

INTERVIEWER INVOLVEMENT IN SAMPLE SELECTION SHAPES THE RELATIONSHIP BETWEEN RESPONSE RATES AND DATA QUALITY

STEPHANIE ECKMAN*
ACHIM KOCH

Abstract Several studies have shown that high response rates are not associated with low bias in survey data. This paper shows that, for face-to-face surveys, the relationship between response rates and bias is moderated by the type of sampling method used. Using data from Rounds 1 through 7 of the European Social Survey, we develop two measures of selection bias, then build models to explore how sampling method, response rate, and their interaction affect selection bias. When interviewers are involved in selecting the sample of households or respondents for the survey, high reported response rates can in fact be a sign of poor data quality. We speculate that the positive association detected between response rates and selection bias is because of interviewers' incentives to select households and respondents who are likely to complete the survey.

Countries differ in the frames they have available and thus in how they select household samples for face-to-face surveys. Several European countries have high-quality person registers that allow selection of individuals. Other countries have household or address registers from which a sample can be selected; interviewers then carry out the final stages of selection. In countries without registers, field staff are used more intensely in selection of households and individuals. Even countries that have registers may choose to use field staff

STEPHANIE ECKMAN is a fellow in the Survey Research Division at RTI International in Washington, DC, USA. ACHIM KOCH is a senior researcher at GESIS – Leibniz-Institute for the Social Sciences in Mannheim, Germany. The authors thank Michael Blohm, Carolina Casas-Cordero, Jill Dever, Frauke Kreuter, and Stanley Presser for helpful suggestions during the writing of the paper, and Curtiss Engstrom and Georg-Christoph Haas for research assistance. Much of the work on this paper was completed while S.E. was a senior researcher at the Institute for Employment Research in Nuremberg, Germany. *Address correspondence to Stephanie Eckman, 701 13th St NW, Suite 750, Washington, DC 20005, USA; email: seckman@rti.org.

doi:10.1093/poq/nfz012

© The Author(s) 2019. Published by Oxford University Press on behalf of the American Association for Public Opinion Research. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

to perform selection. Whenever field staff are involved in surveys, however, we worry about interviewer effects. Interviewers can introduce measurement error when they ask questions and record responses or nonresponse error when they recruit respondents (Durbin and Stuart 1951; Feldman, Hyman, and Hart 1951; Franzen and Williams 1956; Gray 1956; Athey et al. 1960; Bailar, Bailey, and Stevens 1977; O'Muircheartaigh and Campanelli 1999). For a review of interviewer effects in surveys, see West and Blom (2017). This paper explores the impacts that interviewers have on data quality when they are involved in sample selection.

When field staff select cases that they will interview, several incentives may push them to select cases that appear likely to complete the interview. First, selecting likely respondents makes the interviewers' job easier, because they may not have to return multiple times and may not have to persuade anyone to take part. Second, because interviewers are often judged by their response rates, selecting likely-to-respond households or persons can make interviewers look more productive. Third, when interviewers are paid by the completed case, as is the case in many countries, they may maximize their effective hourly wage by selecting likely respondents. Fourth, interviewers are not statisticians and may not understand the reasons behind the directions they are given about how to select cases. They may believe that completing a case quickly is more important than following the complex selection rules. In addition, the instructions we give interviewers may seem straightforward when we are in the office writing manuals, but they can be challenging to apply in the field.

Consider an interviewer selecting households via the random walk method, where she should select every k^{th} household. Is it surprising that she might interfere to select Household $k+1$, where the resident is sitting on the porch and waving, rather than Household k , which is dark and has a scary dog? Alternatively, consider a household selected from a register. The interviewer rosters the family members, and the laptop selects the mother for the interview. Unfortunately, the mother is rarely home, but the stay-at-home father seems willing to complete the interview right away. Once again, it should not be surprising that the interviewer may choose to interview the available father. To get around the programmed selection, the interviewer could perhaps alter the members' ages or back up and run the selection routine again.

Adequate review of selections made by interviewers is often difficult. Commonly used quality control measures, such as backchecks (short telephone re-interviews to verify that an interviewer visited), audio recordings, and timestamps (Koch 1995; American Association for Public Opinion Research 2003; Rosen et al. 2011), do not necessarily detect deviations from protocol. In the first scenario, the phone number recorded for the unit and used for the backcheck would match the unit interviewed, not the unit she should have selected. In the second scenario, backchecks might catch interviewer manipulation of ages in the household roster but would likely not detect if the interviewer simply reran the selection routine until the father was selected. The recording and the timestamps might appear normal in both scenarios.

The combination of interviewer involvement in selection, incentives to keep response rates high, and a lack of review of the selection procedure leads to a dangerous situation. Under such circumstances, a high survey response rate may in fact signal that interviewers have manipulated selection and that the data may not be representative of the population. Both situations described above could lead to artificially high response rates, because the hard-to-contact cases are not recorded as nonrespondents. Both could also introduce bias into the data. When a survey variable correlates with response propensity, and interviewers select cases with high response propensities, they can introduce bias in the data (Eckman 2017).

This paper investigates two research questions:

1. Do sampling methods involving interviewers lead to lower data quality than sampling methods where interviewers are not involved?
2. Does sampling method moderate the relationship between response rates and data quality?

Literature Review

Several other papers have explored the relationship between interviewer degrees-of-freedom in sample selection and data quality. Some of these studies compare samples selected via various methods to external benchmarks (Alt, Bien, and Krebs 1991; Bien, Bender, and Krebs 1997; Koch 1997; Hoffmeyer-Zlotnik 2006). Sodeur (2007) proposes an alternative measure of data quality: the percent of completed cases from gender-heterogeneous couples that are female. Kohler (2007) and Menold (2014) use this measure to investigate the impact of interviewer involvement in selection on data quality. These studies concluded that surveys that allow interviewers more leeway in selection are associated with lower quality. The authors take this result as a clue that undocumented substitution takes place when interviewers are involved in sample selection. Regarding the role of response rates, Blohm (2006) finds no evidence of an association between response rates and data quality, positive or negative. Kohler (2007, 63, fig. 6), in contrast, concludes that “poorly controlled surveys tend to have high response rates.”

Although these results are interesting and suggestive, the studies suffer from drawbacks that we attempt to remedy here. Several use only German data (Alt, Bien, and Krebs 1991; Bien, Bender, and Krebs 1997; Koch 1997; Blohm 2006; Hoffmeyer-Zlotnik 2006; Sodeur 2007). Hoffmeyer-Zlotnik (2006) and Blohm (2006) confound changes in sampling method with time trends. Others use only one measure of data quality, making them vulnerable to criticisms of that measure (Kohler 2007; Sodeur 2007; Menold 2014). None use regression models to control for potentially confounding relationships and to isolate the influence of response rates and sampling methods on data quality, as we aim to do.

Another branch of research has investigated the association between response rates and various measures of data quality, but without considering sampling method. Several studies have found no evidence of a relationship, positive or negative (Koch 1998; Keeter et al. 2000; Merkle and Edelman 2002; Groves and Peytcheva 2008; Blohm and Koch 2015; Sturgis et al. 2017). A reanalysis of the Groves and Peytcheva data by Brick and Tourangeau (2017) found that 25 percent of variance in nonresponse bias is the result of survey-level characteristics. Sampling method could be one of the survey-level variables that drives between-study variation in data quality.

Our research questions bring these two threads of research together to investigate the joint impact of response rates and sampling method on data quality. To address these questions, we use data from the first seven rounds of the European Social Survey (ESS). Pooling data across countries and rounds, we can explore the relationship between response rates and data quality and how it varies across sample selection methods. Because we use just one survey, many other factors, such as survey length and response rate definitions, are held constant.

The European Social Survey

The ESS is a face-to-face survey of behavior and opinions conducted every two years since 2002 in many European countries. Data collection is carried out independently in each country, but the Coordinating Center oversees questionnaire translation, sampling, fieldwork processes, data preparation, and documentation. We use data from the first seven rounds of the ESS. The unit of analysis in our research is the country-round: Belgium in 2002, Poland in 2006, and so on. For each country-round, we record the sampling method used and the response rate obtained. The ESS website provides access to all rounds of response data and detailed methodology reports for each country-round.

SAMPLING METHODS

Face-to-face surveys use one of several methods to select a sample of persons. The methods differ in the degree of involvement of the interviewer or other members of the field staff in the selection process. ESS standards require random probability samples, and countries' sample plans are reviewed by the Coordinating Center (Stoop et al. 2010, chap. 3). In this section, we describe selection methods commonly used in the ESS and other European surveys (Scherpenzeel et al. 2017). We discuss the methods in order from least to most interviewer involvement; table 1 summarizes the discussion.

Individual register: When a register of persons is available, a sample can be selected, often in geographic clusters to reduce interviewer travel costs. An

Table 1. Sampling methods used in European Social Survey, by increasing interviewer involvement

Sampling method	Field staff involvement in selection of:		Collapsed sampling method	Number of country-rounds using this method
	Housing unit	Person		
Individual register	None	None	Individual register	83
Household register	None	Interviewer	Other register	46
Address register	Interviewer	Interviewer		
Listing	Lister*	Interviewer	Walk	27
Random walk	Lister*	Interviewer		

NOTE.—* According to ESS standards, these tasks should not be done by the interviewer.

interviewer working from such a sample is given a list of names and available contact and sometimes demographic information for the selected individuals. Although interviewers could intervene in the selection process by substituting in an unselected person for an uncooperative selected case, such violations of the prescribed procedure could be detected by managers, because the name, age, and gender of the intended respondent is usually known from the register.

Household register: In situations where a register of persons is not available, a register of housing units may be. Following ESS terminology, we call this a household register. The Postcode Address File in the UK and the Postal Service Delivery Sequence File in the United States are household registers (Lynn and Taylor 1995; Iannacchione 2011).

The interviewer is more involved in the respondent selection process in this method than in the individual register method (see table 1). When more than one eligible person lives in the unit, the interviewer selects one using a Kish table or a method such as next birthday (see Gaziano 2005 for details on respondent selection methods). The ESS permits both paper and computerized methods of respondent selection, and the methodology reports do not indicate which method was used.

Address register: An address register is a list of residential buildings. An interviewer working from an address register sample receives a list of selected addresses, which can be single-family or multi-unit buildings. When the selected address contains more than one housing unit, the interviewer should

select one unit at the address using a technique like that used for respondent selection and may also need to select one respondent from the household (table 1).

Listing: When no register is available, field staff can create a frame of households in selected clusters via a process called listing. True listing involves frame creation only: staff members canvass the selected areas and create a frame of all housing units, which is returned to the central office. From this frame, a sample is selected and interviewers visit the selected cases for interviewing. For analysis of undercoverage and variance in listing, see Eckman (2013) and Eckman and Kreuter (2013).

If different staff members perform the listing and interviewing in a given area, this sampling method should appear to the interviewer just like the household register method. Sometimes, however, one staff member does both the listing and interviewing work, which can create incentives for staff to leave uncooperative-looking households off the frame (Eckman 2013; Eckman and Kreuter 2013). ESS guidelines require that the listing and interviewing work be done by different people, a point we return to in the next section.

Random walk: In the random walk method, the interviewer starts at a given spot and follows walking rules, selecting every k th unit and conducting interviews. (See Bauer 2014 for several random walk methods.) Random walk is commonly used in European countries that lack registries, or those where registry access is costly or burdensome (Hoffmeyer-Zlotnik 2006; Bauer 2014, 2016; Scherpenzeel et al. 2017); the method is also common in surveys conducted in Africa (Afrobarometer Network 2017).

Three problems exist with the random walk method. First, the sample selected can vary by interviewer because of different interpretations about which street to choose when turning right, or different methods of ordering apartments within buildings (Häder 2010, 154). Second, the probabilities of selection of the units are generally not equal, yet the method assumes they are (Bauer 2014, 2016). Third, interviewers tend to select households where someone is willing to do the survey (Alt, Bien, and Krebs 1991; Schnell 1997). Samples selected via random walk often do not represent the population well (Alt, Bien, and Krebs 1991; Bien, Bender, and Krebs 1997; Blohm 2006; Hoffmeyer-Zlotnik 2006; Kohler 2007).

The ESS quality guidelines require that countries using either listing or random walk separate the selection task from the interviewing task (Stoop et al. 2010, section 3.3.3). However, the methodology reports suggest that some countries do use the same staff to select households and carry out the interviewing.

Each of the five sampling methods discussed above is used by one or more countries in the first seven rounds of the ESS. We coded the sampling method in each country and round from sampling forms collected by the Coordinating

Center, which are not publicly available (although the methodology reports on the website include a summary for each country and round). However, even the internal reports do not always provide enough information to distinguish the methods, and thus we grouped these five methods into three sets that we could clearly distinguish: individual register, other register (which includes both the address and household register methods discussed above), and walk (which includes both listing and random walk). This collapsing gives us enough cases to compare the different methods while preserving the ordering of methods from low to high interviewer involvement, as shown in [table 1](#). The same three categories are used in [Scherpenzeel et al. \(2017\)](#) and [Maineri et al. \(2017\)](#).

Among the 156 country-rounds we use in our analyses, 83 use individual register, 46 use another register, and 27 use walk sampling methods ([table 1](#)). Eight countries switched from one sampling method to another during these rounds, and we use data from the 45 rounds available for these countries for some of our analyses.

RESPONSE RATES

We use response rates reported in the Round Quality Reports, available on the ESS website ([Matsuo and Loosveldt 2013](#); [Beullens et al. 2014](#)). The rates are calculated according to the American Association for Public Opinion Research's response rate standard (RR1) and are unweighted ([Beullens et al. 2014](#); [American Association for Public Opinion Research 2016](#)). Although the ESS guidelines specify a target response rate of 70 percent, only 20 percent of the country-rounds in our analysis data set achieve this goal. Country-rounds using the walk method meet the target more often (48 percent) than those using the other register method (11 percent) and the individual register method (16 percent).¹

Two Measures of Selection Bias

To be specify what we mean by data quality, we turn to the Total Survey Error framework, which summarizes the sources of error that can cause a survey estimate to deviate from the corresponding population value ([Lyberg and Stukel 2017](#)). Major error sources in this framework are undercoverage error (some persons have no chance to be selected), nonresponse error (not all selected persons participate in a survey), sampling error (because we take a sample

1. Response rates in the different methods capture slightly different things. In the other register and walk methods, the interviewer must select one household member for the interview. This step is not necessary in the individual register method, because a respondent has already been selected. Thus there is an additional point where a refusal can occur in the other register and walk methods, which likely depresses response rates in these methods.

Table 2. Variables used in external benchmark comparison

Variable	Categories
Gender	Female; male
Age	15–24; 25–34; 35–44; 45–54; 55–64; 65–74; 75+
Marital status	Legally married; not legally married
In paid work	Working at least one hour in last week
Nationality	Citizen of country; not citizen
Household size	1; 2; 3; 4; 5+

rather than a census), and measurement error (the response given does not match the true value). Sampling error can be accounted for using traditional standard errors (although one measure developed here does not have a recognized standard error). Measurement error is hopefully not a concern for the demographic variables used in this analysis, as discussed in more detail below. Thus, we focus on undercoverage and nonresponse. Following [Klausch, Hox, and Schouten \(2015\)](#), we refer to these error sources together as *selection bias*.

We have two reasons for combining nonresponse and undercoverage. First, there is often a trade-off between the two. One way to increase response rates is to undercover those with low response propensities ([Eckman and Kreuter 2017](#)). Considering the two error sources jointly protects against these sorts of manipulations. Second, the two errors have similar effects on the data: in the [Groves \(1989\)](#) framework, both are errors of nonobservation. A more complete picture of how well a survey represents its target population can therefore emerge by considering undercoverage and nonresponse together.

We use two measures to evaluate the selection bias in the ESS data. The first benchmarks the ESS against an external data source of high quality, Eurostat's Labour Force Survey. The second uses only internal data derived from the survey itself.

EXTERNAL MEASURE OF SELECTION BIAS

Most of the countries that participate in the ESS also conduct a yearly Labour Force Survey (LFS) for Eurostat to capture trends in employment and labor force participation over time. [Table 2](#) shows the six variables that are captured in both the ESS and LFS. These variables often have been investigated as possible correlates of survey participation ([Groves and Couper 1992](#), section 5.1). [Table 3](#) indicates the countries and years in which both the ESS and the LFS were conducted in a country. In these countries and rounds, we can compare the ESS to the LFS on these six variables to develop a measure of how well the ESS represents the population. [Table 3](#) also indicates the sampling method used in each country-round. These 156 country-rounds make up the current analysis data set.

There are several reasons to expect that LFS data are of sufficient quality to be used as an external benchmark. First, in 40 percent of the countries in [table 3](#), participation in the LFS is mandatory. Second, the LFS uses a much larger sample than the ESS, approximately 140,000 adults per country on average in 2012. Third, the LFS poststratifies samples to population totals, which should reduce both sampling variance and bias because of undercoverage and nonresponse ([Eurostat 2013](#)).

To produce the external measure of selection bias, we use the index of dissimilarity, which often is used to quantify the extent of racial or other forms of residential segregation ([Duncan and Duncan 1955](#); [Sakoda 1981](#)). The index reports the percent of respondents in the ESS who would have to change their categorization for the ESS data to match the LFS data. [Blohm \(2006\)](#), [Ortmanns and Schneider \(2015\)](#), [Biemer et al. \(2017\)](#), and [Gummer \(2017\)](#) similarly use the index to quantify differences between surveys.

To calculate the index, we estimate, for each country-round indicated in [table 3](#) and each variable in [table 2](#), the percent of the population falling into each category. These estimates are weighted by the DWEIGHT variable, which adjusts for the inverse of the selection probability of each respondent.² We then calculate the same estimates from the LFS data, using the COEFF weight, which adjusts for selection and includes a poststratification adjustment to population totals.

Let $p_{C,r,k}^{ESS}$ be the weighted proportion from the ESS of the share of persons in country C and round r who fall into category k of variable p , and $p_{C,r,k}^{LFS}$ be the corresponding proportion estimated from the LFS. Then the index for variable p in country C and round r is:

$$D_{p,C,r} = 0.5 * \sum_k |p_{C,r,k}^{ESS} - p_{C,r,k}^{LFS}| * 100 \quad (1)$$

The use of absolute values ensures that positive and negative differences do not cancel each other out. Multiplying by 100 puts the external and internal measures on approximately the same scale. We do not provide standard errors for the external measure, because there is no agreement in the literature about how to estimate the standard error of $D_{p,C,r}$ ([Mulekar, Knutson, and Champanerkar 2008](#)). In addition, not all countries in the ESS provide a clustering variable that allows us to calculate standard errors that account for the survey design.

We then average these variable-level dissimilarity measures, $D_{p,C,r}$ across all six variables (or as many as are available in a given country-round; see [table 3](#)).

2. The ESS also provides a poststratified weight, but we do not use it here, because our interest is in understanding what response rates can tell us about interviewer effects in sample selection. In this paper, we are not interested in the important question of whether poststratification can reduce the error.

Table 3. Data available from both the European Social Survey and the Eurostat Labour Force Survey, by country and round

Country	ESS round						
	1	2	3	4	5	6	7
	2002	2004	2006	2008	2010	2012	2014
Austria	W ^c	W	W				I
Belgium	I ^c	I	I	I	I	I	I
Bulgaria			W	W	W	W	
Croatia				O	O		
Cyprus			O	O	O	O	
Czech Republic	O	W		O	O	O	O
Denmark ^a	I ^c	I	I	I	I	I	I
Estonia		I	I	I ^b	I	I	I ^d
Finland ^a	I ^c	I	I	I	I ^c	I	I
France	W ^{bc}	W ^b	W	W	O	O	W
Germany	I ^c	I	I	I	I	I	I
Greece	W ^c	W		W	W		
Hungary	I ^c	I	O	I	I	I	I
Iceland ^a		I ^d				I ^d	
Ireland	O ^c	O	O	O	O	O	O
Italy	O ^c					I	
Lithuania					O	O	O
Luxembourg	O ^c	O					
Netherlands	O ^c	O	O	O	O	O	O
Norway ^a	I ^{cd}	I ^d	I ^d	I ^d	I ^d	I ^d	I ^d
Poland	I ^c	I	I	I	I	I	I
Portugal	W ^c	W	W	W	W	W	W
Slovenia	I ^c	I	I	I	I	I	I
Slovakia		I	I	W	W	W	
Spain	O ^c	I	I	I	I	I	I
Sweden ^a	I ^c	I	I	I	I ^d	I ^d	I ^d
Switzerland ^a	O ^c	O	O	O	I	I	I
United Kingdom	O ^c	O	O	O	O	O	O

NOTE.—I = Individual Register; O = Other Register; W = Walk.

^aHousehold size data missing in LFS in all years.

^bMarital status data missing in LFS.

^cNationality data missing in LFS.

^dData on persons 75 and older missing from LFS.

^eMarital status data collected with error in ESS, not used.

$$\bar{D}_{C,r} = \frac{1}{6} \sum_k D_{p,C,r}$$

$\bar{D}_{C,r}$ is a measure of selection bias for country C and round r . Blohm (2006) and Biemer et al. (2017) also use the average index of dissimilarity to summarize survey data quality across variables.

$\bar{D}_{C,r}$ can vary from 0 to 100, where 0 means that the ESS and LFS data match exactly on all categories of the six variables. The higher the number, the larger the difference between the ESS and LFS survey.³

This approach to measuring selection bias relies on several assumptions. The first is that the LFS data are of appropriate quality: we provided arguments for that assertion above. The second is that the population being measured is the same in both surveys. Because the ESS covers only the noninstitutionalized population 15 years and older, we drop the LFS cases that reside in institutions and all persons younger than 15. In some countries and rounds, the LFS includes only people younger than 75 (see table 3), and we drop cases in the ESS to match. The third assumption underlying the external measure is that the demographic variables we can test with this method are related to other variables of interest in the survey, for which external reference data are not available. Unfortunately, high-quality external benchmark data for the more substantive variables collected in the ESS are not available.

The fourth assumption is that measurement error does not bias the comparison between the ESS and LFS surveys, which is satisfied if either measurement error does not exist in these variables or if the error is the same in the two surveys. The wording of the relevant ESS questions is provided in Appendix A. Eurostat does not mandate how questions should be worded, and the wording and language used in the LFS vary by country.⁴ Measurement error should not be a serious problem in the six variables used, because demographic variables are less susceptible than attitudes to measurement error (Sakshaug, Yan, and Tourangeau 2010). Moreover, the variables do not include any that are difficult to measure in a comparable way in a cross-cultural context, such as education (Schneider 2010; Ortmanns and Schneider 2016). However, some measurement error difference between the two surveys may bias our external measure of selection bias.

3. We considered other measures that make use of external data. The R-indicator (Schouten, Cobben, and Bethlehem 2009) requires frame data about each selected household or person, which is not available for most country-rounds. The Kullback–Leibler divergence measure quantifies how similar two distributions are (Kullback and Leibler 1951). The average of this divergence measure across our six variables for each country-round was highly correlated with $\bar{D}_{C,r}$ (Pearson correlation coefficient 0.87).

4. For more information on the LFS questionnaires, see http://ec.europa.eu/eurostat/statistics-explained/index.php/EU_labour_force_survey_-_methodology (accessed March 21, 2019).

INTERNAL MEASURE OF SELECTION BIAS

Our second measure uses only survey data and exploits the fact that 50 percent of the persons in gender-heterogeneous couples are male and 50 percent are female. To calculate the internal measure, we use only those respondents age 15 years and older who live with a partner of the opposite sex in the same household. Among this subset, the chance of being selected is the same for the male and female partners, and thus the weighted respondent sample should be approximately 50 percent female. The proportion might vary because of sampling error, but we can control for that, as discussed below. Alternatively, it could deviate from 50 percent because of differential undercoverage and non-response by gender. If men and women in a country differ in their accessibility and amenability to respond, we could see an overrepresentation of the gender that is easier to interview. If interviewers notice that women (for example) are more likely to respond, they may change their behavior: when selecting housing units, they may select those where a woman is at home; when selecting respondents within a household, they may force the selection of women. Interviewers also might make more attempts to complete the interview with female respondents, believing them to have higher response propensities.

To calculate the internal measure of selection bias, we estimate the percentage of respondents in gender-heterogeneous couples who are female for each country C and round r , using the design weight (DWEIGHT) provided by the ESS. We then estimate the bias as the difference from 50 percent female divided by the standard error of the estimate of the percent female:

$$I_{c,r} = \frac{\% \text{ female}_{c,r} - 50}{\sqrt{50*50/n}}$$

This statistic follows a normal distribution: values outside the range of $[-1.96, 1.96]$ are likely the result of differential nonresponse or undercoverage by gender.⁵

Other studies have used this measure of selection bias (Hoffmeyer-Zlotnik 2006; Sodeur 1997, 2007; Kohler 2007; Menold 2014). Sodeur (2007), Kohler (2007), and Menold (2014) have investigated when this measure is positive (the realized sample includes too many women) and when it is negative (too many men). Our interest, however, is in selection bias generally, and for this reason we take the absolute value of $I_{c,r}$ as our internal measure. Higher values of $|I_{c,r}|$ indicate lower data quality, just as in the external measure, which eases interpretation. Kohler (2007) and Menold (2014) also use the absolute value in some of their analyses.

5. We do not adjust the standard error for clustering, because many countries do not release the clustering indicators necessary to do so and because gender is spatially heterogeneous.

This approach to assessing sample representativity can be used with only one variable, gender, and it refers only to the subset of respondents in gender-heterogeneous couples. The share of such couples in the sample varies from 43 percent to 72 percent across our country-rounds. The external measure has its own shortcomings, as discussed above. Because we use both measures and show that they lead to the same substantive results, our findings are robust to these shortcomings.

EXPLORATION OF THE TWO MEASURES

Figure 1 shows the distributions of the external and internal measures of selection bias and a scatterplot of the relationship between them. The external measure, $\bar{D}_{c,r}$, is shown on the horizontal axis. It ranges from 1.0 to 8.9 across the 156 observations in our data set, but most values are concentrated between two and four. High values indicate that the ESS deviates from the LFS and thus does a poorer job of representing the population.

The internal measure of selection bias, $|I_{c,r}|$, varies from 0.04 to 6.3; again, higher values mean more error. Most country-rounds (62 percent) lie below 1.96, indicating that the deviation of the percentage of women observed in gender-heterogeneous couples may be the result of sampling error and not selection bias.

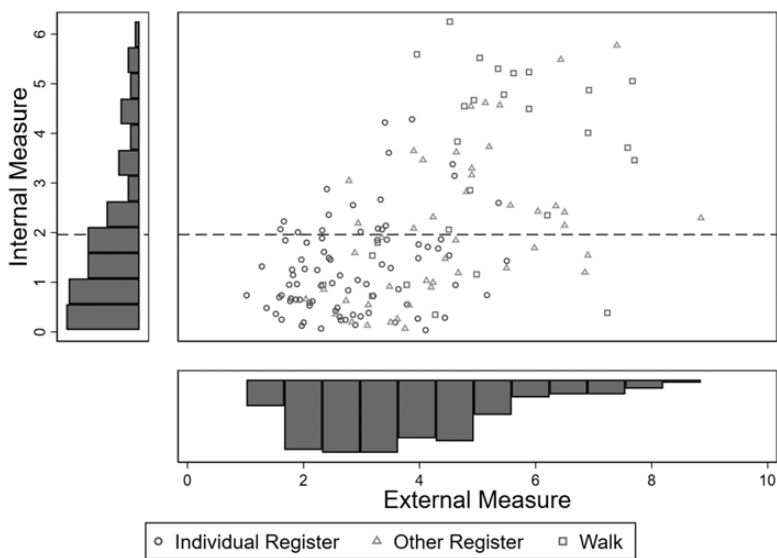


Figure 1. Histograms and scatterplot of two measures of selection bias. Area above dashed line is the rejection region on the internal measure: $|I_{c,r}| > 1.96$.

In the scatterplot in the main body of the figure, each symbol represents one country-round. The Spearman correlation coefficient between the two measures is 0.53.

Results

We now turn to addressing our research questions: whether sampling methods involving interviewers result in lower data quality and how sampling method moderates the relationship between response rates and data quality.

RELATIONSHIP BETWEEN RESPONSE RATES AND SELECTION BIAS

We first plot the two measures of selection bias against the response rate, separately by sample type, exposing an interesting relationship between response rate and data quality. Figure 2 shows separate graphs in each column for each of the three sampling methods: individual register, other register, and walk; the top row corresponds to the external measure and the bottom row to the internal measure.

In these graphs, we are interested in two issues, corresponding to our research questions. First, is there more selection bias when interviewers are involved in selection? That is, does selection bias increase as we move from left to right in each row? Second, we are interested in the slope of the lines, which quantify the relationship between selection bias and response rates. If there is no relationship between response rates and selection bias, relatively flat regression lines should exist in figure 2. If the oft-assumed relationship between bias and response rates holds, downward-sloping lines should appear, reflecting the association between high response rates and less bias.

As shown in the upper left panel, all country-rounds using the individual register sampling method have relatively low external measures, lower than the other methods in the top row of graphs. Furthermore, there is no significant relationship between response rate and error. A simple linear regression line through the points in this graph slopes slightly downward but is not significantly different than zero.

Moving right to the middle graph in the top row, the country-rounds using the other register sampling methods have slightly higher measures on average, meaning the samples are less representative than those selected from individual registers. Again, there is no significant relationship between the variables on the horizontal and vertical axes. In the third panel of the top row, the values are also high. The linear regression line in this graph is upward sloping, and the slope is significant. Higher response rates are associated with *more* selection bias ($\beta = 0.053$; $se = 0.011$) in the walk methods.⁶

6. Standard errors on the regression coefficients depicted in figure 2 account for clustering of country-rounds within countries.

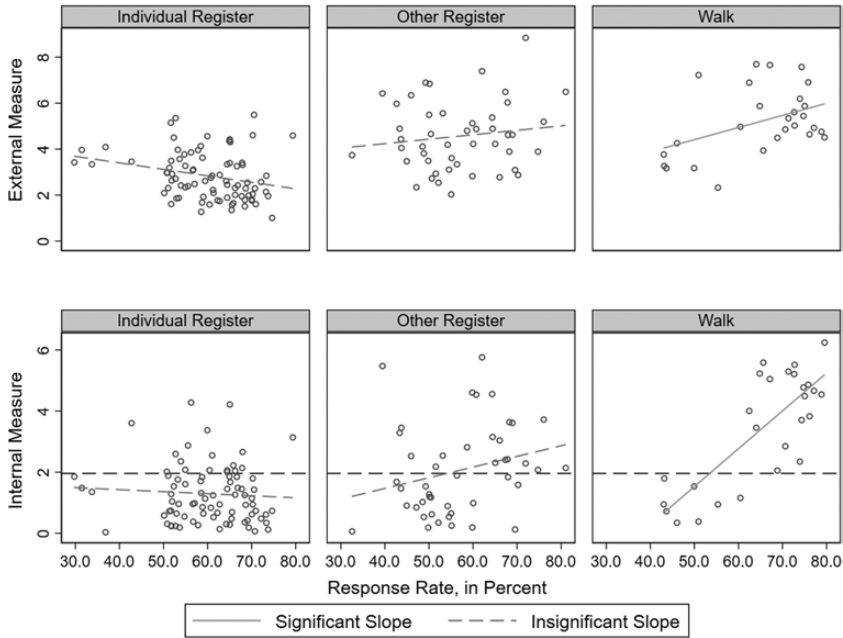


Figure 2. Two measures of selection bias, by response rate and sample type. Area above horizontal dashed line is the rejection region on the internal measure: $|I_{c,r}| > 1.96$.

The results are similar in the bottom row, which shows the relationships between response rates and the internal measure of selection. Among the country-rounds using the individual register method of sample selection, most (78 percent) fall inside the $[0, 1.96]$ region, meaning that the observed deviations from 50 percent female may be explained entirely by sampling error. Among those country-rounds using one of the other register methods, 48 percent show gender ratios that are significantly different from 50 percent. In the third graph of the bottom row, showing surveys using the walk methods of sample selection, 70 percent of the country-rounds deviate significantly from 50 percent female. In this panel, there is again a positive relationship between response rate and the internal measure, suggesting that high response rates are associated with more selection bias ($\beta = 0.12$; $se = 0.0078$). Our two measures of selection bias give the same results.

Figure 2 suggests that interviewers' influence on the selection process is associated with lower data quality and that sampling method moderates the relationship between response rate and selection bias. However, this analysis does not control for confounding variables or for cultural differences between countries that may impact the results. The next section uses regression models to further investigate these findings.

REGRESSION MODELS OF SELECTION ERROR

We ideally would like to estimate the causal effect of sample type and response rate on the two measures of selection bias. However, sample types and response rates are not randomly assigned: they are a function of the resources, culture, and survey environment in the participating countries. Some countries tend to have lower response rates than others, and response rates may also vary over time. Other countries do not have registers and thus cannot use some sampling methods. For these reasons, we cannot eliminate bias in our estimates of the relationship between sample type and response rate on data quality.

Random-effects models are used here to remove as much endogeneity bias as possible. The external and internal measures discussed above are the dependent variables in the models. Because both are continuous measures, linear regression is used. Response rate, sample type, and their interactions are the independent variables in all models. No other country-round characteristics, such as the use of backchecks and switch in data collection contractor, are used, because Menold (2014, section 5.3) found that these variables do not explain data quality. Furthermore, the ESS specifications require a certain level of quality and thus there is little room for variation across country-rounds (European Social Survey 2013, section 7.6).

To control for round- and country-specific influences and isolate the contribution of response rates and sample type on representativity, all models include random intercepts for countries and rounds. Letting C index the countries and r the rounds, the cross-classified multilevel models are:

$$\bar{D}_{C,r} = \alpha_C + \phi_r + \vec{\beta} * \vec{X} + \varepsilon_{C,r} \quad (2)$$

$$\alpha_r \sim N(0, \theta_{ext, countries})$$

$$\phi_r \sim N(0, \theta_{ext, rounds})$$

$$|I_{C,r}| = \delta_C + v_r + \vec{\gamma} * \vec{X} + \zeta_{C,r} \quad (3)$$

$$\delta_C \sim N(0, \theta_{int, countries})$$

$$v_r \sim N(0, \theta_{int, rounds})$$

where the \vec{X} vector includes indicators for the other register and walk sampling methods; indicators for the second and third tertiles of the response rate; and indicators for the interactions of sampling method and response rate tertiles. The response rate tertiles are 29.7 to 56.2 percent; 56.3 to 66.6 percent; and 66.8 to 81.0 percent. There are at least five country-rounds in each tertile in each sampling method. These models were fit in Stata 15.1 with the mixed

command (StataCorp LP 2014). Both models passed the link test for specification error (Pregibon 1980).⁷

Estimated coefficients and confidence intervals are given in figure 3. The two panels correspond to the external (left) and internal (right) measures of selection bias. Coefficients that are significant at the 5 percent level are those where the confidence interval does not cross the vertical line at zero. The case base for both models is the 156 country-rounds shown in table 3.

As reflected in the estimates in the first panel, samples selected via the other register methods have significantly higher selection bias than those using the individual register method (the reference category) when the response rate is held at its reference category, the lowest tertile. In the middle section of the first panel, the association between response rate and the external measure is not statistically significant in the reference sampling method (individual register).⁸ This result contradicts many researchers' expectations that higher response rates are associated with less selection bias. However, it confirms the data in the top left panel of figure 2: in the individual register sampling method, there is no evidence of a relationship, positive or negative, between response rate and data quality.

The bottom set of estimates in this panel displays coefficients on the interaction of sampling method and response rate. In the other register methods, there is no evidence for a positive or negative relationship between response rate and selection bias. In the walk methods, there is evidence of a positive relationship between response rates and the external measure of selection bias: the higher the response rate, the more the ESS data deviates from the LFS data.

In the second panel of figure 3, the results are quite similar. There is no significant relationship between response rate and selection bias in the individual register method. In the other register methods, there is some evidence that country-rounds in the second tertile show more bias than those in the first tertile: a joint test that the coefficients on the second and third tertiles are both 0 rejects ($\chi^2 = 13.3$; $p = 0.0013$). In the walk methods, response rates in the second and third tertiles are again associated with higher selection bias.

The results in figure 3 are largely consistent with those in figure 2, but here the findings are more robust because of the additional controls in the model. The positive relationship between response rates and selection bias in the walk

7. We tested many models, including random slopes on sample type and response rate and different parameterizations of the response rate. The specific method used made no substantive difference to our results, and a likelihood ratio test comparing the more complex models to the simpler models provided no evidence that the complex models explained the relationship better. Models using the variable-level external measure ($D_{p,c,r}$ in equation 1) as the dependent variable, where the variables were nested within countries, came to substantively the same conclusions as the model given in equation 2 and showed the same patterns of significance.

8. The coefficient on the second tertile is $\beta' - 0.58$, $se = 0.295$, just barely significant at the 5 percent level. The coefficient on the third tertile is not significant, and a joint test that both coefficients are 0 fails to reject.

and, to a lesser extent, other register methods is in contrast to the common expectation that high response rates are a signal of data quality.

As a further check on these results, we reran the two models shown in equations 2 and 3 on the 45 country-rounds conducted in countries that changed sampling method at least once during the seven rounds. These results are given in Appendix B. The findings are quite similar to those shown in figure 3: increased selection bias with higher response rates in the walk methods; no relationship between selection bias and response rates in the other methods.

Discussion of Results

The relationships between response rates, sampling method, and selection bias are more complex than previous literature has recognized. When interviewers are not involved in selecting cases for a survey, the present evidence suggests that response rates are unrelated to selection bias. However, when interviewers are involved in sample selection, high reported response rates are associated with more selection bias. There is no evidence in this study that high response rates are a sign of good-quality data. Our findings are robust to alternative model specifications and two measures of selection bias. These results are consistent with the ideas put forth in the introduction, that interviewers may

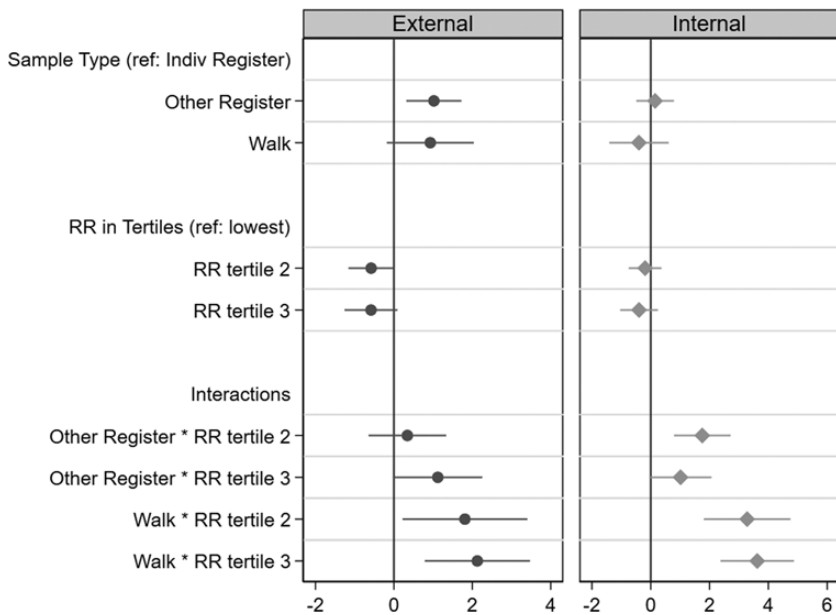


Figure 3. Estimated coefficients and 95 percent confidence intervals from models ($n = 156$). Models also include random round and country intercepts (not shown). Figure created with the coefplot package in Stata (Jann 2013).

manipulate selection of respondents because of incentives to keep response rates high. An interviewer may believe that high response rates are important to job security, or her pay may be tied to her response rates. Our data do not allow us to say for certain that this behavior is taking place, but the behavior is consistent with the results we find.

The methods used in this paper have several limitations, which we have attempted to mitigate as much as possible. First, both our selection bias measures rely on assumptions. The external measure assumes that the LFS data are more accurate than the ESS measures, that the demographic variables for which we have benchmarks are related to other substantive variables, and that differential measurement error does not bias the comparisons. The internal measure assumes that one estimate, calculated on a subset of the sample, is indicative of the selection bias in the full sample. We hope that by using these two measures together, our results are robust. Second, we cannot be sure that we have entirely removed endogeneity from our regression models and thus cannot make causal arguments based on our results. Third, we have used reported response rates in this analysis. Although the ESS tries to standardize response rates across countries, any manipulation in the recording of disposition codes would bias our results. Fourth, we were unable to control for interviewer payment methods, which likely also influence interviewer incentives and data quality. In the ESS, only Finland, Norway, and Sweden pay interviewers by the hour, meaning we cannot separate payment effects from a regional effect.

On the other hand, the positive relationship between response rate and selection bias detected in this study for countries using the walk sampling methods may be underestimated, for two reasons. First, the ESS specifies that selection in the housing unit listing and the random walk methods be performed by someone other than the interviewer, to reduce incentives for manipulation of the selection routine. Other surveys using random walk do not include this instruction; in these surveys, selection bias in the listing and random walk methods may be larger. The second reason our results may underestimate the relationship between response rate and selection bias is the collapsing of the sampling methods shown in [table 1](#). The methodology reports we used to code the sampling method are not detailed enough to allow us to reliably distinguish the listing and random walk methods, and thus we combined them for the purposes of the analyses. The positive relationship between response rates and selection bias may be particularly strong in the random walk method and weaker in the listing method: our analysis could not detect this result.

The most important issue for researchers who rely on survey data is how we can prevent manipulation of selection by interviewers. We recommend using sampling methods that minimize interviewer selection, as far as possible. Improved training and supervision of interviewers could also reduce interference in the selection process. If interviewers did not feel pressured to achieve high response rates, they might allow the selection process to be fully random

and selection bias would be smaller. (However, it is possible that too little pressure on interviewers could result in a different relationship between response rates and quality.) Another idea is to develop improved interviewer metrics to detect undesirable behavior. For example, interviewers who complete many interviews on the first contact, controlling for experience and other relevant variables, might be flagged for additional quality control. We also recommend GIS-based sampling methods that allow checks on some interviewer behavior via GPS devices (Himelein, Eckman, and Murray 2014; Wagner, Olson, and Edgar 2017; Eckman, Himelein, and Dever 2018).

Ultimately, much of the pressure for high response rates that interviewers feel comes from researchers, data users, and journal editors who continue to demand high response rates. More education is needed, at all levels, about the appropriate role of response rates. We hope that by demonstrating that high (reported) response rates do not always indicate data quality, and in fact correlate with increased error in some sampling methods, we can reduce the emphasis on response rates as the main indicator of quality in survey research.

Appendix A. Text of Questions in European Social Survey

The question text below is from Round 6 of the ESS. Some of the questions changed slightly between the waves.

Gender. CODE SEX (Male; Female)

Age. Derived from question on year of birth (F3. And in what year were you born?)

Marital status. Post-coded variable derived from answers to two questions (the set of response options actually used varies between countries):

F6 CARD 49 You just told me that you live with your husband / wife / partner. Which one of the descriptions on this card describes your relationship to them? (Legally married; In a legally registered civil union; Living with my partner (cohabiting)—not legally recognised; Living with my partner (cohabiting)—legally recognised; Legally separated; Legally divorced / Civil union dissolved; Don't know)

IF NOT LIVING WITH A HUSBAND / WIFE / PARTNER OR ARE COHABITING:

F11 CARD 50 This question is about your legal marital status not about who you may or may not be living with. Which one of the descriptions on this card describes your legal marital status now? (Legally married; In a legally registered civil union; Legally separated; Legally divorced / Civil union dissolved; Widowed / Civil partner died; None of these (NEVER married or in legally registered civil union); Don't know)

Work status. Post-coded variable from two questions: F17a (code 01) and F18

F17a CARD 53 Using this card, which of these descriptions applies to what you have been doing for the last 7 days? Select all that apply. (In paid

work (or away temporarily) (employee, self-employed, working for your family business); In education, (not paid for by employer) even if on vacation; Unemployed and actively looking for a job; Unemployed, wanting a job but not actively looking for a job; Permanently sick or disabled; Retired; In community or military service; Doing housework, looking after children or other persons; Other; Don't know)

F17a Can I just check, did you do any paid work of an hour or more in the last seven days? (Yes; No; Don't know)

Nationality. Are you a citizen of [country]? (Yes; No; Don't know)

Household size. F1 Including yourself, how many people including children live here regularly as members of this household? WRITE IN NUMBER: (two digits); Don't know

Appendix B. Results of Regression Models for Countries Changing Sampling Method

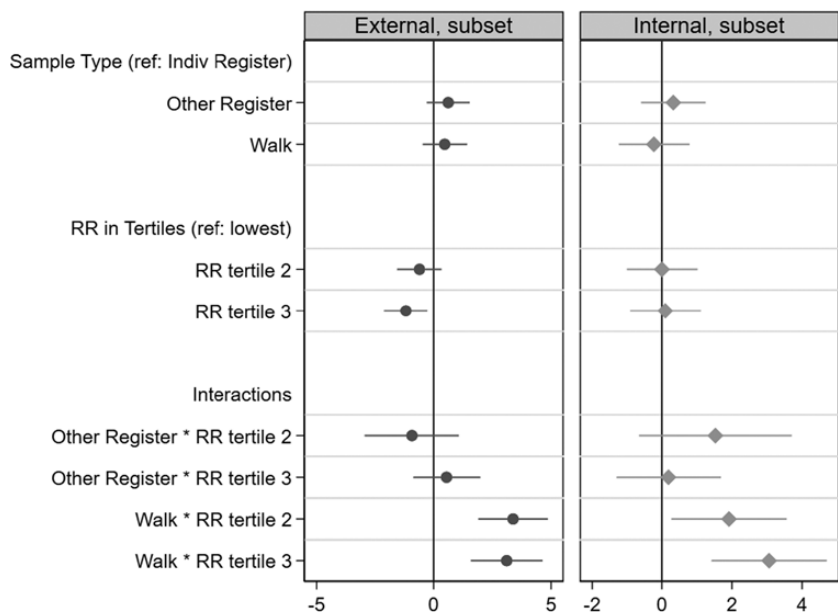


Figure B1. Estimated coefficients and 95 percent confidence intervals from models run on countries changing sampling methods ($n = 45$). Models also include random round and country intercepts (not shown). Figure created with the coefplot package in Stata (Jann 2013).

References

- Afrobarometer Network. 2017. "Afrobarometer Round 7 Survey Manual." Available at http://afrobarometer.org/sites/default/files/survey_manuals/ab_r7_survey_manual_en1.pdf, accessed March 21, 2019.
- Alt, Christian, Walter Bien, and Dagmar Krebs. 1991. "Wie zuverlässig ist die Verwirklichung von Stichprobenverfahren? Random Route versus Einwohnermeldeamtstichprobe." *ZUMA-Nachrichten* 28:65–72.
- American Association for Public Opinion Research. 2003. "Interviewer Falsification in Survey Research: Current Best Methods for Prevention, Detection and Repair of Its Effects." Available at https://www.aapor.org/aapor_main/media/mainsitefiles/falsification.pdf, accessed March 21, 2019.
- . 2016. "Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys." Available at https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf, accessed March 21, 2019.
- Athey, K. R., Joan E. Coleman, Audrey P. Reitman, and Jenny Tang. 1960. "Two Experiments Showing the Effect of the Interviewers Racial Background on Responses to Questionnaires Concerning Racial Issues." *Journal of Applied Psychology* 44:244–46.
- Bailar, Barbara, Leroy Bailey, and Joyce Stevens. 1977. "Measures of Interviewer Bias and Variance." *Journal of Marketing Research* 14:337–43.
- Bauer, Johannes J. 2014. "Selection Errors of Random Route Samples." *Sociological Methods and Research* 43:519–44.
- . 2016. "Biases in Random Route Surveys." *Journal of Survey Statistics and Methodology* 4:263–87.
- Beullens, Koen, Hideko Matsuo, Geert Loosveldt, and Caroline Vandenplas. 2014. "Quality Report for the European Social Survey Round 6 London." Available at https://www.europeansocialsurvey.org/docs/round6/methods/ESS6_quality_report.pdf, accessed March 21, 2019.
- Biemer, Paul P., Joe Murphy, Stephanie Zimmer, Chip Berry, G. Deng, and Katie Lewis. 2017. "Using Bonus Monetary Incentives to Encourage Web Response in Mixed-Mode Household Surveys." *Journal of Survey Statistics and Methodology* 6:240–61.
- Bien, Walter, Donald Bender, and Dagmar Krebs. 1997. "DJI-Familienurvey: Der Zwang, mit Unterschiedlichen Stichproben zu Leben." In *Stichproben in der Umfragepraxis*, edited by Siegfried Gabler and Jürgen H. P. Hoffmeyer-Zlotnik, 33–42. Wiesbaden: Westdeutscher Verlag.
- Blohm, Michael. 2006. "Data Quality in Nationwide Face-to-Face Social Surveys." Available at <https://webarchive.nationalarchives.gov.uk/20160129122503/http://www.ons.gov.uk/ons/about-ons/get-involved/events/events/q2006---european-conference-on-quality-in-survey-statistics-24-26-april-2006/agenda/session-17-wednesday.pdf>, accessed March 21, 2019.
- Blohm, Michael, and Achim Koch. 2015. "Führt eine Höhere Ausschöpfung zu Anderen Umfrageergebnissen? Eine Experimentelle Studie zum ALLBUS 2008." In *Nonresponse Bias. Qualitätssicherung sozialwissenschaftlicher Umfragen*, edited by Jürgen Schupp and Christof Wolf, 85–129. Wiesbaden, Germany: Springer VS.
- Brick, J. Michael, and Roger Tourangeau. 2017. "Responsive Survey Designs for Reducing Nonresponse Bias." *Journal of Official Statistics* 3:735–52.
- Duncan, Otis Dudley, and Beverly Duncan. 1955. "A Methodological Analysis of Segregation Indexes." *American Sociological Review* 20:210–17.
- Durbin, J., and A. Stuart. 1951. "Differences in Response Rates of Experienced and Inexperienced Interviewers." *Journal of the Royal Statistical Society: Series A (General)* 114:163–206.
- Eckman, Stephanie. 2013. "Do Different Listers Make the Same Housing Unit Frame? Variability in Housing Unit Listing." *Journal of Official Statistics* 29:249–59.

- . 2017. "Interviewers' Expectations of Response Propensity Can Introduce Nonresponse Bias in Survey Data." *Statistical Journal of the IAOS* 33:231–34.
- Eckman, Stephanie, Kristen Himelein, and Jill Dever. 2018. "Innovative Sample Designs Using GIS Technology." In *Advances in Comparative Survey Methods: Multicultural, Multinational and Multiregional Context*, edited by Timothy P. Johnson, Beth-Ellen Pennell, Ineke Stoop, and Brita Dorer, 67–92. New York: Wiley.
- Eckman, Stephanie, and Frauke Kreuter. 2013. "Undercoverage Rates and Undercoverage Bias in Traditional Housing Unit Listing." *Sociological Methods and Research* 42:264–93.
- . 2017. "The Undercoverage-Nonresponse Trade-Off." In *Total Survey Error in Practice*, edited by Paul P. Biemer, Edith de Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars Lyberg, Clyde Tucker, and Brady West, 95–113. New York: Wiley.
- European Social Survey. 2013. "Round 7 Specification for Participating Countries." Technical Report.
- Eurostat. 2013. "Labour Force Survey in the EU, Candidate and EFTA Countries: Main Characteristics of National Surveys, 2012." Technical Report, ISSN 2315-0807. Luxembourg: Publications Office of the European Union.
- Feldman, J. J., Herbert Hyman, and C. W. Hart. 1951. "A Field Study of Interviewer Effects on the Quality of Survey Data." *Public Opinion Quarterly* 15:734–61.
- Franzen, Raymond, and Robert Williams. 1956. "A Method for Measuring Error Due to Variance Among Interviewers." *Public Opinion Quarterly* 20:587–92.
- Gaziano, Cecillie. 2005. "Comparative Analysis of Within-Household Respondent Selection Techniques." *Public Opinion Quarterly* 69:124–57.
- Gray, Percy G. 1956. "Examples of Interviewer Variability Taken from Two Sample Surveys." *Journal of the Royal Statistical Society: Series C* 5:73–85.
- Groves, Robert M. 1989. *Survey Errors and Survey Costs*. Hoboken, NJ: John Wiley and Sons.
- Groves, Robert M., and Mick Couper. 1992. *Nonresponse in Household Surveys*. Hoboken, NJ: John Wiley and Sons.
- Groves, Robert M., and Emilia Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis." *Public Opinion Quarterly* 72:167–89.
- Gummer, Tobias. 2017. "Assessing Trends and Decomposing Change in Nonresponse Bias: The Case of Bias in Cohort Distributions." *Sociological Methods & Research* 48:92–115.
- Häder, Michael. 2010. *Empirische Sozialforschung. Eine Einführung*. Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.
- Himelein, Kristen, Stephanie Eckman, and Siobhan Murray. 2014. "Sampling Nomads: A New Technique for Remote, Hard-to-Reach, and Mobile Populations." *Journal of Official Statistics* 30:191–213.
- Hoffmeyer-Zlotnik, J. H. P. 2006. "Stichprobenziehung in der Umfragepraxis: Die unterschiedlichen Ergebnisse von Zufallsstichproben in face-to-face Umfragen." In *Stichprobenqualität in Bevölkerungsumfragen*, edited by Frank Faulbaum and Christof Wolf, 19–36. Bonn: Informationszentrum Sozialwissenschaften.
- Iannacchione, Vincent G. 2011. "The Changing Role of Address-Based Sampling in Survey Research." *Public Opinion Quarterly* 75:556–75.
- Jann, Ben. 2013. *COEFPLOT: Stata Module to Plot Regression Coefficients and Other Results*. Statistical Software Components, Boston College Department of Economics. Available at <https://ideas.repec.org/c/bocode/s457686.html>, accessed March 21, 2019.
- Keeter, Scott, Carolyn Miller, Andrew Kohut, Robert M. Groves, and Stanley Presser. 2000. "Consequences of Reducing Nonresponse in a National Telephone Survey." *Public Opinion Quarterly* 64:125–58.
- Klausch, Thomas, Joop Hox, and Barry Schouten. 2015. "Selection Error in Single- and Mixed Mode Surveys of the Dutch General Population." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178:945–61.

- Koch, Achim. 1995. "Gefälschte Interviews: Ergebnisse der Interviewerkontrolle beim ALLBUS 1994." *ZUMA Nachrichten* 19:89–105.
- . 1997. "ADM-Design und Einwohnermelderegister-Stichprobe- Stichprobenverfahren bei mündlichen Bevölkerungsumfragen." In *Stichproben in der Umfragepraxis*, edited by S. Gabler and Juergen H. P. Hoffmeyer-Zlotnik, 99–116. Weisbaden, Germany: VS Verlag für Sozialwissenschaften.
- . 1998. "Wenn 'mehr' nicht gleichbedeutend mit 'besser' ist: Ausschöpfungsquoten und Stichprobenverzerrungen in Allgemeinen Bevölkerungsumfragen." *ZUMA Nachrichten* 22:66–90.
- Kohler, Ulrich. 2007. "Surveys from Inside: An Assessment of Unit Nonresponse Bias with Internal Criteria." *Survey Research Methods* 1:55–67.
- Kullback, S., and R. A. Leibler. 1951. "On Information and Sufficiency." *Annals of Mathematical Statistics* 22:79–86.
- Lyberg, Lars E., and Diana M. Stukel. 2017. "The Roots and Evolution of the Total Survey Error Concept." In *Total Survey Error in Practice*, edited by Paul P. Biemer, Edith de Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars Lyberg, Clyde Tucker, and Brady West, 1–22. New York: Wiley.
- Lynn, Peter, and B. Taylor. 1995. "On the Bias and Variance of Samples of Individuals: A Comparison of the Electoral Registers and Postcode Address File as Sampling Frames." *Journal of the Royal Statistical Society: Series D* 44:173–94.
- Maineri, Angelica M., Annette Scherpenzeel, Johanna Bristle, Santa-Melissa Pflüger, Sarah Butt, Stephan Zins, Tom Emery, and Ruud Luijckx. 2017. "Evaluating the Quality of Sampling Frames Used in European Cross-National Surveys." Technical Report Deliverable Number: 2.2, Synergies for Europe's Research Infrastructures in the Social Sciences. European Union's Horizon 2020 research and innovation programme GA No: 654221. Available at https://pure.uvt.nl/ws/portalfiles/portal/18507342/Soc_Maineri_Report_Evaluating_the_quality_of_sampling_frames_Open_access_2017.pdf, accessed March 21, 2019.
- Matsuo, Hideko, and Geert Loosveldt. 2013. "Report on Quality Assessment of Contact Data Files in Round 5: Final Report 27 Countries." Available at https://www.europeansocialsurvey.org/docs/round5/methods/ESS5_response_based_quality_assessment_e01.pdf, accessed March 21, 2019.
- Menold, Natalja. 2014. "The Influence of Sampling Method and Interviewers on Sample Realization in the European Social Survey." *Survey Methodology* 40:105–23.
- Merkle, Daniel M., and M. Murray Edelman. 2002. "Nonresponse in Exit Polls: A Comprehensive Analysis." In *Survey Nonresponse*, edited by Don A. Dillman, John L. Eltinge, Robert M. Groves, and Roderick J. A. Little, 243–58. Hoboken, NJ: Wiley-Interscience.
- Mulekar, M. S., J. C. Knutson, and J. A. Champanerkar. 2008. "How Useful Are Approximations to Mean and Variance of the Index of Dissimilarity?" *Computational Statistics and Data Analysis* 52:2098–2109.
- O'Muircheartaigh, Colm, and Pamela Campanelli. 1999. "A Multilevel Exploration of the Role of Interviewers in Survey Non-Response." *Journal of the Royal Statistical Society: Series A* 162:437–46.
- Ortmanns, Verena, and Silke L. Schneider. 2015. "Harmonization Still Failing? Inconsistency of Education Variables in Cross-National Public Opinion Surveys." *International Journal of Public Opinion Research* 28:562–82.
- . 2016. "Can We Assess Representativeness of Cross-National Surveys Using the Education Variable?" *Survey Research Methods* 10:189–210.
- Pregibon, Daryl. 1980. "Goodness of Link Tests for Generalized Linear Models." *Applied Statistics* 29:15–24.
- Rosen, Jeffrey, Joe Murphy, Andy Peytchev, Sarah Riley, and Mark Lindblad. 2011. "The Effects of Differential Interviewer Incentives on a Field Data Collection Effort." *Field Methods* 23:24–36.
- Sakoda, J. M. 1981. "A Generalized Index of Dissimilarity." *Demography* 18:245–50.

- Sakshaug, Joseph, Ting Yan, and Roger Tourangeau. 2010. "Nonresponse Error, Measurement Error, and Mode of Data Collection: Tradeoffs in a Multi-Mode Survey of Sensitive and Non-Sensitive Items." *Public Opinion Quarterly* 74:907–1033.
- Scherpenzeel, Annette, Angelica M. Maineri, Johanna Bristle, Santa-Melissa Pfluger, Ilziya Mindarova, Sarah Butt, Stefan Zins, Tom Emery, and Ruud Luijckx. 2017. "Report on the Use of Sampling Frames in European Studies." Technical Report Deliverable Number: 2.2, Synergies for Europe's Research Infrastructures in the Social Sciences. European Union's Horizon 2020 research and innovation programme GA No: 654221.
- Schneider, S. L. 2010. "Nominal Comparability Is Not Enough: (In-)Equivalence of Construct Validity of Cross-National Measures of Educational Attainment in the European Social Survey." *Research in Social Stratification and Mobility* 28:343–57.
- Schnell, Rainer. 1997. *Nonresponse in Bevölkerungsumfragen: Ausmaß, Entwicklung und Ursachen*. Opladen: Leske + Budrich.
- Schouten, Barry, Fannie Cobben, and Jelke Bethlehem. 2009. "Indicators for the Representativeness of Survey Response." *Survey Methodology* 35:101–13.
- Sodeur, Wolfgang. 1997. "Interne Kriterien zur Beurteilung von Wahrscheinlichkeitsauswahlen." *ZA-Information* 41:58–82.
- . 2007. "Entscheidungsspielräume von Interviewern bei der Wahrscheinlichkeitsauswahl." *Methoden Daten Analysen* 1:107–30.
- StataCorp LP. 2014. *Stata Statistical Software: Release 15*. College Station, TX: StataCorp.
- Stoop, Ineke, Jaak Billiet, Achim Koch, and Rory Fitzgerald. 2010. *Improving Survey Response: Lessons Learned from the European Social Survey*. New York: Wiley.
- Sturgis, Patrick, Joel Williams, Ian Brunton-Smith, and Jamie Moore. 2017. "Fieldwork Effort, Response Rate, and the Distribution of Survey Outcomes: A Multilevel Meta-Analysis." *Public Opinion Quarterly* 81:523–42.
- Wagner, James, Kristen Olson, and Minako Edgar. 2017. "The Utility of GPS Data in Assessing Interviewer Travel Behavior and Errors in Level-of-Effort Paradata." *Survey Research Methods* 11:218–33.
- West, Brady T., and Annelies G. Blom. 2017. "Explaining Interviewer Effects: A Research Synthesis." *Journal of Survey Statistics and Methodology* 5:175–211.