Assessing ESS sample quality by using external and internal criteria[1]

GESIS, Mannheim/Germany, October 2014

Achim Koch (GESIS – Leibniz Institute for the Social Sciences)
Verena Halbherr (GESIS – Leibniz Institute for the Social Sciences)
Ineke A. L. Stoop (SCP – The Netherlands Institute for Social Research/SCP)
Joost W. S. Kappelhof (SCP – The Netherlands Institute for Social Research/SCP)

Contents

---

1. Introduction

The European Social Survey (ESS) is an academically driven cross-national survey that has been conducted every two years across Europe since 2002. The ESS aims to produce high-quality data on social structure, attitudes, values and behaviour patterns in Europe. Much emphasis is placed on the standardisation of survey methods and procedures across countries and over time. Each country implementing the ESS has to follow detailed requirements that are laid down in the "Specifications for participating countries". These standards cover the whole survey life cycle. They refer to sampling, questionnaire translation, data collection and data preparation and delivery. As regards sampling, for instance, the ESS requires that only strict probability samples should be used; quota sampling and substitution are not allowed. Each country is required to achieve an effective sample size of 1,500 completed interviews, taking into account potential design effects due to the clustering of the sample and/or the variation in inclusion probabilities. Regarding data collection, the ESS specifies – among other things – that face-to-face interviewing is the only mode allowed. Targets are set for the response rate (70%) and the noncontact rate (3% maximum). The fieldwork period is specified (September until December of the survey year), the personal briefing of interviewers is required, and a detailed call schedule for the interviewers is laid down.

The purpose of setting these standards is to achieve accurate and comparable survey data. An important aspect of survey quality refers to the quality of the realised samples in terms of representation of the target population. The sample in each ESS country should reflect the target population of the ESS adequately, which means that bias due to nonresponse should be minimised.[2] Up till now, quality control activities in the ESS were mainly directed at compliance with the prescribed data collection procedures. In each survey round, for instance, it is checked whether or not a country achieved the target response rate, whether the interviewers were adequately briefed, whether the call schedule was adhered to, etc. The (implicit) assumption is that a country that follows the ESS survey procedures and achieves a high response rate will also achieve a sample of good quality. In the present paper we take a first step to assessing empirically how "good" the samples actually are. We analyse the socio-demographic sample composition in ESS countries by comparing ESS variable distributions with more accurate benchmark data. We start with comparing ESS data with external benchmark data from the European Union Labour Force Survey (LFS). These analyses are restricted to ESS 5 which was fielded in 2010. Subsequently, we use an internal benchmark, derived from the samples in the ESS countries itself. Here we include data from the first five survey rounds in ESS. With our analyses we pursue two aims. First, we want to provide an indication of the degree of over-/underrepresentation of certain demographic subgroups in ESS samples. Second, we analyse the correlates of over-/underrepresentation, focusing on two basic parameters, namely the response rate achieved and the type of sampling frame used.

2. Assessing socio-demographic sample composition with external benchmark data

The comparison of survey results with independent and more accurate information about the population parameters is a well-known method to analyse sample quality and the degree of nonresponse bias (Groves 2006). For this approach no information at the individual level is required. There needs to be another survey or administrative record system containing

---

[2] As a matter of course, the ESS also requests that sampling error should not exceed a certain level (a minimum effective sample size of 1,500 completed interviews is to be achieved), and over-/undercoverage of certain groups should be avoided in all countries. The focus of the present paper is on the potential negative effect of nonresponse on sample quality.

estimates of variables similar to those being produced from the survey. Then, the survey estimates can be benchmarked with information from the other data source, the so-called gold standard. The difference between estimates from the survey and the other data source can be used as an indicator of bias.

The advantage of this method is that it is in theory relatively simple to implement. Usually, the method is not so expensive since it does not require collecting additional data. The drawback is that normally only a limited set of variables can be compared. In order to draw valid conclusions about nonresponse bias, the benchmark data have to be quite accurate, i.e. they should not be severely affected by, for instance, measurement or nonresponse errors. In addition, the measurement of the relevant variables should match closely between the two data sources (equivalent measurements). Both data sources have to refer to the same target population and also the reference period should be as close as possible. Even if these conditions hold, one has still to be aware that differences between the survey data and the benchmark data might arise from both nonresponse error and sampling error.

It goes without saying that no benchmark information is available for the ESS key survey variables – this is the reason, why the ESS exists! Comparisons have to be restricted to several socio-demographic variables. The results, however, are important beyond these variables. Socio-demographic characteristics are intrinsically important since they are – potentially – related to many attitudes and behaviours. This is the reason, why some of these variables are often used to construct post-stratification weights. From 2014 onwards, post-stratification weights are also provided for the ESS (European Social Survey 2014).

For a cross-national survey like the ESS the most promising candidate to act as a valid standard for such a comparison is the Labour Force Survey (LFS).

2.1. The European Union Labour Force Survey

The European Union Labour Force Survey (LFS) is a large sample survey among residents in private households in Europe.[3] It is an important source for European statistics about the situation and trends in the EU labour market. The LFS is currently fielded in 33 European countries. These include the 28 Member States of the European Union, three EFTA countries (Iceland, which at the same time is an EU candidate country, Norway and Switzerland), and two EU candidate countries, i.e. the Former Yugoslav Republic of Macedonia and Turkey. The sampling units are dwellings, households or individuals depending on the country-specific sampling frames. Each quarter some 1.8 million interviews are conducted throughout the participating countries to obtain statistical information for some 100 variables. The sampling rates in the various countries vary between 0.2% and 3.3%.

The EU LFS is conducted by the National Statistical Institutes across Europe and is centrally processed by Eurostat (for details of national implementation, see Eurostat (2012a)). The National Statistical Institutes of the Member States are responsible for designing national questionnaires, drawing the sample, conducting interviews and forwarding results to the Commission (Eurostat) in accordance with a common coding scheme. As a rule the data are collected by interviewing the sampled individuals directly, but proxy interviews (through a responsible person in the household) are also possible. Moreover, part of the data can also be

---

[3] http://epp.eurostat.ec.europa.eu/portal/page/portal/employment_unemployment_lfs/introduction

supplied by equivalent information from alternative sources, such as e.g. administrative registers (mainly social insurance records and population registers).

Table 1: Basic information on LFS 2010 (23 countries which took also part in ESS 2010)*

	LFS compulsory	Response rate LFS (%)	Response rate ESS (%)	LFS proxy rate among 15-74 year old respondents (%)
BE	Yes	72	53	17
BG	No	82	76	36
CH	Yes	83	53	2
CY	Yes	97	72	32
CZ	No	81	70	47
DE	Yes	98	30	26
DK	No	52	55	4
EE	No	61	56	29
ES	Yes	84	69	53
FI	No	78	59	4
FR	Yes	83	47	31
GR	Yes	86	66	42
HU	No	84	61	44
IE	No	81	60	48
LT	No	84	39	35
NL	No	79	60	49
NO	Yes	85	59	15
PL	No	73	70	37
PT	Yes	84	67	49
SE	No	76	52	3
SI	No	80	64	57
SK	Yes	93	75	35
UK	No	59	56	34

* Source: Eurostat 2012a, 2012b

As already mentioned, we restrict the comparison with the LFS on the fifth survey round of ESS which was fielded in 2010.[4] At the time the analyses were performed, the ESS 2010 provided data for 27 countries.[5] Among the 27 countries, 24 countries also participated in the LFS 2010. Only Israel, Russia and Ukraine were not part of the LFS and had to be excluded from our analyses. In addition, Croatia had to be excluded since this country was not included in the LFS 2010 data (edition 2012) which we used. Table 1 shows response rates both for the ESS and the LFS[6] for the 23 countries included in both data sets. Among the 23 countries,

---

[4] The ESS specifications require fieldwork to take place in each country between September and December of the survey year. Unfortunately, not all countries managed to adhere to this schedule in ESS 5. Among the 23 countries included in our analyses, in nine countries all (or the majority of) interviews were completed only in 2011 (BE, BG, CY, CZ, ES, GR, IE, LT, PT). In footnote 16 we briefly touch upon the question whether this compromises the analyses.

[5] Data from Austria were not yet available.

[6] In the LFS most countries calculate response rates on the household level, only in a minority of countries response rates are calculated on the person level (which is the standard in ESS).

participation in the LFS was mandatory in 10 countries.[7] The LFS response rates vary between 52% (DK) and 98% (DE). Accordingly, the LFS, too, has a severe nonresponse problem in some countries. The consequences for the nonresponse error of the LFS cannot be assessed here. However, two points can be made in favour of still using LFS as a benchmark for the ESS. First, in each country except Denmark, the LFS response rate is (often considerably) higher than the ESS response rate. The difference in response rates between the two surveys varies between 3 and 68 percentage points. On average, the response rate in the LFS is 20 percentage points higher than in the ESS (80% vs. 60%). Second, it has to be taken into account that the LFS data itself are weighted in (nearly) all countries to adhere to the population distribution of sex, age and region (Eurostat 2012b). Accordingly, at least the distributions of these variables should validly reflect the countries' population.

In addition to the question of nonresponse error, the measurement error properties of the LFS might also be queried. At least in one respect it seems debatable, whether the LFS is in fact a more accurate 'gold standard' which should be used as a benchmark for the ESS. This is the issue of proxy interviewing. Whereas in the ESS proxy interviewing is forbidden by the survey specifications, the LFS allows proxy interviewing. As can be seen from Table 1, many countries make use of proxy interviewing to a larger extent. The proportion of proxy interviews varies between 2% (CH) and 57% (SI). On average across all 23 countries, around one third of the interviews are proxy interviews (32%). We cannot empirically assess what this means for the quality of the LFS data. However, it seems justifiable to assume that the basic demographic information which we use for our analyses will not noticeably be impaired by this problem (Köhne-Finster & Lingnau 2009; Zühlke 2008).

2.2. Data and variables

For our analyses we use ESS round 5 data (edition 03)[8] and anonymised EU LFS 2010 data (edition 2012)[9]. Comparisons between ESS and LFS were possible for variables which were either measured in an identical way or, if this was not the case, where the measurements could be recoded to a common standard. [10] This was true for six variables: gender, age, marital status, work status, nationality and household size. Table 2 shows the variables and the respective categories which we distinguish, plus their source variables in ESS and LFS.[11]

---

[7] In all but one of these countries the LFS response rate was 80 percent or higher. The only exception was Belgium with a response rate of 72%.

[8] European Social Survey Round 5 Data (2010). Data file edition 03. Norwegian Social Science Data Services, Norway - Data Archive and distributor of ESS data. The Core Scientific Team (CST) and the producers bear no responsibility for the uses of the ESS data, or for interpretations or inferences based on these uses.

[9] All results and conclusions are those of the authors and not those of Eurostat, the European Commission or any of the national authorities whose data have been used.

[10] The focus here is on comparability between the general standards set in the LFS and the ESS. However, one has to note that the comparability of measurements between countries within the LFS also might be an issue. The LFS sets various standards to ensure that the national surveys provide data that are compatible with the EU definitions. However, the leeway for differences in national questions is certainly larger than in the ESS. Accordingly, the quality report for LFS 2010 states: "As a general conclusion it emerges that, in spite of the progress regarding the adherence to the EU regulations, principles and guidelines (i.e. the explanatory notes), the national questionnaires still largely differ even in the collection of key variables such as WSTATOR (Labour status in the reference week)" (Eurostat 2012b, p. 29).

[11] Originally, we intended to include also the information on the highest level of education successfully completed. Both ESS and LFS use the ISCED classification of educational attainment. However, whereas the ESS documents in detail how the national degrees were mapped into the international standard (see ESS5DataDoc-ReportAppendix_A1_3.0.pdf on the ESS website), the respective information is not available for the LFS.

The ESS interviews persons aged 15 years and over resident within private households, regardless of their nationality, citizenship or language. In order to achieve comparable target populations, we excluded persons under 15 years in the LFS. In addition, persons living in an institutional household (which were surveyed in a few LFS countries) were excluded. In Norway and Sweden, LFS data are only available for persons aged 74 years or younger. For these two countries, we also restricted the ESS analyses to persons 74 years or younger.

Table 2: Variables of the ESS – LFS comparison

Variable	Categories	ESS source variable	LFS source variable
Gender	• Male • Female	gndr	sex
Age	• 15-24 years • 25-34 years • 35-44 years • 45-54 years • 55-64 years • 65-74 years • 75 years and older	agea (recoded)	age (recoded)
Marital status	• Not married • Married (incl. registered partnership)	maritalb (3-6 = 0) (1-2 = 1)	marstat (0-1 = 0) (2 = 1)
Work status	• Not in paid work • In paid work (for at least one hour)	pdwrk + crpdwrk	wstator (3-5 = 0) (1-2 = 1)
Nationality	• National of country • No national of country	ctzcntr	national (non-nationals recoded in one category)
Household size	Respondent lives in household comprising • 1-person • 2-persons • 3-persons • 4-persons • 5 and more persons	hhmmb (recoded)	hhnbpers (recoded)

ESS data were weighted with the design weight (dweight). This weight corrects for differences in selection probabilities between sampling units in a country. The design weights are computed as normed inverse of the inclusion probabilities. LFS data were weighted with the standard weight variable COEFF, as recommended by Eurostat. This weight too corrects for differences in selection probabilities. In addition, it includes a post-stratification adjustment to adapt the LFS data to known population characteristics. In (nearly) all LFS countries data on sex, age and region were used for the adjustment. Several countries used additional information for weighting, like information on unemployment or nationality (see Eurostat 2012a, b).

2.3. Description of ESS-LFS differences

In order to allow for an overview of which groups are over- or underrepresented in the ESS we provide line charts for each variable. Each chart displays at a time the proportions for one category of a variable both for ESS and LFS. Countries are in ascending order according to their value in the LFS. In order to facilitate comparisons between variables, each chart is scaled to show a range of 30 percentage points (however, often on a different 'level').The figures show at a glance the absolute differences between ESS and LFS distributions.[12] It can easily be checked whether the structure of over-/underrepresentation is similar across countries, and whether the size of differences differs between variables. For dichotomous variables (gender, marital status, work status, national), only the proportions for one category are shown. For age and household size one chart is provided for each category of the variables.

The differences between the ESS and LFS shown in the charts can result from sampling error and/or nonresponse error (if we can assume that differences in measurement do not contaminate the comparison). If we wanted to determine whether or not a difference is still within the limits of sampling error, this would require estimating standard errors which take into account the complex sampling design in many countries, both for the ESS and the LFS. Unfortunately, this is neither possible for the ESS nor for the LFS, since the relevant information is not publicly available. In order to provide a rough indication of relevant differences, we will use a rule of thumb and highlight all differences larger than 3 percentage points.[13]

---

[12] This approach does not take into account whether the difference between ESS and LFS refers to a category where, for instance, the LFS reports a proportion of 50% or of 10% only. The alternative would have been to calculate relative differences where the size of the percentage which is used as a standard of comparison is taken into account. An absolute difference of 5 percentage points, for instance, would indicate a relative difference of 10% when the proportion in LFS is 50%, and a relative difference of 50% when the LFS proportion is only 10%. The drawback of using relative differences is that for rather skewed distributions very large relative differences will be calculated. For a dichotomous variable with a 90/10 LFS distribution, for instance, one would receive very different estimates, depending on which category is chosen for the comparison. If the ESS result is 95/5, one might either report a 5.6% relative overrepresentation if the first category is chosen for the comparison, or a 50% relative underrepresentation if the second category is used.

[13] The following thoughts led to the decision to use this criterion. First: We do not take into account sampling errors in the LFS. Due to the rather large sample size sampling errors tend to be low in LFS (see the examples in Eurostat 2012b, p. 15). Additionally, due to post-stratification weighting, the LFS distributions for sex and age reflect population characteristics. Second: As regards the ESS, the analyses of 96 variables carried out by the ESS sampling panel yielded an average effective sample size of 1.400 cases for the ESS 5 countries. All countries (except four) achieved an average effective sample size of 1.000 cases; the lowest effective sample size was 750 cases. When we use an average effective sample size of 1.000 as a basis, any difference from a population value larger than 3.1 percentage points will be significant if the population proportion is around 50% (assuming a significance level of p<.05). If the population proportion is around 10 percent, a difference larger than 1.9 percentage points will be a significant difference (with p<.05).

Gender

Figure 1 shows the percentage of females in the ESS and LFS samples. In around half of the countries the proportion of females in ESS and LFS data are quite similar. In seven countries the proportion of females in the ESS is more than 3 percentage points higher than in the LFS. The largest discrepancy is 11.6 percentage points (Lithuania: 66.0% in ESS vs. 54.4% in LFS). Hence, in a number of countries females tend to be overrepresented in ESS.

Figure 1: Females in ESS and LFS (in %)
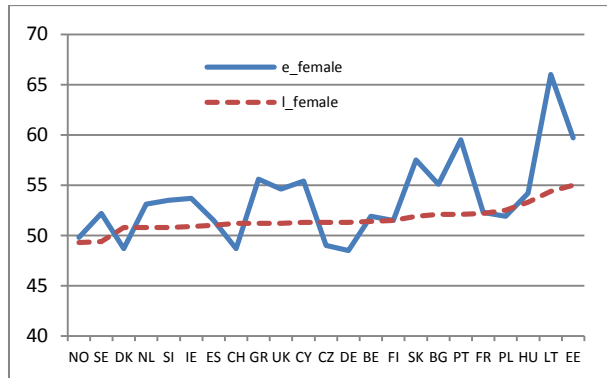
Age

As regards the variable age, we distinguish seven 10-year age groups in the comparison between ESS and LFS (see Figure 2). For the youngest age group, ESS and LFS results do not differ much in around half of the countries. In five countries the proportion of 15-24 years old persons is more than 3 percentage points lower in the ESS than in the LFS (the maximum difference is -4.8 percentage points), in one country it is more than 3 percentage points higher. The percent of 25-34 years old persons is lower in all countries in the ESS.[14] Seven countries show a difference larger than the critical value of 3 percentage points (the maximum difference is -6.8 percentage points in Slovakia). Regarding the 35-44 years old persons, the results of ESS are close to the ones of the LFS. Only in Portugal, the proportion of 35-44 years old persons is considerably lower in the ESS than in the LFS. In the next three age groups (45-54 years, 55-64 years, 65-74 years) the direction of the differences changes. In each of these age groups we find considerable differences in 3-4 countries. Each time, the relevant proportion is larger in the ESS than in the LFS. The largest difference pertains to the 55-64 years old persons in Bulgaria: their proportion is 21.1% in the ESS and 14.6% in the LFS. The picture becomes more varied again for the oldest age group. In two countries, persons 75 years or older are underrepresented by more than 3 percentage points, in another two countries they are overrepresented.

---

[14] Note that it is very unlikely that such a result arises by chance only. The likelihood that such a result could have arisen by chance only is $1:2^{23}$

Figure 2: 10-year age groups in ESS and LFS
(in %)*
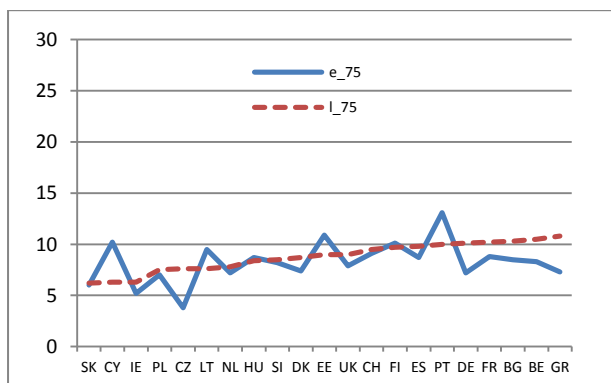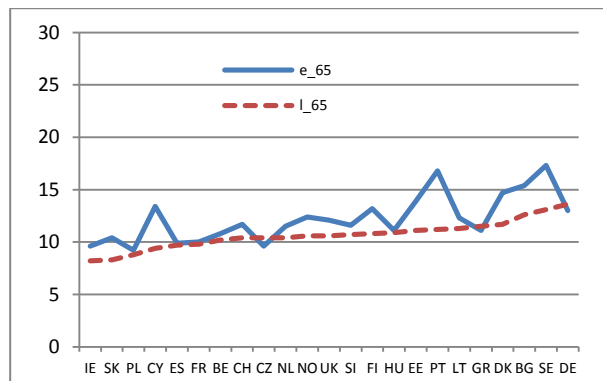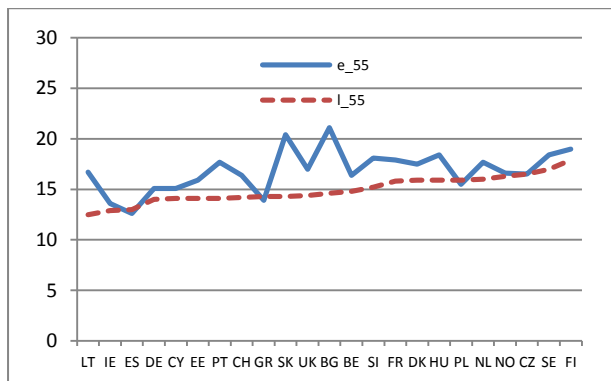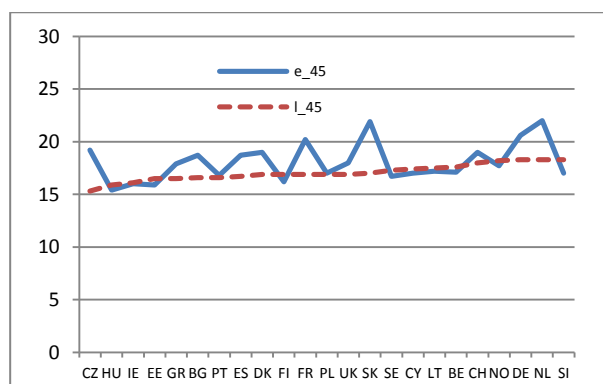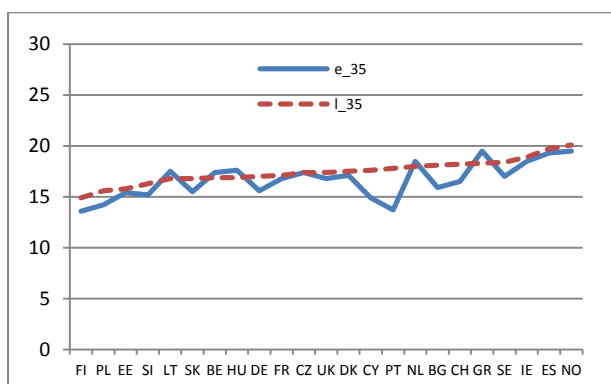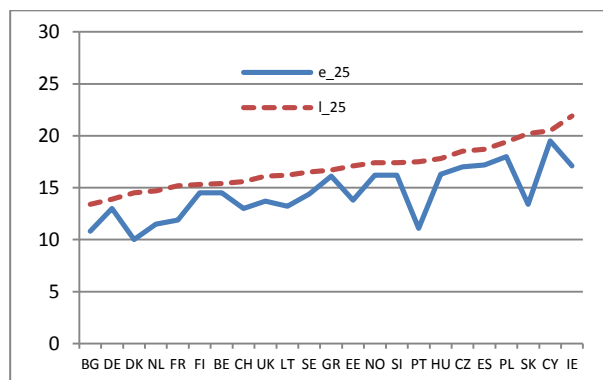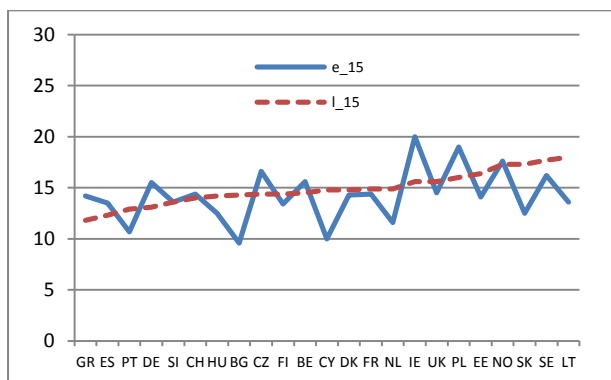
* e_15 = ESS: 15-24 years old,
  l_15 = LFS: 15-24 years old,
  ….
  ….
  e_75 = ESS: 75 years or older,
  l_75 = LFS: 75 years or older

Taken together, we find considerable differences in the age distribution for some ESS countries. By and large, these differences seem to follow a common pattern: There is a tendency to underrepresent the younger age groups, and a tendency to overrepresent the older age groups. Only for the oldest age group (75 years or older), the results vary to a greater extent between the countries. Figure 3 shows the complete age distribution for the three countries which exhibit the largest difference between the ESS and the LFS. These were Slovakia, Portugal and Bulgaria.

Figure 3: Age distribution (10-year age groups) in ESS and LFS for SK, PT and BG (in %)
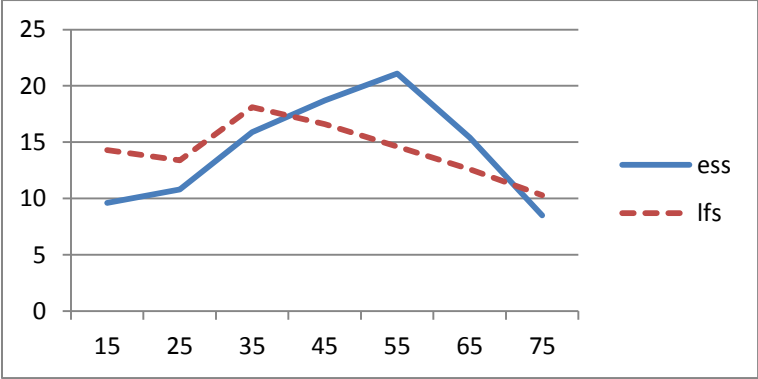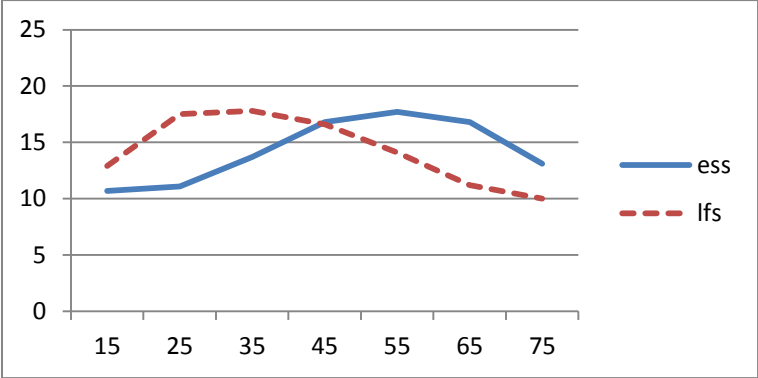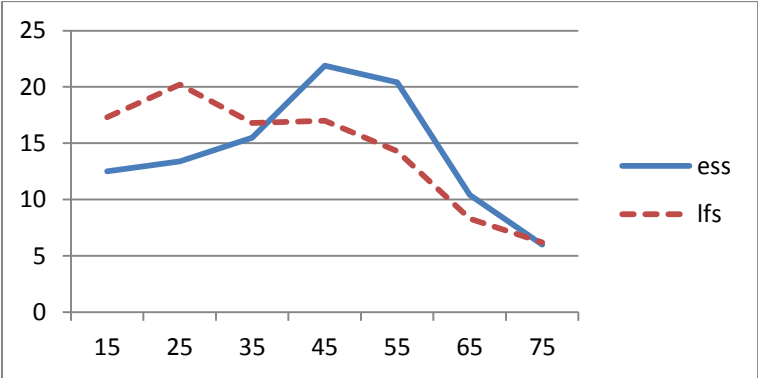
Marital status[15]

As regards the marital status, we can only distinguish between married and non-married persons. Both in the ESS and the LFS, the category 'married persons' includes persons living in a registered partnership. In the majority of ESS countries, the respective results do not differ much from the LFS data (see Figure 4). In eight countries, however, the proportion of married people differs more than 3 percentage points between the ESS and LFS samples. The majority of these countries (six countries) overrepresent married persons in the ESS, two countries underrepresent married persons. The largest discrepancy is observed for the Netherlands. Here, the proportion of married persons is 62.8% in the ESS and 53.5% in the LFS.[16]

Figure 4: Married persons in ESS and LFS (in %)*
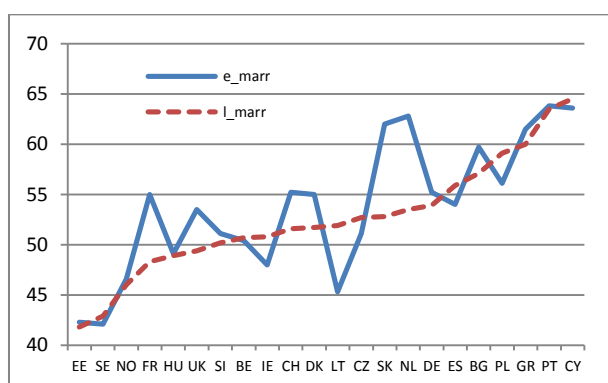
*incl. persons living in a legally registered partnership

---

[15] The comparison of marital status was not possible for Finland, due to a filter error in the ESS data.

[16] In the Netherlands, the proportion of persons living in a registered partnership is highest among all ESS countries (5.8%). It would be interesting to see whether in the LFS this proportion is also that high. Unfortunately, it is not possible to identify these persons in the LFS data. According to the LFS rules they should be classified as "married" in the data set. In the "Explanatory notes" for the LFS it is laid down: "Some countries have a legal framework for registering partnerships (in most countries these are same-sex partnerships and they have a legal status parallel to married couples). Such information has also to be treated in a harmonised way so they should be treated as married ..." (Eurostat p. 17).

Work status

As regards the work status, again a dichotomous distinction can be defined as a common standard for ESS and LFS. We can distinguish persons who have been working for pay for at least one hour in the past seven days from persons who did not. In most countries, the distribution of this variable in the ESS resembles the LFS distribution (see Figure 5). In eight countries, however, the percentage of people in paid work differs substantially (more than 3 percentage points) between the ESS and the LFS. In seven countries, the proportion of people in paid work is lower in the ESS than in the LFS, in one country it is higher. The largest difference is observed in Portugal, where the proportion of people in paid work is 55.2% according to the LFS and 40.0% according to the ESS.[17] Hence, there is a tendency to underrepresent people in paid work in some countries of the ESS.

Figure 5: Persons in paid work in ESS and LFS (in %)

---
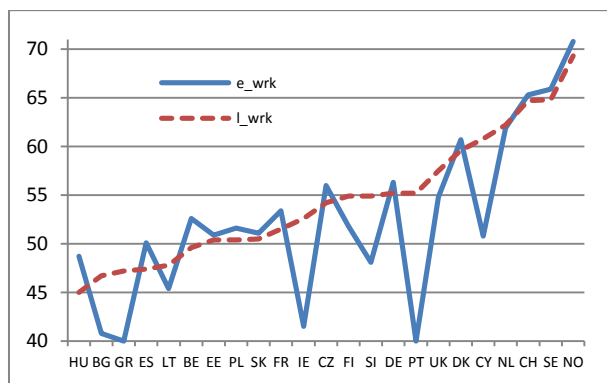
[17] Among the eight countries with large differences, five countries fielded the ESS 5 not in 2010 as requested, but only in 2011 (some countries only started in 2011; others started in 2010, however, the majority of interviews were completed in 2011). As the percentage of people in paid work might have changed between 2010 and 2011, we re-run the analyses with 2011 LFS data. The results remained very similar, albeit the differences became somewhat smaller: Bulgaria: -5.8% (instead of -5.9%); Cyprus: -7.9% (instead of -10.0%); Greece -3.8% (instead of -7.2%); Ireland: - 9.9% (instead of -11.1%); Portugal: -13.5% (instead of -15.1%). In this context a specific point concerning the comparability of measurements between the LFS and the ESS might be mentioned. The LFS is designed as a continuous survey with interviews spread uniformly over all weeks of the survey year. Thus, the LFS provides a measure for the average respective status in the survey year. In the ESS, however, no effort is made to cover all weeks of the survey year during fieldwork. According to the ESS Specifications, countries should field the survey between September and December of the survey year. The comparability between ESS and LFS in this respect might be further improved by using quarterly results from the LFS, which are available in principle, too. The potential drawback, however, are the smaller sample sizes of the quarterly data.

Nationality

As regards the nationality of respondents, we can distinguish between nationals of a country on the one hand (including persons holding a dual citizenship), and non-nationals of that country on the other hand. A special feature of this variable is its rather skewed distribution in some countries. In eight out of the 23 countries the proportion of non-nationals is less than 2 percent (according to the LFS data, see Figure 6). In these countries, as a matter of course, the ESS and LFS rates of non-nationals do not differ much. Among the countries with a higher share of non-nationals we see quite often an underrepresentation of non-nationals in the ESS. In five countries, the percentage of non-nationals in the ESS is more than 3 percentage points below the result of the LFS. The largest difference is observed in Cyprus. According to the LFS, the percent of non-nationals is 18.3%. The percent of non-nationals in the ESS, however, is only 4.0%.[18]

Figure 6: Non-nationals in ESS and LFS (in %)
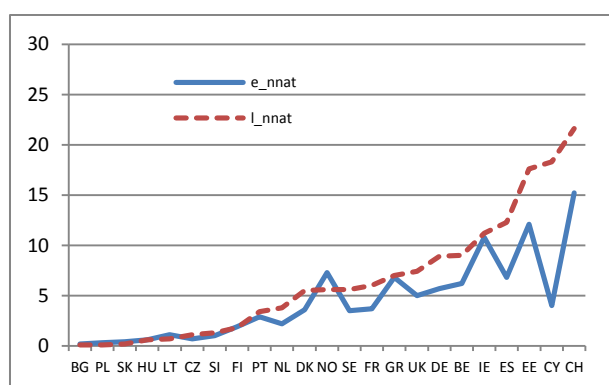
Household size[19]

In order to compare ESS and LFS data on household size, we recoded all persons living in households with 5 or more persons in one category. Figure 7 shows, that there is no clear-cut trend towards over- or underrepresentation among the five different categories of household size. Among all categories, we see both countries for which the proportions in ESS are below and countries for which the proportions in ESS are above the respective level in the LFS. Most likely, one could say that in some countries there is a tendency to overrepresent persons living in 2-person households in the ESS (four countries with a difference larger than 3 percentage points), and a tendency to underrepresent persons living in 4-person households (again, four countries with a difference larger than 3 percentage points). This pattern is a common feature of the three countries with the largest differences in the distribution of household sizes (see Figure 8).

---

[18] Among other things, language problems might have contributed to this result. According to the ESS Documentation report (European Social Survey 2010), more than 6% of the target persons in Cyprus could not be interviewed due to language problems. This was the highest rate among all ESS 5 countries. It seems reasonable, that these persons predominantly were non-nationals.

[19] LFS data on household size is not available for CH, DK, FI, NO and SE. These countries were excluded from the present comparison.

Figure 7: Persons living in households of
different sizes in ESS and LFS
(in %)*
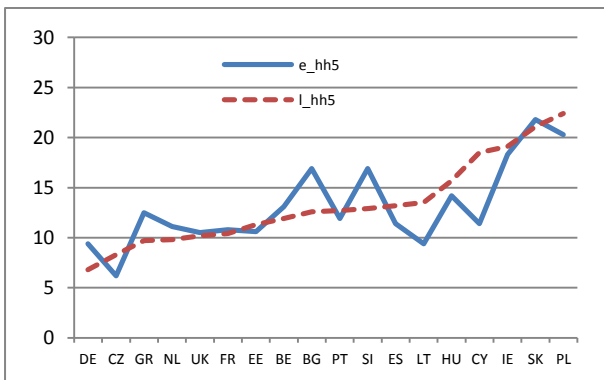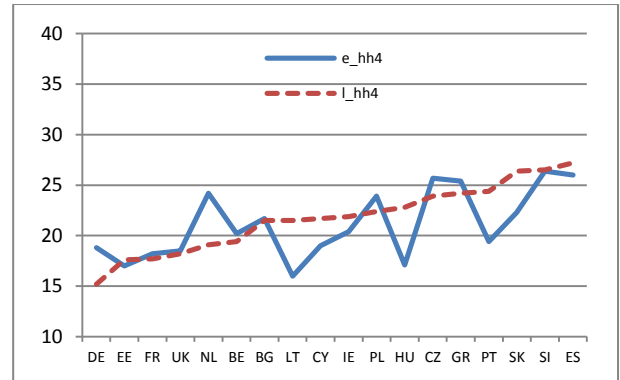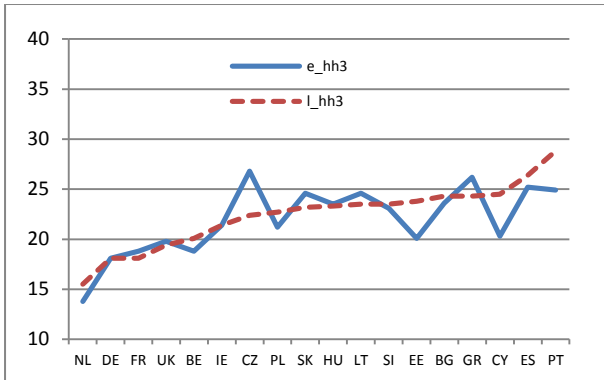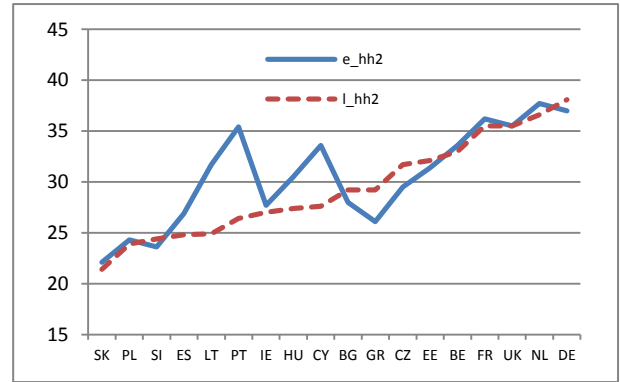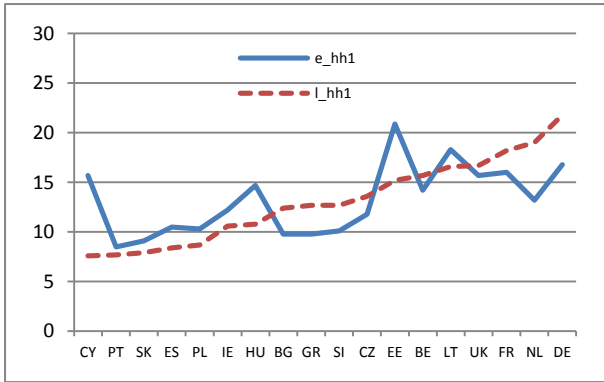
* e_hh1 = ESS: persons living in 1-person-hh,
  l_hh1 = LFS: persons living in 1-person-hh,
  ….
  ….
  e_hh5 = ESS: persons living in 5 or more-person-hh,
  l_hh5 = LFS: persons living in 5 or more-person-hh

Figure 8: Distribution of household size in ESS and LFS for CY, PT and LT (in %)
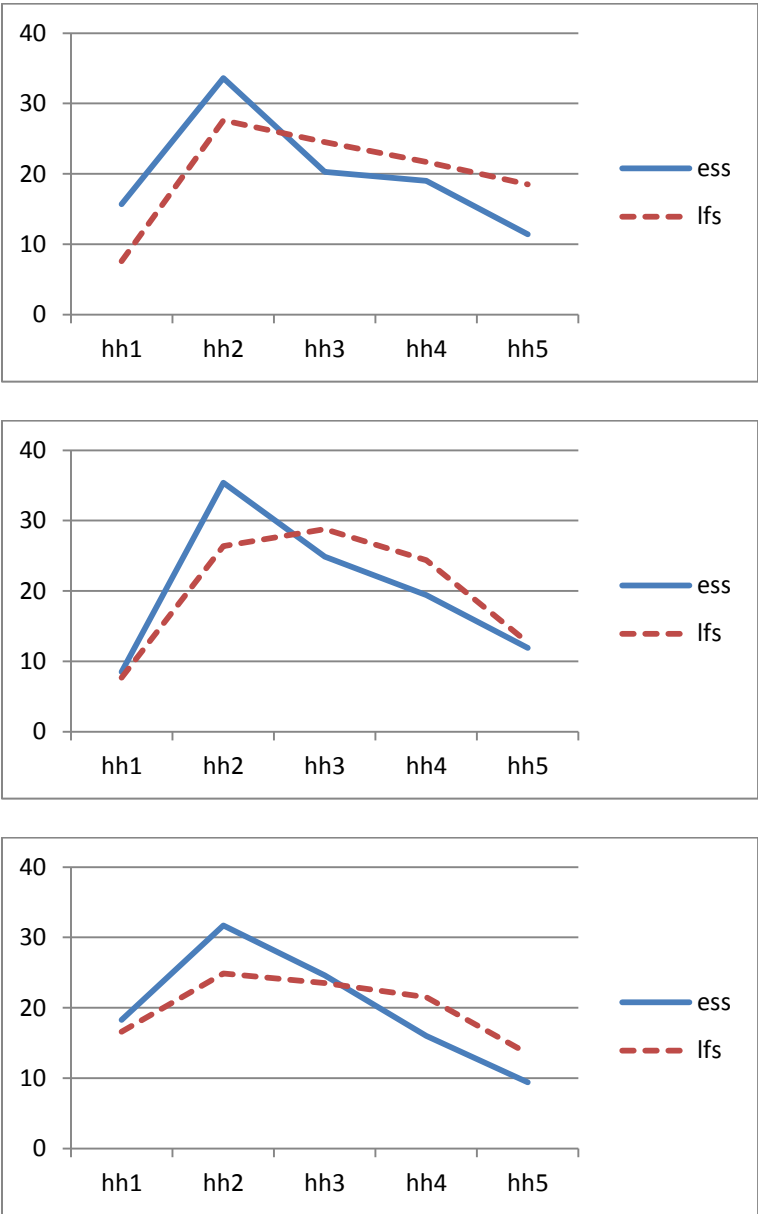
To conclude, we can note that the comparison of socio-demographic variable distributions between ESS and LFS revealed only small differences for a large number of countries. At each variable, however, a few countries exhibited larger discrepancies. For these countries, we might briefly summarise the direction of the differences observed as follows: There is a tendency to overrepresent females, persons in older age groups, married persons, persons not in paid work, persons holding the citizenship of that country and persons living in 2-person households. In the next section, we will no longer take care of the direction of the differences between the ESS and the LFS; our focus will be on the size of the differences and their correlates.

2.4. A summary measure of ESS-LFS differences

In order to arrive at a summary measure for the consistency of ESS and LFS variable distributions we calculate the index of dissimilarity (Duncan & Duncan 1955):

$$D = \tfrac{1}{2} \sum_{i}^{n} |ESSi - LFSi|,$$

with n = number of categories,
ESSi = percentage in category i of ESS,
LFSi = percentage in category i of LFS.

The index of dissimilarity (D) is a measure widely used in research on segregation. The range of the index is between 0 and 100. In the present context, a value of 0 indicates that there is no dissimilarity between the LFS and the ESS in the relative shares of respondents across the categories of a variable. A value of 100 indicates that the two distributions are completely dissimilar (consider, e.g., a dichotomous variable, where the first category comprises 100% in LFS and 0% in ESS, and the second category comprises 0% in LFS and 100% in ESS). The index of dissimilarity measures the percentage of respondents that would need to move between the categories of a variable to produce exactly the same distribution for the two surveys.

Table 3: Index of dissimilarity (D) between ESS and LFS distributions

| Country	Gender	Age	Marital status	Work status	Natio-nality	Household size	mean
BE	0.5	3.7	0.3	3.0	2.8	2.7	2.2
BG	3.0	11.3	2.6	5.9	0.1	4.6	4.6
CH	2.4	4.7	3.6	0.6	6.4		3.5
CY	4.1	8.9	0.9	10.0	14.3	14.0	8.7
CZ	2.3	6.1	1.6	1.8	0.4	6.1	3.1
DE	2.8	5.9	1.3	1.1	3.3	6.2	3.4
DK	2.1	6.7	3.3	1.2	1.9		3.0
EE	4.7	6.5	0.5	0.5	5.5	5.7	3.9
ES	0.5	3.4	1.8	2.6	5.5	4.2	3.0
FI	0.0	4.0		3.1	0.1		1.8
FR	0.1	5.6	6.7	1.9	2.3	2.2	3.1
GR	4.4	4.9	1.5	7.2	0.2	5.9	4.0
HU	0.8	3.6	0.2	3.7	0.0	7.2	2.6
IE	2.8	6.5	2.9	11.1	0.4	2.3	4.3
LT	11.6	7.7	6.6	2.3	0.4	9.6	6.4
NL	2.4	7.1	9.3	0.3	1.6	7.5	4.7
NO	0.4	2.4	0.7	1.5	1.6		1.3
PL	0.6	3.5	3.1	1.3	0.2	3.5	2.0
PT	7.4	12.6	0.3	15.1	0.6	9.7	7.6
SE	2.7	5.7	0.8	1.1	2.1		2.5
SI	2.7	3.9	0.9	6.7	0.3	3.9	3.1
SK	5.6	13.1	9.2	0.6	0.3	4.1	5.5
UK	3.3	5.2	4.2	2.7	2.4	1.1	3.2
mean	2.9	6.2	2.8	3.7	2.3	5.6	3.8

The mean value of D across all variables and countries is 3.8. This means that – on average – less than 4% of respondents in ESS would have to change categories in order to achieve the same distribution as in the LFS. D is highest for the variables age (mean 6.2) and household size (mean 5.6). To some extent, this is the consequence of these two variables having a larger number of categories than the remaining variables. If we, for instance, recode age into three (15-34 years, 35-54 years, 55 years and older) rather than in seven age groups, the average value of D is reduced from 6.2 to 4.8 (what is still higher than the average value of D of the four dichotomous variables). The smallest D refers to the variable nationality. This is – at least in part – related to the skewed distribution of this variable. As already mentioned, in several countries the percentage of non-nationals is rather small, and differences between the two surveys are therefore not very likely to arise in these countries.[20]

The size of D varies both between countries and between variables. At each of the six variables there is at least one country with a D larger than 10. The one exception is the variable marital status, where the maximum dissimilarity is around 9 in two countries. The largest value observed pertains to the variable work status in Portugal with a D of 15.1. On the other hand, at each variable there are usually a few countries with a rather low value of D. Among four of the six variables (gender, marital status, work status, nationality) there are several countries with a D smaller than 1.

The mean value of D across the six variables varies between a low of 1.3 in Norway and a high of 8.7 in Cyprus (see Figure 9).[21] Countries with a rather high average D typically show values well above-average in several variables (see Table 3).

Figure 9: Index of dissimilarity: mean value across six variables

---
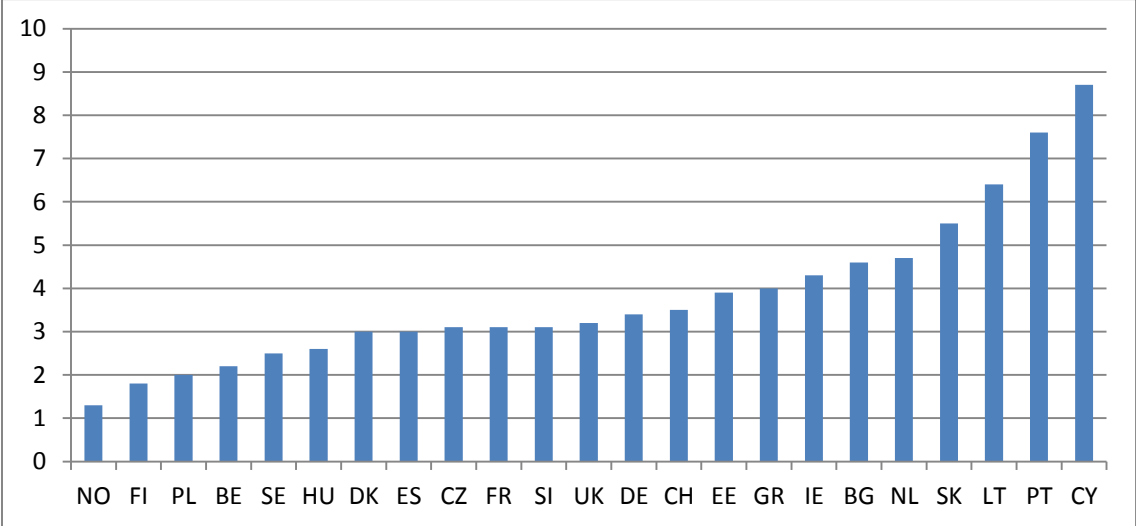
[20] This is reflected in the strong correlation between the index of dissimilarity on the one hand, and the percent of non-nationals according to the LFS on the other hand: Pearson's r between these two variables is .81.
[21] In five countries (CH, DK, FI, NO, SE) the index of dissimilarity for household size is not available. In FI the index of dissimilarity is also missing for marital status. In these countries, the average value of D is based on the remaining five or four variables, respectively.

2.5. Correlates of ESS-LFS differences

The previous section revealed that the size of the differences between the ESS and the LFS – measured by the mean index of dissimilarity across six variables – varies between countries. In the remainder we will briefly analyse whether the average D is related to two basic survey parameters: the response rate achieved and the sample design used.

For several decades the response rate achieved in a survey has been used as a proxy for the quality of the realised sample and the degree of nonresponse bias (Kreuter 2013). With respect to the present analysis, we therefore should expect that higher response rates come along with smaller ESS-LFS differences. Empirically this is not the case. Figure 10 shows a weak positive relationship between the (average) D per country and the response rate (Pearson's r = .20).

Figure 10: Index of dissimilarity by response rate ESS 5 (in %)
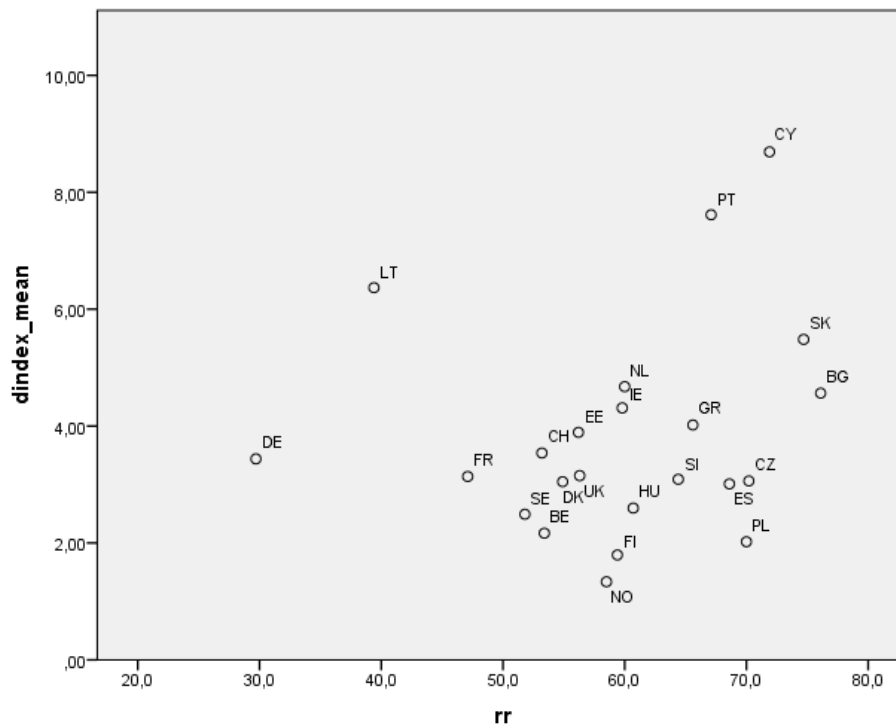
Even more pronounced differences can be found when we turn to the type of sample used (see Figure 11). ESS prefers countries to use a sample of named individuals from a register (European Social Survey 2013). If such a sample is used, the sample is drawn without any involvement of the interviewers. Where a sampling frame of individuals is not available, countries may use a sampling frame of households or addresses. This can take various forms, differing among other things, in the degree of interviewer involvement required. One possibility is that the households/addresses come from a list, like a registry of telephone numbers or the customer directory of an electricity provider. If such a list is not available, the survey organisation and their interviewers have to enumerate the households, usually before fieldwork starts. In all household/address designs, the interviewers have to select a respondent in the household, since the ESS interviews only one person per household. Among the 23 countries included in our analyses, 12 countries used a sample of individuals and 11 countries used a sample of households or addresses in ESS 5. Figure 11 shows the average index of dissimilarity, separately for countries with a sample of individuals and for countries with a household/address sample.

Figure 11: Index of dissimilarity by type of sample in ESS 5
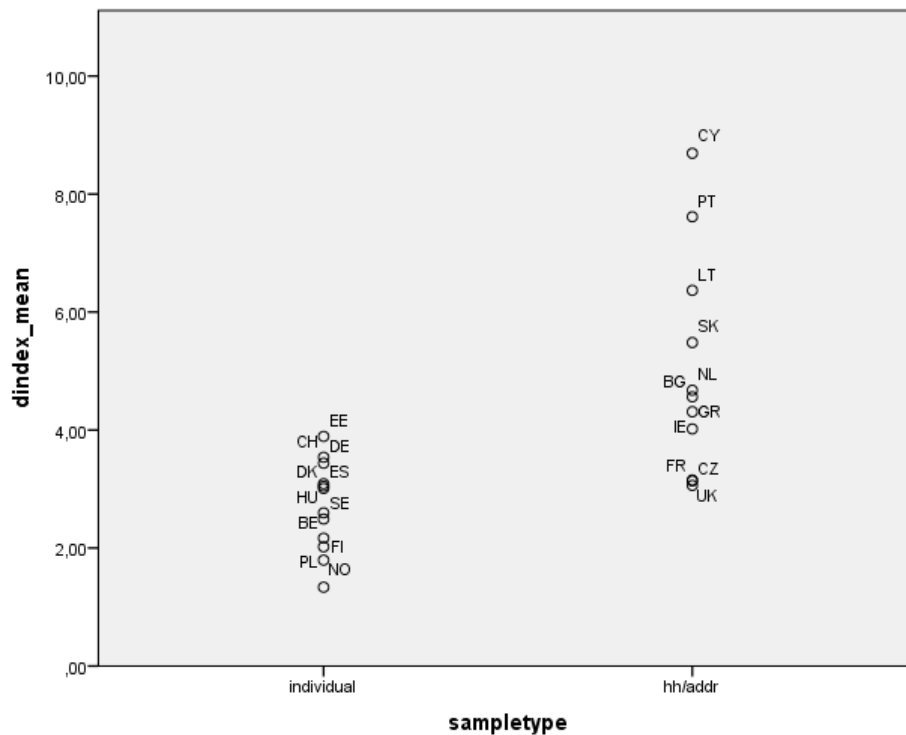
In the group of countries using a sample of households/addresses, the mean size of D across all countries is nearly twice as high (mean = 5.01) as in the group of countries using a sample of individuals (mean = 2.70). In addition, the results vary more strongly among countries with a household/address sample than among countries using a sample of individuals. Countries with a sample of individuals exhibit a D between 1.3 and 3.9, whereas countries with a sample of households or addresses show a D between 3.1 and 8.7. Phrased differently, we might say that all countries with an individual sample show only small or medium differences to the LFS results. Among the countries with a household or address sample, a few countries deviate to a large degree (for instance CY, PT, LT), whereas other countries (for instance CZ, FR, UK) reached a similar (low) level of differences as the countries with an individual sample. Thus, countries with a household/address sample do not necessarily perform less well than countries with an individual sample with respect to the criteria used here: It seems as if the way *how* a sample of households/addresses is implemented plays a decisive role for the size of D.

An interesting pattern reveals itself when the relationship between the response rate and the index of dissimilarity is analysed separately for countries with samples of individuals on the one hand, and countries using an address/household sample on the other hand (see Figure 12). For countries with samples of individuals a weak negative relationship between the response rate and D can be observed. As conventional wisdom suspects, the sample composition of countries with a higher response rate corresponds closer with the LFS data than the sample composition of countries with a lower response rate. The opposite is true for countries with a sample of addresses/households: Higher response rates come along with larger differences to the LFS.[22]

---

[22] However, one has to be aware that the observed relationships rest on a small number of countries only. They can be strongly influenced by including or excluding individual countries (see, for example, the indicated changes in the size of Pearson's r in Figure 12).

Figure 12: Index of dissimilarity by response rate in ESS 5 (in %), separately for countries with samples of individuals and countries with samples of households/addresses

Sample of individuals: Pearson's r = -.33 (if DE excluded: r = -.15)
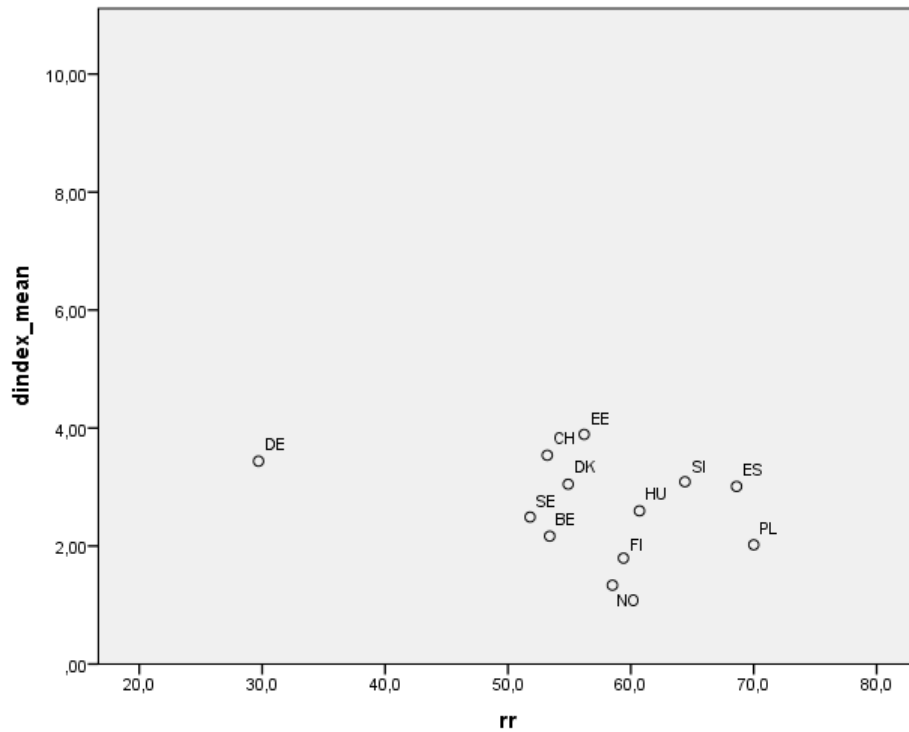
Sample of households/addresses: Pearson's r = +.19 (if LT excluded: r = +.48)
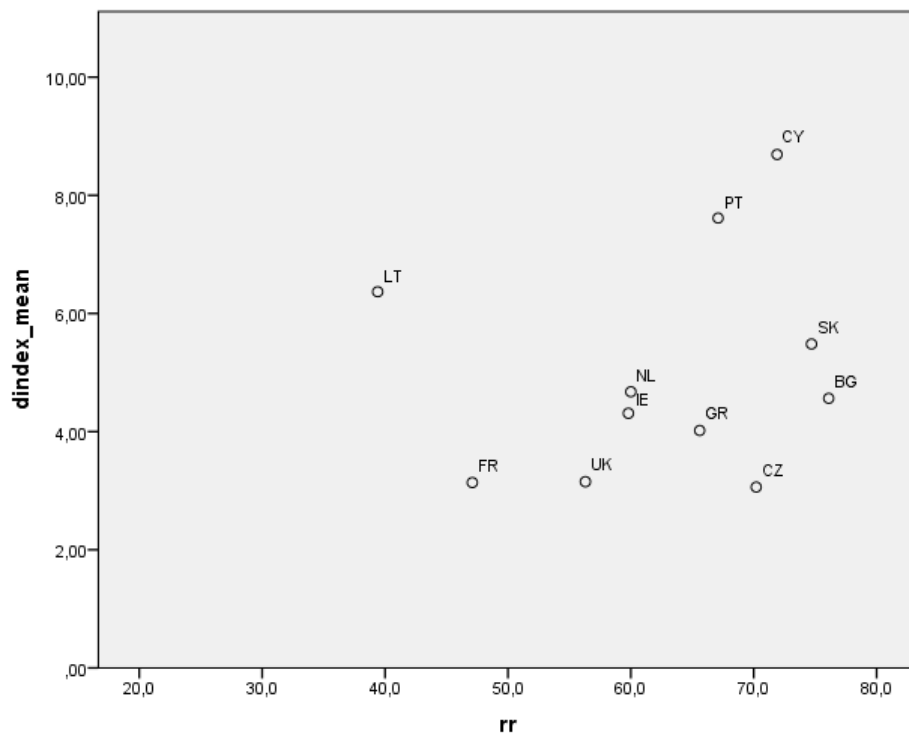
3. Assessing demographic sample composition with internal benchmark data

3.1. Respondents' gender distribution among the subsample of gender heterogeneous couples as an internal quality criterion

The idea of using an internal criterion for evaluating sample quality was brought up by Sodeur (1997). Kohler (2007) developed the idea further and applied it to various international survey programmes. As Kohler (2007: 55) notes: "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." The subgroup which is used in the present analyses consists of households with a gender heterogeneous couple, where one of the partners is interviewed. Among this subgroup, the chance of being selected as a respondent is exactly the same for the male and female partner of the couple. As a consequence, we would expect a proportion of 50% females (and 50% males) among the respondents from these couples. If the respondents' gender distribution deviates from the 50/50 female/male split beyond some acceptable random fluctuation, we might interpret this deviation as an indication of nonresponse bias.

This internal criterion for sample quality has the advantage that the same fixed benchmark (50%) can be applied to all countries. Furthermore, problems as regards the comparability of measurement do not compromise the analyses. On the downside, we have to restrict the analyses to one survey estimate (gender) and to a subgroup of the sample (respondents from gender heterogeneous couples).[23]

Figure 13: % female respondents by country, ESS 1 – 5*
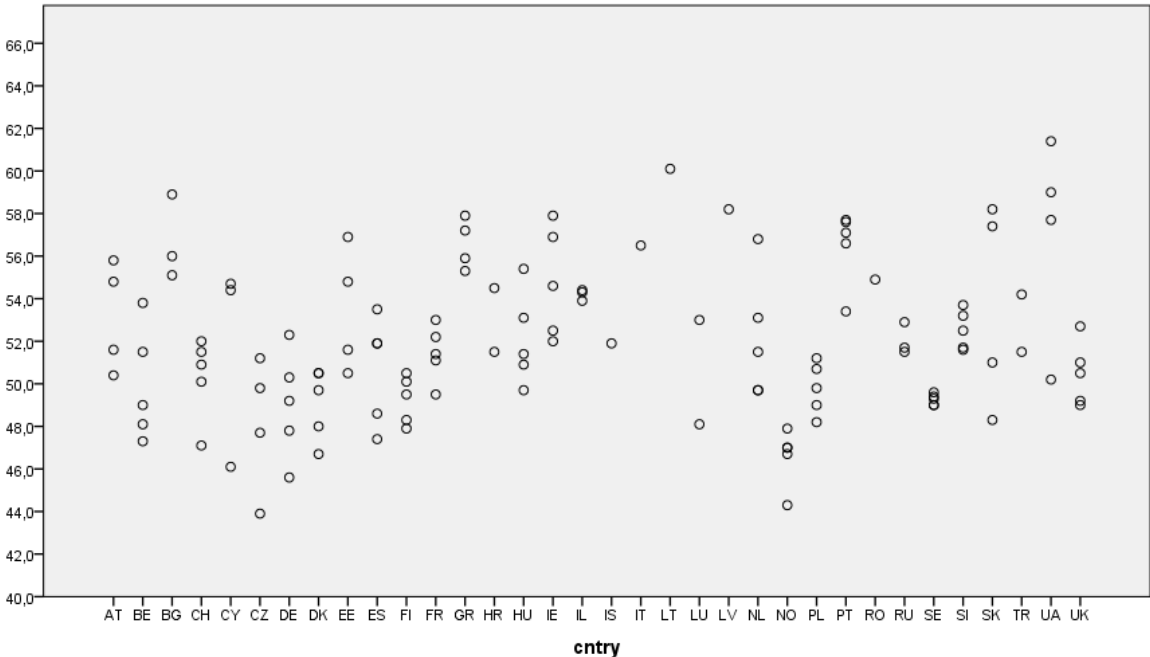
*Subsample of respondents living with a partner of the opposite sex (15 years and older) in the same household

---

[23] On average this subgroup comprises 58% of the cases of the original sample of a country. Among all countries in ESS 1-5, the smallest proportion of the subgroup is 43%, the largest proportion is 68%.

Figure 13 shows the percentage of female respondents among the defined subgroup[24] for all countries in ESS 1 thru 5. 34 different countries participated in at least one of the first five rounds of ESS. 16 countries participated in each of the five rounds, 6 countries participated in 4 rounds, 4 countries participated in 3 rounds, 3 countries participated in 2 rounds, and 5 countries participated in only one round. Accordingly, we have 127 cases in total, each representing a country participating in a specific round. The average percent of females among the 127 cases is 51.9%. This means that there is a slight tendency to overrepresent women when interviewing one person from a gender heterogeneous couple. However, the cases vary a lot in the proportion of women interviewed. The minimum percentage observed is 43.9% (Czech Republic in ESS 1), the maximum 61.4% (Ukraine in ESS 5). Around one third of the values are below 50%, and around two thirds are above 50%.

More formally, bias can be defined as a deviation of the respondents' gender distribution from the 50/50 female/male split divided by the standard error (Kohler 2007):

Bias = (% female − 50) / sqrt [(50 * 50) / n],
with n = number of respondents from gender heterogeneous couples

This measure resembles the Z statistic used for inference procedures on the population mean. In analogy to the practice for the Z-test, values larger than |1.96| can be considered as a significant deviation. Among the 127 valid cases, 49 cases are larger than |1.96| and thus indicate a significant bias. A large majority of these cases (41 cases) indicates a significant overrepresentation of females, only a minority (8 cases) indicates an underrepresentation of females. These results for the internal criterion are therefore in line with the respective results we received for the comparison with the LFS: There, too, the main pattern was an overrepresentation of women in several countries (see section 2.3.).

3.2. Correlates of bias according to the internal criterion

Figure 14 shows for ESS 1-5 how the internal bias criterion is related to the response rate. In this and in the following figures, the absolute value of the bias measure is displayed. This means that we no longer distinguish whether there is an over- or an underrepresentation of women. Figure 14 exhibits that a higher level of bias is observed in countries with above average response rates. There is a weak positive correlation between the response rate and the degree of bias (Pearson's r = .24).

The type of sample also has an impact on the level of absolute bias (see Figure 15). Among the 56 cases which used a sample of individuals the absolute bias is 1.23 on average. In 46 out of the 56 cases the value of the bias indicator is below |1.96| and thus within the limits of random fluctuation. In contrast, among the 71 cases with a household/address sample the average value of the absolute bias is 2.68 – more than twice as high as in the group of cases using a sample of individuals. In 32 out of the 71 cases the bias indicator is below the critical value of |1.96|, whereas in 39 cases it is above this value. Figure 15 also reveals that the size of bias varies more strongly among the cases with a household/address sample than among those with a sample of individuals.

---

[24] Both the respondent and his/her partner have to belong to the target population of the ESS, i.e. they have to be 15 years or older.

Figure 14: Absolute value of internal bias criterion by response rate (in %), ESS 1-5

Figure 15: Absolute value of internal bias criterion by type of sample, ESS 1-5

A closer inspection of the relationship between the type of sample used, the response rate achieved and the internal bias indicator reveals a similar pattern as the one we observed for the external benchmark data (see section 2.5). Among the cases with a sample of individuals there is a weak negative correlation (Pearson's r = -.18) between the size of the bias and the response rate (Figure 16). In contrast, those with a sample of households/addresses reveal a moderate positive correlation (Pearson's r = +.45) between the response rate and the bias indicator. Among these cases, contrary to expectations, the bias is larger when a country reports an above average response rate.

Figure 16: Absolute value of internal bias criterion by response rate (in %) in ESS 1-5,
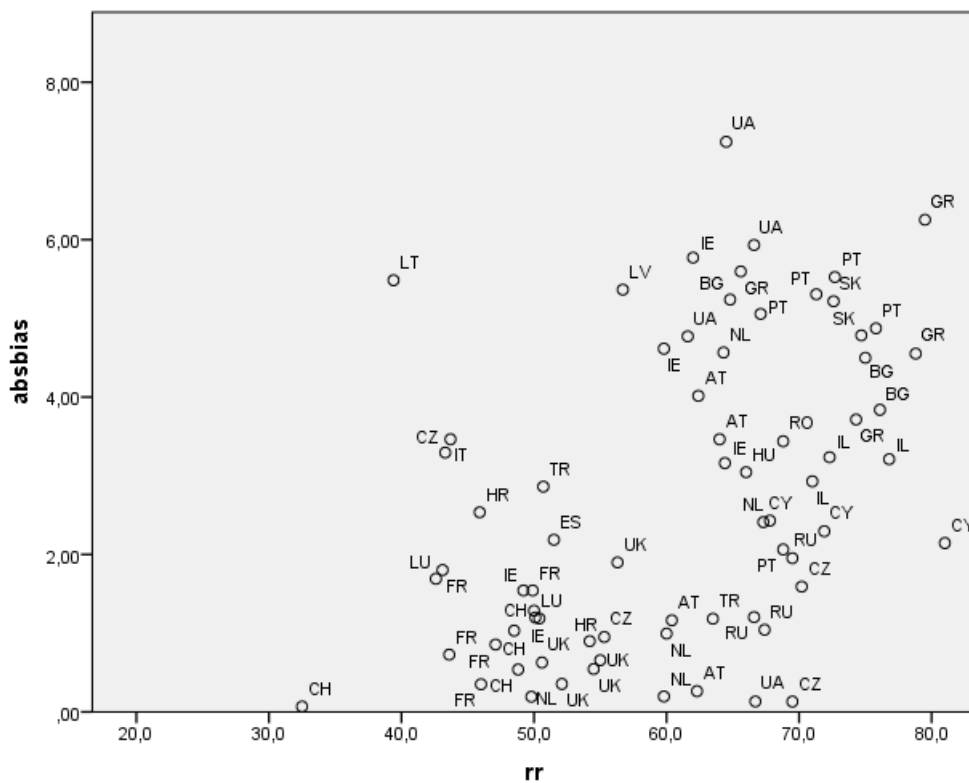separately for countries with samples of individuals and countries with samples of
households/addresses

Sample of individuals: Pearson's r = -.18
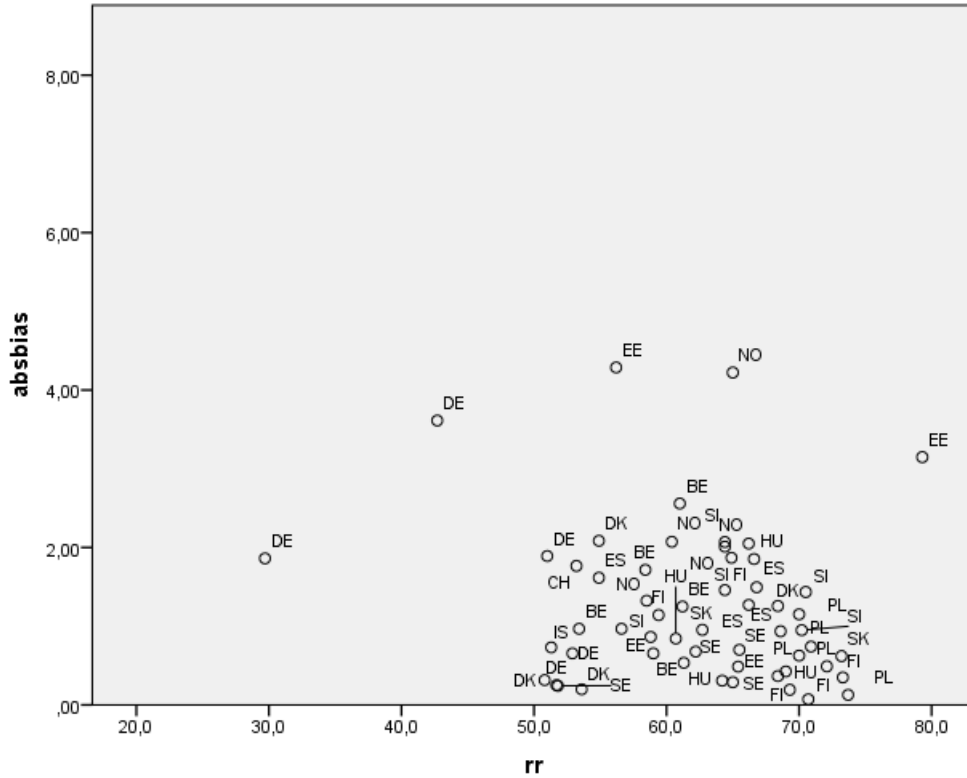
Sample of households/addresses: Pearson's r = +.45

4. Summary and conclusions

This paper used external and internal benchmark data to analyse the socio-demographic sample composition in ESS. The comparison with data from the LFS revealed only small differences for six socio-demographic variables (gender, age, marital status, work status, nationality and household size) for the majority of ESS 5 countries. At the same time, large differences were observed in a number of countries, sometimes only with respect to one variable, sometimes with respect to several of the variables examined. These analyses were complemented by investigating an internally derived sample quality criterion for the ESS rounds 1 to 5, namely deviations from respondents' gender distribution from a 50/50 female/male split in households with a gender heterogeneous couple. Using this criterion, significant differences were observed in a substantial number of cases (in 39% of all countries participating in ESS 1 to 5).

With respect to both the external and the internal quality criteria a weak positive association with the response rate could be observed. This means that, contrary to the usual expectation, larger deviations were observed in countries with high response rates than in countries with low response rates. The type of sample used, however, seemed to be more important in determining the size of the deviation than the response rate. According to both approaches, countries with a sample of individuals exhibited on average a demographic sample composition which was less biased than the one of countries with a household/address sample. Furthermore, there was less variation in our indicators of sample quality among countries with an individual sample. These countries achieved acceptable sample quality more or less across the board, whereas sample quality differed a lot among the countries which used a sample of households/addresses. An interesting pattern emerged when the relationship between response rates and sample quality was analysed separately for countries with sample of individuals on the one hand, and countries with samples of households/addresses on the other hand. Whereas for countries with a sample of individuals sample quality was positively related to the response rate achieved, the direction of this correlation was just the other way round for countries with a sample of households/addresses: Among this group, countries with high response rates – surprisingly – revealed lower sample quality than countries with low response rates.

These results give rise to several questions. First of all, one might ask for the reasons of the differences in sample quality observed. The most obvious explanation will be that countries differ as regards the differential response propensities of socio-demographic subgroups. If, for instance, a certain group is particularly difficult to contact in a country, and/or the efforts to contact this group are below average in that country, then an underrepresentation of this subgroup will occur. Corresponding detailed analyses of the fieldwork processes were not part of the present task. Each ESS country should check its own results and try to decide whether they make sense, given the available insights into the societal conditions and the fieldwork efforts exerted in that country. As the case may be, further country-specific analyses might be considered.

Referring exclusively to potential differences in subgroups' response propensities, however, will probably not provide a sufficient explanation of the full pattern of results observed. This holds in particular with regard to the differences in sample quality we found with respect to the type of sample used and the response rate achieved. From our point of view, a comprehensive explanation also has to take the interviewer and his/her behaviour into account. As mentioned earlier, interviewers play an important role in sample selection when a sample of households/addresses is used. If interviewers do not follow the rules of random

sampling properly – for instance during the listing of households before fieldwork, or during the selection of target persons within households – this might contribute to the patterns of results we observed. If interviewers, for instance, tend to substitute a reluctant male target person by his cooperative wife when selecting a respondent within a household, this will lead to an overrepresentation of women in the final sample (see Kohler 2007, Sodeur 1997). In addition, such a misconduct of interviewers will not only increase sample bias, but will also undermine the reliability of the response rate calculated. Undocumented substitution of reluctant target persons, as this behaviour might be termed, will lead to inflated response rates. This might explain the – counterintuitive – positive correlation between sample bias and response rates which we observed among countries using a sample of households/addresses. In particular countries with high response rates and large biases should check whether their system of quality control back-checks is sufficient to prevent and/or detect such interviewer misbehaviour.

Even though the abovementioned processes presumably are the main factors contributing to the results observed, it may still be true that in some instances other factors also play a role. As regards the comparison between ESS and LFS, in particular the comparability of measurement instruments and classification rules deserves attention. In contrast to the ESS, countries in the LFS have a greater leeway in deciding on how to measure concepts nationally (the dominant philosophy in LFS is 'ex-ante output harmonisation', Körner 2012). Countries have to make sure that the national measurement instruments and their results are bridged properly into the international LFS standard. Since this process is not always well documented, it is difficult or nearly impossible to decide from outside, whether or not all countries actually used equivalent measurement instruments for all variables (for an example on relevant differences see Mikucka & Valentova 2013). Therefore, we cannot totally exclude the possibility that measurement inconsistencies also influenced the results here and there.[25]

Finding out about the causes of our results is but one part of the challenge. The other is to discuss the implications these results may have for the design of future ESS survey rounds. If we assume, that the observed differences can reliably be interpreted as indications of nonresponse bias, an obvious question is whether dedicated efforts to balance response rates for socio-demographic subgroups are desirable at least in some ESS countries. A study of Peytcheva & Groves (2009) might cast doubt as to whether aiming for balanced response rates in demographic subgroups is important. They found that bias in demographic variables is not predictive of the difference between respondents and nonrespondents in substantive variables of the same survey. If we nevertheless came to the conclusion that balancing response rates is a good thing, further questions arise: How should balanced response rates be achieved? Every effort to balance response rates needs additional data on the gross sample to allow for targeted fieldwork efforts. In many surveys – in particular in surveys based on samples of households/addresses – such information is not (routinely) available. And even if the relevant information is available, it still has to be decided how to proceed. For instance: What should be done if balancing is desirable for several variables? Which measures are most appropriate to raise response rates in specific subgroups, etc.? In a nutshell: There are many obstacles to overcome in order to achieve a balanced sample in the end. To complicate things further, one might finally ask whether a balanced sample might not just as well be achieved by simply applying post-stratification (PS) weights. Using PS-weights would be much easier to implement, and just relying on them would definitely be a less expensive way to deal with

---

[25] The authors would welcome it very much if countries would get in touch with them when they have indications that this issue is relevant.

the issue of nonresponse bias (and sampling error) in socio-demographic variable distributions.

Whatever the answers to these questions may be, from our point of view it seems worthwhile to repeat the present analyses with data from another ESS round. Replicating the analyses with new data will show whether the results vary from round to round or whether stable patterns can be observed for individual countries and variables.[26] A broader data base will allow better informed reasoning about the causes and knowledgeable decisions about possible consequences to be drawn. In a replication, it could also be analysed whether the recent introduction of PS-weights in the ESS helps to decrease deviations.

If the present approach is deemed to be useful in general, it might be considered to make these analyses of sample quality a routine part of the quality control procedures which are performed by the CST after each ESS round.

---

[26] In a replication study, the analyses might be slightly refined by selecting the LFS data to be used country-by-country according to the actual year in which the ESS was fielded. This procedure would ensure that in the ESS countries with a delayed fieldwork period those LFS data are used which are closest to the actual fieldwork period.

References:

Duncan, O. D., Duncan, B. (1955): A Methodological Analysis of Segregation Indexes. American Sociological Review, 20, 210-217

European Social Survey (2010): ESS-5 2010 Documentation Report. Edition 3.0. Bergen, European Social Survey Data Archive, Norwegian Social Science Data Services

European Social Survey (2013). Round 7 Specification for Participating Countries. London: Centre for Comparative Social Surveys, City University London

European Social Survey (2014). Documentation of ESS Post-Stratification Weights. www.europeansocialsurvey.org, 25th April 2014

Eurostat (2012a): Labour force survey in the EU, candidate and EFTA countries. Main characteristics of national surveys – 2010. 2012 edition. eurostat Methodologies and Working papers

Eurostat (2012b): Quality report of the European Union Labour Force Survey – 2010. 2012 edition. eurostat Methodologies and Working papers

Groves, R. M. (2006): Nonresponse Rates and Nonresponse Bias in Household Surveys. In: Public Opinion Quarterly, 70, 646-675

Köhne-Finster, S. & Lingnau, A. (2009): Untersuchung der Datenqualität erwerbsstatistischer Angaben im Mikrozensus. Ergebnisse des Projekts "Nachbefragung im Mikrozensus/LFS". Wirtschaft und Statistik, 12/2008, 1067-1088

Körner, Th. (2012): Measuring the Labour Status in Official Statistics: The Labour Force Concept of the International Labour Organisation and its Implementation in the Labour Force Survey. P. 123-138 in: J. H. P. Hoffmeyer-Zlotnik & U. Warner (eds.): Demographic Standards for Surveys and Polls in Germany and Poland: National and European Dimension. GESIS - Leibniz-Institut für Sozialwissenschaften Köln, 2012 (GESIS-Schriftenreihe 10)

Kohler, U. (2007): Surveys from inside: An assessment of unit nonresponse bias with internal criteria. Survey Research Methods, 1, 55-67

Kreuter, F. (2013): Facing the Nonresponse Challenge. The ANNALS of the American Academy of Political and Social Science, 645, 23-35

Mikucka, M. & Valentova, M. (2013): Employed or inactive? Cross-national differences in coding parental leave beneficiaries in European Labour Force Survey data. Survey Research Methods, 7, 169-179

Peytcheva, E. & Groves, R. M. (2009): Using Variation in Response Rates of Demographic Subgroups as Evidence of Nonresponse Bias in Survey Estimates. Journal of Official Statistics, 25, 193-201

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

Zühlke, S. (2008): Auswirkungen von Proxy-Interviews auf die Datenqualität des Mikrozensus. P. 3-10 in: Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen (ed.): Statistische Analysen und Studien NRW, Band 53