Survey Research Methods (2019)
Vol. 13, No. 2, pp. 167-180
doi:10.18148/srm/2019.v13i2.7383

ISSN 1864-3361
http://www.surveymethods.org

Within-household selection of target-respondents impairs demographic representativeness of probabilistic samples: evidence from seven rounds of the European Social Survey

Piotr Jabkowski
Adam Mickiewicz University
Poznan, Poland

Piotr Cichocki
Adam Mickiewicz University
Poznan

This paper examines the demographic representativeness of different types of probabilistic samples based on the results of seven rounds of the European Social Survey. Focusing on the distinction between personal-register and non-personal-register samples, it demonstrates that the latter exhibit systematically larger gender- and age-biases. Expanding upon a "gold standard" evaluation based on external criteria derived from Eurostat population statistics, an internal criteria analysis leads to the conclusion that the inferior quality of surveys involving interviewer-driven within-household selection of target respondents results from interviewer discretion. Such interference brings about the selection of individuals with higher levels of readiness and availability, which superficially improves survey outcome rates while yielding samples of actually inferior quality. The internal-criteria approach provides a straightforward and undemanding way of monitoring representativeness of samples, and proves especially handy when it comes to large cross-country projects, as it requires no data external to the survey results, and allows for comparing surveys regardless of possible differences in sampling frames, sampling design and fieldwork execution procedures.

Keywords: probabilistic samples; survey quality; personal-register samples; non-personal-register samples; interviewer interference; European Social Survey

1 Introduction

Assessing sample quality in cross-country surveys presents unique challenges, given that participating countries may strongly differ in such terms as sophistication of fieldwork procedures, work-ethic or data-protection regimes (Lynn, Häder, Gabler, & Laaksonen, 2007; Menold, 2014; Smith, 2007; Tomescu-Dubrow & Slomczynski, 2016). On the other hand, longitudinal multi-national measurements provide remarkable opportunities for investigating such survey-methodological issues as the impact of sample-type on survey quality (Gaziano, 2005; Kohler, 2007; Menold, 2014). While it is unreasonable to expect multiple parallel runs of single-country surveys implementing different sample-types, this is exactly a collateral advantage of large cross-country projects. It comes especially handy for comparing the impact of sample types on demographic representativeness of surveys. In turn, such investigations are crucial in the pursuit of cross-country equivalence, and they also remain invaluable inputs for discussing the inherent merits of different ways of sampling.

In interviewer-administrated surveys (both PAPI and CAPI), probabilistic samples come in four common varieties: (a) personal-register samples (PRS), (b) household-based samples (HHS), (c) address-based samples (ADS) and (d) non-register samples (NRS) (Lynn et al., 2007; Stoop, Billiet, Koch, & Fitzgerald, 2010). This distinction stems from the level of aggregation of target population units, and has a significant impact on fieldwork procedures. While PRS surveys allow for direct random sampling of individuals identified by name, the non-PRS (HHS, ADS and NRS) require performing within-household selection as a component-part of fieldwork execution. Researchers have a limited capacity for effective control over such a within-unit selection of target-persons, as back-checking whether the right respondent was indeed selected is much more challenging than scrutinising PRS fieldwork compliance. PRS surveys typically present researchers with informational advantages over interviewers, given that on top of the target-respondent names and addresses additional register-derived characteristics (e.g., birthdates) are typically known prior to fieldwork execution and may readily be used for checking its quality. Non-PRS surveys lack such inexpensive and robust control-measures, which gives interviewers higher dis-

Contact information: Piotr Jabkowski, Adam Mickiewicz University, Poznan, Faculty of Social Science, Institute of Sociology, Szamarzewskiego 89c, 60-568 Poznan (E-mail: piotr.jabkowski@amu.edu.pl)

cretionary influence over the process of respondent selection. In turn, this seems likely to result in such quality differences between PRS and non-PRS surveys which would not register on the standard survey outcome rates of response, contact, cooperation and refusal.

Especially in cross-country settings, there is a growing demand for going beyond the standard gauges of survey outcome (Schouten, Cobben, & Bethlehem, 2009). Most notably, doubts abound when it comes to the response rate. It remains a common handbook-invoked indicator of survey quality (Biemer & Lyberg, 2003), even though the actual linkage between the response rate and sample quality has been contested (Curtin, Presser, & Singer, 2005; Groves, 2006; Keeter, Miller, Kohut, Groves, & Presser, 2000). However, in spite of forceful criticism, the jury is still out on the response-rate question: while ample evidence exists that decreasing nonresponse rates need not lead to lower nonresponse bias (Fricker & Tourangeau, 2010; Groves & Peytcheva, 2008), a recent piece of persuasive analysis points to a positive association between representativeness and the response rate (Cornesse & Bosnjak, 2018). From a theoretical point of view, Bethlehem, Cobben, and Schouten (2011) also demonstrated that survey estimates are not affected by response probabilities if (a) response behaviour is not correlated with the target variable, or (b) probabilities of response are equal for all population units. Thus, so far as individuals differ in terms of the propensity to respond, higher response rates should lead to lower magnitude of bias.

Regardless of the hotly disputed merits of the response rate, in practice, its continuing prominence stems mainly from the ready calculability irrespective of sampling frames or fieldwork procedures (Stoop, 2005). This is not always true of many of its proposed replacements. For instance, one such promising alternative metric comes in the form of the Representativeness Indicator, which associates the nonresponse bias with a differential propensity to respond (Luiten & Schouten, 2013; Schouten et al., 2009). R-Indicator remains of little use, however, when it comes to evaluating cross-country surveys (especially those employing different sampling frames), because it requires a uniform set of auxiliary variables in order to estimate the propensity to respond (Schouten et al., 2012). Another alternative approach involves conceiving of bias as the difference between unweighted and weighted estimators (Billiet, Vehovar, Beullens, & Matsuo, 2009). While this assures availability, adding weights may actually increase the Total Survey Error (TSE), if they are not correlated with the propensity to respond (Little & Vartivarian, 2005). Thus, for instance, uncorrelated post-stratification weights would not be effective, and usually such correlation cannot be readily attested with respect to surveys conducted within multi-wave cross-country projects. Furthermore, comparisons can also be made between different types of respondents and non-respondents: the bias can be treated as 1) the difference between cooperative and reluctant respondents (Matsuo, Billiet, Loosveldt, Berglund, & Kleven, 2010), or 2) the difference between survey results and follow-up-studies among non-respondents (Peytchev, Baxter, & Carley-Baxter, 2009). Still, even this approach proves deficient when it comes to comparing respondents and non-respondents in cross-country surveys, as countries are likely to exhibit substantial differences in their procedures of refusal conversions and the ways of implementing follow-up studies (Billiet, Philippens, Fitzgerald, & Stoop, 2007).

Assessments of sample quality in cross-country settings should: 1) rely on variables that are readily available for all countries and sample-types, and 2) correlate with survey quality. Our approach to the cumulative dataset of the European Social Survey (ESS) combines 1) an external evaluation, based on a comparison between the sample distributions of age and gender with official population statistics, with 2) an internal evaluation, based on checking a sub-sample distribution against aprioristic ratios known by definition. We argue against using external criteria other than those strictly demographic, and demonstrate that distributions of census-derived variables remain the only viable grounds for external comparison (in spite of some of their own deficiencies). When it comes to internal criteria, we expand upon the ideas of Kohler (2007), Sodeur (1997), and Menold (2014), by using deviations from the expected 50/50 male-to-female ratio of respondents living in heterosexual couples as a basis for evaluating unit nonresponse bias. Those analyses demonstrate that non-PRS surveys yield samples characterised by systematically inferior rates of demographic representation, and lead to the suggestion that this can be explained in terms of the greater capacity for interviewer impact on survey fieldwork execution.

2 Impact of sample type on sample quality

In line with the TSE approach, sample representativeness should be seen as one of the key factors determining overall survey quality (Groves et al., 2011). Imbalances of demographic representation involve significant deviations from population parameters, such as age and gender. These may arise due to a variety of factors involving frame-quality issues as well as challenges inherent in the mode of fieldwork execution. With respect to frame quality, the principal difficulties of conducting non-PRS surveys may result from (a) within-unit coverage errors, e.g., undercoverage due to omissions from household rosters (Martin, 1999; Tourangeau, Shapiro, Kearney, & Ernst, 1997), (b) population coverage errors, e.g., units missing from non-personal frames such as address- or household-registers due to their lower accuracy and topicality (Eckman & English, 2012). While the population coverage error is independent of respondent availability and readiness to participate in sur-

veys, within-household coverage errors should be correlated with those two characteristics. For instance, female respondents are known to be more available at home, due to a lower labour participation rate (Groves & Couper, 2012; Stoop et al., 2010). Nevertheless, at least in the ESS, women seem to be exhibiting higher refusal rates (Menold, 2014). Likewise, high refusal rates combined with ready availability are characteristic of respondents comprising the oldest age-categories of all genders (Goyder, 1987; Stoop et al., 2010), while youngest respondents exhibit both low availability and low readiness to participate (Voogt & Van Kempen, 2002).

Even though all types of probabilistic sampling are influenced by respondent availability and readiness to participate, the resulting absolute value of the bias of demographic representation is likely to be significantly higher in non-PRS than PRS surveys. This imbalance should also be stable over time, i.e., one would expect this contrast to hold in all surveys rounds. Thus, on the basis of the seven waves of ESS data, the following hypothesis is tested on the basis of census-derived external criteria:

H1: Absolute value of the bias of demographic representation is significantly higher in non-PRS than PRS surveys.

Considering the distinctive features of the two sample types, one such prominent contrast is the interviewer discretion within the respondent-selection process. The fact that within-unit selection is susceptible to manipulation makes non-PRS surveys ripe for overrepresentation of respondents characterised by higher availability and/or readiness to cooperate. Given the absence of control-tools based on the individual-name sampling frame, interviewer manipulation of the selection process constitutes a low-risk strategy combining low probability of detection with low accountability. No direct proof of interviewer interference is obtainable when it comes to historical data in the form of archived surveys. Yet, in spite of there being no hope of catching anyone red-handed, an indirect path remains open for uncovering traces of aggregate misdeeds (Menold, Winker, Storfinger, & Kemper, 2013; Simmons, Mercer, Schwarzer, & Kennedy, 2016). Thus, in order to test hypotheses stemming from such circumstantial narrative, we put forward a statistical meta-analysis of ESS surveys which employs an internal-criteria approach for investigating sample-type effects on survey quality. Note as well that meta-analysis has not been used here as a type of research design but an analytical procedure. Its main purpose is to consider the results of individual surveys in order to combine single-study estimators (the so-called "effect sizes ") into one measure of Effect Size denoted as Overall (sometimes alternately: Weighted).

On the basis of internal criteria, i.e., comparing estimators against aprioristic parameters, it is possible to juxtapose the PRS and non-PRS surveys with respect to their unit nonresponse bias (Koch, 2016; Koch, Halbherr, Stoop, & Kappelhof, 2014). This allows probing for such patterns in the aggregate data that are consistent with interviewer influence over the respondent selection process, while having no other obvious explanations. The first among those patterns consist in the systematic differences between PRS and non-PRS surveys when it comes to the unit nonresponse bias (H2), and the second relates to the direction of relationship between refusal rates and the absolute value of unit nonresponse bias (H3).

When it comes to the gender composition of households of heterosexual couples, a positive unit nonresponse bias amounts to an overrepresentation of women relative to their 50% share in such couples (correspondingly, their underrepresentation would indicate a negative bias). We suggest, expanding upon Menold (2014), that if H2 holds then this can be readily explained by the known gender differences in the propensity to respond within the ESS. Those differences remain independent of the sample-type used, however, they lead to a different Overall Effect Size of unit nonresponse bias in PRS than in non-PRS surveys due to their contrastive characteristics in terms of interviewer discretion.

H2: Overall Effect Size of unit nonresponse bias is positive in non-PRS and negative in PRS surveys.

Apart from its direction, when the bias resulting from unit nonresponse is analysed in its absolute form, it also reveals intriguing contrasts between PRS and non-PRS surveys in the context of their relationship with the refusal rates (H3). In non-PRS surveys, one should expect a significant negative correlation between the refusal rate (REF1) (AAPOR, 2016) and the absolute value of unit nonresponse bias, as interviewer discretion would yield effective interviews in some cases where the properly selected target-respondent was not in fact available. Correspondingly, in PRS surveys, refusal rates should be positively associated with the absolute value of unit nonresponse bias, in line with the expectation that a higher fraction of refusals reduces effective sample quality.

H3: The correlation between REF1 and absolute value of unit nonresponse bias is negative in non-PRS surveys, and positive in PRS surveys

3 The ESS as a tool for investigating sample-type effects

European Social Survey (ESS, 2016) constitutes a well-regarded cross-country survey conducted biennially since 2002. ESS standards are stringent, transparent as well as exhaustively documented, and hence, it has a rich tradition of serving as a test-ground for survey-methodology disputes. Notable topics of such ESS-based studies include enhancing response rates and minimizing nonresponse bias (Billiet et al., 2007; Kreuter & Kohler, 2009; Matsuo et al., 2010), examining mixed mode design and mode effects of data collection (Jäckle, Roberts, & Lynn, 2010; Revilla, 2010; Vannieuwenhuyze, Loosveldt, & Molenberghs, 2010), studying interviewer effects (Beullens & Loosveldt, 2016; Loosveldt

& Beullens, 2013), as well as looking for ways of rectifying measurement error and measurement bias (Coromina & Saris, 2009; Saris & Revilla, 2016; Saris, Satorra, & Coenders, 2004).

Multiple angles of investigation have already been employed to study the sample-type impact on representativeness in the ESS. Notably, Lynn et al. (2007) demonstrated that ADS-, HHS- and NRS-designs have a higher sampling error resulting from unequal selection probabilities. Kohler's 2007 contrastive analysis of PRS and non-PRS surveys was expanded upon by Menold (2014), whose analysis further demonstrated that PRS surveys are consistently associated with higher sample quality. Comparing deviations from the true parameter of 50/50 gender ratio in heterosexual couples, she provided firm evidence that PRS surveys differ substantially from the non-PRS, among whose constituent types (NRS, ADS and HHS) no statistically significant diversity could be attested. Her analysis did also demonstrate that back-checking procedures or interviewer remunerations have a negligible impact on selection bias relative to the sample-type effect. In these two respects, our analysis follows Menold's findings, however, we attempt to go beyond the limitations of her approach, i.e., her making use of absolute values of difference between estimators and parameters without taking into consideration (a) the sign of the value of difference pointing to over- or under-representation, and (b) the between-study-variance of estimators.

The ESS proves especially valuable for exploring possible effects of sample types on survey quality as it comprises both PRS and a variety of non-PRS surveys, while striving to maintain uniformly high methodological standards across different countries and rounds (Stoop et al., 2010). Therefore, the ESS use of non-PRS sampling should not be construed as indicative of lower research standards, which are in fact acknowledged to be on top of the current state of the art (Mohler, 2007). For instance, Kohler's 2007 comparison of five major cross-country projects (Eurobarometer, European Quality of Life Survey, ESS, European Value Study, International Social Survey Programme) concluded that the ESS exhibited the most rigid and reliable standards. Furthermore, its approach to sampling also involves a comprehensive consideration of cross-country equivalence of effective sample-sizes with a precise estimation of design effects (Häder & Lynn, 2007). On top of that, published ESS fieldwork documentation is exceptionally exhaustive. This allows for precise consideration of all relevant steps in the fieldwork execution of different sample types.

While it seems safe to assume that observed sample-type effects do not constitute country-specific artefacts of differential methodological effort in the ESS project, yet, some degree of suspicion would perhaps be warranted that those effects may at least to some extent come about through the confounding influence of country characteristics. In particular, the necessity to opt for non-PRS designs due to the unavailability of usable registers could be construed as indicative of underlying country-specific challenges to the research-quality. In order to assure absolute certainty that countries exert no confounding influence, any analysis would require the randomisation of sample-type assignment, which is simply not available. Still, there are a number of good reasons for assuming that country-characteristics do not act as confounding variables within the ESS project. Firstly, in geographical terms, it should be noted that although ESS country-assignment of sample frames does exhibit some vague geographical patterning along the North—South axis, this arrangement abounds with exceptions and does not correlate neatly with any of the standard regional subdivisions of European countries. Thus, PRS is used in all Nordic countries, some countries of Western Europe, e.g. Austria, Belgium, Germany, Switzerland, as well as most Central and East-European countries, e.g. Hungary, Poland and Slovenia. Non-PRS, on the other hand, is frequently implemented in Southern and Eastern Europe as well as in Israel and Turkey. However, it is easy enough to point out notable exceptions, e.g., non-PRS is also implemented in some countries of Western Europe, and conversely, Spain and Slovenia employ PRS. On top of that, non-PRS surveys are also implemented in countries know for high methodological standards, such as the Netherlands and the UK. Secondly, referring to ESS survey outcomes, there are no significant differences between non-PRS and PRS surveys with respect to rates of response (RR2: 57.6% vs. 60.7%), refusal (REF1: 28.3% vs. 23.8%), contact (CON1: 90.9% vs. 92.2%) or cooperation (COOP2: 63.5% vs. 65.8%). Accordingly, no significant differences could be attested when it comes to the fraction of surveys utilising respondent-incentives (66.7% vs. 63.2%) as well as the ratio of respondents selected for (27.4% vs. 26.6%) or confirmed through (68.9% vs. 67.2%) back-checking.

An additional step has been taken to assure that the use of non-PRS is not correlated with some underlying deficiency in methodological stringency at the country level. In contrast to the approach prevalent in most methodological studies of the ESS, the present analysis only includes such countries which took part in at least 6 out of 7 rounds, so as to leave out those with patchy track-records. Exclusion of irregular participants allows for focusing on cases with stable, highly standardised and systematically implemented sampling and fieldwork execution procedures, which are likely less stable and reliable among irregularly participating countries. While the overall number of ESS-participating countries varies from 20 to 30 across different rounds, with 36 having participated in at least 1, the set included in the analysis comprises 19 countries. Of the included cases, 53 implemented non-PRS and 77 PRS-surveys. In spite of range-reduction, the countries taken into consideration stem from all the main regions of Europe. Additionally, although our analysis does not en-

Table 1
Sample types of the ESS participating countries(i) included in the analysis

Country	ESS1-2002	ESS2-2004	ESS3-2006	ESS4-2008	ESS5-2010	ESS6-2012	ESS7-2014
AT	non-PRS	non-PRS	non-PRS	non-PRS	non-PRS	-	PRS
BE	PRS	PRS	PRS	PRS	PRS	PRS	PRS
CH	non-PRS	non-PRS	non-PRS	non-PRS	non-PRS	PRS	PRS
CZ	non-PRS	non-PRS	-	non-PRS	non-PRS	non-PRS	non-PRS
DE	PRS	PRS	PRS	PRS	PRS	PRS	PRS
DK	PRS	PRS	PRS	PRS	PRS	PRS	PRS
EE	-	PRS	PRS	PRS	PRS	PRS	PRS
ES	non-PRS	PRS	PRS	PRS	PRS	PRS	PRS
FI	PRS	PRS	PRS	PRS	PRS	PRS	PRS
FR	non-PRS	non-PRS	non-PRS	non-PRS	non-PRS	non-PRS	non-PRS
GB	non-PRS	non-PRS	non-PRS	non-PRS	non-PRS	non-PRS	non-PRS
HU	PRS	PRS	non-PRS	PRS	PRS	PRS	PRS
IE	non-PRS	non-PRS	non-PRS	non-PRS	non-PRS	non-PRS	non-PRS
NL	non-PRS	non-PRS	non-PRS	non-PRS	non-PRS	non-PRS	non-PRS
NO	PRS	PRS	PRS	PRS	PRS	PRS	PRS
PL	PRS	PRS	PRS	PRS	PRS	PRS	PRS
PT	non-PRS	non-PRS	non-PRS	non-PRS	non-PRS	non-PRS	non-PRS
SE	PRS	PRS	PRS	PRS	PRS	PRS	PRS
SI	PRS	PRS	PRS	PRS	PRS	PRS	PRS

Countries are labelled according to ISO31166-1 Source: Tabulation based on (ESS, 2002, 2004, 2006, 2008, 2010, 2012, 2014)

compass mode effects, it should be noted that PAPI-based interviews were employed in 38 cases, while the remaining 92 surveys utilised CAPI. All the following analyses of ESS data implement design (dweight) but not post-stratification weights (pweight), as the latter would make distributions of age and gender similar to that of the population, which would mask actual over- or under-representation of particular categories of respondents in the sample.

4 Methods

4.1 External criteria of sample representativeness

Demographic representativeness of surveys may be investigated by comparison with more accurate estimates of population characteristics, i.e., the so-called "gold standards ". Groves (2006) argues in favour of such a method, as it relies on comparing independent estimates of survey items while requiring no individual-level information. Census-type data constitute a perfect fit for such "gold standard " assessments, as they are reliably measured within well-defined categories. While there may exist no external data on substantive survey findings, such highly credible estimates are readily available when it comes to some demographic characteristics. Although consistency of sample composition with known "gold standards " of such population parameters as age or gender does not in itself preclude other biases of key survey findings, it does constitute a confidence-booster with respect to overall survey quality (Voogt & Van Kempen, 2002).

Established "gold standard " evaluations of ESS samples usually rely on comparisons with the European Union Labour Force Survey (LFS). For instance, Koch (2016), and earlier Koch et al. (2014), examined demographic representativeness of ESS rounds 5 and 6 by comparing the survey distributions of gender, age, marital status, work status, nationality and household size, with relevant LFS-obtained characteristics. These analyses demonstrated that non-PRS fared systematically worse than PRS surveys on indices of dissimilarity. Note, however, that LFS itself is also a survey well-known to have considerable nonresponse and measurement problems in some countries. Therefore, its utilisation as a source of externally-known, true benchmark parameters seems somewhat thorny. For instance, LFS employs four data gathering modes: F2F interviews, CATI, CAWI and self-administered questionnaires, and no information on possible mode-effects is made available. Furthermore, its nonresponse rates range from 2,1% in Germany to 79,4% in Luxembourg. In addition, LFS also exhibits country-specific coverage issues, e.g., 12% of sample-frame units in Poland do not belong to the population, as they consist of uninhabited dwellings, ones that have only seasonal inhabitants, or have been transformed into non-residential properties (cf. Eurostat, 2015). Therefore, even if one were to maintain that LFS measurements remain somehow superior, any benchmarking of ESS estimates against them seems to be more

of a "silver" than "gold" standard.

Given the uncertain superiority of LFS over ESS sample quality, an external evaluation of the ESS would require an indisputably more golden standard, which is in fact available in the form of Eurostat population data (Eurostat, 2017). The Statistical Office of the European Union publishes annual demographic figures on the basis of submissions from National Statistical Institutes, whose validity is verified through multiple control procedures (Eurostat, 2018). The following assessment takes gender and age as external benchmarks, so as to bypass possible concerns over measurement errors to which other constructs such as marital status, work status, nationality or household size seem more exposed.

Prior to benchmarking of the ESS demographic bias against Eurostat data, two obvious questions regarding data equivalence had to be addressed. Since the ESS target-population only includes 15+ individuals, the comparisons necessitated an exclusion of the under-15 age-category from Eurostat data. Additional adjustments resulted from the timing of ESS fieldwork execution. While Eurostat demo pjangroup variable provides annual population parameters set on January 1st, ESS fieldwork lasts for a number of months, and sometimes may span over two calendar-years. Changes in population structure over short time-periods may seem miniscule, yet, they are not statistically negligible. In order to obtain precise comparisons across different timings of fieldwork execution the following adjustments have been made: (a) if ESS fieldwork took place in a middle of the year the reference date for Eurostat data was set on the 1st of January of this year, (b) if fieldwork took place at the turn of the year, the reference date was the 1st of January of the year when ESS fieldwork was finished.

Based on Eurostat age and gender data, we construct three measures of demographic imbalance:

1. female misrepresentation (% of females in a study—% of females in population aged 15+);

2. youth misrepresentation (% of respondents aged 15–24 in a study—% of population aged 15–24 in population aged 15+);

3. elderly misrepresentation (% of respondents aged 75+ in a study—% of population aged 75+ in population aged 15+).

Thus, in line with the TSE approach (Biemer, 2010), for each study the imbalances are defined as the difference between the estimated value and the population parameter: $\widehat{p_i} - p_i$. As true gender and age proportions are known, the standard error of each estimator is equal to $\mathrm{SE}_i = \sqrt{\frac{p_i(1-p_i)}{n_i}}$, where p_i is the true fraction of gender and age categories in the 15+ population, and n_i is the total number of respondents in each of the 130 different surveys included in the analysis.

In order to make comparisons between multiple studies, we define a measure of absolute bias of demographic representation ($|\mathrm{dem}_{\mathrm{bias}}|$) as an absolute value of the deviation of a survey estimate from its underlying true parameter value divided by the standard error. The result is statistically significant (at $p = 0.05$), if $|\mathrm{dem}_{\mathrm{bias}}| > 1.96$.

$$\left|\mathrm{dem}_{\mathrm{bias},i}\right| = \frac{\left|\widehat{p_i} - p_i\right|}{\sqrt{p_i(1-p_i)/n_i}} \quad .$$

4.2 Internal criteria of sample quality

Especially in cross-country surveys, benchmarking sample quality against external criteria suffers from uncertainty concerning the actual accuracy, reliability and uniformity of any supposedly "gold " standards. Additionally, external evaluations allow for estimating demographic imbalances, but not other sample properties, such as unit nonresponse bias. Given such inherent limitations, Sodeur (1997) put forward the idea of internal criteria of representativeness, reliant on evaluating the composition of specific sub-samples against values known by definition. His analysis, originally performed on German General Social Survey (ALLBUS) datasets, focused on gender composition of heterosexual two-person households, whose expected gender ratio is 50/50 (as the sub-sample of respondents living together with a heterosexual partner comprises individuals with equal chances of selection). In Sodeur's original analysis, the observed deviations from this theoretical benchmark involved 1) overrepresentation of young females (explained as an effect of higher availability due to lower labour-market participation and higher rates of stay-at-home parenting) as well as 2) corresponding underrepresentation of older women (accounted for by reference to differences in social roles conditioning readiness to cooperate). More recently, Kohler (2007) applied internal criteria to compare the quality of cross-country surveys, which allowed for bypassing country-specific quality-differences in the reliability of statistical information available for use as external criteria.

Our implementation of internal-criteria evaluation of ESS sample quality focuses on the same data-set comprising 130 ESS surveys on which the external-criteria analysis was performed. It required a prior separation of a sub-sample of heterosexual couples, which excluded all singles, partners not sharing a household, homosexual partners, heterosexual partners living together with other relatives or unrelated individuals belonging to the target population. The analysis focuses exclusively on overall gender distribution, as when age comes into play it necessitates additional dubious assumptions—for instance, the fraction of under-35 females living in heterosexual couples need not be 50%—whatever it in fact is the case in any given society poses an empirical question not something that can be known a priori. Theoretically, one could even imagine a society whose customs would allow 18+ females to marry and only treat 35+ males as eligible, whereby the expected gender ratio of under-35

Table 2
Variables of the ESS and Eurostat comparison

Variable	ESS variable	Eurostat variable	ESS survey estimator(i)	Eurostat population parameter
Gender	gndr	demo pjangroup sex	% of females	% of females n population 15+
Age: 15-24 years old	agea (recoded)	demo pjangroup age	% of respondents aged 15-24 years old	% of population aged 15-24 in population 15+
Age: 75 years and older	agea (recoded)	demo pjangroup age	% of respondents 75 years and older	% of population aged 75+ in population 15+

Table 3
The basic characteristics of ESS1-ESS7 sub-sample separation process

Reasons for excluding respondents from ESS1-ESS7 cumulative dataset	*n*
Step 1: One-person households or households with unknown number of members	51, 425
Step 2: Respondent does not live with husband/wife/partner or there is no information about it	51, 554
Step 3: Respondent lives with: a) parents/parents-in-law, b) other relative, c) other non-relative	8, 525
Step 4: Respondent lives with homosexual partner	1, 613
Step 5: Respondent declared cohabitation in one household with two or more husbands/wives	216
Step 6: No data about: a) gender of respondent/partner, b) relationship with other residents	531
Step 7: Respondent lives in a household with children aged 15 years and older	33, 445
Total number of respondents excluded from ESS1–ESS7 cumulative dataset	147, 309
Total number of cases included in analysis	103, 770

persons living in heterosexual couples would be 100% female.

The sub-sample of n=103,770 has been selected from the entire ESS1-ESS7 dataset, which more than halved the original total number of 251,079 units. Even if such separation reduces the sample-size, it constitutes the only way to calculate the difference between observed proportion of females in households of heterosexual partners and the true proportion of females in such household-types (% of females in a sub-sample of a study—50%). This difference constitutes the TSE, which is known within statistical meta-analysis as Single-Study Effect Size (ESi) (Borenstein, Hedges, Higgins, & Rothstein, 2011). Consistently with the measures of misrepresentation based on external criteria, ESi is calculated independently for each survey, as a difference between survey estimate $\widehat{p_i}$ and the true parameter value p_i (equal by definition to 0.5). Values of ESi are then used (a) to calculate Overall Effect Size (OES) (in order to evaluate H2), and then to (b) estimate the absolute value of the unit nonresponse bias (for evaluation of H3).

Overall Effect Size of unit nonresponse bias. In order to arrive at a quantitative comparison of PRS and non-PRS surveys that encompasses both within- and between-study variance of single-study effect sizes, we make use of the analytical procedure of meta-analysis that incorporates those estimators into one measure of Overall Effect Size. There are two common, yet, conceptually different approaches to calculating OES: (a) The fixed effect model, and (b) the random effect model. The former assumes existence of one true effect for all studies, i.e., all factors impacting the effect are assumed equal and the variability of single-study effects only then stems from the random error within each survey. On the other hand, the random effect model postulates many different true effects, which seems especially appropriate when different factors influence true effect size in each survey. From this perspective, the measures of single-study effects estimate not one true effect size but different true effects. Hence, the between-study variability of single-study effects occurs not only due to within-study variance but also because of between-study variance, which should be therefore incorporated in weights.

The random effect model of meta-analysis is much more conservative than the fixed one. Hence, much higher deviations from underlying true parameters must be attested to recognize OES as statistically significant. This model also seems much more suitable when attempting a meta-analysis

of ESS studies, since they provide information about true effects for all particular survey populations. In a random effect model, weighted OES is computed separately for PRS and non-PRS surveys in accordance with the following formula:

$$\widehat{\mathrm{OES}}^{*} = \frac{\sum_{i=1}^{k} w_i^{*} \cdot \mathrm{ES}_i}{\sum_{i=1}^{k} w_i^{*}}$$

Parameter k is the number of included studies and estimator ES_i stands for single-study effect size, i.e., the difference between the observed and the true proportion of females in two-person heterosexual households. For each study, weights w_i^{*} are computed as reciprocals of the variance of $\widehat{\mathrm{OES}}^{*}$, i.e., $V_i^{*} = V_i + \tau^2$, where V_i is within-study variance and τ^2 denotes between-study variance. Following Borenstein et al. (2011), τ^2 is defined as the maximum of two values $\left\{\frac{Q-\mathrm{df}}{c}; 0\right\}$, where:

$$\mathrm{Q} = \frac{\sum_{i=1}^{k} \left(\mathrm{ES}_i - \widehat{\mathrm{OES}}\right)^2}{V_i} \quad ,$$

with

$$\widehat{\mathrm{OES}} = \sum_{i=1}^{k} V_i^{-1} \cdot \mathrm{ES}_i / \sum_{i=1}^{k} V_i^{-1} \quad ,$$

$$df = k - 1 \quad \text{, and}$$

$$c \sum_{i=1}^{k} V_i^{-1} - \frac{\sum_{i=1}^{k} V_i^{-2}}{\sum_{i=1}^{k} V_i^{-1}} \quad ..$$

Standard error of $\widehat{\mathrm{OES}}^{*}$ allows for computing confidence intervals, as well as z-value and p-values to verify whether there are any significant deviations of observed fractions from true parameters in PRS and non-PRS surveys.

Absolute value of unit nonresponse bias. In order to verify whether PRS and non-PRS surveys exhibit contrastive correlation patterns between unit nonresponse bias and refusal rate (H3), a measure of absolute value of unit nonresponse bias was defined:

$$\left|\mathrm{unit}_{\mathrm{bias},i}\right| = \frac{\left|\widehat{p_i} - 0.5\right|}{\sqrt{0.25/n_i}}$$

As the expected proportion of females is equal to 0.5, the variance of female ratio estimator is equal to $0.25/n_i$, whereby n_i is the total number of respondents in each extracted sub-sample. Note that $|\mathrm{unit}_{\mathrm{bias}}|$ constitutes an internal analysis equivalent of $|\mathrm{dem}_{\mathrm{bias}}|$ used in the context of external evaluation (Kohler, 2007).

5 Results

5.1 External criteria assessment

In order to test whether $|\mathrm{dem}_{\mathrm{bias}}|$ is indeed significantly higher in PRS than non-PRS surveys, it is first necessary to exclude the possibility of significant influence on the part of another fixed factor—ESS round. Therefore, we specify the following General Linear Model (GLM), where the inclusion of the two-way interaction allows for testing the assumption that the relationship between $|\mathrm{dem}_{\mathrm{bias}}|$ and type of sample is ESS-round invariant:

$$|\mathrm{dem}_{\mathrm{bias}}| = \text{constant} + \text{type of sample} + \text{ESS round} + \text{type of sample} \cdot \text{ESS round}$$

Prior to conducting GLM (univariate MANOVA), the non-orthogonality of type_of_sample and ESS_round was attested. Hence, it would make a difference whether a given factor was tested on its own or together with another correlated factor, as this would affect the p-values, because the inclusion of an extra factor reducing the variance of error would lead to an increase in the F-ratios. A contingency table of type_of_sample and ESS_round demonstrated that the distribution of sample types over ESS rounds deviated from what would be expected if they were entirely unrelated. Even though Pearson Chi-Square test did not provide a statistically significant result ($\chi^2 = 1.272$; $df = 6$; p-value= 0.973), the fact that $\chi^2 > 0$ means that the two variables are not orthogonal. This necessitates their treatment as correlated within the GLM model, as their effects (type II of sums of squares) would be adjusted for each other.

Table 4 presents GLM summary characteristics:

1. F-ratios demonstrating the significance of main factors (type of sample and ESS round) and of the effect of interaction between them;

2. mean values of $|\mathrm{dem}_{\mathrm{bias}}|$ with corresponding standard errors within categories selected according to sample type and ESS round;

3. the proportion of variance explained by the GLM.

The GLM analysis leads to the conclusion that there is a strong, statistically significant main effect associated with type of sample: for all three dependent variables average $|\mathrm{dem}_{\mathrm{bias}}|$ is consistently higher for non-PRS samples and F-ratios prove statistically significant. Therefore, GLM analysis supports the contention that the absolute bias of demographic representation is significantly higher in non-PRS than PRS samples (H1). Furthermore, gender and age imbalances are stable over time as no effect of ESS round was attested on any of the dependent variables, and no significant interaction was found to exist between type of sample and ESS round.

5.2 Internal criteria assessment

The results of statistical meta-analysis of 130 ESS surveys are presented in Table 5, and in Figure 1. They allow for testing whether OES of unit nonresponse bias is positive in non-PRS and negative in PRS surveys. Table 5. displays mean values of OES, standard errors of respective means and

Table 4
GLM Univariate analysis of between-subject effects on gender and age imbalance

Fixed factors	Coef.	Female %	Coef.	Aged 15–24 %	Coef.	Aged 75+ %	# of countries
Type of sample							
non-PRS	2.19	0.254	2.65	0.272	3.19	0.336	53
PRS	1.47	0.143	1.69	0.145	2.01	0.180	77
F(1;125)	7.669**		10.834**		7.585**		
ESS Round (time)							
ESS1-2002	1.54	0.285	2.15	0.416	2.67	0.438	18
ESS2-2004	2.17	0.449	1.53	0.272	3.15	0.442	19
ESS3-2006	1.55	0.362	2.28	0.355	2.17	0.515	18
ESS4-2008	1.58	0.377	2.28	0.326	2.17	0.413	19
ESS5-2010	1.83	0.378	2.34	0.419	2.62	0.565	19
ESS6-2012	1.89	0.411	1.75	0.338	2.14	0.505	18
ESS7-2014	1.83	0.304	2.49	0.520	2.39	0.498	19
F(6;125)	0.543		0.776		0.593		
Type of sample · ESS Round							
F(6;125)	1.229		0.653		0.680		
R2	13.0%		14.1%		12.1%		

** $p < 0.01$

Table 5
OES by type of survey sample based on internal criteria of representativeness

Type of sample	# of studies	OES	SE(OES)	z-value
PRS	77	−0.008	0.003	−2.87**
non-PRS	53	0.013	0.003	3.73***

** $p < 0.01$ *** $p < 0.001$

z-values for each type of sample. In turn, Figure 1 displays a box-and-whisker-plot with $\widehat{OES}^{*}$, standard error of mean (boxes) and 95% confidence intervals of mean (whiskers). Note that the OES is estimated individually for PRS and non-PRS, and significance tests are also performed separately for each group of surveys. Therefore, it is not the difference between the two mean values of OES that is important, but the significant deviation from the known parameter of each of them.

The results of meta-analysis fall in line with hypothesis H2: non-PRS surveys have a significant positive impact on unit nonresponse bias (z-value= 3.73; p-value< 0.001), in contrast to the significantly negative impact of PRS surveys (z-value= −2.87; p-value< 0.01). This amounts to a significant over-selection of women in non-PRS, and their corresponding under-selection in PRS surveys.

Having attested a significant OES-contrast between PRS and non-PRS, the subsequent step: the hypothesis that the absolute value of unit nonresponse bias for the two sample types is significantly correlated with refusal rates of the ESS surveys included in the meta-analysis. Figures 2 and 3 present scatter-plots contrasting the distributions of $|\text{unit}_{\text{bias}}|$ and REF1 for PRS and non-PRS sub-samples of ESS studies.

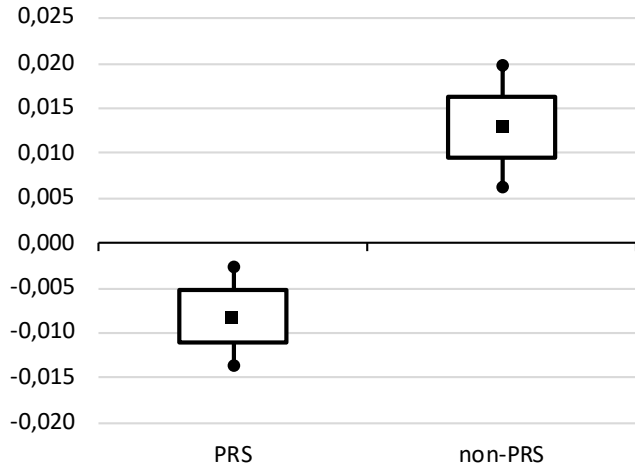

Figure 1. Box-and-whisker-plots of OES by type of sample

PRS and non-PRS surveys clearly exhibit distinct patterns of relationship between refusal rates and unit nonresponse bias in line with H3. PRS surveys exhibit a signif-

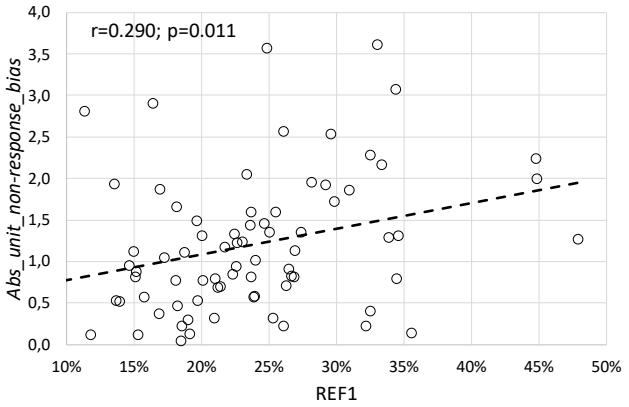

Figure 2. Correlation between refusal rate (REF1) and $|\text{unit}_{\text{bias}}|$ in PRS surveys (Note: The variance of errors in the model is constant; Breush-Pagan test: p-value = 0.173)

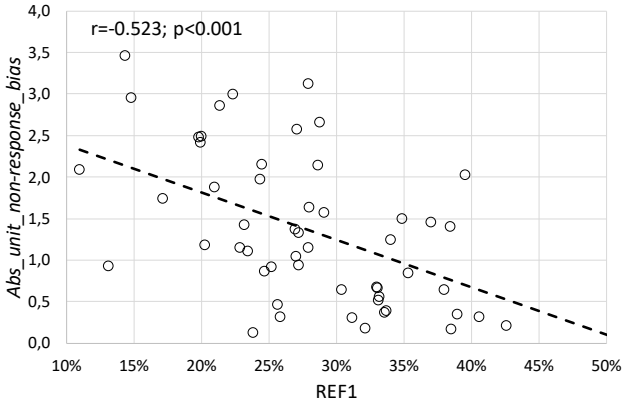

Figure 3. Correlation between refusal rate (REF1) and $|\text{unit}_{\text{bias}}|$ in non-PRS surveys (Note: The variance of errors in the model is constant; Breush-Pagan test: p-value = 0.606)

icant, moderately positive relationship between REF1 and $|\text{unit}_{\text{bias}}|$ ($r = 0.29$; p-value< 0.05), falls in line with conventional wisdom—higher refusal rates correlate with unit nonresponse bias. On the other hand, for non-PRS the reverse is true—decreasing refusal rates are associated with increasing unit nonresponse bias: the corresponding relationship is significant, strong and negative ($r = -0,523$; p-value< 0.001). It is not the strength of association that really matters here, however, but rather its direction: negative correlation in the case of non-PRS casts a troubling shadow on the practise of using REF1 as a quality measure in such type of surveys.

6 Discussion

The external evaluation of ESS surveys proved that the absolute value of the bias of demographic representation is significantly higher in non-PRS than PRS cases (in line with H1). This falls in line with the mainstream view that holds PRS to be the preferred fieldwork option. Our main point, however, is that the contrast can be explained by inherent characteristics of the two types of samples. Specifically, we contend that the inferior demographic representativeness of non-PRS surveys results from higher interviewer discretion leading to aggregate over-selection of respondents that tend to be more available or likely to respond. The state of affairs that one should expect to find if this suggestion is correct is exactly what H2 and H3 describe. Conversely, it seems hard to come by other plausible explanations why such systematic contrast between PRS and non-PRS surveys could be brought about by other means.

Within the framework of internal-criteria assessment, systematic deviations from the 50/50 gender ratio can be explained in terms of interaction between (a) sample type characteristics pertaining to interviewer discretion, and (b) gender-specific differences in respondent availability and readiness to cooperate. Starting from the latter, irrespective of sample type, due to differential labour-market participation, women tend to be more often available at home. This is true in the ESS (Stoop et al., 2010), as well as more generally (Groves & Couper, 2012). On the other hand, while only partial evidence exists for a universally lower female propensity to cooperate in surveys (Stoop, 2005), existing evaluations of ESS data had clearly showed that within this project women are consistently exhibiting higher refusal rates (Menold, 2014). Thus, with respect to the ESS, it is fully warranted to make the assumption that female respondents are more available at home and, yet, less willing to cooperate. Therefore, a negative unit nonresponse bias would be expected in PRS surveys, given that availability plays a negligible role in the particular case of the ESS due to the fact that its fieldwork execution involves exceptionally intensive efforts to establish contact with selected respondents (Engel, Jann, Lynn, Scherpenzeel, & Sturgis, 2014; Stoop, Koch, Halbherr, Loosveldt, & Fitzgerald, 2016). This would only leave the refusal to cooperate as a factor of influence. In non-PRS surveys, however, higher female availability does play a role, as their contact rate is not defined in relation to a specific individual but on the basis of having established contact with anyone in the target household (Koen, Loosveldt, Vandenplas, & Stoop, 2018). Since establishing contact with male respondents tends to be more challenging, some hard-to-reach males would be replaced with their easier-to-reach household counterparts (such females that prove willing to cooperate). Conversely, females unwilling to cooperate are likely to terminate the interview at the stage of within-household selection, thus, neutering the effect of higher propensity to refuse and leading to a significant positive unit nonresponse bias.

Our analysis follows Menold (2014), who demonstrated that an overrepresentation of women in spite of higher refusal rates can be seen as indicative of interviewer influence through the selection of individuals with higher levels

of readiness and availability. Thus, an overrepresentation of females in a sub-sample of heterosexual couples can be plausibly explained by interviewer interference with the process of selection. Her analysis demonstrates as well that the opposite cases of female underrepresentation stem from lower propensity to respond. Still, her analysis does not implement those insights to the evaluation of sample types, which she performs based on an absolute difference between the observed and true fraction of females, which misses the interaction between availability and readiness to participate: as both significant over- and under-representation of females would indicate a unit nonresponse bias, while only the former would involve target-respondent substitutions. However, our analysis goes further in two crucial respects: (a) we investigate the sign of bias and propose an explanation for it, and (b) we probe the relationship between refusal rates and the unit-nonresponse bias (H3).

That the correlation between refusal rates and absolute value of unit nonresponse bias is positive in PRS surveys seems in line with common sense expectations: higher incidence of non-cooperating respondents should depress sample representativeness. Crucially however, it is the reverse relationship within non-PRS surveys that is both surprising and plausibly indicative of an underlying problem. If higher recorded refusal rates tend to be associated with lower absolute values of unit nonresponse bias, then this seems to cast a shadow of doubt on the manner in which they were recorded. Again, the most straightforward scenario generating such results involves interviewer discretion: substitutions of less-available males with their household-present female partners would distort refusal rates, lead to gender misrepresentation, and produce an inversely proportional relationship between them. Therefore, in the case of non-PRS surveys the use of refusal rates as metrics for evaluating the quality of fieldwork execution should be approached with a pinch of salt.

7 Conclusions

The demonstration, on the basis of external census-based criteria, that non-PRS fare consistently worse than PRS surveys when it comes to demographic representativeness of survey samples follows the findings of established evaluations of ESS measurements. This quality contrast is far from ESS-specific, yet, the fact that it occurs in spite of high methodological sophistication, procedural stringency and ample funding points to the existence of inherent, and probably insurmountable limitations of non-PRS surveys. Furthermore, we show that a key incorrigible factor responsible for the superiority of sampling from personal registers consists in the interviewers not being able to influence the process of respondent selection. Such interference is hard to spot after the fact, as seasoned interviewers could easily manipulate within-household selection protocols in subtle ways so as to select the person actually available at the time of contact, and even if individual manipulations were to be uncovered they can be easily blamed on honest mistakes. Still, in the aggregate, the sum of those substitutions leaves a detectible footprint: gender imbalances in survey samples.

Making use of internal criteria for evaluating survey quality is a welcome addition to the hitherto far more prevalent approaches reliant on external benchmarking. Internally focused checking requires no external data-sources, and can be performed without restriction on all types of probabilistic samples. In this sense, investigating unit nonresponse bias can be a useful tool of survey quality control, so far as those surveys include collection of data on the composition of respondent households. Thus, this procedure is readily applicable to most major cross-country surveys that routinely record such characteristics. In principle, other aprioristic quantitative relations than the 50/50 male-to-female ratios in heterosexual couples are possible, but this is the only one so far that has a credible track-record.

Our choice of gender bias detection on the basis of internal criteria is motivated by the fact that given the higher availability of females at home they constitute a probable target for substitution. The statistical meta-analysis of ESS surveys clearly showed that the Overall Effect Size for the former is indeed significantly positive: females tend to be overrepresented, while the latter actually exhibit small yet significant female underrepresentation. Furthermore, within PRS, the samples' higher absolute values of unit nonresponse bias tend to be associated with higher refusal rates, with the reverse being true of non-PRS surveys. The only feasible explanation for those differences seems to be higher interviewer discretion in terms of selecting target persons in non-PRS cases. If interviewers choose to bend the rules of within-household selection by over-selecting stay-at-home cooperative individuals, this would have a significant differential impact on both the unit nonresponse bias and refusals rates. Thus, if interviewers tend to substitute an unavailable male target person by his easier-to-reach female partner, then in a sub-sample of respondents living as heterosexual couples a significant overrepresentation of females will occur. Undocumented substitutions of reluctant respondents by someone else within the same household will decrease the values of refusal rates, explaining why in non-PRS surveys lower refusal rates coexist with higher absolute values of unit nonresponse bias.

Acknowledgments

This work was supported by the grant "Comparative analysis of the quality of survey samples in the cross-national studies on the basis of external and internal criteria of representativeness: survey archivisation and meta-bases of results" awarded by the (Polish) National Science Centre (No. 2017/01/X/HS6/01304). We would like to thank the anonymous reviewers for useful comments and remarks.

References

AAPOR. (2016). Standard definitions: Final dispositions of case codes and outcome rates for surveys. American Association for Public Opinion Research. Retrieved from http://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf

Bethlehem, J., Cobben, F., & Schouten, B. (2011). Nonresponse and representativity. In J. Bethlehem, F. Cobben, & B. Schouten (Eds.), *Handbook of nonresponse in household surveys* (pp. 178–208). New York: John Wiley % Sons.

Beullens, K. & Loosveldt, G. (2016). Interviewer effects in the european social survey. *Survey Research Methods*, *10*(2), 103–118.

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, *74*(5), 817–848.

Biemer, P. P. & Lyberg, L. E. (2003). *Introduction to survey quality*. New York: John Wiley & Sons.

Billiet, J., Philippens, M., Fitzgerald, R., & Stoop, I. (2007). Estimation of nonresponse bias in the european social survey: Using information from reluctant respondents. *Journal of Official Statistics*, *23*(2), 135.

Billiet, J., Vehovar, V., Beullens, K., & Matsuo, H. (2009). Non-response bias in cross-national surveys: Designs for detection and adjustment in the ESS. *ASK. Research & Methods*, *18*, 3–43.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. New York: John Wiley & Sons.

Cornesse, C. & Bosnjak, M. (2018). Is there an association between survey characteristics and representativeness? a meta-analysis. *Survey Research Methods*, *12*(1), 1–13.

Coromina, L. & Saris, W. E. (2009). Quality of media use measurement. *International journal of public opinion research*, *21*(4), 424–450.

Curtin, R., Presser, S., & Singer, E. (2005). Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly*, *69*(1), 87–98.

Eckman, S. & English, N. (2012). Creating housing unit frames from address databases: Geocoding precision and net coverage rates. *Field Methods*, *24*(4), 399–408.

Engel, U., Jann, B., Lynn, P., Scherpenzeel, A., & Sturgis, P. (2014). *Improving survey methods: Lessons from recent research*. New York: Routledge.

ESS. (2016). ESS cumulative dataset: Rounds 1–7. Norway: NSD— Norwegian Centre for Research Data. Retrieved from http://www.europeansocialsurvey.org/downloadwizard/

Eurostat. (2017). Population (demo_pop)—eurostat metadata. Retrieved from http://ec.europa.eu/eurostat/cache/metadata/en/demo_pop_esms.htm#annex1491295136130

Eurostat. (2018). *Methodology for data validation 2.0*. Brussels: Eurostat.

Fricker, S. & Tourangeau, R. (2010). Examining the relationship between nonresponse propensity and data quality in two national household surveys. *Public Opinion Quarterly*, *74*(5), 934–955.

Gaziano, C. (2005). Comparative analysis of within-household respondent selection techniques. *Public Opinion Quarterly*, *69*(1), 124–157.

Goyder, J. (1987). *The silent minority: Nonrespondents on sample surveys*. Boulder, CO: Westview Press.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public opinion quarterly*, *70*(5), 646–675.

Groves, R. M. & Couper, M. P. (2012). *Nonresponse in household interview surveys*. New York: John Wiley & Sons.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). *Survey methodology*. New York: John Wiley & Sons.

Groves, R. M. & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias a meta-analysis. *Public opinion quarterly*, *72*(2), 167–189.

Häder, S. & Lynn, P. (2007). How representative can a multination survey be. In R. Jowell, C. Roberts, R. Fitzgerald, & G. Eva (Eds.), *Measuring attitudes cross-nationally: Lessons from the European Social Survey* (pp. 33–53). London: Sage.

Jäckle, A., Roberts, C., & Lynn, P. (2010). Assessing the effect of data collection mode on measurement. *International Statistical Review*, *78*(1), 3–20.

Keeter, S., Miller, C., Kohut, A., Groves, R. M., & Presser, S. (2000). Consequences of reducing nonresponse in a national telephone survey. *Public opinion quarterly*, *64*(2), 125–148.

Koch, A. (2016). *Assessment of socio-demographic sample composition in ESS round 6*. European Social Survey, GESIS Mannheim.

Koch, A., Halbherr, V., Stoop, I. A., & Kappelhof, J. W. (2014). *Assessing ESS sample quality by using external and internal criteria*. European Social Survey, GESIS Mannheim.

Koen, B., Loosveldt, G., Vandenplas, C., & Stoop, I. (2018). Response rates in the European Social Survey: Increasing, decreasing, or a matter of fieldwork efforts? *Survey Methods: Insights from the Field*. Retrieved from https://surveyinsights.org/?p=9673

Kohler, U. (2007). Surveys from inside: An assessment of unit nonresponse bias with internal criteria. *Survey Research Methods*, *1*(2), 55–67.

Kreuter, F. & Kohler, U. (2009). Analyzing contact sequences in call record data. potential and limitations of sequence indicators for nonresponse adjustments in the European Social Survey. *Journal of Official Statistics*, *25*(2), 203.

Little, R. J. & Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, *31*(2), 161–168.

Loosveldt, G. & Beullens, K. (2013). ''how long will it take?" an analysis of interview length in the fifth round of the European Social Survey. *7*(2), 69–78.

Luiten, A. & Schouten, B. (2013). Tailored fieldwork design to increase representative household survey response: An experiment in the Survey of Consumer Satisfaction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(1), 169–189.

Lynn, P., Häder, S., Gabler, S., & Laaksonen, S. (2007). Methods for achieving equivalence of samples in cross-national surveys: The European Social Survey experience. *Journal of Official Statistics*, *23*(1), 107.

Martin, E. (1999). Who knows who lives here? within-household disagreements as a source of survey coverage error. *63*(2), 220–236.

Matsuo, H., Billiet, J., Loosveldt, G., Berglund, F., & Kleven, bibinitperiod. (2010). Measurement and adjustment of non-response bias based on non-response surveys: The case of Belgium and in the European Social Survey round 3. *Survey Research Methods*, *4*(3), 165–178.

Menold, N. (2014). The influence of sampling method and interviewers on sample realization in the European Social Survey. *Survey Methodology*, *40*(1).

Menold, N., Winker, P., Storfinger, N., & Kemper, C. (2013). A method for ex-post identification of falsifications in survey data. In N. Menold, P. Winker, & R. Porst (Eds.), *Survey standardization and interviewers' deviations-impact, reasons, detection and prevention* (pp. 25–48). Frankfurt am Mein: PL Academic Research.

Mohler, P. (2007). What is being learned from the ess? In R. Jowell, C. Roberts, R. Fitzgerald, & G. Eva (Eds.), *Measuring attitudes cross-nationally: Lessons from the European Social Survey* (pp. 157–68). London: Sage.

Peytchev, A., Baxter, R. K., & Carley-Baxter, L. R. (2009). Not all survey effort is equal reduction of nonresponse bias and nonresponse error. *Public Opinion Quarterly*, *73*(4), 785–806.

Revilla, M. (2010). Quality in unimode and mixed-mode designs: A multitrait-multimethod approach. *Survey Research Methods*, *4*(3), 151–164.

Saris, W. E. & Revilla, M. (2016). Correction for measurement errors in survey research: Necessary and possible. *Social Indicators Research*, *127*(3), 1005–1020.

Saris, W. E., Satorra, A., & Coenders, G. (2004). 8. a new approach to evaluating the quality of measurement instruments: The split-ballot MTMM design. *Sociological Methodology*, *34*(1), 311–347.

Schouten, B., Bethlehem, J., Beullens, K., Kleven, bibinitperiod, Loosveldt, G., Luiten, A., ... Skinner, C. (2012). Evaluating, comparing, monitoring, and improving representativeness of survey response through r-indicators and partial r-indicators. *International Statistical Review*, *80*(3), 382–399.

Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, *35*(1), 101–113.

Simmons, K., Mercer, A., Schwarzer, S., & Kennedy, C. (2016). Evaluating a new proposal for detecting data falsification in surveys. *Statistical Journal of the IAOS*, *32*(3), 327–338.

Smith, T. W. (2007). Survey non-response procedures in cross-national perspective: The 2005 ISSP non-response survey. *Survey Research Methods*, *1*(1), 45–54.

Sodeur, W. (1997). Interne Kriterien zur Beurteilung von Wahrscheinlichkeitsauswahlen. *ZA-Information/Zentralarchiv für Empirische Sozialforschung*, *41*, 58–82.

Stoop, I. (2005). *The hunt for the last respondent: Nonresponse in sample surveys*. The Hague: Sociaal en Cultureel Planbureau.

Stoop, I., Billiet, J., Koch, A., & Fitzgerald, R. (2010). *Improving survey response: Lessons learned from the European Social Survey*. New York: John Wiley & Sons.

Stoop, I., Koch, A., Halbherr, V., Loosveldt, G., & Fitzgerald, R. (2016). Field procedures in the European Social Survey round 8: Guidelines for enhancing response rates and minimising nonresponse bias. Technical report. London: ESS ERIC Headquarters. Retrieved from http://www.europeansocialsurvey.org/docs/round8/

Tomescu-Dubrow, I. & Slomczynski, K. M. (2016). Harmonization of cross-national survey projects on political behavior: Developing the analytic framework of survey data recycling. *International Journal of Sociology*, *46*(1), 58–72.

Tourangeau, R., Shapiro, G., Kearney, A., & Ernst, L. (1997). Who lives here? survey undercoverage and household roster questions. *Journal of Official Statistics*, *13*(1), 1–18.

Vannieuwenhuyze, J., Loosveldt, G., & Molenberghs, G. (2010). A method for evaluating mode effects in mixed-mode surveys. *Public Opinion Quarterly*, *74*(5), 1027–1045.

Voogt, R. J. & Van Kempen, H. (2002). Nonresponse bias and stimulus effects in the Dutch National Election Study. *Quality & Quantity*, *36*(4), 325–345.