Klinke, Dennis et al.

Social media sampling is an effective way to access hard to survey populations and low prevalence groups

International Journal of Social Research Methodology

Provided in Cooperation with:
WZB Berlin Social Science Center

Suggested Citation: Klinke, Dennis et al. (2025) : Social media sampling is an effective way to access hard to survey populations and low prevalence groups, International Journal of Social Research Methodology, ISSN 1464-5300, Taylor & Francis, London, Iss. Latest Articles, pp. 1-20, https://doi.org/10.1080/13645579.2025.2564866

This Version is available at:
https://hdl.handle.net/10419/331212

INTERNATIONAL JOURNAL OF SOCIAL RESEARCH METHODOLOGY
https://doi.org/10.1080/13645579.2025.2564866

Routledge
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Social media sampling is an effective way to access hard to survey populations and low prevalence groups

Dennis Klinke[a,b], Jannes Jacobsen[c], Manuel Dierse[b], Thorsten Faas[d], Denis Gerstorf[a], Hannah Helal[b], Swen Hutter[b,e], David Schieferdecker[f], Hanna Schwander[g], Christian von Scheve[b] and Jule Specht[a]

[a]Humboldt-Universität zu Berlin, Faculty of Life Sciences, Department of Psychology, Berlin, Germany; [b]Freie Universität Berlin, Department of Political and Social Sciences, Institute of Sociology, Berlin, Germany; [c]German Center for Integration and Migration Research (DeZIM), Cluster "Data-Methods-Monitoring", Berlin, Germany; [d]Freie Universität Berlin, Department of Political and Social Sciences, Otto-Suhr-Institute of Political Science, Berlin, Germany; [e]WZB Berlin Social Science Center, Center for Civil Society Research, Berlin, Germany; [f]Freie Universität Berlin, Department of Political and Social Sciences, Institute for Media and Communication Studies, Berlin, Germany; [g]Humboldt-Universität zu Berlin, Department of Social Sciences, Berlin, Germany

ABSTRACT

Non-probability samples have become increasingly popular despite some criticism. This paper examines social media sampling as a tool for accessing hard-to-survey populations. Our study targeted individuals in Germany using ads on Facebook and X, yielding 4,590 respondents. We compared these samples with high-quality probability samples (SOEP, ESS) through measurement equivalence analysis of shared measures between samples. Results show that our social media sampling strategy yielded effective sample sizes for our target population that exceeded those from SOEP and ESS by ratios between 2:1 and 5:1. Our findings suggest that non-probability sampling can be a viable method for researchers examining relational patterns among variables in hard-to-survey populations. Because we observe varying levels of measurement equivalence, rigorous methodological strategies for post-hoc analyses are recommended. We propose measurement equivalence analysis as a post-hoc assessment strategy to quantify the analytical effectiveness of the employed sampling strategy.

ARTICLE HISTORY
Received 24 October 2024
Accepted 10 September 2025

KEYWORDS
Probability sampling; non-probability sampling; social media sampling; survey methodology; measurement invariance

Introduction

The survey-based research community has witnessed a growing interest in non-probability-based sampling methods in recent years (Cornesse et al., 2020). This surge in attention can be attributed to several factors: First, non-probability-based sampling methods have become increasingly cost-effective (Dutwin & Buskirk, 2017; Sakshaug

CONTACT Dennis Klinke klinkede@hu-berlin.de Humboldt-Universität zu Berlin, Faculty of Life Sciences, Department of Psychology, Berlin, Germany

Supplemental data for this article can be accessed online at https://doi.org/10.1080/13645579.2025.2564866.

et al., 2019, 2020; Zack et al., 2019), facilitated by the emergence of large companies offering access to respondents at minimal expense. Second, traditional probability-based approaches are facing challenges due to the rapid decline in response rates (Curtin et al., 2005), which raises concerns about their ability to yield unbiased parameter estimates (Brick & Tourangeau, 2017; Peytchev et al., 2010). If small population groups are of interest, the relationship between response rates and costs poses particularly high hurdles for researchers.

Recent comparisons of probability vs. non-probability approaches concluded that, despite all limitations, probability sampling yields better data quality, e.g. regarding coverage bias, non-response error, and measurement error, than non-probability-based sampling (e.g. Lavrakas et al., 2022). However, these studies usually compared the two strategies for the general population. In some instances, researchers are not aiming at inference for the general population but aim at certain sub-populations. These subpopulations may be *hard to survey* (Tourangeau, 2014), because they are *hard to identify*, like the LGBTQI* community (e.g. Kühne & Zindel, 2020) or *hard to persuade* like immigrants (e.g. Poetzschke, 2022). Additionally, sub-populations may be *hard to survey* (Tourangeau, 2014) simply due to their size relative to the total population. In such cases, the advantages of probability-based sampling can only be achieved with unrealistically large sample sizes (an illustrative power analysis for various sample sizes and effect strengths can be found in the Online Supplementary Materials). Therefore, under what conditions can non-probability sampling yield data that are sufficiently comparable to probability-based samples for hard to survey populations (Tourangeau, 2014) and thus inferentially adequate for a plethora of social science research aims? The present study aims to shift the debate from the usefulness of non-probability studies for general population surveys to hard-to-survey populations, while discussing pitfalls and post-hoc strategies to assess accuracy and enhance the credibility of such samples.

We present an original survey targeting highly politicized individuals as a fit-for-purpose application for non-probability sampling. After laying out why non-probability sampling was seen as a promising approach for our research goals, we answer our research question by comparing responses to several items in our survey to the results from those surveys from which these items were drawn, specifically the internationally established probability-based German Socio-economic Panel (SOEP) and European Social Survey (ESS). Aside from descriptive statistics for sociodemographics, we tested measurement equivalence between our own non-probability sample and the SOEP and ESS studies for three measures: personality, affect, and participation. We show that social media sampling can effectively reach and oversample a specific target population. Furthermore, our findings suggest that non-probability sampling can be a viable method for specific subpopulations if rigorous methodological strategies are employed and some level of homogeneity for the target population can be expected.

Why (non-)probability sampling?

The attractiveness of probability sampling for researchers lies in the central limit theorem. The central limit theorem states that as the sample size approaches infinity, the probability distribution of parameters will closely approximate the normal distribution, thus allowing researchers to quantify confidence in their sample (Kohler et al., 2019).

Hence, for descriptive measurements of populations with non-homogenous units or average treatment effects for heterogeneous populations, probability samples remain the only way for reliable measurements.

Having said this, probability samples have shortcomings too: Collecting such high-quality samples comes with significant costs and time commitment, and given the research aims, probability samples may be practically unfeasible when following the *total survey error paradigm* (Biemer, 2010). To circumvent cost considerations for individual researchers, large collaborative projects such as the ESS, SOEP, and the European Values Survey make probability samples accessible to researchers worldwide. However, even with these high-profile surveys, resources are limited, and questionnaire development is often a closed shop. This makes it difficult for researchers to use such studies for highly specialized research projects. In addition, these large-scale probability surveys focus mostly on the general population. Researchers interested in special subpopulations that are *hard to survey* (Tourangeau, 2014) because they are *hard to sample*, *hard to identify*, *hard to find or contact*, *hard to persuade* or *hard to interview* (Tourangeau, 2014), may find such general population surveys of limited use for their purposes because the subpopulation they are interested in is inadequately, be it qualitatively or quantitatively, represented, hence creating *frame errors* (Biemer, 2010) such as noncoverage. Thus, it may be theoretically possible but practically unfeasible to create a survey frame that neatly captures the target population and allows for probability sampling. There exists, for example, no register of immigrants without residence permits to draw from, leaving researchers with the choice of resorting to general population surveys that may or may not address the impediments that make their target population *hard to survey* in the first place, or venturing out to non-probability samples. Consequently, research on some *hard-to-survey* groups cannot be carried out adequately based on these large studies. Researchers interested in such subpopulations or non-mainstream variables are increasingly using non-probability samples to overcome these issues.

Various studies have shown that respondents in typical non-probability samples differ from respondents in probability samples (Zack et al., 2019) and, thus, measures in non-probability samples differ from corresponding measures in probability samples (Einarsson et al., 2022; Roulin, 2015; Zack et al., 2019). Regarding reliability, MacInnis et al. (2018) implemented a replication study with a set of 50 measures of 40 benchmark variables captured via probability- and non-probability sampling strategies. Their main result was that probability samples, regardless of whether interviewed by telephone or internet, are most accurate, notwithstanding substantial drops in response rates compared to earlier studies. Opt-in samples via the internet showed the worst accuracy, not even improved by common post-stratification strategies (MacInnis et al., 2018). Regarding validity, Einarsson et al. (2022) employed a latent factor model to assess the measurement equivalence of six multi-item attitude measures in Germany and Austria between probability- and non-probability samples and found only two of them to reach full measurement equivalence. Similarly to MacInnis et al. (2018), poststratification methods were not a reliable strategy to improve measurement equivalence and weighting affected coefficients in unpredictable ways (see also Pasek, 2016). Finally, Dutwin and Buskirk (2017) found the mean absolute bias for non-probability samples was twice as

high and more varied as for low-response probability samples – after weighting and/or matching.

Even though some evidence suggests that some measures are more robust than others (see e.g. the discussion about point estimates and relations between variables by Pasek, 2016) and weighting can improve the reliability of estimates (Cornesse et al., 2020; Wang et al., 2015), it is safe to say that *ceteris paribus*, one needs to be more cautious when analyzing and interpreting non-probability samples. While some argue researchers should generally stick to probability samples (e.g. Baker et al., 2013; Cornesse et al., 2020), there are certainly arguments for the use of non-probability samples, as Baker et al. (2013) address in their 'fit for purpose' argument.

One such purpose may be the research of a small or hard-to-reach population, which typically is not sufficiently represented in general population studies. As most of the debate centers on general population surveys, it remains unanswered whether the criticism holds empirically true for such purposes. We argue that the shortcomings of non-probability samples should be less severe for parameter estimation if the target population is small and/or reasonably homogenous. If the number of observations of a small target population in a general population survey is prohibitively small or accessing respondents in strict probabilistic ways is unfeasible, non-probability sampling may be fit for purpose.

Research aim

For this study, we aim to empirically test whether non-probability sampling can be an effective alternative to probability sampling in the case of *hard-to-survey* populations. A non-probability sampling strategy for a *hard-to-survey* population may be considered effective if: (i) when oversampling of the target population yields an effective sample size that enables more robust inferential statistical analyses with enhanced statistical power compared to a probability-based sampling strategy (quantitative); and (ii) when the effective sample yields sufficient measurement equivalence, loadings equivalence that is, to ensure the validity of the inferential statistics (qualitative). The *hard-to-survey* population in our case is politically active members of civil society. Such individuals are a highly relevant object of study in the social sciences. Their beliefs, attitudes, motivations, and behaviors are crucial to understanding processes of social cohesion and political contestation. They can be found in established probability surveys such as the SOEP and ESS, but oftentimes in such small numbers that subgroup analyses are impossible. Thus, they are considered *hard to sample* (Tourangeau, 2014) due to their characteristic as a *minor domain*, comprising 1 to 10 percent of the general population (Kalton, 2009, 2014; Kish, 1987; Tourangeau, 2014). We recruited political activists via a campaign on social media where political discourse, mobilization, and organization are widespread (McClain et al., 2024). We then compared our sample of political activists to two established high-quality probability samples focusing on (a) demographic composition and (b) measurement equivalence of core variables.

Method

Data sets

Non-probability sample

We recruited a non-probability sample as part of the research project 'Social Cohesion and Civil Society. Interaction Dynamics in Times of Disruption'. Following a shift in civil society research that acknowledges potentially disruptive, even detrimental aspects of uncoerced, collective engagement (Grande, 2022), we were interested in the politically contentious aspects of civil society and thus focused on three hot-topic issues in contemporary German politics: climate change and the environment, migration and integration, and renting and housing. Interested in politically active members of civil society with strong opinions on these issues, we ran a visually aided ad campaign on social media, namely the platforms X,[1] and Facebook (see Appendix for pictures of all ads). Since users first react to visual cues when scrolling through their feed (Kühne & Zindel, 2020), we used stock pictures with blatant symbolizations of these political issues, such as a stamp showing the word 'Asylum'. 'Some people are interested in different societal topics. Which of the following topics are you interested in?' was implemented as a filter item in our questionnaire, where individuals were barred from further participation when they indicated no interest in either of the three aforementioned issues (more than one issue could be named).

Using the targeting of predefined audience offered by the platforms can help to show the ads only to users who have a higher likelihood to belong to the subpopulation, however the accuracy of the targeting significantly varies from variable to variable (for Facebook, see Grow et al., 2022; Sances, 2019). Therefore, we only relied on the platforms targeting to address individuals of legal age, currently living in Germany, as well as their interest in climate change and the environment, migration and integration, and renting and housing. The platforms did not offer to target users for their interest in topics related to rent or housing politics. For this issue, we instead targeted only users living in metropolitan areas, where rent is a pressing issue.[2] Our CAWI-Survey, running for six weeks from August to September 2021, generated 4,806 respondents.

Established probability samples for comparison

As there is no access to the true parameters of the population, we use surveys that are recognized for their quality, namely SOEP, and ESS. All samples of SOEP are multi-stage random samples which are regionally clustered. The sample units (households) are selected either by random-walk or via public registers. The SOEP is commonly used as a reference dataset for samples not covering the full population of interest (Siedler et al., 2009). The SOEP Core for 2019 alone consists of a gross sample of 15,339 households out of which 12,481 participated. The net sample for 2019 alone consists of 20,842 adults, 1,319 youths, and 1,672 children.

The ESS is a cross-national survey that has been conducted biennially in over 30 European countries since 2002. Countries must follow general guidelines, such as representativeness for all persons aged 15 and over, strict probability sampling, and a minimum sample size of 1,500. In Germany, the data collection of round 9 was carried out from 29 August 2018, to 4 March 2019, through computer-assisted face-to-face interviews and achieved a response rate of 27.6%. The survey utilized a multistage

probability sampling procedure based on a stratified two-stage design. Municipalities were chosen with a probability proportional to the size of the population aged 15 and above. During the second stage of sampling, 44 individuals per PSU, aged 15 years or older, were selected using random sampling. The sample size achieved for Germany is 2,358.

We provide descriptive statistics comparing our nonprobability sample to SOEP and ESS on sociodemographic variables, as well as those variables that we intentionally adopted from SOEP and ESS in our survey that capture engagement. All statistics provided by SOEP and ESS are weighted according to their recommendations.[3] Additionally, statistics will be provided for Facebook and X separately, because, even when differences in user demographics are set aside, the advertising algorithms remain a black box (Zindel, 2023), warranting caution in analysis.

Measures

For the comparison of our non-probability sample recruited via social media with high-quality probability samples of the ESS and SOEP, we measured a variety of indicators ranging from socio-demographics over psychological variables to behaviors specific to highly politicized individuals. Importantly, we had to rely on variables for which we can reasonably assume temporal stability. The data from our original survey, the SOEP and the ESS were collected within a span of 3 years.

To start with, we asked respondents about their *socio-demographic background*. This included their age, gender, and education. Respondents also identified the state and the size of the city they currently live in (less than 5,000, 5,000 to 20,000, 20000 to 100,000, and more than 100,000) , and the state in which they live. In terms of socio-economics, respondents indicated how many people they share their household with (categorized as under 14, between 14 and 17, and 18+), their joint disposable household income (< 1,000 €, 1,000–1,999€, 2,000–3,999€, 4,000–5,999€, >6,000€), and their current employment status (full-time, part-time, self-employed, training/school/university, unemployed, other). In addition, we were interested in relationship status as well as their migration history (i.e. whether both of their parents were born in Germany or not). The wording of the items ensured a direct comparability to the ESS and SOEP.

Our survey contained two measures of *political participation* that were utilized for the *post hoc* construction of the comparative samples. First, we included an item battery that is used in the ESS. Respondents were asked: 'There are different ways of improving things in Germany or help prevent things from going wrong. During the last 12 months, have you done any of the following?' Respondents then indicated either yes or no for six items: contacted a politician or government official; worn or displayed a campaign badge/sticker; signed a petition; took part in a lawful public demonstration; boycotted certain products; or posted or shared anything about politics online.[4] Second, we included two items from the SOEP. Respondents were asked to report how often they took part in various activities in their leisure time. Among the activities were 'participating in political parties, municipal politics, citizens' initiatives' and "doing volunteer work in clubs, associations, or social services." Respondents chose between 'daily', 'at least once per week', 'at least once per month' or 'seldom or never.'

Finally, we contained two psychological measures. First, we included the tested and validated *Big Five* Inventory-SOEP (BFI-S). This short measure uses 15 items to measure the personality traits neuroticism, extraversion, openness, conscientiousness, and agreeableness, consisting of 15 items (Schupp & Gerlitz, 2008). Second, we used the *Affective Well-Being* measure (Entringer et al., 2022). In four items, respondents indicated how often they felt upset, afraid, happy, and sad in the past 4 weeks, with answers ranging on a scale from 1 'very rarely' to 5 'very much.'

Subsamples, weighting, and analysis

We compare our non-probability sample to the probability samples of the ESS and SOEP in two steps. We started with a simple comparison of the sample compositions in terms of demographics, psychological variables, and participation. Subsequently, we compared the measurement equivalence of three specific constructs: the BFI-S, Affective Well-Being, and Political Participation.

We applied the same recruiting strategy for Facebook and X, yet decided to compare the subsamples from Facebook and X separately. This way, we could account for biases that result from differences in the user demographics between the two platforms. Moreover, platform algorithms decided which users were exposed to our ads. The parameters of the algorithms remain a black box, but they are most likely platform-specific (Zindel, 2023).

For the tests of measurement equivalence, we *post hoc* created comparable subsamples of politically active members of civil society with similar levels of participation across all samples. For the comparison to the SOEP, we reduced our Social Media Sample and the SOEP sample to respondents that indicated in both items that they participated at least once a week. This leaves us with 165 respondents in the SOEP sample and 1,327 in our sample (Facebook: $N = 595$; X: $N = 688$). For the comparison with the ESS, we reduced our Social Media Sample and the ESS sample to those respondents who indicated on a majority of the Political Participation items (at least 4 out of 6) that they had participated in the past 12 months. This leaves only 152 respondents from the ESS, while 2,524 (Facebook: $N = 850$; X: $N = 1192$) respondents from our sample meet this criterion.

ESS and SOEP samples are weighted according to their recommendations. For our sample, we provide unweighted as well as weighted data. Sample weights for respondents of our own sample were calculated by a raking method (using 'weightipy', Nelson et al., 2023) based on gender, education, and age grouped by region (see OSM for data). Weighting was performed after the initial sample was filtered by recruitment platform and engagement variables.

Whereas the bivariate comparison of the composition of the samples is straightforward, the comparison of measurement equivalence warrants further explanation. Latent variable frameworks are a popular and well-established approach to test for measurement equivalence (Davidov et al., 2014). A latent variable consists of observed variables, whose strength of relationship to the latent variables is measured by factor loadings. We implemented a multigroup confirmatory factor analysis (initially suggested by Jöreskog, 1971). This approach has the benefit that it can differentiate between different

levels of measurement equivalence, and researchers can perform analyses of nested models that iteratively become more stringent.

We followed Einarsson et al. (2022) in their approach and constructed five nested models: (1) the *configural model* states that the underlying structure of the factors that reflect the latent variable is equivalent, while still allowing the factor loadings to differ; (2) the *loadings model* fixes the factor loadings across the surveys, forcing them to be equal; (3) the *intercept model* furthermore sets the intercepts to be equal across the surveys; (4) the *means model* additionally sets the mean of the latent variable to be equivalent across the surveys; (5) the *residual model* inherits all the restrictions of the models above, while setting the variance of residuals to be equal across the surveys.

Configural equivalence would mean a similar structure of relationships of variables between the samples, e.g. that age relates to participation in both samples. Loadings equivalence allows for meaningful comparisons of regression coefficients between samples, which would allow for statements like 'age has a stronger effect on salary in the probability sample than in the non-probability sample.' At intercepts equivalence, the intercepts of regression models can be compared, too; absolute, not just relative differences can be inferred. Means and residual equivalence would mean the average levels of the dependent variable are the same across samples after controlling for predictors, eliminating fears of bias between samples, and that unexplained variance is equivalent between samples, indicating equal predictive power of regression models. We argue that, depending on the research aim, loadings equivalence can ensure qualitatively effective sampling, as point estimates should generally be treated with caution (see Morey et al., 2016 for discussion of the false confidence often induced by confidence intervals). For many research objectives, establishing relationships such as 'each additional year of age corresponds to a 2% decrease in participation' that hold regardless of sampling methods constitutes a sufficient level of equivalence.

As for the cut-off values for the comparative fit measures, we follow Chen (2007) in his suggestions derived from a variety of Monte Carlo simulations to assess the sensitivity of fit indices, from which we use the comparative fit index (CFI) as well as the root mean square error of approximation (RSMEA). As Δ CFI, *ceteris paribus*, tends to be higher when sample sizes differ greatly, Chen (2007) suggests a Δ CFI ≤ -0.005 supplemented by a Δ RSMEA ≥ 0.01 to determine measurement noninvariance.

Results

Quantitative effectiveness

To assess the quantitative effectiveness of our non-probability sampling strategy, we first report the relative oversampling of the target population in our Social Media Sample compared to the competing probability samples. Additionally, we provide descriptive statistics regarding the composition of the samples. Statistically significant differences between respondents recruited via Facebook and X, on the one hand, and the probability samples of the ESS and SOEP, on the other hand, are accentuated.

From 20,842 adult respondents in the SOEP, only 165 respondents (0.7%[5]) answered the BFI-S, the Affective Well-Being instrument, *and* fulfilled our *post hoc* criteria for being a member of the target population. This compares to 1,321 respondents out of our

initial Social Media Sample of 4,806 respondents (27.6%), yielding a sizable oversampling in relative as well as absolute terms. From 2,358 respondents in the ESS, only 152 respondents (6.4%) answered the Political Participation instrument *and* fulfilled our *post hoc* criteria. This compares to 2,524 respondents out of our initial Social Media Sample (52.5%), similarly yielding a significant oversampling of the target population. The importance of such an oversampling varies from case to case; however, *ceteris paribus* larger samples are preferable, especially when sub-group analyses are intended. Practically, in our case, this means that if we were to construct a standard linear multiple regression with five predictors and a two-sided alpha error probability of 0.05,[6] based on the ESS sample, statistical power could vary between 0.2 and 0.9 (effect size $\rho^2 = 0.02$ and $\rho^2 = 0.1$, respectively) when analyzing only respondents that took part in a demonstration or between 0.12 and 0.72 (effect size $\rho^2 = 0.02$ and $\rho^2 = 0.1$, respectively) when analyzing only respondents who wore a badge in the past 12 months. Using our sample, the very same analyses would yield statistical power of >0.99 in either case (for further details see OSM).

Table 1. Descriptive statistics for the variables of interest between our sample, SOEP, and ESS.

	Our NP-sample		SOEP	ESS
Item	Facebook	X		
Age	$\bar{x} = 53.38$ (±0.59) $\sigma = 12.87$	$\bar{x} = 47.07$ (±0.53) $\sigma = 13.27$	$\bar{x} = 51.61$ (±0.24) $\sigma = 17.56$	$\bar{x} = 50.34$ (±0.74) $\sigma = 18.44$
Gender (1 = M, 0 = F)	$\bar{x} = 0.70$ (±0.02) $\sigma = 0.46$	$\bar{x} = 0.65$ (±0.02) $\sigma = 0.48$	$\bar{x} = 0.49$ (±0.01) $\sigma = 0.50$	$\bar{x} = 0.49$ (±0.02) $\sigma = 0.50$
Education				
Primary	30.8 % (±2.12)	15.4 % (±1.43)	25.2 % (±0.58)	30.9 % (±1.87)
Secondary	35.8 % (±2.2)	35.3 % (±1.89)	42.9 % (±0.67)	42.8 % (±2)
Tertiary	33.4 % (±2.2)	49.3 % (±1.98)	31.8 % (±0.63)	26.3 % (±1.78)
Household Income <1000€	6.5 % (±1.13)	5.2 % (±0.88)	3.4 % (±0.25)	
1000€-1999€	18.1 % (±1.77)	12.7 % (±1.32)	16.2 % (±0.5)	
2000€-3999€	40.7 % (±2.26)	36.4 % (±1.9)	42 % (±0.67)	
4000€-5999€	23.2 % (±1.94)	28.1 % (±1.78)	24.1 % (±0.58)	
>6000€	11.6 % (±1.47)	17.7 % (±1.51)	14 % (±0.47)	
NUTS1 Region†	Mo = 10 $H = 2.46$ $D = 0.89$	Mo = 10 $H = 2.40$ $D = 0.89$	Mo = 10 $H = 2.44$ $D = 0.89$	Mo = 10 $H = 2.42$ $D = 0.88$
Political Participation	$\bar{x} = 3.99$ (±0.05) $\sigma = 1.16$	$\bar{x} = 4.07$ (±0.04) $\sigma = 1.13$	$\bar{x} = 4.81$ (±0.01) $\sigma = 0.55$	
Volunteering	$\bar{x} = 3.55$ (±0.06) $\sigma = 1.31$	$\bar{x} = 3.68$ (±0.05) $\sigma = 1.27$	$\bar{x} = 4.31$ (±0.02) $\sigma = 1.13$	
Contacted Politician	$\bar{x} = 0.51$ (±0.02) $\sigma = 0.50$	$\bar{x} = 0.49$ (±0.02) $\sigma = 0.50$		$\bar{x} = 0.17$ (±0.02) $\sigma = 0.37$
Worn a Badge	$\bar{x} = 0.30$ (±0.02) $\sigma = 0.46$	$\bar{x} = 0.30$ (±0.02) $\sigma = 0.46$		$\bar{x} = 0.06$ (±0.01) $\sigma = 0.23$
Signed Petition	$\bar{x} = 0.69$ (±0.02) $\sigma = 0.46$	$\bar{x} = 0.72$ (±0.02) $\sigma = 0.45$		$\bar{x} = 0.36$ (±0.02) $\sigma = 0.48$
Demonstrated	$\bar{x} = 0.30$ (±0.02) $\sigma = 0.46$	$\bar{x} = 0.33$ (±0.02) $\sigma = 0.47$		$\bar{x} = 0.09$ (±0.01) $\sigma = 0.28$
Boycotted Products	$\bar{x} = 0.68$ (±0.02) $\sigma = 0.47$	$\bar{x} = 0.71$ (±0.02) $\sigma = 0.45$		$\bar{x} = 0.37$ (±0.02) $\sigma = 0.48$
Post on Social Media	$\bar{x} = 0.88$ (±0.01) $\sigma = 0.32$	$\bar{x} = 0.87$ (±0.01) $\sigma = 0.34$		$\bar{x} = 0.19$ (±0.02) $\sigma = 0.39$

† Statistical difference is calculated by Mann-Whitney U tests. Only the State variable shows no significant difference for individuals recruited via Facebook when compared to ESS and SOEP; individuals recruited via X do differ significantly.
Statistics: *Mo* refers to the modal value, *H* to Entropy and *D* to Gini-Simpson Index.
Margin of Error: Margins of error are calculated on a 95% confidence level and are added in parentheses.

Table 2. Goodness of fit between our sample, SOEP, and ESS.

Measure	Weight	Model	X			Facebook		
			DoF	RMSEA	CFI	DoF	RMSEA	CFI
Big Five	Unweighted	Configural	8	0.041	0.977	8	0.051	0.968
		Loadings	12	**0.047**	**0.954***	12	0.045	0.962
		Intercepts	16	0.064	0.886	16	**0.048**	**0.944***
		Means	17	0.061	0.889	17	0.049	0.937
		Residuals	22	0.055	0.886	22	0.047	0.924
	Weighted	Configural	8	0.046	0.971	8	0.065	0.949
		Loadings	12	0.041	0.966	12	0.060	0.934
		Intercepts	16	**0.061**	**0.901****	16	0.057	0.921
		Means	17	0.058	0.905	17	0.057	0.917
		Residuals	22	0.051	0.903	22	0.054	0.903**
Affective Well-Being	Unweighted	Configural	2	0.000	1.000	2	0.105	0.984
		Loadings	5	0.000	1.000	5	0.076	0.979
		Intercepts	9	**0.082**	**0.948*****	9	**0.101**	**0.934*****
		Means	10	0.146	0.818	10	0.165	0.805
		Residuals	14	0.156	0.710	14	0.176	0.691
	Weighted	Configural	2	0.052	0.995	2	0.090	0.989
		Loadings	5	0.036	0.995	5	0.071	0.982
		Intercepts	9	**0.091**	**0.938*****	9	**0.092**	**0.947*****
		Means	10	0.144	0.827	10	0.155	0.831
		Residuals	14	0.154	0.722	14	0.170	0.717
Political Participation	Unweighted	Configural	16	0.093	0.724	16	0.092	0.680
		Loadings	21	0.086	0.693	21	0.091	0.584
		Intercepts	27	0.084	0.619	27	**0.097**	**0.399*****
		Means	28	0.084	0.606	28	0.097	0.380
		Residuals	34	**0.092**	**0.430*****	34	0.112	0.000
	Weighted	Configural	16	0.128	0.576	16	0.130	0.700
		Loadings	21	0.095	0.693	21	**0.136**	**0.568*****
		Intercepts	27	0.092	0.633	27	0.133	0.475
		Means	28	0.091	0.626	28	0.130	0.476
		Residuals	34	**0.095**	**0.503*****	34	0.133	0.339

Goodness of fit for the nested, latent models constructed from the BFI-S (Big Five Inventory-SOEP) and four items measuring recent positive and negative affect (Affective Well-Being) for analysis of measurement equivalence between our sample and SOEP as well as six items capturing participatory behaviour (Political Participation) for comparison between our sample and ESS. Fit measures are reported for participants recruited via X and Facebook separately.
Weighted/unweighted indicates whether our sample was weighted. Sample weights were calculated with a raking procedure based on age, gender, education, and region.
Significant differences based on CFI and RMSEA are highlighted in boldface; p-values of CFI differences.
$^{*}p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$.

Table 1 shows detailed descriptive statistics for sociodemographic variables (Age, Gender, Education, Household Income, and NUTS1 Region) as well as for those variables that were used *post hoc* to construct the samples of politically active members of the civil society. As expected, with the exception of the regional composition of respondents between the Facebook sample and both, SOEP, and ESS, all other indicators are statistically significantly different from one another (Mann-Whitney U test).

Not only were respondents from our sample statistically different from SOEP and ESS, we found statistically significant differences between recruitment platforms as well. For example, respondents recruited via Facebook had a higher middle age (53.38 ± 0.59) compared to SOEP (51.61 ± 0.24) and ESS (50.34 ± 0.74), while respondents recruited with X were significantly younger (47.07 ± 0.53).

Regarding the set of variables that were used to post-hoc construct the samples of our target population, we find a consistent pattern of sizable increases in the share of politically active members of the civil society, with significant but minor differences between recruiting platforms. On the items taken from the SOEP, 14% (±1.6) of Facebook sample and 13% (±1.33) of X sample indicated that they were politically active at least once a week, compared to 1.9% (±0.19) in the SOEP sample. The difference was less marked, but still significant for volunteering at least once a week (Facebook: 21% ±1.97; X: 20% ±1.58; SOEP: 14% ±0.47). On the items from the ESS, respondents in our non-probability samples were more likely to say they have contacted politicians (Facebook: 0.51 ± 0.02; X: 0.49 ± 0.02, ESS: 0.17 ± 0.02), worn a badge (Facebook; 0.30 ± 0.02; X: 0.30 ± 0.02; ESS: 0.06 ± 0.01), signed a petition (Facebook: 0.69 ± 0.02; X: 0.72 ± 0.02; ESS: 0.36 ± 0.02), attended a public demonstration (Facebook: 0.30 ± 0.02; X: 0.33 ± 0.02; ESS: 0.09 ± 0.01), boycotted certain products (Facebook: 0.68 ± 0.02; X: 0.71 ± 0.02; ESS: 0.37 ± 0.02), and posted on social media (Facebook: 0.88 ± 0.01; X: 0.87 ± 0.01; ESS: 0.19 ± 0.02).

Qualitative effectiveness: comparing measurement equivalence

The Political Participation measure showed equivalence on the 'means level' in the X sample (see Table 2). Specifically, the X sample showed lower variance for the *social media* item (0.029 vs. 0.166 in the SOEP) that even persisted in the weighted sample (0.033 vs. 0.166 in the SOEP). The Facebook sample yielded equivalence only on the 'loadings level', with higher intercepts for the items for *contacted politician* (0.79 vs. 0.64 in the SOEP) and *social media* (0.99 vs. 0.80 in the SOEP). The measurement equivalence of the Facebook sample further decreased when weighting was applied, yielding only equivalence on the 'configural level'.

For the measure of Affective Well-Being, the 'loadings model' fitted the best, regardless of whether individuals were recruited via Facebook or X or whether we weighted the samples or not. Notably, the intercepts for *anger* (Facebook unweighted: 3.45; Facebook weighted: 3.39; X unweighted: 3.45; X weighted: 3.43 vs. 2.56 in SOEP), *fear* (2.16; 2.15; 2.19; 2.15 vs. 1.67 in SOEP), and *sad* (2.63; 2.59; 2.53; 2.51 vs. 2.15 in SOEP) were significantly higher than for the SOEP sample across all subsamples, regardless of whether we weighted the sample. *Hope* (3.38; 3.39; 3.48; 3.51 vs. 3.81 in SOEP) showed smaller intercepts across all samples.

When it comes to the BFI-S, the X sample showed equivalence on the 'configural level' while the Facebook sample showed a marginal improvement with equivalence on the 'loadings level'. This non-invariance was presumably due to the loading of *conscientiousness*, which was substantially different between the X sample and the probability sample (0.268 vs. 0.733 in SOEP). Regarding the intercepts of the Facebook sample, we found higher scores for *openness* (5.17 vs. 4.99 in SOEP), *agreeableness* (4.90 vs. 4.78 in SOEP), and *neuroticism* (4.13 vs. 3.78 in SOEP). Weighting the samples increased measurement equivalence in both cases, yielding equivalence on the 'loadings level' for the weighted X sample and full measurement equivalence for the weighted Facebook sample. The weighted X sample showed higher intercepts for *openness* (5.19 vs. 4.99 in SOEP) and *neuroticism* (4.02 vs. 3.78); the *extraversion* intercepts were lower (4.70 vs. 4.91 in SOEP).

In summary, the best fitting models were in 2 out of 12 cases the 'configural model', in 7 cases the 'loadings model', in 2 cases the 'means model'. Only one case showed full measurement equivalence.

Discussion

We provide corroborating indications that, in the case of *hard to survey* populations, a non-probability-based sampling strategy can yield effective samples. However, success remains impossible to predict *a priori.*

Regarding the quantitative effectiveness of our non-probability sampling, compared to the benchmarks of two high-quality probability samples, SOEP, and ESS, we were able to oversample members of our target group in relative and absolute terms by ratios ranging from 2:1 to 5:1. Thus, our social media advertising campaign was highly efficient in recruiting respondents from our target population, who are otherwise *hard to sample (Tourangeau, 2014)*. By choosing this approach, we were able to increase the effective sample size tenfold, depending on the guiding research question. This leads to significant improvements in statistical power, especially in subgroup analyses.

The potential benefits derived from quantitative effectiveness are only useful insofar they withstand rigorous qualitative scrutiny, because flawed, unsound statistical inference does not improve in quality just because the sample sizes are increased. Thus, we employed a measurement equivalence analysis to assess the qualitative effectiveness of our sampling strategy.

We calculated measures of fit for three instruments: the BFI-S, the Affective Well-Being measure, and the Political Participation measure . For each instrument, we computed measures of fit using both weighted and unweighted samples, and analyzed respondents recruited via X and Facebook separately, resulting in 12 distinct fit measures. Out of these 12 fit measures, 10 achieve loadings equivalence, which would allow for meaningful comparison of regression coefficients between samples or standardized coefficients between models, thus achieving our threshold for qualitative effectiveness.

However, no measure shows consistently good results. While the BFI-S measure shows full measurement equivalence for the weighted Facebook sample, other samples merely reach the loadings or configural level. The Affective Well-Being measure is consistent across subsamples, but only achieves non-invariance at the loading level, thus not allowing meaningful inter-sample comparisons, drastically hampering researchers' analytical capabilities. The Political Participation measure from ESS does reach near full measurement equivalence for the X subsamples, but only configural or loadings level for the Facebook samples, hence showing inconsistent performance across platforms. Apparently, the instrument quality has no consistent or predictable effect on measurement equivalence across platforms. While the BFI-S and the Political Participation measure are high-quality instruments, the Affective Well-Being measure in the SOEP shows poorer item-retest correlations (0.46, .49, .51, .46), whereas the scores for the BFI-S (.64, .53, .64, .57, .62) show higher reliability (Entringer et al., 2022). Yet, the Affective Well-Being measure neither shows the best nor the poorest performance, further highlighting the unpredictability of non-probability sampling approaches, including varying performance of the same sampling strategy on different platforms.

In line with and in continuation of this finding, we did not observe that measurement equivalence was better on one platform or the other. Regarding the BFI-S measure, the Facebook samples, especially after weighting, perform better than the X sample, but it performs worse on the Political Participation measure. Likely, both samples fail full measurement equivalence due to our social media sampling strategy, as for X as well for Facebook, the non-invariance is caused by one item of the Political Participation measure, measuring whether social media was used in the past 12 months for political reasons. Although it seems obvious at first glance to explain the differences in the equivalence between Facebook and X sample with differences between the party identities, a closer examination shows counterintuitive results. While the X sample shows an underestimation of the conscientiousness item and the Facebook sample overestimates intercepts, e.g. of openness and agreeableness, expected differences in party identity are inverted.[7]

Weighting showed mixed results. While in some cases, weighting improved measurement equivalence, in other cases no improvement or even reduced measurement equivalence can be found. We find no patterns, as the performance of weighting does neither depend on the measure nor the platform.

These findings fall in line with the consensus emerging from studies such as Berrens et al. (2003), Blom et al. (2017), Brüggen et al. (2016), Chan and Ambrose (2011), Chang and Krosnick (2009), MacInnis et al. (2018), and Yeager et al. (2011). That is, generally probability samples show more accurate estimates than non-probability samples, at least when it comes to general population surveys, and poststratification procedures do not reliably reduce biases in non-probability samples and can even act in unforeseeable ways, just as in our case. Even so, these studies focused on general population surveys, while we were particularly interested in *hard to survey* populations. This is a key difference, as it may tilt the weight when researchers consider competing sampling strategies.

Researchers may follow a *Survey Quality Framework* (Biemer, 2010) that gives researchers about nine different dimensions to consider when designing their surveys. Aside from the *accuracy* of a survey, defined as the mean squared error of the survey parameters compared to the true parameter, other considerations like *relevance*, *timeliness*, or *accessibility* come into consideration. If we focus on dimensions that are potentially impacted by non-probability sampling strategies, *credibility*, *comparability*, *completeness*, and *accuracy* should be considered. While the completeness may be sufficed with quantitative efficiency, which is achievable as we have demonstrated, the comparability, credibility, and accuracy of the data are intertwined dimensions that rely on each other, as data with low accuracy will likely be considered not trustworthy, and vice versa, as will data that may be accurate but not comparable in terms of demographics or spatiality. By assessing qualitative effectiveness, researchers can address these considerations in the researcher community. This framework should also guide researchers on their decision on whether to employ non-probability sampling strategies in the first place, as it allows researchers to find the sampling strategy that, given fixed constraints such as *time*, *cost*, and *relevance*, that reduces the Total Survey Error (Groves & Lyberg, 2010).

Our analysis showed that measurement equivalence sufficient for statistical analysis required by fairly standard research aims can be achieved, but also showed that *ex ante* analyses are guesswork at best and cannot replace *post-hoc* evaluation strategies. As a deeper analysis of the non-invariance of the BFI-S measure between platforms shows, non-probability sampling is inherently unpredictable, even when researchers are aware

of biases in the samples. Therefore, researchers need to be aware of unknown biases as well as unexpected effects in their analyses, even if some biases are known.

In sum, we have shown that a non-probability sampling strategy can yield quantitatively and qualitatively effective samples, and thus can be a reasonable choice for researchers to conduct empirical studies. We operationalize qualitative effectiveness through loadings-level equivalence. While adequate for studying variable relationships, this approach cannot support point estimation or population inference. Our implementation with a *hard-to-survey* populations leads us to recommend this method exclusively for such contexts. However, *a priori* guidelines for researchers on when non-probability sampling might be fit for purpose cannot be derived from our case. Hence, we call for further investigation into the usefulness of non-probability samples for fit-for-purpose designs. Kohler et al. (2019) laid out scenarios in which researchers might turn to non-probability sampling. Our findings invite researchers in various fields to identify areas of research for which such assumptions are expected to hold, based on the current state of research, and to use the measurement equivalence test to test such assumptions.

We argue that, where no *a priori* guidelines exist, researchers need to fall back on *a posteriori* quantification to achieve these goals of *credibility*, *comparability*, *completeness,* and sufficient *accuracy* for their data. Based on what we learned, we encourage researchers who rely on non-probability sampling to design their surveys, from recruitment to item selection, in a way that allows for post hoc assessment of measurement equivalence. This includes thoughtful selection of items that are present in recognized probability samples, to allow for meaningful post hoc composition of the target population, as well as comparability with respect to sociodemographic variables and substantive item batteries or scales. Even if researchers implement new or niche items that are not part of established probability samples, assessing measurement equivalence for neighboring concepts can lend credibility to their research.

The lack of a high-quality benchmark sample may present itself as a serious limitation for this approach. In such cases, we believe the guiding principle should be transparency, as this is a viable path to ensure credibility, as laid out under the Survey Quality Framework. Hence, researchers should still aim to include tools they see fit for *post-hoc* evaluations for qualitative effectiveness of their sampling strategy.

Our own survey was designed in such a way, that *a posteriori* quantification of measurement quality compared to established probability samples was viable. By choosing SOEP, we ensured that comparisons of our target population were feasible, even if it only makes up a small proportion of the total population, as SOEP operates with an uncommonly large sample. By choosing ESS, which *inter alia* focuses on political activism, we ensured that we could quantify measurement equality for concepts related to political activism, thus informing the quality of original items not present in established probability samples. Furthermore, after *post hoc* generating the SOEP and ESS samples for comparison, we were left with 165 and 152 respondents, respectively. This highlights a potential pitfall of the measurement equivalence strategy, as it inevitably relies on the quality of the gold standard sample used for comparison. Thus, significance tests for the measurement equivalence analyses and a critical assessment of the quality of the gold standard sample are required. However, our advertising strategy only covered three polarized issues and thus did not cover the whole landscape of civic and political engagement. Depending on how representative our selection of issues for the landscape

of political activism was, this could impact our measurement equivalence analyses. Furthermore, the images used in our ad-campaign probably introduced various self-selection biases that, too, pose a risk of impacting our analysis. For example, we used six photographs associated with three political issues that constitute a substantial part of the contemporary German political landscape but do not encompass all relevant topics, thus creating potential for non-coverage. Moreover, interviews were conducted using various data collection modes. While we relied on CAWI, the ESS implemented CAPI and CAMI, and the SOEP employed CAPI, PAPI, SELF, MAIL, and CAWI. Although mode differences can substantially affect measurements, the multitude of differences between and within surveys precludes the assessment of potential implications for our survey. Nonetheless, researchers should not disregard such considerations when designing their surveys. However, the implementation of our suggested comparative method – the measurement equivalence analysis – implicitly addresses these considerations, as it precisely quantifies the degree to which the measures are comparable, notwithstanding differences in sampling strategies and data collection modes. Lastly, targeting algorithms on social media platforms remain a black box for researchers. We wanted to reach politically active members of the civil society, yet we see significant differences between the sampled respondents of Facebook and X. These algorithmic peculiarities have the potential to impede the researcher's ability to effectively sample their target population.

Since non-probability samples are not going anywhere soon, researchers should focus on how to make the most of them by finding ways and cases where they can be used, making sure that they are used in ways that can be assessed post hoc for fit for purpose. This transparency will allow the research community to assess where non-probability samples are best used and where researchers should steer clear of non-probability sampling strategies and make better use of other forms of quantitative or qualitative science.

Notes

1. At the time of our survey, the platform was still called 'Twitter'.
2. The performance metrics are part of the Online Supplementary Material.
3. That is by 'phrf' for SOEP and 'anweight' for ESS.
4. The items are coded as 'contplt', 'badge', 'sgnptit', 'pbldmn', 'bctprd', and 'pstplonl' in ESS9.
5. This stark discrepancy stems from the SOEP's structure, which comprises multiple sub-samples that receive different questionnaires with varying measures.
6. Calculations have been made in GPower 3.1.
7. Traits such as openness and agreeableness are highly correlated with green and left ideology, whereas conscientiousness and neuroticism are linked with right-wing ideology (Gerber et al., 2011, 2012). However, these patterns are not reflected in our data. Respondents recruited via Facebook identify with the party sitting on the right-wing spectrum (AfD) at a higher proportion: 32% of the Facebook sample report an AfD and 19% a Green-Party identification, a party sitting on the green-left spectrum. For respondents recruited via X, 33% report identifying with the Green-Party and only 16% do so for the AfD.

Acknowledgments

The authors would like to thank Zaza Zindel and Simon Kühne for their comments.

The code for this paper was assisted by AI. After the code was written and reviewed by the lead author, it was reviewed using Claude 3.5 Sonnet for errors and consistency. Afterwards, any changes suggested by the AI were subsequently reviewed and checked for errors by the lead author.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by a grant from the Berlin University Alliance Grand Challenge Initiative Social Cohesion Grant Number: [111_MC-SocCoh-4].

Notes on contributors

Dennis Klinke is a doctoral student at the Institute of Psychology at Humboldt-Universität zu Berlin, where he researches affective polarization. He previously worked as a research assistant at FU Berlin and was a guest researcher at the WZB Berlin Social Science Center. Klinke holds both Bachelor's and Master's degrees in Political Science from FU Berlin.

Jannes Jacobsen is Deputy Scientific Director and Head of the "Data-Methods-Monitoring" cluster at the German Centre for Integration and Migration Research (DeZIM). He studied Social Sciences and Philosophy at Leipzig University and Sociology at FU Berlin, completing his PhD at Humboldt-Universität zu Berlin. Prior to DeZIM, Jacobsen worked at the Socio-Economic Panel (SOEP) at the German Institute for Economic Research (DIW) and the Institute of Sociology at FU Berlin.

Manuel Dierse holds a Bachelor's degree in Political Science from FU Berlin. He worked as a student assistant on a Berlin University Alliance-funded project researching affective polarization at FU Berlin.

Thorsten Faas is Professor of Political Science at FU Berlin's Otto Suhr Institute and heads the Center for Political Sociology of Germany. Currently holding the Hannah Arendt Visiting Chair at University of Toronto (2025-26), his research focuses on elections, political communication, and voter behavior. After studying in Bamberg and LSE, he earned his PhD in 2008 (DVPW best dissertation prize, 2011). Previously professor at Mainz (2012-17) and Mannheim (2009-12), Faas serves on boards of the German Political Science Association and Electoral Studies Society, regularly analyzes elections in German media, and co-hosts the podcast "unter 3."

Denis Gerstorf is Professor and Chair of Developmental and Educational Psychology at Humboldt-Universität zu Berlin . A lifespan developmental scholar, he studies how everyday lives and developmental trajectories are shaped by contexts and interactions across different time scales. After studying psychology at FU Berlin (1997-2001) and completing his PhD there (2004), he held positions at Max Planck Institute, University of Virginia (postdoc), and Penn State (2007-2011). With 200+ publications and millions in research funding, Gerstorf serves as editor for Psychology and Aging and Gerontology, chairs the Berlin Aging Study-II consortium (since 2015), and holds a research fellowship at SOEP

Hannah Helal studies Sociology at FU Berlin and holds a Bachelor's degree in Social Sciences from the University of Cologne. She worked as a student assistant on a Berlin University Alliance-funded project researching affective polarization at FU Berlin, and as a student assistant at the German Centre for Integration and Migration Research (DeZIM).

Swen Hutter is Professor of Political Sociology at FU Berlin and Director of the Center for Civil Society Research at WZB. His research focuses on civil society, protest politics, party competition, and citizens' engagement. After studying in Zurich and Växjö, he earned his PhD from Munich University (2011, best dissertation prize). Author of "Protesting Culture and Economics in Western Europe" (2014) and co-editor of several Cambridge University Press volumes, his work appears in leading journals including Social Forces and European Journal of Political Research. He currently directs large-scale research projects funded by major German foundations.

David Schieferdecker is a postdoctoral research fellow at the Institute for Media and Communication Studies at FU Berlin. His research focuses on political communication processes, particularly regarding social identities and prosocial behavior. He holds an M.A. from Halle-Wittenberg (2011) and a PhD from the University of Mannheim (2018). Schieferdecker has conducted research in South Africa and Israel with DAAD support. His work appears in leading journals including Journal of Communication, Journal of Health Communication, and Journal of Clinical Medicine. He has received multiple best paper awards from the International Communication Association and funding from BMBF and Baden-Württemberg.

Hanna Schwander is Professor of Political Sociology and Social Policy at HU Berlin's Department of Social Sciences. Her research spans social policy, political inequality, party-voter alignments, and political economy of climate change. After studying in Zurich (2002-2008), she earned her PhD there (2012) with a thesis on labor market vulnerability politicization. Following positions at Bremen, EUI Florence, Oxford, Essen-Duisburg, and Hertie School, she joined HU Berlin. Her DFG/SNSF-funded work appears in Journal of Politics, Political Science Research & Methods, and Oxford University Press. She convenes the ECPR Standing Group on Political Economy and participates in RTG "Dynamics."

Christian von Scheve is Professor of Sociology at FU Berlin. His research focuses on the social and cultural constitution of emotion, particularly group-based and collective emotions in political contexts. After studying sociology in Hamburg (1994-2002, PhD 2007), he held positions at Vienna University and FU Berlin's Cluster of Excellence "Languages of Emotion." Currently executive board member of Research Centers "Affective Societies" (SFB1171) and "Intervening Arts" (SFB1512), and DIW research fellow since 2014. His DFG/BMBF-funded work earned the René König Award. He co-edits Soziale Welt and serves on editorial boards including Emotion Review.

Jule Specht is Professor of Personality Psychology at Humboldt-Universität zu Berlin. Her research focuses on political psychology, personality development, subjective well-being, and coping with life events. After studying psychology in Münster (2005-2010, PhD 2011), she held positions at Leipzig, FU Berlin (2012-2016), and Lübeck. She served as President of Die Junge Akademie (2017-2018) and received the Berliner Wissenschaftspreis (2014). Her DFG/Volkswagen Foundation-funded research complements extensive public engagement through popular science books, columns in Psychologie Heute and Tagesspiegel.

ORCID

Dennis Klinke http://orcid.org/0009-0000-7785-4993
Jannes Jacobsen http://orcid.org/0000-0003-4358-0458
Thorsten Faas http://orcid.org/0000-0003-3192-1155
Denis Gerstorf http://orcid.org/0000-0002-2133-9498
Swen Hutter http://orcid.org/0000-0002-1107-1213
David Schieferdecker http://orcid.org/0000-0003-2376-0929
Hanna Schwander http://orcid.org/0000-0002-0352-7620
Christian von Scheve http://orcid.org/0000-0003-4296-6623
Jule Specht http://orcid.org/0000-0003-2840-697X

Data availability statement

Data and code are available at https://osf.io/7rwk2/, the DOI: 10.17605/OSF.IO/7RWK2

References

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., & Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, *1*(2), 90–143. https://doi.org/10.1093/jssam/smt008

Berrens, R. P., Bohara, A. K., Jenkins-Smith, H., Silva, C., & Weimer, D. L. (2003). The advent of internet surveys for political research: A comparison of telephone and internet samples. *Political Analysis*, *11*(1), 1–22. https://doi.org/10.1093/pan/11.1.1

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, *74*(5), 817–848. https://doi.org/10.1093/poq/nfq058

Blom, A., Herzing, J., Cornesse, C., Sakshaug, J., Krieger, U., & Bossert, D. (2017). Does the recruitment of offline households increase the sample representativeness of probability-based online panels? Evidence from the German Internet Panel. *Social Science Computer Review*, *35*(4), 498–520. https://doi.org/10.1177/0894439316651584

Brick, J. M., & Tourangeau, R. (2017). Responsive survey designs for reducing nonresponse bias. *Journal of Official Statistics*, *33*(3), 735–752. https://doi.org/10.1515/jos-2017-0034

Brüggen, E., van den Brakel, J., & Krosnick, J. (2016). Establishing the accuracy of online panels for survey research. *Statistics Netherlands*. https://www.cbs.nl/-/media/_pdf/2016/15/2016-dp04-establishing-the-accuracy-of-online-panels-for-survey-research.pdf

Chan, P., & Ambrose, D. (2011). Canadian online panels: Similar or different? *Vue* , 16–20. http://www.mktginc.com/pdf/VUE%20JanFeb%202011021.pdf

Chang, L., & Krosnick, J. A. (2009). National surveys via RDD telephone interviewing versus the internet. *Public Opinion Quarterly*, *73*(4), 641–678. https://doi.org/10.1093/poq/nfp075

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., Struminskaya, B., & Wenz, A. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, *8*(1), 4–36. https://doi.org/10.1093/jssam/smz041

Curtin, R., Presser, S., & Singer, E. (2005). Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly*, *69*(1), 87–98. https://doi.org/10.1093/poq/nfi002

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, *40*(1), 55–75. https://doi.org/10.1146/annurev-soc-071913-043137

Dutwin, D., & Buskirk, T. D. (2017). Apples to oranges or Gala versus Golden Delicious? *Public Opinion Quarterly*, *81*(S1), 213–239. https://doi.org/10.1093/poq/nfw061

Einarsson, H., Sakshaug, J. W., Cernat, A., Cornesse, C., & Blom, A. G. (2022). Measurement equivalence in probability and nonprobability online panels. *International Journal of Market Research*, *64*(4), 484–505. https://doi.org/10.1177/14707853221085206

Entringer, T., Griese, F., Zimmermann, S., & Richter, D. (2022). Soep scales manual (updated for SOEP-Core v37). *Soep survey papers*, *1217*.

Gerber, A. S., Huber, G. A., Doherty, D., & Dowling, C. M. (2011). The big five personality traits in the political arena. *Annual Review of Political Science*, *14*(1), 265–287. https://doi.org/10.1146/annurev-polisci-051010-111659

Gerber, A. S., Huber, G. A., Doherty, D., & Dowling, C. M. (2012). Personality and the strength and direction of partisan identification. *Political Behavior*, *34*(4), 653–688. https://doi.org/10.1007/s11109-011-9178-5

Grande, E. (2022). Civil society, cleavage structures, and democracy in Germany. *German Politics*, 1–20. https://doi.org/10.1080/09644008.2022.2120610

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, *74*(5), 849–879. https://doi.org/10.1093/poq/nfq065

Grow, A., Perrotta, D., Del Fava, E., Cimentada, J., Rampazzo, F., Gil-Clavel, S., Zagheni, E., Flores, R. D., Ventura, I., & Weber, I. (2022). Is Facebook's advertising data accurate enough for use in social science research? Insights from a cross-national online survey. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185*(Supplement_2), S343–S363. https://doi.org/10.1111/rssa.12948

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409–426. https://doi.org/10.1007/bf02291366

Kalton, G. (2009). Methods for oversampling rare subpopulations in social surveys. *Survey Methodology*, *35*(2), 125–141.

Kalton, G. (2014). Probability sampling methods for hard-to-sample populations. In R. Tourangeau, B. Edwards, T. P. Johnson, K. M. Wolter, & N. Bates (Eds.), *Hard-to-survey populations* (1st ed. pp. 401–423). Cambridge University Press.

Kish, L. (1987). *Statistical design for research* (1st ed.). Wiley. https://doi.org/10.1002/0471725196

Kohler, U., Kreuter, F., & Stuart, E. A. (2019). Nonprobability sampling and causal analysis. *Annual Review of Statistics and Its Application*, *6*(1), 149–172. https://doi.org/10.1146/annurev-statistics-030718-104951

Kühne, S., & Zindel, Z. (2020). Using Facebook and Instagram to recruit web survey participants: A step-by-step guide and application. *Survey Methods: Insights from the Field (SMIF)*. https://surveyinsights.org/?p=13558

Lavrakas, P. J., Pennay, D., Neiger, D., & Phillips, B. (2022). Comparing probability-based surveys and nonprobability online panel surveys in Australia: A total survey error perspective. *Survey Research Methods*, 241–266. https://doi.org/10.18148/SRM/2022.V16I2.7907

MacInnis, B., Krosnick, J. A., Ho, A. S., & Cho, M.-J. (2018). The accuracy of measurements with probability and nonprobability survey samples: Replication and extension. *Public Opinion Quarterly*, *82*(4), 707–744. https://doi.org/10.1093/poq/nfy038

McClain, C., Anderson, M., & Gelles-Watnick, R. (2024). How Americans navigate politics on TikTok, X, Facebook and Instagram. https://www.pewresearch.org/internet/2024/06/12/how-americans-navigate-politics-on-tiktok-x-facebook-and-instagram/

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123. https://doi.org/10.3758/s13423-015-0947-8

Nelson, G., Buchhammer, A., Eaglestone, A., Griffiths, J., Müller, K., Sigurðsson, B. H., & Freysson, G. (2023). *Weightipy* [Graphic]. https://pypi.org/project/weightipy/

Pasek, J. (2016). When will nonprobability surveys mirror probability surveys? Considering types of inference and weighting strategies as criteria for correspondence. *International Journal of Public Opinion Research*, *28*(2), 269–291. https://doi.org/10.1093/IJPOR/EDV016

Peytchev, A., Ridenhour, J., & Krotki, K. (2010). Differences between RDD telephone and ABS mail survey design: Coverage, unit nonresponse, and measurement error. *Journal of Health Communication*, *15*(sup3), 117–134. https://doi.org/10.1080/10810730.2010.525297

Pötzschke, S., Weiß, B., Hebel, A., Piepenburg, J. G., & Popek, O. (2022). Geflüchtete aus der Ukraine – Erstedeskriptive Ergebnisse einer Onlinebefragung in Deutschland und Polen. *GESIS - Leibniz-Institut fürSozialwissenschaften*. https://doi.org/10.34879/gesisblog.2022.60

Roulin, N. (2015). Don't throw the baby out with the bathwater: Comparing data quality of crowdsourcing, online panels, and student samples. *Industrial and Organizational Psychology*, *8*(2), 190–196. https://doi.org/10.1017/iop.2015.24

Sakshaug, J., Wiśniowski, A., Ruiz, D. A. P., & Blom, A. (2019). Supplementing small probability samples with nonprobability samples: A Bayesian approach. *Journal of Official Statistics*, *35*(3), 653–681. https://doi.org/10.2478/jos-2019-0027

Sakshaug, J., Wiśniowski, A., Ruiz, D. A. P., & Blom, A. (2020). Combining Scientific and Non-scientific Surveys to Improve Estimation and Reduce Costs. In Rudas, Tamás, Péli, Gábor (Eds.), *Pathways Between Social Science and Computational Social Sciences* (pp. 71–93). Springer, Cham. https://doi.org/10.1007/978-3-030-54936-7_4

Sances, M. W. (2019). Missing the target? Using surveys to validate social media ad targeting. *Political Science Research and Methods*, *9*(1), 215–222. https://doi.org/10.1017/psrm.2018.68

Schupp, J., & Gerlitz, J.-Y. (2008). Big Five inventory-SOEP (BFI-S). *Zusammenstellung Sozialwissenschaftlicher Items Und Skalen (ZIS)*. https://doi.org/10.6102/ZIS54

Siedler, T., Schupp, J., Spiess, C. K., & Wagner, G. G. (2009). The German socio-economic panel (SOEP) as reference data set. *Schmollers Jahrbuch*, *129*(2), 367–374. https://doi.org/10.3790/schm.129.2.367

Tourangeau, R. (2014). Defining hard-to-survey populations. In R. Tourangeau, B. Edwards, T. P. Johnson, K. M. Wolter, & N. Bates (Eds.), *Hard-to-survey populations* (1st ed. pp. 3–20). Cambridge University Press.

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, *31*(3), 980–991. https://doi.org/10.1016/j.ijforecast.2014.06.001

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, *75*(4), 709–747. https://doi.org/10.1093/poq/nfr020

Zack, E. S., Kennedy, J., & Long, J. S. (2019). Can nonprobability samples be used for social science research? A cautionary tale. *Survey Research Methods*, 215–227. https://doi.org/10.18148/SRM/2019.V13I2.7262

Zindel, Z. (2023). Social media recruitment in online survey research: A systematic literature review. *Methods, Data, Analyses: A Journal for Quantitative Methods and Survey Methodology (MDA)*, *17*(2), 207–248. https://doi.org/10.12758/mda.2022.15