One also would expect that reachability differences affect unit nonresponse bias only for less well controlled surveys. Back checking regulations would be expected to decrease unit nonresponse bias, and hidden nonresponse to be more widespread in less well controlled surveys. Most of the results of this study fit to the explanation that defective interviewer behaviour is responsible for the observed unit nonresponse biases. Sampling methods matter, but not because some sampling methods are more "random" than others, but because some sam-

[14] The correlation for SRS (r =0.32, p =0.37) is not interpreted because of the very small variance of the response rates of the these samples.

pling methods offer better opportunities to control interviewers' behaviour than others.

What can be learned from an analysis such as the one presented here? The main message is that survey methodology is important. In the European context some excuses for low sample quality do not suffice, or at least do not suffice anymore. There is no excuse for obtaining poor sample quality due to surveys being conducted in less developed countries or in countries that present various challenges to survey research. If one achieves a sample with low sample quality, the results presented here suggest that the cause lies in aspects of the survey methodology, especially those that make defective interviewers' behaviour less controllable.

Acknowledgements

Thanks to Jens Alber, Frauke Kreuter, Rainer Schnell and Christoph Welzel for their critical and encouraging remarks. Moreover I like to thank Susan Kenney and Scott Fricker for copy editing, and Magdalena Luniak for careful assistance.

References

Ahrend, D. (2003). *The Quality of Life Survey. On behalf of the European Foundation for the Improvement of Living and Working Conditions.* (Fieldwork Technical Report of Intomart GfK, Hilversum)

Alber, J., Anderson, R., Delhy, J., Domansky, H., Fahey, T., Keck, W., et al. (2004). *Quality of Life in Europe. First Results of a New Pan-European Survey*. Luxembourg: Office for Official Publications of the European Communities.

Alber, J., Fahey, T., & Saraceno, C. (Eds.). (2007). *Social Conditions in an Enlarged Europe*. Oxford: Routledge.

Berk, R. A. (2004). *Regression Analysis. A Constructive Critique*. Thousand Oaks: Sage.

Biemer, P. B., & Lyberg, L. E. (2003). *Introduction to Survey Quality*. Hoboken: Wiley.

Cleveland, W. S. (1994). *The Elements of Graphing Data.* (2 ed.). Summit: Hobart Press.

Elliot, D. (1993). The Use of Substitution in Sampling. *Survey Methodology Bulletin*, *33*, 8–11.

European Social Survey. (2004). *ESS Documentation Report 2002/03: The ESS Data Archive, Ed. 5.1.* (http://ess.usd.nib.no/2003/ESS1DataDocReport_e04_1.pdf)

European Social Survey. (2006). *ESS2-2004 Documentation Report: The ESS Data Archive, Ed. 2.0.* (http://ess.nsd.uib.no/index.jsp)

Fahey, T., Nolan, B., & Whelan, C. T. (2003). *Monitoring Quality of Life in Europe*. Luxembourg: Office for Official Publications of the European Communities.

Groves, R. M. (2004). *Survey Errors and Survey Costs*. New York: Wiley.

King, G., Keohane, R. O., & Verba, S. (1994). *Designing Social Inquiry*. Princeton: Princeton University Press.

Kish, L. (1965). *Survey Sampling*. New York: Wiley.

Klein, S., & Harkness, J. (2004). *ISSP Study Monitoring 2002. Report to the ISSP General Assembly on Monitoring Work Undertaken for the ISSP by ZUMA, Germany.* (ZUMA Methodenbericht 2004/10)

Kohler, U. (2007). Quality Assessment of European Surveys. Towards an Open Method of Coordination for Survey Data. In J. Alber, T. Fahey, & C. Saraceno (Eds.), *Social Conditions in the Enlarged Version.* Roudledge.

Lieberson, S. (1985). *Making it Count. The Improvement of Social Research and Theory*. Berkely: University of California Press.

Schnell, R. (1994). *Graphisch gestützte Datenanalyse*. München u. Wien: Oldenbourg.

Schnell, R. (1997). *Nonresponse in Bevölkerungsumfragen. Ausmaß, Entwicklung, Ursachen.* Opladen: Leske und Budrich.

Schnell, R., & Kreuter, F. (2005). Separating Interviewer and Sampling-Point Effects. *Journal of Official Statistics*, *3*, 389–410.

Sodeur, W. (1997). Interne Kriterien zur Beurteilung von Wahrscheinlichkeitsauswahlen. *ZA-Information*, *41*, 58-82.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading: Addison-Wesley.

Winship, C., & Morgan, S. (1999). The Estimation of Causal Effects from Obervational Data. *Annual Review of Sociology*, *25*, 659–707.

Wolf, C., & Hoffmeyer-Zlotnik, J. H. (2003). How to Measure Sex/Gender and Age. In J. H. Hoffmeyer-Zlotnik & C. Wolf (Eds.), *Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables* (pp. 259–266). New York: Kluwer Academic.