Survey Research Methods (2024)
Vol. 18, No. 2, pp. 79-97
https://doi.org/10.18148/srm/2024.v18i1.8145
European Survey Research Association

The Role of Interviewer Observations in Obtaining Representative Data in Repeated Cross-National Studies

Hafsteinn Einarsson[1] · Alexandru Cernat[1] · Natalie Shlomo[1]
[1]University of Iceland

Collecting representative data from surveys can be challenging in a survey climate where response rates are declining and costs are rising, but such data are essential for making valid inferences in social research. These issues can be compounded when conducting cross-national surveys that are repeated over time. This paper utilises the European Social Survey (ESS) contact form data to explore cross-national associations with nonresponse over time. Using sample frame data and interviewer observations, a form of paradata where interviewers record characteristics of the household and its neighbourhood, we construct representativity indicators (R-indicators) for each country over nine survey rounds and identify variables that contribute to non-representativeness. We find that interviewer observations produce evidence of non-representativeness, even after controlling for frame information, consistently and without large cross-national variations. In addition, they are stable over time, despite the trend of declining response rates. Consistent associations with the propensity to respond can be leveraged in several ways in survey practice, and we discuss how interviewer observations may be utilised in tasks such as fieldwork monitoring, to inform case prioritisation schemes where units associated with underrepresented characteristics are targeted (including in adaptive and responsive survey designs), and in post-survey adjustment in repeated cross-national surveys. These methods may reduce the risk of differential survey errors occurring in such surveys.

Keywords: nonresponse; paradata; R-indicator; 3MC surveys; European Social Survey; cross-country comparisons

1 Introduction

Multinational, multiregional, and multicultural (3MC) surveys can advance our understanding of cultural differences, as well as how they change over time. While 3MC surveys are essential for social research, collecting such data adds new layers of complexity to every facet of the survey process (Harkness et al. 2010; Johnson et al. 2018; Lyberg et al. 2021; Survey Research Center 2016). Inferences from 3MC surveys rest on the assumption that they are collected in comparable measurement conditions in each group, (Johnson 1998; Smith 2018) and failing to do so can result in differential survey errors, where data collected in one or more participating countries may be biased to a greater degree than others (Couper and de Leeuw 2003). The fact that many cross-national surveys are repeated over time adds to the complexity because of the implicit assumption that nonresponse mechanisms are stable over time. As relates to nonresponse errors, this is difficult to ensure in the present survey climate where response rates have been declining in most countries (de Leeuw et al. 2018; Wagner and Stoop 2019). It is therefore essential to examine whether 3MC survey data suffer from nonresponse errors, and whether these occur at different rates in participating countries and/or over time.

In analyses of 3MC survey data, the absence of differential survey errors is often assumed, without consideration for the possibility that different selection mechanisms may be at work in different groups. We have limited evidence to support the notion that differential nonresponse bias does not occur due to the scarcity of equivalently measured auxiliary variables, as such information can be hard to come by in 3MC contexts (Lyberg et al. 2021). Auxiliary variables refer to data present for both respondents and nonrespondents (e.g., from the sampling frame, prior survey rounds, paradata, and other sources). In this paper we explore the use of interviewer observations, a form of paradata that can be collected for both respondents and nonrespondents (Kreuter 2013; Olson 2013). Given the limited number of

Corresponding author: Hafsteinn Einarsson, University of Iceland, Reykjavík, Iceland (Email: hbe@hi.is)

auxiliary variables available in cross-national surveys, turning to interviewer observations may be helpful in a range of survey procedures aiming to improve the representativeness of the data, such as fieldwork monitoring, informing changes to survey design or in post-survey adjustment.

In this paper, we use interviewer observations to model survey representativeness over time in the European Social Survey (ESS), a repeated, cross-national survey. Our measures of representativeness are R-indicators, describing the variation in subgroup response propensities when all predictors are categorical (Schouten et al. 2009). However, equivalent R-indicator comparisons means that the same underlying response model, link function and predictors are required across all countries and over time (Schouten et al. 2012). To this end we will use the ESS Contact Form Data which registers the outcome of each contact attempt, as well as a number of other variables, including interviewer observations. We examine whether this information relates to target variables in the survey and to survey response. R-indicators are computed separately for each country-round combination to understand how survey nonresponse affects cross-country and over time comparisons with respect to representativeness. We find that interviewer observations are associated with nonresponse in most country-round combinations and that the strength of this association exceeds that of sample frame variables. Nevertheless, as they are strongly associated with both response propensities and target variables, interviewer observations fail to meet the criteria for informing nonresponse weighting adjustments. We conclude by discussing implications for survey practice.

2 Background and research questions

2.1 Nonresponse in repeated cross-national surveys

Cross-national survey research is guided by the ‘principle of equivalence’ (Jowell 1998). This entails pursuing the highest degree of comparability of survey data, by standardising survey designs across participating countries (Stoop et al. 2010). To investigate whether differential nonresponse bias occurs researchers often turn to analysing contact data (Blom et al. 2010). If contact data are measured under comparable conditions, they can be used to investigate whether differential nonresponse associations occur across countries or over time.

Meaningful comparisons of representation in cross-national surveys are challenging, as they depend on the methodological choices made in participating countries. Considerable methodological variation can make comparisons of traditional survey quality estimates, such as response rates, unfeasible (Smith 2007, 2018). Allowing for some cross-national variation, the general trend identified by the literature indicates that response rates are falling over time (e.g. Beullens et al. 2018; de Heer 1999; de Leeuw et al. 2018). However, much less is known regarding trends in representativeness, as relates to sample composition, between countries and over time.

The main challenge that arises when comparing sample quality in repeated cross-national surveys is that few variables are measured, let alone made publicly available, for the full sample (Jabkowski and Cichocki 2019). One method of studying sample quality involves using internal criteria, such as the expected 50/50 gender split (Sodeur 1997). This method has been utilized to explain country variations in unit nonresponse on the gender variable, highlighting the importance of sample type, as samples drawn from registers of individual persons tend to produce more representative samples than other sample types (Jabkowski and Kołczyńska 2020; Kohler 2007; Menold 2014). Jabkowski and Cichocki (2019) also examine sample quality by external criteria, where sample distributions from the ESS are compared to external sources of “gold standard” data (Groves 2006), such as census, register or high-quality survey data. However, they discourage external evaluations relying on survey data, as surveys which employ rigorous survey methodology are likely to suffer from correlated error sources.

Other studies on cross-national nonresponse errors have focused on reluctant respondents and on paradata. Billiet et al. (2007) found some indications of differences between co-operative, easy-to-convert, and hard-to-convert respondents in patterns that varied by country, which is indicative of differential errors. However, extrapolating this information to those who refuse participation may be unfeasible. Billiet et al. (2009) devise several methods for assessing and adjusting for nonresponse errors using ESS data, including examinations of bias on interviewer observations as well as comparing reluctant and cooperative respondents. Kreuter et al. (2007) explore the use of interviewer observations in weighting but find the strength of the association too weak to correct for nonresponse bias. Finally, Stoop et al. (2010) provides a number of lessons learned on improving survey response from the first rounds of the ESS. While many studies have approached the topic of representation cross-national surveys, there are still many open questions.

2.2 Modelling the representativeness of repeated cross-national surveys

In recent years, survey research has increasingly moved from a narrow focus on response rates to indicators of sample composition or direct proxies of nonresponse bias

(Wagner 2012). The R-indicator (1; Schouten et al. 2009, 2011, 2012),

$$R = 1 - 2S(\rho), \tag{1}$$

is an increasingly popular example of such indicators, where S(ρ) is the standard deviation of estimated response propensities ρ, or the related coefficient of variation (CV), which standardises the variance of response propensities (Schouten and Shlomo 2017).

R-indicators estimate survey representativeness through the variation in response propensities across strata formed by auxiliary variables. By decomposing the influence of predictors of response, partial R-indicators (or partial CVs) allow researchers to identify the variables most strongly associated with non-representativeness and identify strata which are underrepresented in the responding sample (Schouten et al. 2011; Shlomo et al. 2012). Here we focus on unconditional partial R-indicators (P_U) by variables,

$$P_U\,(X_k) = \sqrt{\frac{1}{N}\sum_{h=1}^{H} n_h(\overline{\rho}_h - \overline{\rho})^2} \tag{2}$$

which measures the variation between response categories H of variable $X_{k.}$ Partial R-indicators decompose the variance of the response propensities and take on a value between 0 and 0.5, where high partial R-indicators indicate that the variable is contributing to the lack of representativeness (Schouten et al. 2011).

In fieldwork, cases can be targeted to improve the representativeness as measured by the R-indicators (or CVs). While this does not necessarily reduce nonresponse bias, it can reduce the likelihood of worst-case scenarios occurring, as the R-indicator provides an upper bound for the absolute bias of a sample mean (Maximal Absolute Bias; MAB; Chun et al. 2018; Roberts et al. 2020; Schouten et al. 2009). However, to accurately estimate MAB, R-indicators must be correctly specified, by including important predictors of the response indicator (Nishimura et al. 2016). In cross-national surveys where input harmonisation is prioritised, R-indicators can be useful monitoring tools in coordinating data collection and can inform fieldwork interventions aiming at balancing sample composition.

Identifying auxiliary variables of equivalent quality to model response propensities can be a daunting task in cross-national surveys, where the sampling frames can be incompatible and other variables may be collected inconsistently. In such cases it may be useful to turn to paradata recorded in the process of survey data collection. Examples of paradata include call records, contact data, interviewer information and interviewer observations. Importantly, paradata can be collected on both respondents and nonrespondents and can therefore inform fieldwork interventions (Couper 1998; Kreuter 2013). Paradata can play an important role in data collection, but quality issues such as missing data, measurement error, and low correlations with outcomes of interest may reduce their utility (Olson 2013). Therefore, meticulous record keeping of paradata is crucial if they are to be used to inform fieldwork monitoring and interventions.

2.3 Mechanisms linking interviewer observations on residential characteristics and nonresponse bias

Interviewer observations contain information about sampled units, e.g. describing the neighbourhood, housing unit, or contact person, regardless of participation outcome (Kreuter et al. 2010; Olson 2013). Several surveys collect information on the characteristics of neighbourhoods, as these could be useful proxy measures of urbanicity and population density, which in turn can be associated with crime rates, general trust, and social cohesion (Groves and Couper 1998; Kreuter et al. 2007; Olson 2013). Therefore, they may be particularly relevant in crime surveys or surveys on educational attainment (Olson 2013). These variables have been associated with survey participation (Kreuter et al. 2007), as well as contactability and cooperation (Matsuo et al. 2010). Neighbourhood observations could also function as an indirect measure of socioeconomic status as undesirable neighbourhood characteristics may be associated with key survey measures like income and educational attainment.

Similarly, observations on the sampled housing unit are often collected in surveys. These can include observations on the condition of the building, whether the structure is single or multi-unit, and whether barriers to access (e.g. locked gates, entry phones) are present. These characteristics are often associated with contactability and cooperation in surveys. In particular, they are often associated with survey measures on socio-economic status such as income and home ownership (Olson 2013).

The survey literature provides mixed evidence on the utility of interviewer observations, as they may be weakly associated with response propensities and/or survey variables and can suffer from measurement error. As relates to representation, on the one hand, Kreuter et al. (2010) analysed several surveys and found weak correlations between auxiliary variables (including interviewer observations such as neighbourhood characteristics from the ESS) and target variables. Similarly, West et al. (2014) found that interviewer observations on household income and benefit status in a longitudinal study were weakly correlated to response propensities (due to low attrition rates), and only slightly improved estimates of key variables when prior-wave reports on the same features were included. On the other hand, Sinibaldi et al. (2014) find that interviewer observations better predict household income and receipt of unemployment benefits compared to commercial microdata from multiple sources (e.g. government records, surveys) on small areas or households. Similarly, Billiet et al. (2009) found that interviewer observations from the ESS neighbourhood characteristics form were in many cases related to response propensities and often correlated with education levels.

As relates to measurement, while complex interviewer observations can be hard to validate (West 2013), accuracy can be improved through training (West and Kreuter 2015, 2018; West and Li 2019) and validating interviewer observations at low costs through images or online search engines (Diego-Rosell et al. 2020) and may further improve the reliability of interviewer observations. Sinibaldi et al. (2013) show agreement rates of 97% between interviewer observations and self-reported census data on a type of housing unit question, but only 93% for noncontact cases. This suggests that while interviewers can in most cases accurately observe characteristics such as building type or neighbourhood conditions, if these variables are to play a central role in survey design, further validation would be desirable.

Given the mixed results in terms of data quality and nonresponse bias mitigation, it is not evident *a priori* that interviewer observations are suitable for reducing differential nonresponse bias in cross-national surveys. However, the utility of interviewer observations must be balanced with the limited number of auxiliary variables available for the full sample. Therefore, if one or more interviewer observations are related to response propensities, their potential uses in monitoring and/or informing changes in survey design should be studied.

2.4 Utilising interviewer observations in repeated cross-national studies

Interviewer observations may prove useful in a range of applications when mitigating the detrimental effects of nonresponse in repeated cross-national surveys. Response propensity models including these variables can be used as a tool for centralised fieldwork monitoring (Briceno-Rosas et al. 2020), where central co-ordinators may alert national agencies that they run the risk of producing differential survey errors by underrepresentation in categories of interviewer observations (or other auxiliary variables). Then, in turn, they can be used at the national level to inform case prioritisation schemes (Peytchev et al. 2010) during data collection. Increased focus on these underrepresented groups may also lead to advancements of targeted methods for achieving co-operation. Moving from standardised to targeted approaches may increase co-operation rates with underrepresented groups (Groves et al. 2000; Lynn 2017). Using interviewer observations in monitoring fieldwork cases or targeted approaches in data collection may result in more balanced responding samples and may reduce nonresponse bias.

While an association between auxiliary variables and survey response may be useful on its own, the greatest utility arises when they also correlate with target variables in the survey. If this correlation does not exist, the increased variance of weighted survey means will not be matched with a corresponding reduction in bias (Little and Vartivarian 2005). However, collecting a balanced response set can also be a goal on its own, as this reduces variances due to varying survey weights and produces more efficient estimates (Zhang 2022; Zhang and Wagner 2022). Auxiliary variables can also be used to inform changes to survey design. Responsive (Groves and Heeringa 2006) and adaptive (Schouten et al. 2017; Wagner 2008) survey designs (RASD) may be considered as a means to reduce nonresponse bias and survey costs. RASDs have not received much attention in the context of 3MC surveys, as real-time monitoring data collection and implementing fieldwork interventions is difficult (Lyberg et al. 2021). One example comes from Beullens et al. (2018), who explored the potential for case prioritisation in the ESS and identified it as a viable method to reduce nonresponse bias. Nevertheless, given the speed of change in survey practice in recent years, implementing aspects of RASD in 3MC surveys in the coming years does not seem unfeasible. Should that be the case, identifying auxiliary variables which can inform data collection by consistently identifying non-representativeness is essential.

2.5 Research questions

To be useful in understanding non-response in repeated cross-national surveys, auxiliary variables, such as interviewer observations, should consistently be related to response propensities. To examine whether this is the case, and whether the associations are stable across participating countries and over time, we model R-indicators using interviewer observations and demographic variables from the ESS.

RQ1: *Can interviewer observations be used in response propensity models to produce “actionable” R-indicators for data collection monitoring in the ESS?*

R-indicators can be considered “actionable” if they can inform case prioritisation protocols which improve sample balance compared to uniform assignment or based solely on the demographic variables which are currently used to create survey weights.

RQ2: *Are R-indicators stable between countries and over time?*

As comparisons of survey data between countries and over time rest on the assumption of equivalent representativeness, multilevel regression models for change (Singer and Willett 2003), which predict R-indicator scores are presented. These models can be used to investigate if there are systematic trends in R-indicators over time and whether countries deviate from the average rate of change over time.

3 Data and methods

3.1 Data

The European Social Survey is a cross-national, face-to-face survey, which has been fielded biennially since 2002, with new cross-sectional samples for each round. Thirty-nine countries have participated in at least one round, and fifteen have participated in all ten rounds (European Social Survey 2023). One benefit of analysing data from the ESS is its reliance on input harmonisation, which entails using the same procedures to collect survey data, including the use of standardised questionnaires, modes, sampling frames, and fieldwork procedures (Lynn 2003). This strong focus on sample comparability is accompanied by high quality survey documentation (Jabkowski 2023; Jabkowski and Kołczyńska 2020).

Secondary analysis of ESS fieldwork is possible through the publicly available contact form datasets, which include detailed outcomes of each contact attempt made during the fieldwork period (European Social Survey 2021; Stoop et al. 2003). The ESS sampling units vary by country, with some sampling individuals and others sampling households or addresses. Information on nonrespondents is more detailed in countries where individuals are sampled, as the sample frame contains some unit characteristics, which cannot be registered for nonrespondents at the contact attempt in household or address based samples (Stoop et al. 2010). Three types of survey weights are produced for the ESS: design weights (for differential selection probabilities), post-stratification weights (to adjust for nonresponse) and population size weights to estimate totals. Post-stratification weights are calculated using two dimensions; region and GA/GAE (gender, age, and education) using estimates from external sources, such as the Labour Force Survey. While the ESS collects interviewer observations, they are not presently used to calculate survey weights (Lynn and Anghelescu 2018).

The ESS contact form data (European Social Survey European Research Infrastructure (ESS ERIC) 2012a, 2016, 2018a, b, c, d, e, f, g, 2020a) for rounds 1–10 contains a total of 242 country-round combinations. Data is available for both respondents and nonrespondents, and variables in the dataset include: country, round, type of sample, interviewer information, time, mode and outcome of contact attempts, information on the sampled unit (age, gender, household size (only available for individual sample frames)), presence of a listed telephone number, and interviewer observations. Variables derived from interviewer observations are qualitative assessments made by interviewers at the first contact attempt which describe characteristics of the residence and its neighbourhood.

For the purposes of the present analysis, the publicly available dataset is limited in three important aspects. First, design weights are not included in the publicly available contact form data. While design weights can be found in other ESS datasets for respondents, no information is available for nonrespondents. Therefore, this analysis is conducted on unweighted data. High variation in design weights could affect estimates, but Appendix Table 5 shows that there is low variation of design weights for respondents, reducing the risk of bias in our estimates based on unweighted data. Second, the demographic variables included are limited to gender and age, and are only available for nonrespondents where sample frames are individuals and are sampled in rounds 6–10 (European Social Survey 2021). Finally, not all ineligible categories are represented by variables in the dataset (e.g. deceased, emigrated), therefore our response rates will not perfectly match those in ESS quality reports (e.g. Beullens et al. 2014, 2015; Wuyts and Loosveldt 2019). Despite these limitations, the contact form data is a valuable resource for nonresponse research, as it allows for full-sample analysis on data collected in cross-national surveys with equivalent survey designs over time.

3.2 Measures and Methods

From the onset, the ESS contact forms have included the Neighbourhood Characteristics form (Stoop et al. 2003; Wuyts and Loosveldt 2019), which provide the data used in our analysis. These variables had been theorised to be linked to survey items on social involvement and trust (Kreuter et al. 2010). Because we analyse country-rounds separately, using the original categories results in strata with few cases. Therefore, we restrict our analysis of interviewer observations to two dichotomous variables: type of house, where 1 denotes a multi-unit structure, and neighbourhood characteristics, where 1 indicates the presence of undesirable characteristics. The latter is a composite indicator of the presence of one or more undesirable characteristics: litter, vandalism or the building being in a bad physical condition. A fifth interviewer observation variable, barriers to access (e.g., entry phone), is only available for rounds 5–9 and is therefore excluded from our analysis. Two demographic variables, age and gender, are included for countries with individual sample frames in rounds 6–10. The dependent variable is an indicator of survey response after ineligibles have been excluded, equivalent to response rate type 1 (AAPOR 2016). An overview of variables included in response propensity models (before and after recoding) can be found in Appendix Table 1.

To assess whether our interviewer observation variables meet Little and Vartivarian's (2005) criteria for nonresponse weighting adjustments, we merge the contact form datasets to their corresponding survey data of the respondents (European Social Survey European Research Infrastructure (ESS ERIC) 2012b, 2018h, i, j, k, l, m, 2020b, 2021, 2023) and conduct *t*-tests on the relationship between interviewer observations and three survey variables based on the existing literature. Kreuter et al. (2010) found relationships between neighbourhood observations in the ESS round 1 and survey variables measuring social involvement and general trust to be too weak for nonresponse adjustment. Here, we pool data for each country to study associations in the full ESS dataset. The survey variables examined have been hypothesized to be associated with interviewer observations. These include a measure of general trust (*'Most people can be trusted or you can't be too careful'; 0–10 scale)* as well as two key demographic variables education (in years; 0–60), and household income deciles (0–10). As we are conducting multiple comparisons, Bonferroni adjustments are used to set the relevant α-levels for six comparisons in 36 countries ($\alpha = 0.05 / (6*36) = 0.00023$).

A detailed breakdown of test statistics can be found in Appendix Table 6, showing that a link between interviewer observations and survey variables can often be established. What is also evident is that the type of variable matters, as there is mixed evidence for a relationship between interviewer observations and trust (23 out of 72 comparisons), but a significant relationship is quite often identified for education (53 out of 72 comparisons) and household income (49 out of 72 comparisons).

While significant differences in means can often be identified, Appendix Fig. 1 shows Cohen's d effect sizes for these comparisons, with most significant effects being small or moderate. This supports the notion, as Kreuter et al. (2010) had shown, that these interviewer observations are likely to be insufficient for informing nonresponse weighting adjustments under Little and Vartivarian's (2005) criteria in most instances. However, an association between interviewer observations and response propensities could be leveraged in fieldwork monitoring under RASD as this could affect the distribution of education variables, which are used to inform post-stratification weights in the ESS. Therefore, it is possible that using interviewer observations to inform fieldwork efforts could reduce the variance of ESS weights, which would in turn produce more efficient estimates.

To address RQs 1 and 2, R-indicators are calculated using complete cases (no item missingness) as follows: Of the 242 country-round combinations included in the datasets (see Appendix Table 3 for sample information, including response rates for each country-round combination), 47 are excluded due to missing data on interviewer observations (less than 80% complete cases; 38 country-rounds) or on the response indicator (2% or more; 9 country-rounds). Finally, five country-rounds are excluded from further analyses where demographic variables are included, due to high levels of missing data on demographic variables. An overview of the 197 country-round combinations analysed (and the 44 country-rounds which include demographic variables) can be found in Fig. 1.

Response propensities and R-indicators (with 95% confidence intervals) are estimated separately for each country-round combination using logistic regression models provided by the Representative Indicators for Survey Quality project (de Heij et al. 2015). All analyses are conducted on unweighted data in R 4.0.4 (R Core Team 2021).

Response propensities are first estimated for the probability of individual *i* in country *j* responding in round *t,* using the interviewer observations only model (3) for all eligible combinations (197 country-rounds). The dependent variable is an indicator of response, while the independent variables are indicators of living in a multi-unit building and the presence of one or more undesirable neighbourhood characteristics. An interaction effect between the independent variables is also included. We proceed to analyse data for countries sampling individual persons in rounds 6–10 (43 country-rounds). There, the "full" model (4) adds predictors age and gender (and their interaction effect) to the interviewer observations model. We produce R-indicators using the estimated response propensities from each model. To address RQ1, we consider an indicator 'actionable' if t-tests show that the R-indicators are significantly different from the value 1 (using a multiple-comparison correction based on the Bonferroni correction), as this would indicate that

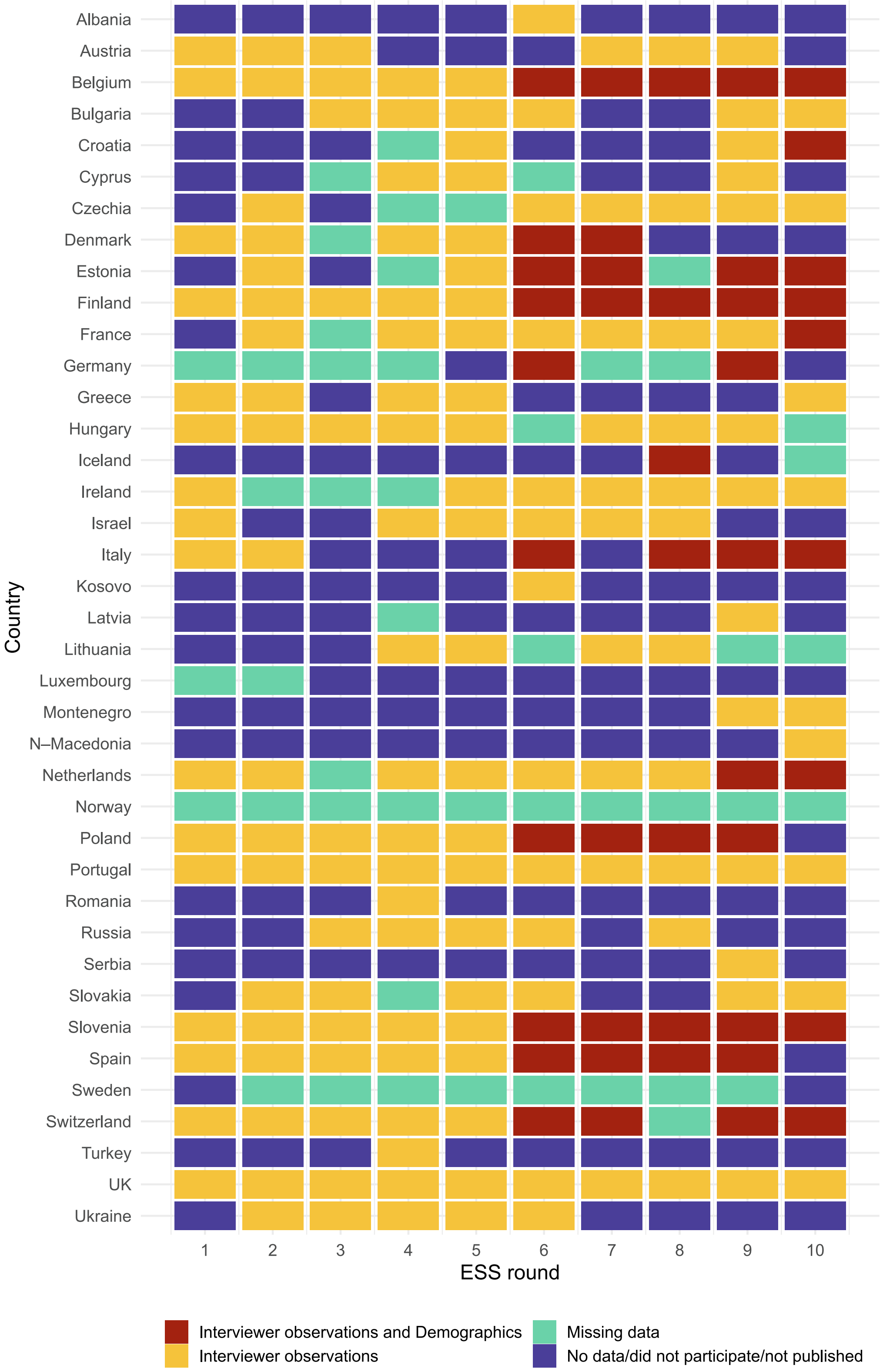

Fig. 1

Country-rounds included in the European Social Survey Contact Form Data by information available. Red country-rounds are included in all analyses, yellow country-rounds are included only in the interviewer observations only model. Green country-rounds are excluded due to missing data, and no data is available for blue country-rounds

Table 1

Descriptive statistics for ESS rounds 1–10 (complete cases (individual or household sample frame types), unweighted sample)

		Interviews	
Characteristic	Overall	*n*	%
Type of housing			
Multi-unit	260,465	140,307	54
Not multi-unit	370,634	221,559	60
Neighbourhood characteristics (Vandalism/Litter/Bad condition)			
None	404,581	229,485	57
One or more	226,518	132,381	58
Type of sample			
Household/Address	375,651	216,994	58
Individual person	255,448	144,872	57
Gender[a]			
Female	88,189	48,382	55
Male	82,437	43,975	53
Age[a]			
Under 35	41,157	22,786	55
35–49	40,487	22,221	55
50–64	35,878	20,775	58
65+	36,278	19,632	54
Overall	631,099	361,899	57

197 Country-rounds, average sample size = 3156. Breakdown by variable can be found in Appendix Table 2. Chi-square tests for all breakdown variables significant ($p < 0.001$)

[a] Information on age and gender is only available for all units in countries with individual sampling frames in rounds 6–10, a total of 43 country-round combinations

case prioritisation for the purposes of informing RASD is possible with the potential to have a more balanced sample.

$$\begin{aligned}&\text{Interviewer observations model :}\\&\text{Logit}\left(p_{\text{ijt}}\right) = \log(p_{\text{ijt}}/(1-p_{\text{ijt}}))\\&= \beta_0 + \beta_1 X_{\text{multi-unit,ijt}} +\\&\beta_2 X_{\text{neighbourhood,ijt}} + \beta_3 X_{\text{multi-unit*neighbourhood,ijt}},\end{aligned} \tag{3}$$

$$\begin{aligned}&\text{“Full” model :}\\&\text{Logit}(p_{\text{ijt}}) = \log(p_{\text{ijt}}/(1-p_{\text{ijt}}))\\&= \beta_0 + \beta_1 X_{\text{multi-unit,ijt}} +\\&\beta_2 X_{\text{neighbourhood,ijt}} + \beta_3 X_{\text{gender,ijt}} + \beta_4 X_{\text{age,ijt}} +\\&\beta_5 X_{\text{multi-unit*neighbourhood,ijt}} + \beta_6 X_{\text{gender*age,ijt}},\end{aligned} \tag{4}$$

To address RQ2, we estimate multilevel regression models for change (Singer and Willett 2003), where each row corresponds to a country-round, using R-indicators (R_{jt}) as the dependent variable from the interviewer observations only model (3) (estimated separately for each country-round, rounds 1–10; $N = 197$). These multilevel regression models enable the estimation of change in time t and estimate country j variability while controlling for survey characteristics.

The unconditional means model in (5) includes an overall intercept at the first round analysed (γ_{00}), a random intercept for each country (ζ_{0j}) and a residual term (ϵ_{jt}), which are assumed to be normally distributed with a mean of 0. The unconditional change model in (6) adds survey round as a fixed effect (γ_{10}) and a random slope for each country over time (ζ_{1j}). Finally, the conditional change model in (7) introduces survey characteristics—type of sample frame (individual person or household/address), response rates and response rates squared[1]—as controls. Random effects from models including random slopes are assumed to follow a bivariate normal distribution, with both random

[1] R-indicators are products of variances, and therefore have parabolic shapes, so low response rates (and very high response rates) can produce high R-indicators. However, in the typical range of 30-70% response rates for many social surveys an association between high response rates and high R-indicators should manifest if they are related.

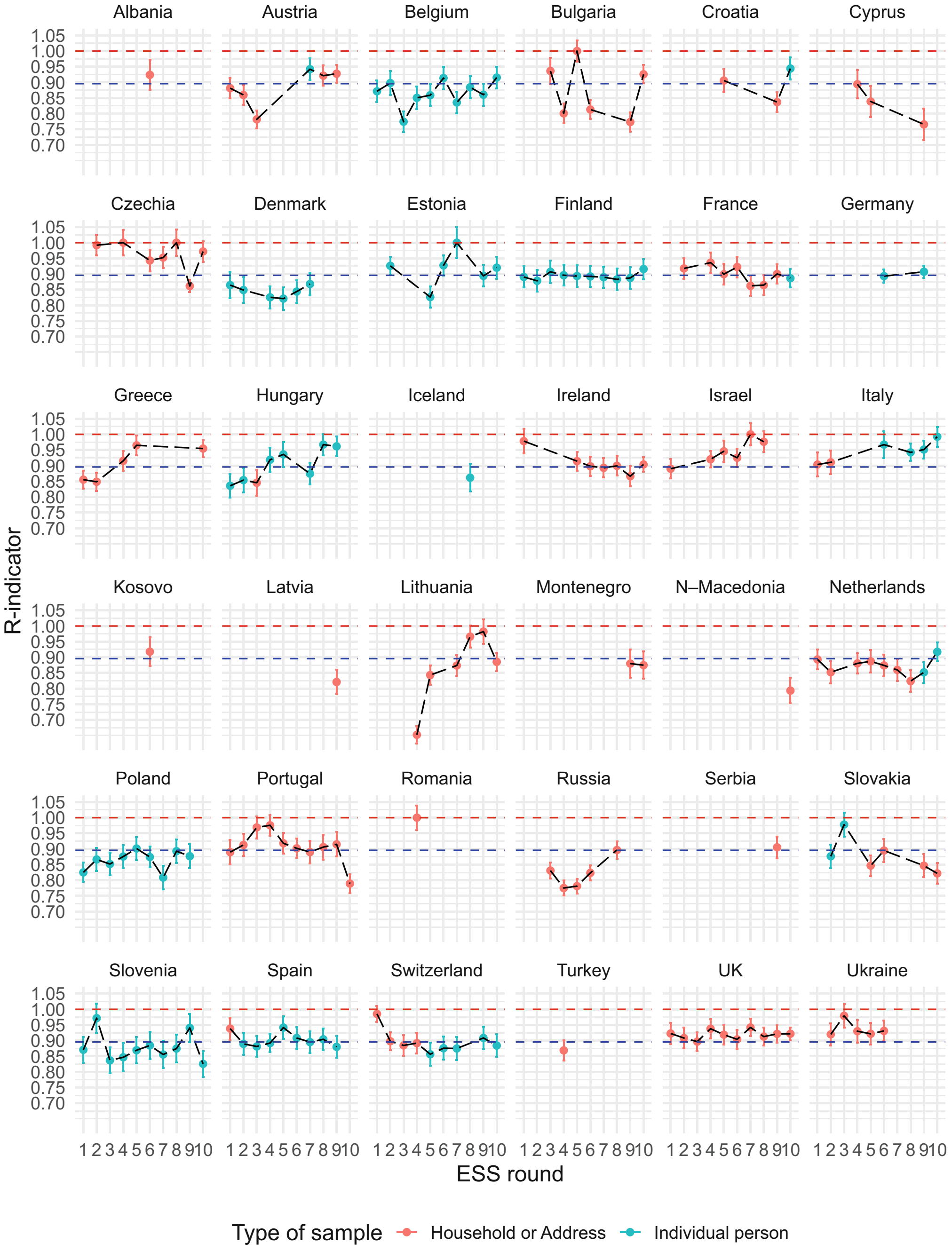

Fig. 2

R-indicators over time, based on models using only interviewer observations. Dashed blue line indicates overall mean

intercepts and random slopes having means of 0 and that residuals are normally distributed with a mean of 0, unknown variances and unknown covariance.

Unconditional means model :

$$R_{jt} = \gamma_{00} + \zeta_{0j} + \epsilon_{jt} \tag{5}$$

Unconditional change model :

$$\mathrm{R}_{\mathrm{jt}} = \gamma_{00} + \gamma_{10}\mathrm{Round}_{\mathrm{jt}} + \zeta_{0\mathrm{j}} + \zeta_{1\mathrm{j}}\mathrm{Round}_{\mathrm{jt}} + \epsilon_{\mathrm{jt}} \tag{6}$$

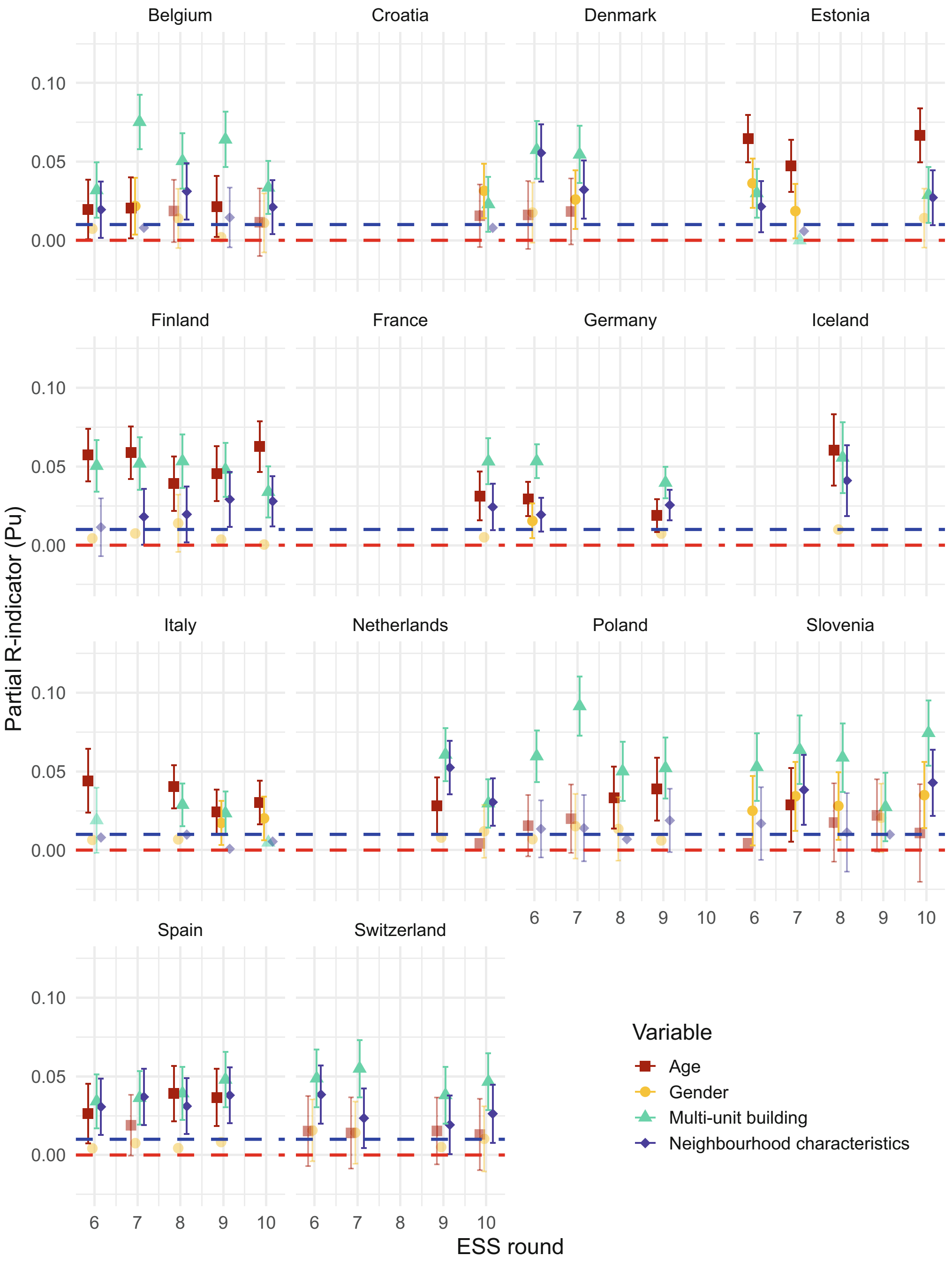

Fig. 3

Partial-R indicator by variables for countries sampling individual persons in ESS rounds 6–10, full model. If estimate of the lower bound of the confidence interval is below 0.00 (highlighted by a dashed red line in the plots) points are made transparent. Because standard errors for partial indicators (Pu) are calculated by dividing by the variance of response propensities, extreme values will occur if the variance is close to zero, therefore, confidence intervals for variables with Pu smaller than 0.01 (dashed blue line) are not shown. Appendix Fig. 2 shows Pus for interviewer observations only model (round 1–10)

Conditional change model :

$$R_{\mathrm{jt}} = \gamma_{00} + \gamma_{10}\mathrm{Round}_{\mathrm{jt}} + \gamma_{20}\mathrm{Sampletype}_{\mathrm{jt}} + \gamma_{30}\mathrm{Response\,rate}_{\mathrm{jt}} + \gamma_{40}\mathrm{Response\,rate}^2_{\mathrm{jt}} + \zeta_{0j} + \zeta_{1j}\mathrm{Round}_{\mathrm{jt}} + \epsilon_{\mathrm{jt}} \tag{7}$$

In analysing the multilevel models, we turn to the intraclass correlation coefficient (ICC; Singer and Willett 2003) to show how much of the variation in R-indicators is explained by differences between countries (*j*) as opposed to individual country-rounds (*jt*).

4 Results

Descriptive statistics for the full unweighted dataset which includes all countries and rounds in the ESS contact form data, after excluding cases with missing data on the interviewer observations, are found in Table 1. They show that 57% of complete cases lead to a fully completed interview. For the combined dataset, all characteristics show significant differences between categories. Those who live in multi-unit housing respond at lower rates than others (54% vs 60%), while the difference in response rates by the presence or absence of undesirable neighbourhood characteristics (57% vs 58%) is smaller. However, analyses of the combined data may hide variation between countries and/or over time.

To address RQ1, we explore whether interviewer observations (type of building and neighbourhood characteristics) produce 'actionable' R-indicators for the 197 country-rounds. 'Actionable' R-indicators mean that t-tests show that the R-indicator is significantly different from 1 which shows a lack of representativeness and the need for introducing interventions in fieldwork and informing case prioritisation strategies. Fig. 2 (with the full list of coefficients in Appendix Table 4) shows that a simple model with only two variables derived from interviewer observations in model (3) produces some evidence of non-representativeness in a large majority of country-rounds, as R-indicators are 'actionable' in 160 of 197 country-rounds after applying a Bonferroni adjustment for multiple comparisons (α = 0.05/197 = 0.00025). Furthermore, it shows that R-indicators cluster around the overall mean (0.893, SD = 0.053) regardless of sample type (Household/Address frame = 0.896, SD = 0.059—Individual person frame = 0.889, SD = 0.043). This shows that the average R-indicator score is high with little variation indicating a stable relationship across countries and over time using the same set of predictors.

In general, response propensity models with more auxiliary variables should be expected to produce lower R-indicators as there is more variability in the response propensities. This is supported by our results as the 'full' model (4) produces 'actionable' (α = 0.05/43 = 0.0011) information in each of the 43 country-rounds analysed (mean = 0.867, SD = 0.031). This suggest that including both demographic characteristics and interviewer observations could inform case prioritisation protocols in each country-round combination where individuals are sampled.

In the next stage of the analysis, we explore the effects of interviewer observations relative to demographic variables in model (4) using partial R-indicators (2; denoted Pu). These variable types can be compared by studying partial R-indicators from the "full" model (4) for countries with individual sampling frames in rounds 6–10, a total of 43 country-rounds. As shown in Fig. 3 (Pu's for the interviewer observations only model (3) are shown in Appendix Fig. 2), interviewer observations reliably produce higher Pu's, indicating a comparatively larger contribution to a lack of representativeness, relative to demographic variables. However, as evidenced by the number of instances where confidence intervals overlap, the relative contributions of specific variables are often not statistically significant from each other. Nevertheless, in each analysed country-round combination, monitoring partial R-indicators during fieldwork could have informed interventions based on one or both interviewer observations which may have increased the overall R-indicator and improved the representativeness in the sample.

To address RQ2, which deals with differences in R-indicators between countries and across time, we return to the output of the interviewer observations only model (3), R-indicators for each country-round combination, to produce multilevel regression models in (5), (6), and (7) where R-indicators are the dependent variable (Table 2). These models can be used to formally test the amount of variation in R-indicators across country and time. The unconditional means model (5), which includes only a random intercept, does not indicate important differences between countries. The random effects illustrate this, as between country variation is very low (0.001), which is not surprising given the low levels of overall variation. The ICC for the unconditional means model indicates that around 23% of the overall variation in R-indicators is explained by differences between countries, with Appendix Fig. 3 showing that only four countries differ from the average trend (Russia, Czechia, Italy, and Israel).

The multilevel models indicate that fixed effects for survey characteristics do not predict R-indicators. In the unconditional change model in (6), the fixed effect for survey round is not significant, and we fail to reject the null hypothesis of no change in R-indicators over time. The conditional change model in (7) addresses the relationship between survey characteristics and R-indicators. Apart from the significant effect for response rates squared (included to account for the expected parabolic distribution of R-indicators), a lack of significant fixed effects indicates that

Table 2

Multilevel linear regression models predicting R-indicators using only interviewer observations, ESS rounds 1–10. Variances of random effects refer to overall variance (σ^2), variance in means between countries (τ_{00}), variance in slopes (τ_{11}), and covariance between the country random intercept and country slope (ϱ_{01})

	Unconditional means model				Unconditional change model				Conditional change model			
		95% C.I.				95% C.I.				95% C.I.		
Predictors	Estimates	Lower	Upper	*p*	Estimates	Lower	Upper	*p*	Estimates	Lower	Upper	*p*
Intercept	0.892	0.881	0.904	*<0.001*	0.892	0.868	0.916	*<0.001*	0.884	0.857	0.911	*<0.001*
Round	–	–	–	–	0.000	–0.004	0.004	0.875	0.001	0.003	0.005	0.587
Type of sample: Individual person	–	–	–	–	–	–	–	–	0.008	0.013	0.029	0.440
Response Rate	–	–	–	–	–	–	–	–	0.085	0.048	0.218	0.210
Response Rate^2	–	–	–	–	–	–	–	–	0.141	0.038	0.245	*0.008*
Random Effects												
σ^2			0.002				0.002				0.002	
τ_{00}			0.001				0.003				0.003	
τ_{11}			–				0.000				0.000	
ρ_{01}			–				–0.857				–0.866	
ICC			0.234				0.440				0.485	
N			36				36				36	
Observations			197				197				197	
logLik			303.183				302.357				299.022	

Reference groups: Type of sample: Household/Address

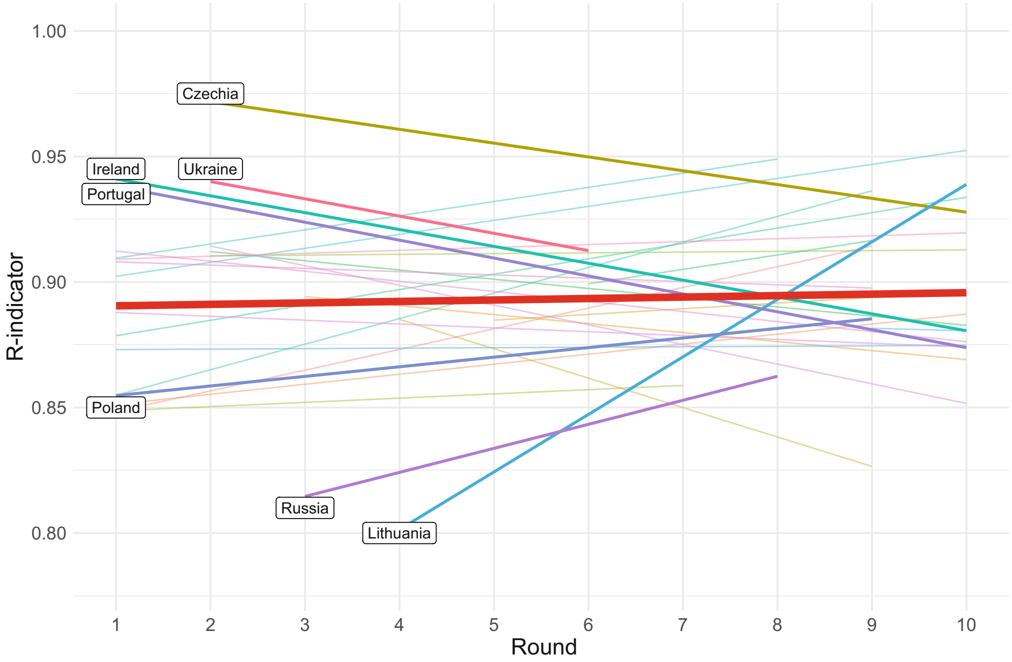

Fig. 4

Predicted R-indicator, conditional change model, red line denotes grand mean. Transparent lines denote countries with non-significant slopes, opaque lines denote significant slopes (Czechia, Ukraine, Portugal, Lithuania (The magnitude of the slope for Lithuania can largely be attributed to an outlier value in round 4 (R = 0.65), the first round in which Lithuania participated and may be indicative of data management issues.), Ireland, Poland, and Russia)

our measure of representativeness is not related to response rates, the type of sample used, and that they do not change substantively over time. However, when survey round is included as a predictor in the unconditional (6) and conditional change models (7), the ICC goes up to 0.440 and 0.485 respectively. This suggests that around half of the overall variation in these models can be explained by differences between countries. This could indicate important differences between countries but must also be viewed in terms of the low levels of overall variation.

The residuals from all three models deviate from the normal distribution at the tails, which is not remedied by a transformation on the response variable. This may be a feature of the nonlinearity of R-indicators or indicate omitted coefficients or that country-rounds with unusually low R-indicators suffer from serious data quality issues. Predicted R-indicators from the conditional change model are shown in Fig. 4 (Appendix Fig. 6 for the unconditional change model, Appendix Figures 3–5 show estimated coefficients for random effects), showing that seven of 36 countries (Czechia, Ireland, Lithuania, Ukraine, Portugal, Poland, and Russia) have slopes which deviate significantly from the overall trend.

5 Discussion

This paper explored the feasibility of using interviewer observations in response propensity models which in turn could be utilised to monitor fieldwork and inform interventions through the calculation of R-indicators or used in post-survey weighting adjustments. Interviewer observations are weakly associated with target variables in the ESS but are in many instances moderately associated with socio-economic variables (income and education). The fact that interviewer observations can predict differences in variables among respondents for some, but not all participating countries, may be indicative of differential nonresponse bias in repeated cross-national surveys, if nonrespondents within the same classes hold similar attitudes. Nevertheless, interviewer observations are not suitable for informing nonresponse weights, but could play a role in informing fieldwork efforts. Balancing the responding sample on interviewer observations could reduce nonresponse bias by improving representativity assuming that nonrespondents in the group have similar characteristics to the respondents.

To assess survey representativeness as relates to interviewer observations (RQ2), R-indicators were calculated based on a response propensity model using only two dichotomous variables (type of house; presence of undesirable neighbourhood characteristics). These R-indicators produce evidence of low, but statistically significant, levels of non-representativeness in a large majority of country-rounds in the ESS (RQ1; accounting for multiple comparisons). The consistency with which this association is identified suggest that interviewer observations can be leveraged in response propensity modelling in the ESS and other cross-national surveys. In addition, we find that interviewer observations generally produce higher partial R-indicators than the demographic characteristics (age, sex) currently used to inform post-stratification weights in the ESS showing that they contribute more to a lack of representativeness.

Finally, no clear trends emerge when analysing R-indicators across countries and over time (RQ2), indicating that samples are mostly equivalent with respect to representativeness. This suggest a significant relationship between interviewer observations and survey response should be expected in most participating countries in the ESS in each round, which raises the question of how this relationship can be leveraged to produce high quality survey data. Prior research has shown a weak relationship between response rates and nonresponse bias (Groves and Peytcheva 2008). Our results are consistent with these findings, as multilevel models predicting R-indicators do not show a significant relationship with response rates and other survey characteristics. The lack of cross-national variation in the relationship between interviewer observations and survey response can be seen as a positive for the ESS, as it indicates a high degree of comparability of ESS data across countries with regards to representativeness on these paradata variables.

In the background section, several applications of how an association between auxiliary variables and survey response were discussed, including fieldwork monitoring, informing the implementation of targeted fieldwork protocols (e.g., RASDs), and post-survey adjustment strategies. The consistency of the association between interviewer observations and the response indicator, and the instances where they are associated with target variables provides a basis for further research e.g., through simulation studies. This topic is further explored in Einarsson et al. (2023). In practice, interviewer observations and other auxiliary variables could be monitored centrally to flag sample imbalances in participating countries. Given the recent announcement that the ESS is moving to a mixed-mode data collection protocol (European Social Survey 2022), the lessons of prior rounds could be incorporated in an adaptive design protocol, where response propensities are predicted to assign high-propensity respondents to cost-effective modes, and low-propensity respondents to modes which are associated with high response rates.

This study has utilized interviewer observations, under the assumption that limited demographic and target variables are available for non-respondents. These variables are among the few covariates available to survey researchers for analyses of nonresponse cross-nationally and over time in 3MC surveys. Interviewer observations are subject to measurement error, but collecting the variables discussed in this paper should be possible with low error rates. Nevertheless, if they are to inform fieldwork interventions, standardisation of interviewer training techniques across countries, as well as external validation methods should be considered. Of course, relying solely on interviewer observations to inform fieldwork interventions is unlikely to be considered in practice, and a greater choice of covariates should be considered when informing changes to survey design.

Aside from the issues relating to interviewer observations already discussed, there are important limitations to the inferences that can be drawn from this study. Design weights for non-respondents are not included in the publicly available dataset, and while many of the countries analysed use uniform design weights or have very limited variation in them (see Appendix Table 5), this may affect some country-rounds. Several country-rounds had to be excluded from our analysis due to missing data. Finally, if response propensity models are misspecified, such as by omitting important predictors, the estimated R-indicator will underestimate the upper bound of the potential nonresponse bias (MAB; Nishimura et al. 2016). Our analysis has focused on top-level models using few variables which do produce some ‘actionable’ information, but which do not include important predictors of response. Therefore, these R-indicators cannot be considered estimates of the true MAB. Including more predictors of nonresponse could produce indicators that allow for more discrimination in outcomes, which would in turn lead to better case prioritisation protocols and a closer approximation of the true MAB.

This paper has brought together insights from several topics within survey research: the utility of interviewer observations (i.e., paradata), using R-indicators to understand the data collection process, as well as conducting cross-country comparisons on the representativeness of the sample over time. While more research is needed to understand the relationship between interviewer observations and survey variables, this research shows that interviewer observations may aid in our understanding of nonresponse errors, and may play a role in fieldwork monitoring, informing adaptive survey designs or in post-survey adjustment in repeated cross-national surveys.

Acknowledgement This work was supported by the University of Manchester and by the Icelandic Research Fund grant no 239630-051. Replication materials can be found in the supplementary files of this paper and at https://osf.io/scw39/. The European Social Survey’s inte-

grated contact form data files and survey data are publicly available at https://ess-search.nsd.no.

References

AAPOR (2016). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (9th edn.). American Association for Public Opinion Research.

Beullens, K., Matsuo, H., Loosveldt, G., & Vandenplas, C. (2014). *Quality report for the European Social Survey, round 6, London*

Beullens, K., Loosveldt, G., Denies, K., & Vandenplas, C. (2015). *Quality Matrix for European Social Survey, round 7, London*

Beullens, K., Loosveldt, G., Vandenplas, C., & Stoop, I. (2018). Response rates in the European social survey: increasing, decreasing, or a matter of fieldwork efforts? *Survey Methods: Insights from the Field*. https://doi.org/10.13094/SMIF-2018-00003.

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

Billiet, J.B., Philippens, M., Fitzgerald, R., Stoop, I.A. (2007). Estimation of nonresponse bias in the European Social Survey: using information from reluctant respondents. Journal of Official Statistics, 23, 135–162.

Blom, A. G., Jäckle, A., & Lynn, P. (2010). The use of contact data in understanding cross-national differences in unit nonresponse. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. P. Mohler, B. E. Pennell & T. W. Smith (Eds.), *Survey methods in multicultural, multinational, and multiregional contexts* (pp. 333–354). Hoboken: John Wiley & Sons. https://doi.org/10.1002/9780470609927.ch18.

Briceno-Rosas, R., Butt, S., & Kappelhof, J. (2020). Improving central monitoring of fieldwork in cross-national surveys: the case of the fieldwork management system in the European social survey. *Survey Methods: Insights from the Field*. https://doi.org/10.13094/SMIF-2020-00004.

Chun, A. Y., Heeringa, S. G., & Schouten, B. (2018). Responsive and adaptive design for survey optimization. *Journal of Official Statistics*, *34*, 581–597. https://doi.org/10.2478/jos-2018-0028.

Couper, M. (1998). Measuring survey quality in a CASIC environment. In *Proceedings of the survey research methods section of the American statistical association* (pp. 41–49).

Couper, M. & De Leeuw E. (2003). Nonresponse in cross-cultural and cross-national surveys. In J.A. Harkness, M. Braun, B. Edwards, T.P. Johnson, L. Lyberg, P.P. Mohler, B. E. Pennell T.W. Smith (Eds.), Survey methods in multicultural, multinational, and multiregional contexts, 157–177. Hoboken: John Wiley Sons. https://doi.org/10.1002/9780470609927.ch18.

Diego-Rosell, P., Nichols, S., Srinivasan, R., & Dilday, B. (2020). Assessing community wellbeing using Google street-view and satellite imagery. In *Big data meets survey science* (pp. 435–486). Wiley. https://doi.org/10.1002/9781118976357.ch15.

Einarsson, H., Cernat, A., & Shlomo, N. (2023). Adaptive designs in repeated cross-national surveys: a simulation study. *Journal of Survey Statistics and Methodology*, *0*, 1–26. https://doi.org/10.1093/jssam/smad038.

European Social Survey (2022). ESS announces change to data collection methodology. https://www.europeansocialsurvey.org/about/singlenew.html?a=/about/news/essnews0130.html

European Social Survey (2023). Participating countries. https://www.europeansocialsurvey.org/about/participating-countries

European Social Survey (2021). *Integrated contact form data files*. London: ESS ERIC Headquarters, Centre for Comparatve Social Surveys, City University London.

European Social Survey European Research Infrastructure (ESS ERIC) (2012a). *ESS2—data from contact forms, edition 3.2 [Data set]*. Sikt—Norwegian Agency for Shared Services in Education and Research. https://doi.org/10.21338/ess2cfe03_2.

European Social Survey European Research Infrastructure (ESS ERIC) (2012b). *ESS2—integrated file, edition 3.6 (Italy not included) [Data set]*. Sikt—Norwegian Agency for Shared Services in Education and Research. https://doi.org/10.21338/ESS2E03_6.

European Social Survey European Research Infrastructure (ESS ERIC) (2016). *ESS8—data from contact forms, edition 3.0 [Data set]*. Sikt—Norwegian Agency for Shared Services in Education and Research. https://doi.org/10.21338/ess8timee01.

European Social Survey European Research Infrastructure (ESS ERIC) (2018l). *ESS6—integrated file, edition 2.4 [Data set]*. Sikt—Norwegian Agency for Shared Services in Education and Research. https://doi.org/10.21338/ESS6E02_4.

European Social Survey European Research Infrastructure (ESS ERIC) (2018m). *ESS7—integrated file, edition 2.2 [Data set]*. Sikt—Norwegian Agency for Shared

Services in Education and Research. https://doi.org/10.21338/ESS7E02_2.

European Social Survey European Research Infrastructure (ESS ERIC) (2018a). *ESS3—data from contact forms, edition 1.1 [Data set]*. Sikt—Norwegian Agency for Shared Services in Education and Research. https://doi.org/10.21338/ess3cf_ed1_1.

European Social Survey European Research Infrastructure (ESS ERIC) (2018b). *ESS4—data from contact forms, edition 2.1 [Data set]*. Sikt—Norwegian Agency for Shared Services in Education and Research. https://doi.org/10.21338/ess4cf_e02_1.

European Social Survey European Research Infrastructure (ESS ERIC) (2018c). *ESS1—data from contact forms, edition 1.0 [Data set]*. Sikt—Norwegian Agency for Shared Services in Education and Research. https://doi.org/10.21338/ess1cfe01.

European Social Survey European Research Infrastructure (ESS ERIC) (2018d). *ESS5—data from contact forms, edition 2.1 [Data set]*. Sikt—Norwegian Agency for Shared Services in Education and Research. https://doi.org/10.21338/ess5cfe02_1.

European Social Survey European Research Infrastructure (ESS ERIC) (2018e). *ESS6—data from contact forms, edition 2.0 [Data set]*. Sikt—Norwegian Agency for Shared Services in Education and Research. https://doi.org/10.21338/ess6cfe02.

European Social Survey European Research Infrastructure (ESS ERIC) (2018f). *ESS7—data from contact forms, edition 2.1 [Data set]*. Sikt—Norwegian Agency for Shared Services in Education and Research. https://doi.org/10.21338/ess7cfe02_1.

European Social Survey European Research Infrastructure (ESS ERIC) (2018g). *ESS9—data from contact forms, edition 3.0 [Data set]*. Sikt—Norwegian Agency for Shared Services in Education and Research. https://doi.org/10.21338/ess9cfe03.

European Social Survey European Research Infrastructure (ESS ERIC) (2018h). *ESS1—integrated file, edition 6.6 [Data set]*. Sikt—Norwegian Agency for Shared Services in Education and Research. https://doi.org/10.21338/ESS1E06_6.

European Social Survey European Research Infrastructure (ESS ERIC) (2018i). *ESS3—integrated file, edition 3.7 (Latvia and Romania not included) [Data set]*. Sikt—Norwegian Agency for Shared Services in Education and Research. https://doi.org/10.21338/ESS3E03_7.

European Social Survey European Research Infrastructure (ESS ERIC) (2018j). *ESS4—integrated file, edition 4.5 (Austria and Lithuania not included) [Data set]*. Sikt—Norwegian Agency for Shared Services in Education and Research. https://doi.org/10.21338/NSD-ESS4-2008.

European Social Survey European Research Infrastructure (ESS ERIC) (2018k). *ESS5—integrated file, edition 3.4 (Austria not included) [Data set]*. Sikt—Norwegian Agency for Shared Services in Education and Research. https://doi.org/10.21338/ESS5E03_4.

European Social Survey European Research Infrastructure (ESS ERIC) (2020a). *ESS10—data from contact forms, edition 1.0 [Data set]*. Sikt—Norwegian Agency for Shared Services in Education and Research. https://doi.org/10.18712/ess10cf.

European Social Survey European Research Infrastructure (ESS ERIC) (2020b). *ESS8—integrated file, edition 2.2 [Data set]*. Sikt—Norwegian Agency for Shared Services in Education and Research. https://doi.org/10.21338/ESS8E02_2.

European Social Survey European Research Infrastructure (ESS ERIC) (2021). *ESS9—integrated file, edition 3.1 [Data set]*. Sikt—Norwegian Agency for Shared Services in Education and Research. https://doi.org/10.21338/ESS9E03_1.

European Social Survey European Research Infrastructure (ESS ERIC) (2023). *ESS10 integrated file, edition 3.1 [Data set]*. Sikt—Norwegian Agency for Shared Services in Education and Research. https://doi.org/10.21338/ess10e03_1.

Groves, R.M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*, 646–675. https://doi.org/10.1093/poq/nfl033.

Groves, R.M., & Couper, M.P. (1998). *Nonresponse in household interview surveys*. New York: Wiley. https://doi.org/10.1002/9781118490082.

Groves, R.M., & Heeringa, S.G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *169*, 439–457. https://doi.org/10.1111/j.1467-985X.2006.00423.x.

Groves, R.M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public Opinion Quarterly*, *72*, 167–189. https://doi.org/10.1093/poq/nfn011.

Groves, R.M., Singer, E., & Corning, A. (2000). Leverage-saliency theory of survey participation. *Public Opinion Quarterly*, *64*, 299–308. https://doi.org/10.1086/317990.

Harkness, J.A., Braun, M., Edwards, B., Johnson, T.P., Lyberg, L., Mohler, P.P., Pennell, B.E., & Smith, T.W. (2010). *Survey methods in multicultural, multi-*

national, and Multiregional contexts. Hoboken. https://doi.org/10.1002/9780470609927.

de Heer, W. (1999). International response trends: results of an international survey. *Journal of Official Statistics*, *15*, 129–142.

de Heij, V., Schouten, B., & Shlomo, N. (2015). *RISQ manual 2.1. Tools in SAS and R for the computation of R-indicators, partial R-indicators and partial coefficients of variation*

Jabkowski, P. (2023). Increase in the quality of methodological documentation of cross-national pan-European multi-wave surveys over the last 40 years—a research note. *International Journal of Social Research Methodology*, *26*, 817–824. https://doi.org/10.1080/13645579.2022.2097394.

Jabkowski, P., & Cichocki, P. (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. https://doi.org/10.18148/srm/2019.v13i2.7383.

Jabkowski, P., & Kołczyńska, M. (2020). Sampling and fieldwork practices in europe: analysis of methodological documentation from 1,537 surveys in five cross-national projects, 1981–2017. *Methodology*, *16*, 186–207. https://doi.org/10.5964/meth.2795.

Johnson, T.P. (1998). Approaches to equivalence in cross-cultural and cross-national survey research. *ZUMA-Nachrichten Spezial*, , 1–40.

Johnson, T.P., Pennell, B.-E., Stoop, I.A.L., & Dorer, B. (Eds.). (2018). *Advances in comparative survey methods*. Hoboken: Wiley. https://doi.org/10.1002/9781118884997.

Jowell, R. (1998). How comparative is comparative research? *American Behavioral Scientist*, *42*, 168–177. https://doi.org/10.1177/0002764298042002004.

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

Kreuter, F. (2013). *Improving Surveys with Paradata*. Hoboken: Wiley. https://doi.org/10.1002/9781118596869.

Kreuter, F., Lemay, M., & Casas-cordero, C. (2007). Using proxy measures of survey outcomes in post-survey adjustments: examples from the European Social Survey (ESS). In *2007 JSM proceedings: papers presented at the joint statistical meeting* (pp. 3142–3149).

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T.M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R.M., & Raghunathan, T.E. (2010). Using proxy measures and other correlates of survey outcomes to adjust for non-response: Examples from multiple surveys. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *173*, 389–407. https://doi.org/10.1111/j.1467-985X.2009.00621.x.

de Leeuw, E.D., Hox, J.J., & Luiten, A. (2018). International nonresponse trends across countries and years: an analysis of 36 years of labour force survey data. *Survey Methods: Insights from the Field*. https://doi.org/10.13094/SMIF-2018-00008.

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

Lyberg, L., Pennell, B.-E., Hibben, K.C., de Jong, J., Behr, D., Fitzgerald, R., Granda, P., Guerrero, L.L., Gyuzalyan, H., Johnson, T., Kim, J., Mneimneh, Z., Moynihan, P., Robbins, M., Schoua-Glusberg, A., Sha, M., Smith, T.W., Stoop, I., Tomescu-Dubrow, I., Zavala-Rojas, D., & Zechmeister, E.J. (2021). *AAPOR/WAPOR task force report on quality in comparative surveys*

Lynn, P. (2003). Developing quality standards for cross-national survey research: five approaches. *International Journal of Social Research Methodology*, *6*, 323–337. https://doi.org/10.1080/13645570210132848.

Lynn, P. (2017). From standardised to targeted survey procedures for tackling non-response and attrition. *Survey Research Methods*, *11*, 93–103. https://doi.org/10.18148/srm/2017.v11i1.6734.

Lynn, P., & Anghelescu, G. (2018). *European social survey round 8 weighting strategy*. https://doi.org/10.1017/s0960116318000143.

Matsuo, H., Billiet, J., Loosveldt, G., & Malnar, B. (2010). *Response-based quality assessment of ESS round 4: results for 30 countries based on contact files*

Menold, N. (2014). The influence of sampling method and interviewers on sample realization in the european social survey. *Survey Methodology*, *40*, 105–123.

Nishimura, R., Wagner, J., & Elliott, M. (2016). Alternative indicators for the risk of non-response bias: a simulation study. *International Statistical Review*, *84*, 43–62. https://doi.org/10.1111/insr.12100.

Olson, K. (2013). Paradata for nonresponse adjustment. *Annals of the American Academy of Political and Social Science*, *645*, 142–170. https://doi.org/10.1177/0002716212459475.

Peytchev, A., Riley, S., Rosen, J., Murphy, J., & Lindblad, M. (2010). Reduction of nonresponse bias in surveys through case prioritization. *Survey Research Methods*, *4*, 21–29. https://doi.org/10.18148/srm/2010.v4i1.3037.

R Core Team (2021). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Roberts, C., Vandenplas, C., & Herzing, J. M. E. (2020). A validation of R-indicators as a measure of the risk of bias using data from a nonresponse follow-up survey. *Journal of Official Statistics*, *36*, 675–701.

Schouten, B., & Shlomo, N. (2017). Selecting adaptive survey design strata with partial R-indicators. *International Statistical Review*, *85*, 143–163. https://doi.org/10.1111/insr.12159.

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

Schouten, B., Shlomo, N., & Skinner, C. (2011). Indicators for monitoring and improving representativeness of response. *Journal of Official Statistics*, *27*, 231–253.

Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N., & Skinner, C. (2012). Evaluating, comparing, monitoring, and improving representativeness of survey response through R-indicators and partial R-indicators. *International Statistical Review*, *80*, 382–399. https://doi.org/10.1111/j.1751-5823.2012.00189.x.

Schouten, B., Peytchev, A., & Wagner, J. (2017). *Adaptive survey design*. Boca Raton: CRC Press.

Shlomo, N., Skinner, C., & Schouten, B. (2012). Estimation of an indicator of the representativeness of survey response. *Journal of Statistical Planning and Inference*, *142*, 201–211. https://doi.org/10.1016/j.jspi.2011.07.008.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: modeling change and event occurrence*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195152968.001.0001.

Sinibaldi, J., Durrant, G. B., & Kreuter, F. (2013). Evaluating the measurement error of interviewer observed paradata. *Public Opinion Quarterly*, *77*, 173–193. https://doi.org/10.1093/poq/nfs062.

Sinibaldi, J., Trappmann, M., & Kreuter, F. (2014). Which is the better investment for nonresponse adjustment: purchasing commercial auxiliary data or collecting interviewer observations? *Public Opinion Quarterly*, *78*, 440–473. https://doi.org/10.1093/poq/nfu003.

Smith, T. W. (2007). Survey non-response procedures in cross-national perspective: the 2005 ISSP non-response survey. *Survey Research Methods*, *1*, 45–54. https://doi.org/10.18148/srm/2007.v1i1.50.

Smith, T. W. (2018), *Improving multinational, multiregional, and multicultural (3MC) comparability using the total survey error (TSE) paradigm*, *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*, (T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, and B. Dorer, eds.), Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118884997.

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

Stoop, I., Devacht, S., Billiet, J., Loosveldt, G., & Philippens, M. (2003). *The development of a uniform contact description form in the ESS*. 14th International Workshop on Household Survey Nonresponse, Leuven. (pp. 1–8).

Stoop, I., Billiet, J., Koch, A., & Fitzgerald, R. (2010). *Improving survey response: lessons learned from the European social survey*. Chichester: Wiley. https://doi.org/10.1002/9780470688335.

Survey Research Center (2016). *Guidelines for best practice in cross-cultural surveys*. Ann Arbor: Survey Research Center, Institute for Social Research, University of Michigan. http://ccsg.isr.umich.edu/

Wagner, J. (2012). A comparison of alternative indicators for the risk of nonresponse bias. *Public Opinion Quarterly*, *76*, 555–575. https://doi.org/10.1093/poq/nfs032.

Wagner, J. R. (2008). *Adaptive survey design to reduce nonresponse bias*. Ann Arbor: University of Michigan. PhD thesis

Wagner, J., & Stoop, I. A. L. (2019). Comparing Nonresponse and Nonresponse biases in multinational, multiregional, and multicultural contexts. In T. P. Johnson, B.-E. Pennell, I. A. L. Stoop & B. N. J. U. S. A. Dorer Hoboken (Eds.), *Advances in comparative survey methods: multinational, multiregional, and multicultural contexts* (pp. 807–833). John Wiley. 3MC.

West, B. T. (2013). An examination of the quality and utility of interviewer observations in the National Survey of Family Growth. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *176*, 211–225. https://doi.org/10.1111/j.1467-985X.2012.01038.x.

West, B. T., & Kreuter, F. (2015). A practical technique for improving the accuracy of interviewer observations of respondent characteristics. *Field Methods*, *27*, 144–162. https://doi.org/10.1177/1525822X14549429.

West, B. T., & Kreuter, F. (2018). Strategies for increasing the accuracy of interviewer observations of respondent features: evidence from the US national survey of family growth. *Methodology*, *14*, 16–29. https://doi.org/10.1027/1614-2241/a000142.

West, B. T., & Li, D. (2019). Sources of variance in the accuracy of interviewer observations. *Sociological Methods and Research*, *48*, 485–533. https://doi.org/10.1177/0049124117729698.

West, B. T., Kreuter, F., & Trappmann, M. (2014). Is the collection of interviewer observations worthwhile in an economic panel survey? New evidence from the German labor market and social security (PASS) study. *Journal of Survey Statistics and Methodology*, *2*, 159–181. https://doi.org/10.1093/jssam/smu002.

Wuyts, C., & Loosveldt, G. (2019). *Quality matrix for the European social survey, round 8: overall fieldwork and data quality report*. London.

Zhang, S. (2022). Benefits of adaptive design under suboptimal scenarios: a simulation study. *Journal of Survey Statistics and Methodology*, *10*, 1048–1078. https://doi.org/10.1093/jssam/smab051.

Zhang, S., & Wagner, J. (2022). The additional effects of adaptive survey design beyond post-survey adjustment: an experimental evaluation. *Sociological Methods & Research*, *0*, 1–34. https://doi.org/10.1177/00491241221099550.