Multi-Project Assessments of Sample Quality in Cross-National Surveys: The Role of Weights in Applying External and Internal Measures of Sample Bias

Jabkowski, Piotr; Cichocki, Piotr; Kołczyńska, Marta

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Jabkowski, P., Cichocki, P., & Kołczyńska, M. (2023). Multi-Project Assessments of Sample Quality in Cross-National Surveys: The Role of Weights in Applying External and Internal Measures of Sample Bias. *Journal of Survey Statistics and Methodology*, *11*(2), 316-339. https://doi.org/10.1093/jssam/smab027

*Journal of Survey Statistics and Methodology (2023) **11**, 316–339*

MULTI-PROJECT ASSESSMENTS OF SAMPLE QUALITY IN CROSS-NATIONAL SURVEYS: THE ROLE OF WEIGHTS IN APPLYING EXTERNAL AND INTERNAL MEASURES OF SAMPLE BIAS

PIOTR JABKOWSKI*
PIOTR CICHOCKI
MARTA KOŁCZYŃSKA

This paper examines existing methods of evaluating sample quality, showing that their practical utility and applicability to large-scale cross-project comparisons depends on whether they require auxiliary individual-level data. Among those methods that do not demand any such additional data, we differentiate between two approaches that rely on (i) external criteria, that is, comparisons of sample estimates to benchmarks derived from external population statistics, and (ii) internal criteria, that is, comparisons of subsample estimates to a theoretically derived aprioristic value. Our analyses demonstrate the advantages and limitations of both approaches based on an evaluation of 1,125 national surveys carried out in Europe between 2002 and 2016 within four survey projects: the Eurobarometer, European Quality of Life Survey, European Social Survey, and International Social Survey Programme. We show that the prevailing absence of design weights in cross-national survey datasets severely limits the applicability of external criteria evaluations. In contrast, using internal criteria without any weights proves acceptable because incorporating design weights in calculations of internal sample quality has only minor consequences for estimates of sample bias. Furthermore, applying internal criteria, we find that around 75 percent of

PIOTR JABKOWSKI is an Associate Professor and PIOTR CICHOCKI is an Assistant Professor, Faculty of Sociology, Adam Mickiewicz University, Szamarzewskiego 89C, 60-568 Poznań, Poland. MARTA KOICZYŃSKA is an Assistant Professor, Department of Research on Social and Institutional Transformations, Institute of Political Studies of the Polish Academy of Science, Polna 18/20, 00-625 Warsaw, Poland. This work was supported by grants awarded by the National Science Centre, Poland (grant numbers 2018/31/B/HS6/00403 and 2019/32/C/HS6/00421). The study design and analysis were not preregistered

*Address correspondence to Piotr Jabkowski, Faculty of Sociology, Adam Mickiewicz University, Szamarzewskiego 89C, 60-568 Poznań, Poland; E-mail: piotr.jabkowski@amu.edu.pl.

doi: 10.1093/jssam/smab027 Advance access publication 3 October 2021

samples in the four analyzed projects are not significantly biased. We also identify surveys with extremely high sample bias and investigate its potential sources. The paper concludes with recommendations regarding future research, which are directed at secondary data users, as well as producers of cross-national surveys.

KEYWORDS: Cross-national surveys; External and internal criteria of sample bias; Sample quality; Secondary analysis; Weights.

Statement of Significance

The paper examines methods for assessing sample quality available for screening data from cross-national survey projects by secondary data users. We demonstrate that the mainstream approach to such assessments in terms of external criteria, that is, benchmarking survey outcomes against known population characteristics, requires design weights. As many cross-national surveys do not publish design weights, we argue for broader use of an alternative approach—internal criteria assessments. We demonstrate that they can be used for evaluating sample bias without applying weights and that variables they require are commonly present in survey datasets. Using internal criteria, we examine sample bias in four European cross-national survey projects and discuss potential sources of bias among the outlying surveys flagged by the screening method.

1. INTRODUCTION

Researchers performing secondary analyses of cross-national surveys would benefit from easy-to-apply methods for assessing data quality because such methods would allow them to make informed choices when selecting data for analysis. Sample quality, that is, the extent to which the sample represents the specified target population (Alter and Hershfield 2014), constitutes a crucial component of Total Survey Error (Groves and Lyberg 2010), with a potential impact on the accuracy of sample-based inferences. Other aspects of Total Survey Error are no less important, most notably measurement quality (Groves et al. 2011; Alter and Hershfield 2014; Pennell, Cibelli Hibben, Lyberg, Mohler, and Worku 2017); however, indicators of sample quality are unique in that they can be used to screen surveys from cross-national multiwave comparative studies (henceforth, projects) to flag suspicious samples, which merit additional attention before their incorporation into substantive comparisons.

Although multiple methods of assessing sample quality exist, most require additional information beyond what is routinely made publicly available, either regarding the details of the survey process or information about nonrespondents. As a result, the sample quality assessments have typically been carried

out within a given survey project to ascertain the comparability of survey estimates (Beullens, Matsuo, Loosveldt, and Vandenplas 2014; Koch, Halbherr, Stoop, and Kappelhof 2014), as well as to provide methodological lessons for future research (e.g., Lynn, Häder, Gabler, and Laaksonen 2007; Smith 2007; Stoop, Billiet, Koch, and Fitzgerald 2010). While cross-project assessments have been much less common, the few studies featuring such comparisons have found systematic differences in sample quality between survey projects and pointed to potential reasons for this variation (Kohler 2007; Ortmanns and Schneider 2016). More comprehensive cross-project analyses would allow for comparing the overall quality of survey programs and yield further insights into the impact of implementing various sampling and fieldwork procedures on sample representativeness.

Our analysis covers four cross-national projects carried out in Europe between 2002 and 2016: the Eurobarometer (EB), European Quality of Life Survey (EQLS), European Social Survey (ESS), and International Social Survey Programme (ISSP). It pursues three main research goals. The first goal is to explore the viability of existing methods for evaluating sample quality in multiwave cross-project assessments. We start by describing the relevant approaches and identify those that do not require individual-level auxiliary information. The two most promising methods are discussed in more detail, namely those based on internal and external criteria. The second goal is to analyze the impact of design and poststratification weights on internal and external criteria assessments. We find that external criteria assessments critically rely on the availability of design weights. Conversely, we demonstrate the workability of internal criteria assessments without weights, which—because most cross-national survey projects still do not provide design weights—makes them the more practical and accessible method for secondary data users and the one that is applicable to cross-project assessments. The third goal is to demonstrate the application of internal criteria assessments to the screening of multi-project cumulative datasets and the identification of suspicious samples. A review of the survey documentation for these selected samples enables us to identify potential reasons for excess bias. In pursuit of these goals, the paper also provides an overview and discussion of the weight availability among cross-national survey projects.

2. APPROACHES TO EVALUATING SAMPLE QUALITY

Sample quality assessment procedures should lead to direct, quantitative measures of representativeness (Lyberg and Biemer 2008). Direct measures enable analyses of the determinants of sample quality, such as the relative importance of country-specific characteristics *vis-à-vis* those of the survey process. Furthermore, secondary data users typically have limited access to survey-internal materials; therefore, sample assessment procedures—to be broadly

applicable—must rely on information that is routinely made publicly available. Thus, regarding the applicability of established approaches to a cross-project setting, the crucial distinction stands between (i) procedures that require data on the entire drawn sample, including nonrespondents, and (ii) those that do not demand any additional individual-level data. We focus on the latter.

Procedures relying on individual-level auxiliary data prove impractical at scale. For example, Groves (2006) compares response rates (RRs) across subgroups, which requires information that is not typically available in survey documentation because the RR is provided, in the best case, only for the entire sample. Similarly, making use of information from the sampling frame or matched data is only possible with additional data sources. More recent approaches of this kind include the R-indicator, which compares the variance in the propensity to participate in the survey among respondents and nonrespondents (Schouten et al. 2012). Furthermore, methods based on the concept of a balanced response set (Särndal 2011) rely on indicators of balance, distance, and variability (Lundquist and Särndal 2013) and compare distributions of characteristics among respondents and the entire drawn sample. Such information is typically unavailable when performing secondary analyses.

On the other hand, because RR somehow remains among the most common survey quality metrics, it has a fair chance of being disclosed in the documentation. However, longitudinal and cross-project comparisons of reported RRs are pointless due to the flexibility observed in the application of the definitions of survey outcome rates by various survey teams and research organizations. For instance, the American Association for Public Opinion Research's definition of RR1 (AAPOR 2016) is consistently used by the ESS (Beullens et al. 2014, pp. 17–18), while the EQLS excludes noncontacts from the denominator in its RR calculations (EQLS 2003, pp. 2–3). In the ISSP, some surveys provide an RR number without explaining how it was derived or provide different numbers in different documents; for example, the study description of the ISSP's 2011 wave recorded a 60.7 percent RR in Sweden (ISSP 2012d), while the Study Monitoring report mentions a value of 59.8 percent (ISSP 2013). It has also been a longstanding practice of the EB not to publish RRs, which led to a well-publicized controversy over data quality and the potential underestimation of Euroscepticism when a Danish newspaper obtained the EB's recorded RRs in 2019 (Larsen 2019). Additionally, in some sample designs, such as quota samples or certain random route samples, and in surveys that allow proxy reports or substitutions, RRs are of limited utility.

There are three principal approaches that require no auxiliary individual-level information. The first method relies on external criteria, that is, comparisons of sample estimates to "gold standard" benchmarks from external sources, such as comparing the sample's proportion of female respondents to population data (Koch 2016; Eckman and Koch 2019). The second method in this group involves recourse to internal criteria, where an estimate from a specific subsample is compared to a parameter known by definition. The model

application compares the proportion of women in the subsample of two-person heterosexual households to the 50:50 ratio (Kohler 2007; Menold 2014; Eckman and Koch 2019; Jabkowski and Cichocki 2019). The third method relies on the comparison of weighted and unweighted estimators (Billiet, Matsuo, Beullens, and Vehovar 2009; Peytchev, Presser, and Zhang 2018; Sakshaug and Antoni 2019).

All three methods are directly related to the concept of Total Survey Error (Biemer 2010) in that they compare a sample estimate to a true value of the parameter. Thus, (i) in gold-standard evaluations, the external population statistic is the true parameter value; (ii) in the internal criteria approach, survey estimates are compared against an a priori true value; and (iii) in comparisons of weighted and unweighted estimators, the weighted estimator is conceived of as the true value. The last approach is not useful for this paper's purposes because it cannot be applied to assessing sample quality in terms of gender composition. Because gender is routinely incorporated in weighting procedures, comparing weighted and unweighted estimators of gender would merely constitute an imperfect and indirect application of external criteria. Although the common goal of applying weights is adjusting sample distributions of demographic characteristics to match population distributions, both the selection of the attributes to correct for and the methodologies employed to calculate the weights differ substantially across survey projects. Thus, in what follows, we only consider the two remaining approaches, that is, based on external and internal criteria (defined in sections 4.2 and 4.3, respectively).

Within the Total Survey Error framework, both external and internal assessments of sample quality primarily address representation errors (Groves and Lyberg 2010; Groves et al. 2011), which are also sometimes referred to as respondent selection issues, that is, sampling error, coverage error, and nonresponse error at the unit level (Weisberg 2009). External assessments constitute a direct measure of sample representativeness, defined as the difference between the sample estimate and the population parameter, which corresponds to the sum of coverage, sampling, and nonresponse error. Internal criteria, on the other hand, rely on theoretical knowledge about the properties of specific subpopulations. As such, rather than capturing some aspect of sample bias in its entirety, internal criteria assessments probe the integrity of the sample by exploiting this a priori knowledge and point to potential irregularities in the survey process.

3. EMPIRICAL BASE

3.1 Survey Projects under Assessment

Our analysis encompasses four major cross-national survey projects conducted in Europe since 2002. This year coincides with the start of the ESS and EQLS,

the extension of country coverage in the ISSP and EB, as well as significant advances in cross-national survey methodology. The integrated dataset is comprised of 1,125 individual-country surveys spread over forty-one waves. The four projects were selected based on their cross-national comparative focus, academic prominence, and established longitudinal track records with multiple survey waves. In addition to providing valuable data for substantive analyses, all four projects are objects of methodological research (Bauer 2016; Vandenplas and Loosveldt 2017; Höhne and Lenzner 2018). For each national sample, bias is calculated according to internal and external criteria, following the procedures described in sections 4.2 and 4.3. Table 1 provides basic information about the four survey projects, while the Supplementary data online (Jabkowski, Cichocki, and Kołczyńska 2021) contain brief profiles of the projects, references to the original data files, and materials enabling the replication of all procedures.

3.2 Heterogeneity of Weights

Internal and external criteria for sample quality are used to evaluate the sample composition with regard to selected respondent characteristics. Applying either procedure thus requires the consideration of weighting. Cross-national surveys may feature design weights, poststratification weights and population size weights. Population weights play no role in the evaluation of sample quality because all observations within a country sample have the same value. Design weights adjust for unequal probabilities of being drawn into the sample resulting from the sampling design. They are used when (i) sampling households

Table 1. Survey Projects Selected for Comparative Consideration

Project acronym	Project name	Time scope	Number of waves[b]	Number of national surveys
EB	Eurobarometer[a]	2002–2016	15	462
EQLS	European Quality of Life Survey	2003–2016	4	125
ESS	European Social Survey	2002–2017	8	199
ISSP	International Social Survey Programme	2002–2015	14	339
Total			41	1,125

NOTE.—[a]Autumn waves of the Standard Eurobarometer and the Candidate Countries Eurobarometer.
[b]The EB and ISSP include pre-2002 waves, which were excluded from the analysis. The scope of the analysis has also been restricted to surveys conducted in European countries, even if a particular project boasts broader coverage.

rather than individuals, (ii) using stratification with unequal probabilities of selection between strata, and (iii) intentionally oversampling certain subpopulations because of specific research objectives (Pfeffermann 1996). Computing design weights requires a knowledge of the selection probabilities at all stages of the sampling process, which explains why only probability samples provide design weights while, for instance, quota samples do not. Poststratification weights adjust the composition of the achieved sample to that of the population, with the primary purpose of correcting for sampling and nonresponse errors (see Lynn et al. 2007). Poststratification weights typically account for basic sociodemographic characteristics, for which reliable external sources exist, such as gender and age and, sometimes, also region, urbanicity, education, and economic status (Zieliński, Powałko, and Kołczyńska 2018).

Of the four selected projects, three—the ESS, EQLS, and EB—have standardized procedures regarding the calculation and availability of weights. The ESS consistently provides both design and poststratification weights, except for three national surveys (out of 199, i.e., 1.5 percent) published without weights at the time of writing [round 3 in Latvia and Romania (ESS 2006) and round 4 in Lithuania (ESS 2008)]. Design weights correspond to the inverse of the inclusion probabilities, which are scaled in a way that maintains the net sample size. Poststratification weights in the ESS take into account both design factors and "the distribution of the cross-classification of age group, gender, and education in the population and the marginal distribution for region in the population" (ESS 2014, p. 1). The EQLS includes poststratification weights and, as of wave 3, also design weights. Its poststratification weights adjust for age crossed with gender, as well as region, urbanization level, and household size. In wave 3, the EQLS added design weights adjusting for household size in most countries (except for Hungary, Malta, Slovenia, Sweden, and Iceland), as well as a "final national weight," which contains design and poststratification components. The design weights available in round 4 of the EQLS account for the unequal selection probabilities stemming from the overall sampling design. A separate weighting variable, the "final weight," combines poststratification factors and design factors as available. The EB provides poststratification weights adjusting for gender, age, region, and locality size and, in some countries, also additional factors, but it does not provide design weights. The ISSP exhibits much more diversity regarding data weighting, with one of the few rules being that the published data contain only one weight variable. The type of weight available and the procedures used for computing weights are part of the country study descriptions, not the centralized per-wave methods report. The quality of the descriptions ranges from having no information at all (ISSP 2012b) through the two-word phrase "poststratification weighting" (ISSP 2012c) to detailed descriptions (ISSP 2012a), making the weights challenging to classify. Of the ISSP surveys we analyzed, none provide design weights, 61.1 percent (207 out of 339) provide poststratification weights, and

the remaining 132 surveys provide no weights at all. Figure 1 shows which weights are available in the national surveys under analysis.

The heterogeneity of approaches to setting and reporting weights creates considerable obstacles for secondary data users. Some surveys do not include any weights. Others make some weights available, albeit with incomplete information about how the particular values were calculated, and even if documentation explicitly and exhaustively explains the weighting procedures, they may prove challenging to compare across projects. For instance, a comprehensive documentation review of twenty-two cross-national survey projects from around the world demonstrated that out of 1,721 national surveys 1,035 provided some weighting factors, among which 450 contained only poststratification weights, eighty-eight only design weights, 237 combined design and poststratification weights, and 260 some weights, whose precise nature could not be determined based on the available documentation (Slomczynski et al. 2017).

For quality evaluations, the inconsistent availability of design weights constitutes a major drawback, as they need to be implemented for data from complex sample designs. In the case of the four projects, weight availability seems to constitute a characteristic of projects and project waves and to be a consequence of the project leadership's decisions and priorities, as well as being associated with the type of sample to some extent. Thus, ISSP, EB, and two

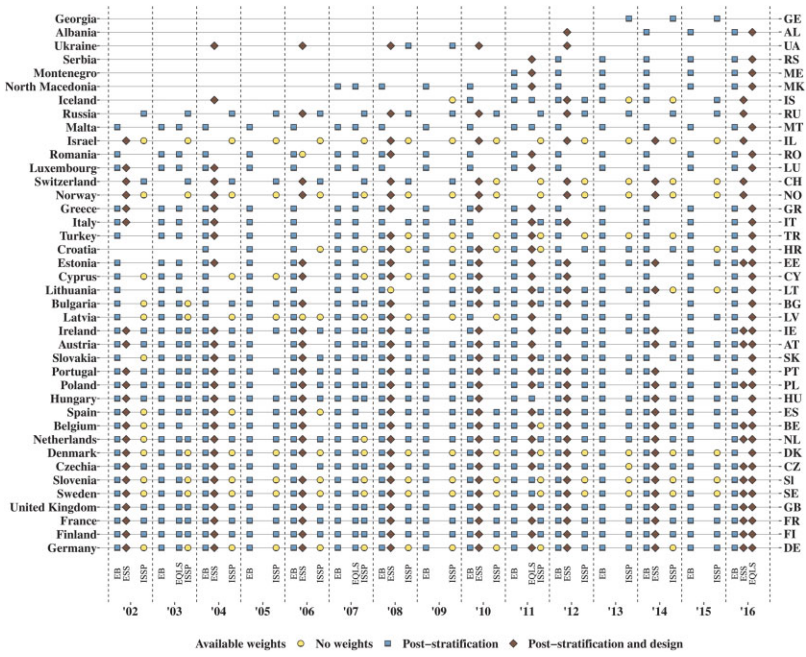

Figure 1. Types of Weights Provided in the Four Survey Projects.

early waves of EQLS have no design weights and are dominated by multistage random route samples, for which the calculation of design factors is not straightforward. Simultaneously, the multistage probability samples in these projects and waves do not provide design weights. While the lack of design weights is no doubt concerning, limiting the data only to surveys that provide them would mean a substantial restriction of the scope of cross-national analyses.

4. METHODS

4.1 External Criteria for Sample Quality

Our application of external criteria for representativeness relies on gender only for the sake of the simplicity of the demonstration, although we note that external criteria can be applied with regard to other individual characteristics and combinations of such. Most applications of external criteria in the literature rely on gender or age (Groves and Peytcheva 2008; Struminskaya, Kaczmirek, Schaurer, and Bandilla 2014; Kobilanski, Pizzolitto, and Seligson 2019) because of their omnipresence in questionnaires, straightforward measurement, and low item nonresponse, as well as the availability of reasonably reliable population statistics in most countries of the world. Other sociodemographic characteristics, such as education (Ortmanns and Schneider 2016), household size, or employment status (Koch 2016), do not share these desirable properties and are used less frequently. It is worth remembering that representativeness in terms of age, gender, or other sociodemographics does not correct for biases concerning other characteristics (Voogt and Van Kempen 2002).

The gender variable is well suited to large-scale analyses because it involves almost no measurement error (Kohler 2007). Nevertheless, upon close inspection, the survey measurement of gender does involve some relatively minor complications. First, gender can be either assessed by the interviewer or declared by the respondent. While the former seems prevalent, the latter may be the case in self-administered surveys, and discerning which survey mode has been employed with each survey requires a review of the survey documentation. Furthermore, some recent surveys feature an "other" gender (e.g., the Americas Barometer 2016 in Canada, also planned in the ESS), although none of the surveys within the scope of this analysis include nonbinary gender options.

Benchmark population data were obtained from the UN Department of Economic and Social Affairs, which compiles population statistics and creates projections and is—to the best of our knowledge—the only data source on the composition of our target populations by sex and exact age in years (UN 2019). While there is an established tradition of benchmarking European survey outcomes against those of the European Union Labour Force

Study (e.g., Koch 2016), the LFS is itself a probability survey with notable nonresponse and measurement problems in some countries. Hence, it may constitute a gold standard for comparisons focused on such extended characteristics as education level or work status, but when it comes to strictly demographic variables, census-type data constitute a firmer benchmark. On top of that, the LFS is limited to the European Union, so relying on it would require eliminating all those European countries that are not EU member states from the analysis.

The information about the population regarding exact age in years that the UN dataset provides is necessary because we calculate external bias for the age range from eighteen to seventy-four. Surveys differ in their definitions of target populations with regard to age, so we chose this age range to increase comparability. As a consequence, the survey variable measuring respondent age is necessary, which eliminates two surveys (ISSP 2007 from the Netherlands and ISSP 2015 from Denmark) that do not provide this variable.

In line with the TSE approach (Biemer 2010), we calculate the difference between the sample estimate for the proportion of women and the corresponding population parameter: $\widehat{p}_i - p_i$. Because true gender proportions are known, the standard error of each estimator is equal to $\mathrm{SE}_i = \sqrt{p_i(1-p_i)/n_i}$, where p_i is the true fraction of women in the country-year in which survey i was conducted and n_i is the total number of respondents in that survey. For comparisons between multiple studies, we define a measure of sample bias according to external criteria (external bias) as the absolute value of the deviation of a survey estimate from the corresponding true parameter value divided by the standard error (the deviation is statistically significant at $\alpha = 0.05$ if external $\mathrm{bias}_i > 1.96$):

$$\text{external bias}_i = \frac{\widehat{p}_i - p_i}{\sqrt{p_i(1-p_i)/n_i}}.$$

4.2 Internal Criteria of Sample Quality

The internal criteria approach is based on evaluating the composition of specific subsamples against values known by definition and originates from Sodeur (1997). Kohler (2007) applied internal criteria to compare sample bias across cross-country surveys. This internal procedure requires a prior separation of a subsample of heterosexual couples living in two-person households within each survey. For such subsamples, we calculate the difference between the proportion of females in the selected subsample of survey i and the true proportion of females in such subpopulations (percentage of females in a subsample of a survey—50 percent). The measure of sample bias according to internal criteria (internal bias) is then defined as follows:

$$\text{internal bias}_i = \frac{\widehat{p}_i - 0.5}{\sqrt{0.25/n_i}}.$$

Because the expected proportion of females is equal to 0.5, the variance of the female ratio estimator is equal to $0.25/n_i$, where n_i is the total number of respondents in each extracted subsample. The result is statistically significant at $\alpha = 0.05$ if internal $\text{bias}_i > 1.96$.

Internal criteria assessments necessitate selecting a subset of respondents living in two-person households with heterosexual spouses or partners in each national sample. Out of the four surveys under consideration, only the ESS and EQLS consistently implement a complete household roster, enumerating all household members and providing their genders and relationships with the respondent. Given that such detailed information about the household composition is not provided by the other two studies and proves rather exceptional among cross-national survey projects, a cross-project application of internal criteria requires pragmatic workarounds to ensure alignment with the available variables. The strictest and best approach requires variables related to (i) the respondent's gender, (ii) household size (number of eligible individuals), (iii) the respondent's relationships with other household members, and (iv) the genders of other household members. In the more lenient approach, it is sufficient to have data about the household size and marital status of the respondent, assuming that a great majority of respondents who are married or in civil unions and live in two-person households actually live with their spouse or partner and that, for a great majority of such respondents, their spouse is of the opposite sex. The lenient approach enables including data from the EB and ISSP. Our evaluation follows the lenient path for all four projects to improve the comparability of the results. Because the ESS and EQLS enable both approaches, we calculated internal representativeness according to both the strict and lenient approaches for these two projects and found the results to be correlated at 0.96 and 0.93 within the two projects, respectively. Thus, acknowledging the imperfections of the lenient approach, we argue that this level of accuracy is sufficient for the purpose of flagging suspicious samples. At the same time, we note that legislative changes enabling civil unions and marriages between homosexual couples in a growing number of European countries will likely reduce the accuracy of the lenient approach to calculating internal sample representativeness. Thus, the absence of information about the genders of household members in surveys will make the application of internal criteria increasingly problematic.

Two EB waves (58.1, 60.1), three ISSP surveys (Slovenia—2003, Turkey—2012, Hungary—2007), and two ESS surveys (round 4 in Estonia and round 5 in Finland) were excluded from the internal assessments because they did not feature questions on household size or marital status.

5. RESULTS

5.1 Impact of Design Weights

In this section, we examine the role of weights when applying measures of internal and external bias to evaluate sample quality. In principle, the application of design weights is advisable whenever one is dealing with survey data collected via complex sampling. Assessments in terms of both external and internal criteria should ideally be performed on data weighted by design factors because comparing unweighted data could prejudice the evaluations against sampling designs with unequal selection probabilities (Horvitz and Thompson 1952). At the same time, secondary data users must face the grim reality that, as demonstrated above, most cross-national survey projects do not provide separate variables with design factors.

Given that design weights are typically unavailable, it is vital to evaluate the impact of not using them on the assessment of sample bias in terms of external and internal criteria. The ESS—the only project that consistently reports design weights in all waves—provides empirical input for use in such considerations. Figure 2 presents a comparison of external bias (upper panel) and internal bias (lower panel), with design weights not used in estimation and included in estimation, where each point represents a national survey and violin plots are employed to visualize distributions of bias within each ESS wave. Violin plots use kernel density estimation to highlight the areas with higher concentrations of datapoints (Wickham 2016); wider sections represent higher approximate frequencies of data points (Hintze and Nelson 1998). The figure includes all samples from all waves except simple random samples, whose design weights are by default equal to 1. Three ESS surveys were also excluded—round 3 in Latvia and Romania and round 4 in Lithuania—because they did not report design weights (see section 3.2). Furthermore, three samples from Slovakia (rounds 4–6) were also filtered out because they constituted extreme outliers in terms of internal bias (for a discussion, see section 6).

The upper panel of figure 2 shows that the application of design weights has a notable effect on bias according to external criteria. Regarding internal criteria (lower panel of figure 2), assessments remain unaffected by the lack of weighting, both in terms of the shapes of the distributions and the median values. We calculated the difference between the value of bias with design weights and without weights for each survey. For external criteria, the mean of these differences equals −0.28 and the median is −0.08. The corresponding differences for internal-criteria assessments have a mean of 0.02, and both median values are equal.

The observation that design weights have almost no impact on the results of internal criteria evaluations falls in line with theoretical expectations, which may seem intuitive but are nevertheless useful to test. Consider the expected influence of design weights on various types of samples. For simple random

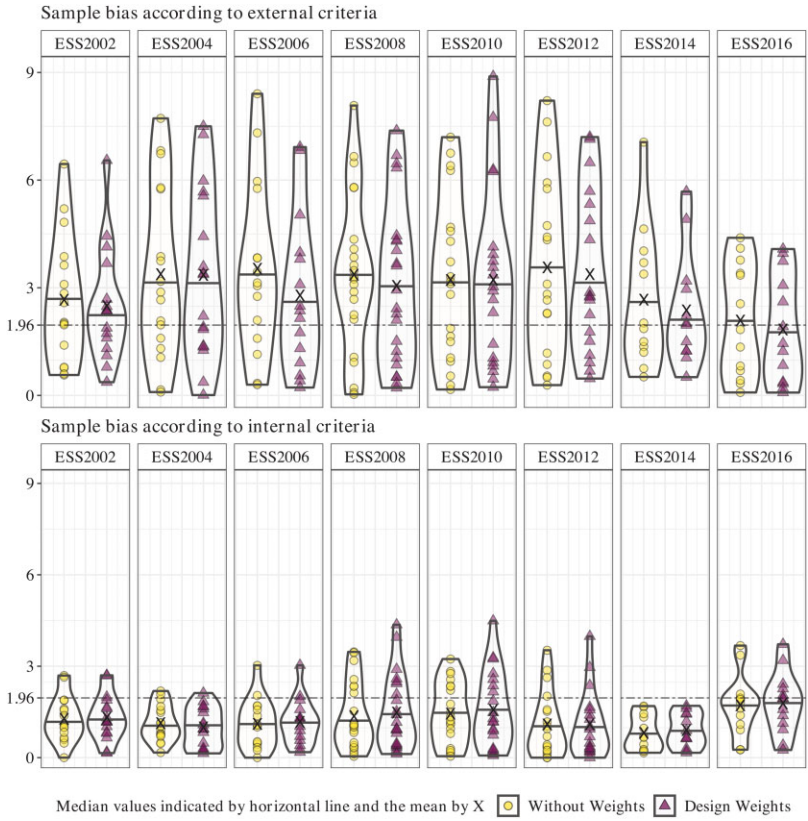

Figure 2. Effects of Design Weights on Measures of External and Internal Bias.

samples, the design weights equal 1. Hence, their absence makes no difference. In household-based samples, address-based samples, and nonregister samples, design weights mainly correct for (i) unequal chances of selection among households or (ii) unequal chances of within-household selection for target respondents. Regarding (i), the absence of design weights has no impact on the estimation of the 50/50 ratio, because irrespective of the chances of selecting two-person heterosexual households among all households, it is only the quality of within-household selection that the internal procedure takes into account. Regarding (ii), the fact that only two-person households fall within the scope of the analysis makes the correction for unequal chances unnecessary. The only potentially problematic scenario relates to individual register samples with complex designs and unequal selection probabilities, which may lead to a situation in which women, for instance, would have higher selection probabilities by design. While not having the design weight is a reason for concern, it is likely not a major one given that the impact of design weights on

the difference between the absolute biases that are adjusted and unadjusted for design in multistage individual-based samples—at least in the ESS—is negligible.

5.2 Impact of Poststratification Weights

We now turn to the examination of the role of poststratification weights on sample bias measured according to internal and external criteria for representativeness. Unlike design weights, poststratification weights are ubiquitous in cross-national survey projects, and indeed, 990 out of 1,125 surveys under consideration contain some form of poststratification weighting factor. Thus, comparing sample bias according to internal and external criteria for samples with and without poststratification weights is possible for all four survey projects. Figure 3 presents the results of this analysis. The left pane refers to

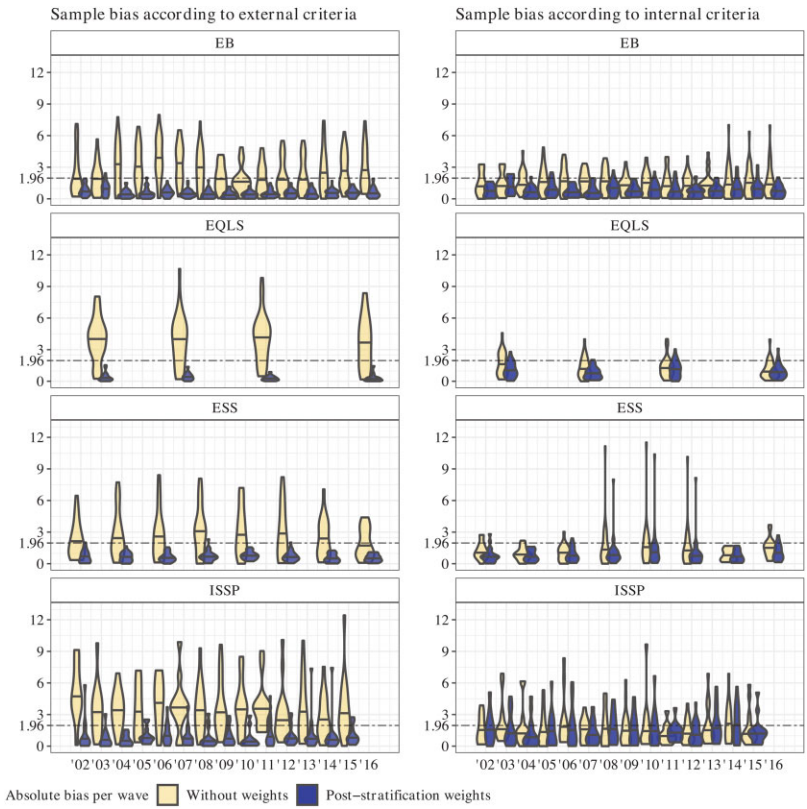

Figure 3. Effects of Poststratification Weights on Measures of External and Internal Bias.

external criteria and demonstrates that applying poststratification weights (dark shade) results in substantially reduced bias as compared to unweighted data (light shade) across all projects and waves. The right pane compares bias according to internal criteria for weighted and unweighted samples.

The differential impact of poststratification weights on the results of external and internal assessments is apparent. Regarding external criteria, weighting produces a substantial reduction in sample bias across all four projects; on the other hand, its impact on the results of internally focused assessments proves visibly smaller and less uniform. For each survey, we calculate the difference between bias with poststratification weights and without weights, which represents the reduction of external bias due to the application of poststratification weights. We summarize these differences by project in table 2.

The outsized bias reduction in external criteria assessments brought about by poststratification weights is easy to explain. Poststratification weights are calculated to adjust the sample proportions of selected social and demographics groups to their population proportions, as represented in the census or other high-quality data. Because virtually all poststratification weights account for gender, weighted sample proportions will match those from the population. Calculating external bias on data weighted using poststratification weights thus yields an indicator not of sample quality but rather the quality of the data processing procedures employed to produce the poststratification weights.

Even though the application of poststratification weights results in smaller reductions of internal bias, they still constitute a sizable distortion. The magnitude of their impact would remain somewhat unpredictable because it would depend on the degree to which the bias in the subsample of two-person heterosexual households differs from that in the entire sample. Given that design weights are routinely unavailable and only internal criteria may be correctly applied without design weights (as explained in section 5.1), using internal criteria without weights remains the preferred approach to cross-project sample-quality assessments.

Table 2. Within-Project Differences between Bias with and without Poststratification Weights

Project name	External criteria		Internal criteria	
	Mean	Median	Mean	Median
Eurobarometer	−2.02	−1.59	−0.70	−0.47
European Quality of Life Survey	−3.49	−3.58	−0.33	−0.26
European Social Survey	−2.01	−1.68	−0.37	−0.30
International Social Survey Programme	−2.71	−2.25	−0.18	−0.14

6. INTERNAL CRITERIA-BASED SCREENING FOR OUTLYING SAMPLES

Having established the preference for the internal criteria assessment of unweighted data, we now apply this procedure to measuring sample quality in our selection of surveys. The box plots in figure 4 present the distribution of internal bias across the four projects, with the jittered dots representing individual surveys. According to these plots, the third quartile of all project distributions of absolute bias is close to the cutoff point value of 1.96 (2.13 for the EB, 1.84 for the EQLS, 1.69 for the ESS, and 1.98 for the ISSP). Thus, in all projects, around 75 percent of the surveys exhibit bias that does not or only barely reaches statistical significance at the standard level of alpha = 0.05.

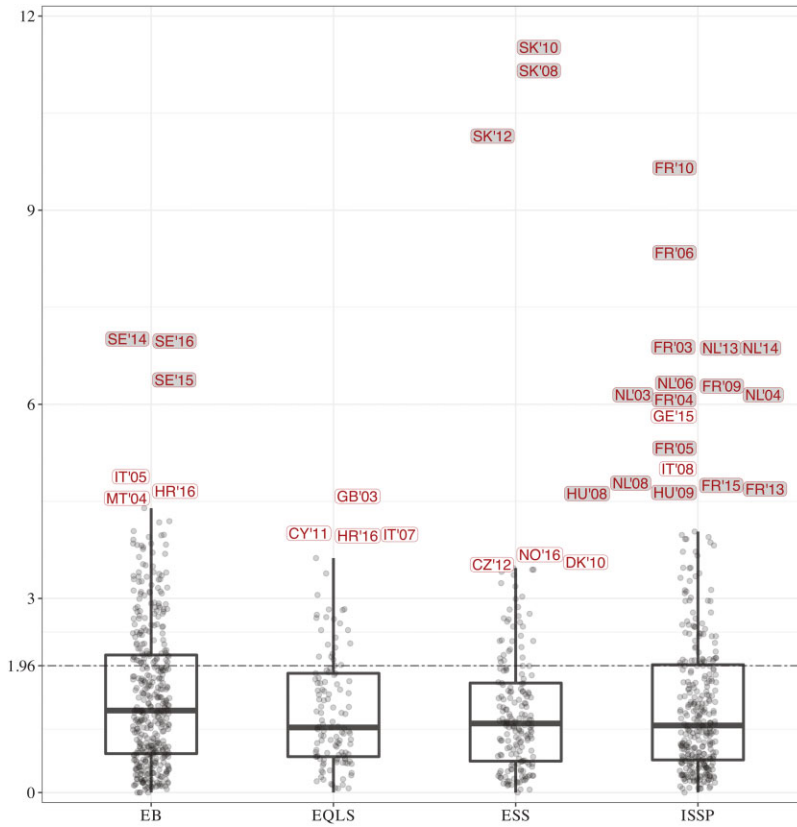

Figure 4. Internal Criteria with No Weights: The Outliers. Surveys are labeled with the country name and abbreviated year: CY = Cyprus, CZ = Czechia, DK = Denmark, FR = France, GB = United Kingdom, GE = Georgia, HR = Croatia, HU = Hungary, IT = Italy, MT = Malta, NL = the Netherlands, NO = Norway, SK = Slovakia, and SE = Sweden.

Conversely, the remaining quarter of surveys in each project suffer from significant bias. Among them, the outliers, that is, surveys with bias exceeding 1.5 for the interquartile range within each project, are of special concern.

The outliers fall into two classes: the occasional and the persistent. Occasional outliers, marked with white labels, are comprised of instances in which each country has, at most, one biased sample within each project. For example, even though three of the four projects contain one outlying sample from Italy, these are singular instances within each project and thus cannot be construed as a pattern. On the other hand, some persistent outliers stand out; these are marked with gray labels on the plot. These cases include multiple outlying surveys from the same country within a project. Our analysis identified four such baffling clusters of surveys from Sweden in the EB; Slovakia in the ESS; and France, the Netherlands, and Hungary in the ISSP.

Internal bias may result from nonresponse or within-household selection. While we cannot determine which is at fault in the cases we identified, the survey documentation seems to provide specific clues. A systematic analysis of the determinants of sample bias is beyond the scope of this paper, but we were nevertheless able to identify elements of sampling design and fieldwork execution that may be responsible for the enormous biases. We note that our investigation only covered the remaining three projects because of the lack of survey-level documentation in the Eurobarometer. Thus, the three Swedish samples in the EB (waves 82.3, 84.3, and 86.2) seem definitely problematic, yet there is no obvious way of finding out how and why.

The ISSP surveys in France (e.g., ISSP 2012c) and the Netherlands (e.g., ISSP 2014) rely on self-administered mail questionnaires, and hence, their within-household selection stage is performed by the household members themselves. Mail surveys are known to struggle with selection bias when they allow for the selection of the target respondent by household members (Lavrakas 2008, pp. 808–809). Interestingly, study descriptions from Dutch surveys include an analysis of the representativeness of the realized samples for the synthetic population constructed from all members of sampled households, as reported in the household roster in the questionnaire. However, these analyses, while generally reassuring, failed to encompass gender. Furthermore, in the Netherlands, ISSP modules tend to be carried out two at a time and are often attached to another survey; hence, the identical bias was registered in the cases of samples from the Netherlands in ISSP waves 2003 and 2004, as well as 2013 and 2014.

Hungarian surveys in the ISSP switched from personal-register samples to household samples starting with the 2007 wave (ISSP 2006, p. 43; 2008). In waves 2002–2006, sample bias according to internal criteria was below the significance threshold. The sample from the 2007 wave of the ISSP was excluded from our analysis due to a lack of the necessary survey variables. Samples from the 2008 and 2009 waves both exhibit internal bias of approximately 4.6;

although bias in subsequent Hungarian ISSP samples declined somewhat, it exceeded 2.0 in all surveys until 2015.

The most eye-grabbing case of extreme internal bias, however, comes from the Slovakian surveys in the ESS. Not only do they exhibit the highest bias out of all the evaluated surveys across all projects but they also show a marked shift in sample bias over time: in rounds 2 and 3, internal bias amounted to 0.56 and 0.94, while in the subsequent rounds from 4 to 6, it jumped to 11.16, 11.52, and 10.15, respectively. A closer investigation of the survey documentation reveals that the ESS in Slovakia instituted a substantial change in sample design between rounds 3 and 4. The project switched from relying on individual registers in rounds 2 and 3 to an area listing sample (nonregister sample with household enumeration) in rounds 4–6. Rounds 2 and 3 relied on the Central Register of Citizens (maintained by the Ministry of the Interior), which boasts an almost 100 percent coverage of residents, excluding homeless persons and institutionalised populations. The sample was drawn systematically from the extract of residents of the appropriate age (ESS 2004, p. 162); a similar procedure was applied in ESS round 3 (ESS 2006, p. 162). Following round 4, sampling relied on a stratified multistage nonregister design. It involved an initial selection of municipalities in each stratum with a probability proportional to population size. Within each selected municipality, starting addresses were drawn from a database of landline and mobile phone numbers. After the enumeration of households in the given street or block, ten households were selected for approach by the interviewer. Within the household, the respondent was selected using the Kish grid, that is, only from among respondents who had lived in their households for at least six weeks (ESS 2008, p. 242); similar procedures were applied in rounds 5 and 6 (ESS 2010, p. 175; 2012, p. 179). This change dramatically increased interviewer discretion in regard to respondent selection. Personal register samples give interviewers minimal wiggle room in the choice of target respondent, especially given the fieldwork control measures employed in the ESS. Sample designs in which the interviewer is responsible for the within-household selection of the target respondent incentivizes the selection of individuals who are perceived as more likely to complete the survey, especially in the context of pressure on high outcome rates (Eckman and Koch 2019).

In addition to the dramatic increase in internal bias, the switch of the sample design in Slovakia coincided with a change in the direction of deviation from the 50/50 gender ratio. Based on internal criteria, samples from rounds 2 and 3 oversampled men, while the later samples saw a sizeable overrepresentation of women, who constituted as much as 70–76 percent of respondents living in two-person households with their heterosexual partners. Such deviations fall in line with earlier findings that nonpersonal-register sample designs are prone to overrepresent women (Jabkowski and Cichocki 2019). Because women tend to be present more often in the household (Stoop et al. 2010), they are more readily available to respond, which is likely to increase selection bias when the

interviewer is responsible for selecting the target respondent in the household (Eckman and Koch 2019).

7. SUMMARY AND CONCLUSIONS

Secondary data users tend to approach existing survey data in good faith, treating surveys as if they were based on high-quality representative samples of respective populations. Practical limitations somewhat necessitate such credence because the documentation made available by many projects fails to provide even the most basic information about sampling and fieldwork execution (Kołczyńska and Schoene 2017; Jabkowski and Kołczyńska 2020). Furthermore, for data users focused on substantive questions, a pursuit of methodological issues may constitute a distracting diversion of resources and attention. While some degree of trust seems necessary to proceed with secondary data use, blind faith in data quality inhibits improvements in methodological standards and—perhaps more importantly—impairs the accuracy of empirical findings based on the data of unverified quality. This paper aims to assist survey analysts who wish to perform preliminary sample quality screenings on data derived from cross-national survey projects.

Appropriate sample-quality assessment procedures should not rely on data on nonrespondents or other information not routinely available in survey documentation. In this respect, most prominent approaches involve evaluations based on either external or internal criteria of representativeness. External criteria assessments rely on the availability of design weights, which remains far from common, and the external benchmarking of samples without weights would be prejudiced against complex sample designs. While most cross-national surveys feature some form of poststratification weights, they must not be used when performing external criteria evaluations because they adjust sample estimates of crucial demographic variables to their distributions in the population. Conversely, we demonstrate that internal criteria may be used without weights, which makes them a broadly and readily applicable tool for large-scale assessments of sample quality. While internal criteria demand no data in addition to the datasets themselves, they do require the presence of a specific set of variables characterizing the respondent's household. Therefore, a strict application of internal criteria is not always possible, yet we demonstrate the feasibility of a more lenient approach, that is, capturing the deviation in the gender ratio among subsamples of two-person households inhabited by couples for which the relevant demographic variables are typically present in survey datasets.

Internal criteria already have an established track record in the field of survey methodology (Kohler 2007; Menold 2014). Recent studies have found, for example, that samples based on individual registers tend to have lower sample bias than those involving the within-household selection of target respondents

(Eckman and Koch 2019; Jabkowski and Cichocki 2019). Systematic analyses of other factors, including, for instance, the survey mode, the degree of centralization in terms of the oversight of the survey process in individual countries, or the characteristics of fieldwork agencies, could inform standards and guidelines for cross-national survey research. Our internal criteria assessment of data from four major European cross-national survey projects (the ESS, EQLS, ISSP, and EB) revealed that, in about 75 percent of national surveys, the bias is within acceptable bounds. Moreover, we demonstrated the utility of internal criteria as a procedure for detecting survey samples with apparent quality issues. In order to identify the factors underlying these patterns of outsize bias, we reviewed the documentation of the outlying surveys found in the ISSP (France, Hungary, and Netherlands) and ESS (Slovakia). In the ISSP, the outliers were found to have been either postal surveys (in France and Netherlands), where the sampling involved a simple random sample of addresses and respondent selection within the household was performed by the household members themselves, or newly adopted household samples with household selection via a random route procedure (in Hungary). In the intriguing case of Slovakia in the ESS, an abrupt deterioration in sample quality starting with round 4 followed a change in sampling design from a complex individual register sample to multistage sampling with household enumeration via random route procedures.

Internal criteria can only be applied to the distribution of gender within narrow subsamples of two-person households inhabited by couples. Therefore, any bias on the gender distribution does not necessarily mean that there is a commensurate bias on other variables, and bias in the selected subsample does not necessarily mean that there is bias in the rest of the sample. Thus, internal criteria assessments need to be interpreted with caution and in the context of other quality indicators. Furthermore, external criteria evaluations may incorporate a range of other characteristics for which both reliable survey measurements and population benchmarks exist. If design weights are available, external evaluations enable an arguably more comprehensive assessment of sample quality.

Internal and external criteria assessments represent different approaches to measuring sample bias. Hence, when applied to the same survey dataset, one may suggest the presence of bias while the other does not. In our view, even one cautionary sign warrants further thorough examination of the flagged surveys in search of potential sources of the bias, as well as caution when pursuing substantive analyses. At the same time, an indication of sample bias with either approach does not automatically disqualify a survey. In this context, it is useful to remember that the presence of sample bias on demographic variables does not inevitably translate into bias on other variables, such as reported behaviors or attitudes. However, the effectiveness of poststratification adjustments in reducing sample bias depends on whether adjustment variables, such as gender, are related both to the propensity to respond and to the characteristic

of interest (Peytcheva and Groves 2009). When screening datasets, secondary data users are encouraged to evaluate various quality indicators pertaining to both sample bias and other aspects of survey quality for all available data sources. The decision about the choice of a particular dataset is best made by weighing the evidence these quality indicators provide in the context of specific research goals, taking into account factors such as the subject of the research, types of analyses, and consequences of erroneous results.

Since our emphasis in this paper is on cross-project comparisons, we abstain from bias explorations according to external criteria on those surveys that include design weights (ESS and the last wave of EQLS). This avenue is left for future research, together with more systematic investigations into the impact of various elements of the survey process on sample quality. Regarding recommendations for data collection organizations, our analyses suggest the enduring need for the better documentation of cross-national surveys. Survey organizations should improve their published descriptions of sampling and weights, on top of that, they should also supply survey variables enabling calculations of sample quality indicators. Notably, those suggestions fall in line with the recent recommendations provided by AAPOR and the World Association for Public Opinion Research in their joint report on quality in comparative surveys (AAPOR and WAPOR 2021). Publication of design weights is of particular importance in this respect, as they are necessary not only for calculating external-criteria sample bias but also for correctly applying statistical methods developed for data from simple random samples when dealing with complex sample designs. To facilitate internal criteria assessments, on the other hand, the necessary published minimum includes the respondent's gender, marital status, and household size, while their strict application requires additional information about the gender of the respondent's partner. Given the overall effort of conducting sustained longitudinal surveys across multiple countries, providing such additional information constitutes a small cost with a sizeable gain.

SUPPLEMENTARY MATERIALS

Supplementary materials are available online at https://osf.io/f9g7s/.

REFERENCES

AAPOR (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 8 July 2021.

AAPOR and WAPOR (2021), "AAPOR/WAPOR Task Force Report on Quality in Comparative Surveys," Available at https://wapor.org/wp-content/uploads/AAPOR-WAPOR-Task-Force-Report-on-Quality-in-Comparative-Surveys_Full-Report.pdf. Accessed 8 July 2021.

Alter, A. L., and H. E. Hershfield (2014), "People Search for Meaning When They Approach a New Decade in Chronological Age," *Proceedings of the National Academy of Sciences of the United States of America*, 111, 17066–17070.
Bauer, J. J. (2016), "Biases in Random Route Surveys," *Journal of Survey Statistics and Methodology*, 4, 263–287.
Beullens, K., H. Matsuo, G. Loosveldt, and C. Vandenplas (2014), Quality Report for the European Social Survey, Round 6, London: European Social Survey ERIC.
Biemer, P. (2010), "Total Survey Error: Design, Implementation, and Evaluation," *Public Opinion Quarterly*, 74, 817–848.
Billiet, J., H. Matsuo, K. Beullens, and V. Vehovar (2009), "Non-Response Bias in Cross-National Surveys: Designs for Detection and Adjustment in the ESS," *ASK. Research & Methods*, 18, 3–43.
Eckman, S., and A. Koch (2019), "Interviewer Involvement in Sample Selection Shapes the Relationship between Response Rates and Data Quality," *Public Opinion Quarterly*, 83, 313–337.
EQLS (2003), "European Quality of Life Survey, 2003. Fieldwork Technical Report," UK Data Archive Study Number 5260.
ESS (2004), *ESS2—2004 Documentation Report, Edition 3.7*, European Social Survey ERIC.
ESS (2006), *ESS3—2006 Documentation Report, Edition 3.7*, European Social Survey ERIC.
ESS (2008), *ESS4—2008 Documentation Report, Edition 5.5*, European Social Survey ERIC.
ESS (2010), *ESS5—2010 Documentation Report, Edition 4.2*, European Social Survey ERIC.
ESS (2012), *ESS6—2012 Documentation Report, Edition 2.4*, European Social Survey ERIC.
ESS (2014), *Weighting European Social Survey Data*, European Social Survey ERIC.
Groves, R. M. (2006), "Nonresponse Rates and Nonresponse Bias in Household Surveys," *Public Opinion Quarterly*, 70, 646–675.
Groves, R. M., F. J. Fowler, Jr, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2011), *Survey Methodology*, New York: John Wiley & Sons.
Groves, R. M., and L. Lyberg (2010), "Total Survey Error: Past, Present, and Future," *Public Opinion Quarterly*, 74, 849–879.
Groves, R. M., and E. Peytcheva (2008), "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis," *Public Opinion Quarterly*, 72, 167–189.
Hintze, J. L., and R. D. Nelson (1998), "Violin Plots: A Box Plot-Density Trace Synergism," *The American Statistician*, 52, 181–184.
Horvitz, D. G., and D. J. Thompson (1952), "A Generalisation of Sampling without Replacement from a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685.
Höhne, J. K., and T. Lenzner (2018), "New Insights on the Cognitive Processing of Agree/Disagree and Item-Specific Questions," *Journal of Survey Statistics and Methodology*, 6, 401–417.
ISSP (2006), *Hungary. ISSP 2006—Role of Government IV. Study Description*, GESIS.
ISSP (2008), *Hungary. ISSP 2008—Religion III. Study Description*, GESIS Datenarchiv.
ISSP (2012a), *Bulgaria. ISSP 2011—Health. Study Description*, GESIS Datenarchiv.
ISSP (2012b), *Croatia. ISSP 2011—Health. Study Description*, GESIS Datenarchiv.
ISSP (2012c), *France. ISSP 2011—Health. Study Description*, GESIS Datenarchiv.
ISSP (2012d), *Sweden. ISSP 2011—Health. Study Description*, GESIS Datenarchiv.
ISSP (2013), *International Social Survey Programme, Study Monitoring 2011. Health*, GESIS Datenarchiv.
ISSP (2014), *Citizenship II. Netherlands. Study Description*, GESIS Datenarchiv.
Jabkowski, P., and P. Cichocki (2019), "Within-Household Selection of Target-Respondents Impairs Demographic Representativeness of Probabilistic Samples: Evidence from Seven Rounds of the European Social Survey," *Survey Research Methods*, 13, 167–180.
Jabkowski, P., P. Cichocki, and M. Kołczyńska (2021), "Replication materials: Multi-Project Assessments of Sample Quality in Cross-National Surveys: The Role of Weights in Applying External and Internal Measures of Sample Bias," Available at https://osf.io/f9g7s.
Jabkowski, P., and M. Kołczyńska (2020), "Sampling and Fieldwork Practices in Europe: Analysis of Methodological Documentation from 1,537 Surveys in Five Cross-National Projects, 1981-

2017," *Methodology. European Journal of Research Methods for the Behavioral and Social Sciences*, 16, 186–207.
Kobilanski, F. S., G. Pizzolitto, and M. Seligson (2019), *Sample Substitutions in the AmericasBarometer 2016/17*, Vanderbilt University.
Koch, A. (2016), *Assessment of Socio-Demographic Sample Composition in ESS Round 6*, European Social Survey ERIC.
Koch, A., V. Halbherr, I. Stoop, and J. Kappelhof (2014), *Assessing ESS Sample Quality by Using External and Internal Criteria*, European Social Survey ERIC.
Kohler, U. (2007), "Surveys from inside: An Assessment of Unit Nonresponse Bias with Internal Criteria," *Survey Research Methods*, 1, 55–67.
Kołczyńska, M., and M. Schoene (2017), "Survey Data Harmonisation and the Quality of Data Documentation in Cross-National Surveys," in *Advances in Comparative Survey Methods: Multicultural, Multinational and Multiregional (3MC) Contexts*, eds. T. P. Johnson, B. Pennell, I. A. Stoop, and B. Dorer, pp. 963–984, New York: Wiley.
Larsen, E. G. (2019), "Eurobarometer and Euroscepticism," Available at https://erikgahner.dk/2019/eurobarometer-and-euroscepticism/. Accessed 8 July 2021.
Lavrakas, P. J. (2008), *Encyclopedia of Survey Research Methods*, Thousand Oaks, CA: Sage Publications.
Lundquist, P., and C.-E. Särndal (2013), "Aspects of Responsive Design with Applications to the Swedish Living Conditions Survey," *Journal of Official Statistics*, 29, 557–582.
Lyberg, L. E., and P. P. Biemer (2008), "Quality Assurance and Quality Control in Surveys," in *International Handbook of Survey Methodology*, eds. E. d. Leeuw, J. J. Hox and D. Dillman, pp. 421–441, London: Lawrence Erlbaum Associates.
Lynn, P., S. Häder, S. Gabler, and S. Laaksonen (2007), "Methods for Achieving Equivalence of Samples in Cross-National Surveys: The European Social Survey Experience," *Journal of Official Statistics*, 23, 107–124.
Menold, N. (2014), "The Influence of Sampling Method and Interviewers on Sample Realisation in the European Social Survey," *Survey Methodology*, 40, 105–123.
Ortmanns, V., and S. L. Schneider (2016), "Can we Assess Representativeness of Cross-National Surveys Using the Education Variable?," *Survey Research Methods*, 10, 189–210.
Pennell, B.-E., K. Cibelli Hibben, L. Lyberg, P. P. Mohler, and G. Worku (2017), "A Total Survey Error Perspective on Surveys in Multinational, Multiregional, and Multicultural Contexts," in *Total Survey Error in Practice*, eds. P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. Clyde Tucker, and B. T. West, pp. 179–202, New York: Wiley.
Peytcheva, E., and R. M. Groves (2009), "Using Variation in Response Rates of Demographic Subgroups as Evidence of Nonresponse Bias in Survey Estimates," *Journal of Official Statistics*, 25, 193–201.
Peytchev, A., S. Presser, and M. Zhang (2018), "Improving Traditional Nonresponse Bias Adjustments: Combining Statistical Properties with Social Theory," *Journal of Survey Statistics and Methodology*, 6, 491–515.
Pfeffermann, D. (1996), "The Use of Sampling Weights for Survey Data Analysis," *Statistical Methods in Medical Research*, 5, 239–261.
Sakshaug, J. W., and M. Antoni (2019), "Evaluating the Utility of Indirectly Linked Federal Administrative Records for Nonresponse Bias Adjustment," *Journal of Survey Statistics and Methodology*, 7, 227–249.DOI: 10.1093/jssam/smy009.
Särndal, C.-E. (2011), "Dealing with Survey Nonresponse in Data Collection, in Estimation," *Journal of Official Statistics*, 27, 1–21.
Schouten, B., J. Bethlehem, K. Beullens, Ø. Kleven, G. Loosveldt, A. Luiten, K. Rutar, N. Shlomo, and C. Skinner (2012), "Evaluating, Comparing, Monitoring, and Improving Representativeness of Survey Response through R-Indicators and Partial R-Indicators," *International Statistical Review*, 80, 382–399.
Slomczynski, K. M., J. C. Jenkins, I. Tomescu-Dubrow, M. Kołczyńska, I. Wysmułek, O. Oleksiyenko, P. Powałko, and M. W. Zieliński (2017), "SDR 1.0 Master Box," *Harvard Dataverse*. 10.7910/DVN/VWGF5Q. Accessed 8 July 2021.

Smith, T. W. (2007), "Survey Non-Response Procedures in Cross-National Perspective: The 2005 ISSP Non-Response Survey," *Survey Research Methods*, 1, 45–54.

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

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

Struminskaya, B., L. Kaczmirek, I. Schaurer, and W. Bandilla (2014), "Assessing Representativeness of a Probability-Based Online Panel in Germany," in *Online Panel Research*, eds. M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick and P. J. Lavrakas, pp. 61–85. Wiley Online.

UN (2019), *World Population Prospects 2019, Online Edition. Rev. 1*, United Nations Population Division.

Vandenplas, C., and G. Loosveldt (2017), "Modeling the Weekly Data Collection Efficiency of Face-to-Face Surveys: Six Rounds of the European Social Survey," *Journal of Survey Statistics and Methodology*, 5, 212–232.

Voogt, R. J., and H. Van Kempen (2002), "Nonresponse Bias and Stimulus Effects in the Dutch National Election Study," *Quality and Quantity*, 36, 325–345.

Weisberg, H. F. (2009), *The Total Survey Error Approach: A Guide to the New Science of Survey Research*, University of Chicago Press.

Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, New York: Springer.

Zieliński, M. W., P. Powałko, and M. Kołczyńska (2018), "The Past, Present, and Future of Statistical Weights in International Survey Projects: Implications for Survey Data Harmonization," in *Advances in Comparative Survey Methods: Multicultural, Multinational and Multiregional (3MC) Contexts*, eds. T. P. Johnson, B. Pennell, I. A. Stoop, and B. Dorer, pp. 1035–1052, New York: Wiley.