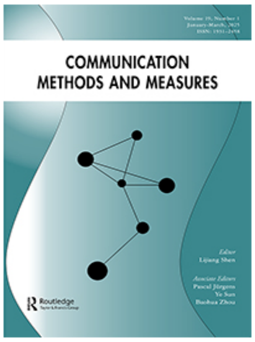

Communication Methods and Measures

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hcms20

Uncovering Digital Trace Data Biases: Tracking Undercoverage in Web Tracking Data

Oriol J. Bosch, Patrick Sturgis, Jouni Kuha & Melanie Revilla

COMMUNICATION METHODS AND MEASURES
https://doi.org/10.1080/19312458.2024.2393165

Uncovering Digital Trace Data Biases: Tracking Undercoverage in Web Tracking Data

Oriol J. Bosch[a,b,c,d], Patrick Sturgis[c], Jouni Kuha[c], and Melanie Revilla[e]

[a]Leverhulme Centre for Demographic Science, Nuffield Department of Population Health, University of Oxford, Oxford, UK; [b]Nuffield College, University of Oxford, Oxford, UK; [c]Department of Methodology, The London School of Economics and Political Science, London, UK; [d]Research and Expertise Centre for Survey Methodology, Universitat Pompeu Fabra, Barcelona, Spain; [e]Institut Barcelona Estudis Internacionals (IBEI), Barcelona, Spain

ABSTRACT
Digital trace data is an increasingly popular alternative to surveys, often considered as the gold standard. This study critically assesses the use of web tracking data to study online media exposure. Specifically, we focus on a critical error source of this type of data, tracking undercoverage: researchers' failure to capture data from all the devices and browsers that individuals utilize to go online. Using data from Spain, Portugal, and Italy, we explore undercoverage in online panels and simulate biases in online media exposure estimates. We show that undercoverage is highly prevalent when using commercial panels, with more than 70% of participants affected. Additionally, the primary determinant of undercoverage is the type and number of devices used, rather than individual's characteristics. Moreover, through a simulation study, we demonstrate that web tracking estimates are often substantially biased. Methodologically, the paper showcases how auxiliary survey data can help study web tracking errors.

Introduction

In the age of the internet, measuring how people behave and what they consume online is crucial for academic, policy, and commercial researchers. Whether they are studying the use and potential effects of fertility apps (Rampazzo et al., 2022); gender inequalities in access and participation in online platforms (Kashyap et al., 2020); or how filter bubbles shape news media exposure (Cardenal et al., 2019), high-quality measures of online behaviors are essential. Although respondent self-reports have traditionally been used as the main instrument to measure online behaviors (Gonzalez-Bailon & Xenos, 2022), there have long been good reasons to doubt their accuracy (Parry et al., 2021). These relate, most notably, to social desirability bias and recall error, as well as the cognitive burden they place on respondents, which can have negative impacts on response rates and sample composition. Consequently, researchers have long been searching for alternative ways of measuring online behaviors.

In this context, the collection of *digital trace data* has become prominent in recent years. This type of data records the interactions of users with specific digital systems (Howison et al., 2011), such as online transaction systems, telecommunication networks, websites, social media platforms, smartphone apps, sensors built on wearable devices, and digital devices (Stier et al., 2019). Given the

CONTACT Oriol J. Bosch o.bosch-jover@lse.ac.uk Leverhulme Centre for Demographic Science, Nuffield Department of Population Health, University of Oxford, 42-43 Park End Street, Oxford OX1 1JD, UK
Supplemental data for this article can be accessed online at https://doi.org/10.1080/19312458.2024.2393165

"objective" and granular nature of these digital traces, some have advocated for the possibility of using them to enhance or substitute selfreports (Revilla, 2022). Indeed, recent studies have already begun to treat digital trace data as the de facto gold standard when measuring online behaviors (Araujo et al., 2017; Scharkow, 2016), with some authors recommending substituting survey self-reports for digital traces when measuring what people do and consume online (Konitzer et al., 2021).

Nonetheless, digital trace data is itself subject to a wide range of different errors, which may introduce bias to estimates and substantive conclusions (Amaya et al., 2020; Bosch & Revilla, 2022b; Sen et al., 2021). Probably the most concerning of these errors is *tracking undercoverage*. It occurs when researchers fail to capture data from certain digital systems that individuals utilize to engage in specific online behaviors. For instance, when investigating an individual's interaction with harmful content on social media platforms, researchers must gather data from all the various social media platforms and accounts that the person employs. Failure to track data from all relevant digital systems leads to incomplete measurements. This, in turn, can introduce bias to the estimates produced with digital trace data, if the observed behaviors differ from participant's true ones. This issue arises because digital trace data is generated at the level of the digital systems individuals employ, rather than at the individual level. Therefore, to obtain individual level measurements, it is crucial to collect information from all pertinent digital systems and devices utilized by each respondent.

In this paper, we focus on understanding the impact of tracking undercoverage for one of the most widely used approaches for collecting individual-level digital trace data: web trackers. These technologies, also known as meters (Revilla et al., 2021), can be installed on participants' browsing devices with their consent. Meters enable researchers to track the traces left by participants while interacting with their devices, such as visited URLs, apps, timestamps, and sometimes even HTML content. We consider the consequences of tracking undercoverage when studying online media exposure, a behavior that is well suited to measurement with web trackers.

The remainder of this article proceeds as follows: in the next section, we review the literature on measurement of media exposure and web trackers. In the third section, we set out how tracking undercoverage can introduce bias to survey estimates. Next, in the fourth section, we describe the data and variables used, and in the fifth section, we describe our analytical approach to estimating bias due to tracking undercoverage. In the final two sections, we present the results and conclude.

Measuring online media exposure with web trackers

Survey self-reports, while widely used for studying media use, face significant and well known challenges (Chang & Krosnick, 2003; Price & Zaller, 1993; Prior, 2009). Research has shown that participants tend to overstate their media exposure, due to the complexities of the recall task. This has negative consequences on the accuracy of media exposure measures, evidenced by the low levels of agreement between self-reported and more "objective" measures of media exposure (Araujo et al., 2017; Jurgens et al., 2019; Parry et al., 2021; Prior, 2009; Scharkow, 2016).

To overcome these challenges, social scientists have made significant efforts to develop alternative approaches for measuring media exposure, either online or offline, that do not rely on participants' memory. For TV media exposure, people meters have been in use since the mid-80s, allowing to record – by a combination of passive tracking and manual imputation from users – when media is being consumed on a TV and by whom (Napoli, 2005). People meters, hence, have allowed for decades to gather semi-passive individual level data on media exposure without having to rely on traditional self-reports (Milavsky, 1992). Online media exposure is also routinely measured, both for commercial and academic purposes, through other means beyond self-reports. At the aggregated level, for example, some researchers have defended the use of third-party audience measurement to understand the online behavior of individuals (Taneja, 2016). This kind of data consists of aggregated measures at the online entities level (e.g. website), collected from a panel of individuals tracked with specific technologies (Taneja et al., 2017), kept by companies such as comScore, Nielsen or GfK. Focusing on individual-level data, while public application programming interfaces (He & Tsvetkova, 2023) and data donations (Bosch et al., 2024; Ohme et al., 2023), have been proposed as options to replace survey self-reports, the most common approach to collected individual-level data on online media exposure is through the use of web tracking technologies. For example, substantive researchers have used meters to quantify the prevalence of dubious media exposure during elections (Guess et al., 2020), the overlap in political media diets between partisans (Guess, 2021), and the extent to which online news environments are segmented by age groups (Mangold et al., 2021). Another salient area of inquiry has been the degree to which social media and other sites serve as intermediaries for online media exposure (Cardenal et al., 2019; Jurgens & Stark, 2022; Scharkow et al., 2020; Stier et al., 2021).

However, meters are not immune to errors themselves. Decades ago, researchers already showed that TV people meters faced many limits, mainly linked with the complexities of tracking multi-TV households, and the challenges of making participants fully comply with the necessary tasks (see Milavsky, 1992). Meters designed for tracking online behaviors are facing similar criticisms (Jurgens et al., 2019; Revilla et al., 2017). Specifically, for researchers to obtain complete data on what an individual does online, they must track respondent behaviors across all relevant digital systems they use to connect to the Internet (similar to multi-TV households). How this is achieved, however, can vary a lot depending on the meters available as well as the number and type of devices used by respondents. Figure 1 provides an illustration of device undercoverage. Here, the target individual browses the Internet using several browsers (sometimes more than one within a device) and connecting through different networks. Collecting data from all devices can be achieved in different ways. Participants may be asked to install tracking apps in all their devices. Alternatively, if it is not possible for one or more devices, they can ask participants to install tracking plug-ins (i.e., VPNs) in each of the browsers they use on those untracked devices. Or a proxy can be manually configured to collect data at the network level. In practice, most tracking projects require a combination of different meters and tracking approaches, as exemplified in Figure 1. This is because there is rarely a onesize-fits-all tracking technology available: PCs are better tracked with browser plug-ins, Android devices with tracking apps, and iOS devices can (generally) only be tracked with proxies. However, achieving the goal of full coverage can be very challenging (Bosch & Revilla, 2022a). For instance, some participants are not willing to install all the required technologies on all their devices, or the meters used by researchers might not be installable in some types of devices/browsers.

Research has shown that between 53% (Spain, Revilla et al., 2021) and 68% (United States, Pew Research Center, 2020) of participants in web tracking surveys do not have the meters installed in all the devices they use to go online. Hence, web tracking data likely only captures some but not all the online behaviors. Nonetheless, so far, most research comparing self-reported and metered data measures of the same concepts has treated metered data as the *de facto* gold standard, considering differences between these measures as attributable to errors in the self-reports. For example, research has shown that metered measures of Internet use and media exposure tend to be substantially lower compared to self-reports (Araujo et al., 2017; Scharkow, 2016) and correlate only modestly with digital trace data measures (Parry et al., 2021). Researchers have concluded that these differences derive from errors in the side of self-reports. For instance, Parry et al. (2021, p. 1541) concluded "self-report measures of media use may not be a valid stand-in for more objective measures." While Ernala et al., (2020, p. 10) went as far as recommending "using logging applications rather than self-reports where feasible and appropriate, treating self-reports as noisy estimates rather than precise values." This approach of considering web tracking data as the gold standard for (online) media research directly contradicts the original scholarship that proposed using TV meter data in academic research, which carefully asked scholars to understand the caveats and potential errors in these new sources of data (Milavsky, 1992; Webster, 2005).

Some research, however, has approached web-tracking data from a more critical perspective. A Pew Research Center (2020) study, for example, found that participants with uncovered devices had a higher probability of a mismatch between self-reported and metered data measures (i.e., self-reported measures being higher than metered data) than those fully covered. This could suggest that differences between measures computed with self-reports and web takers might not be due only to

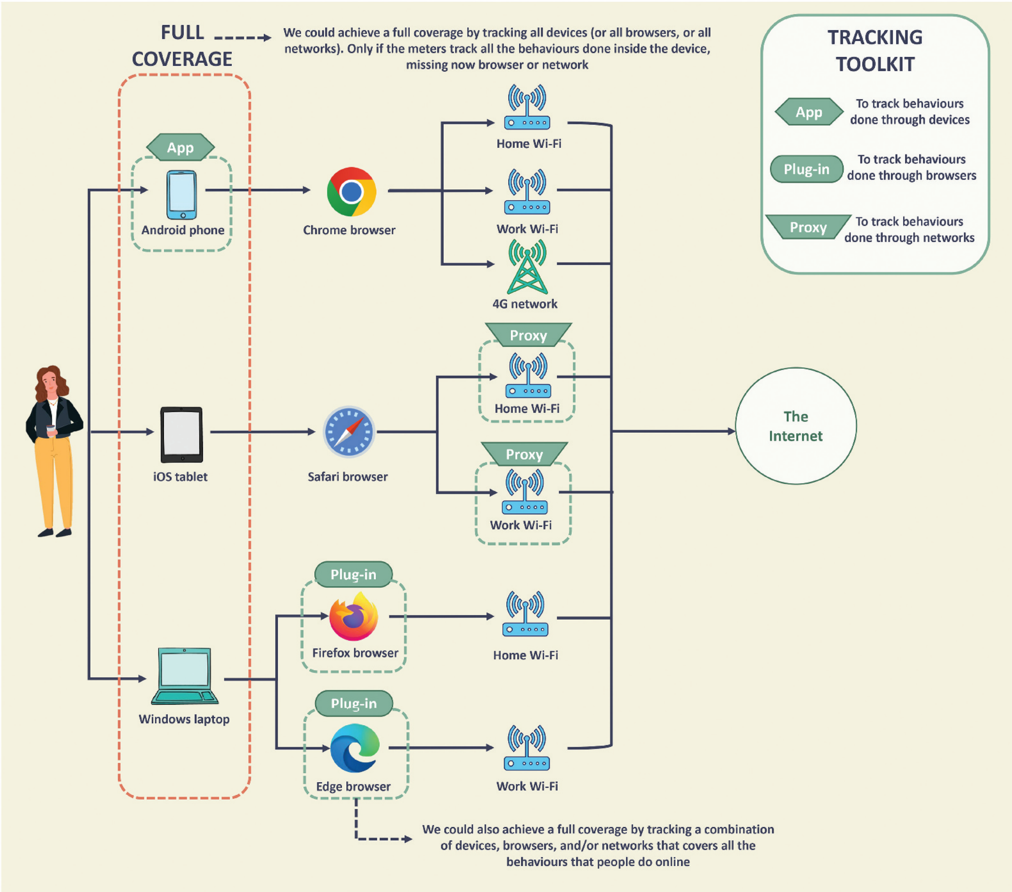

Figure 1. Diagram to illustrate the impact of device undercoverage.

self-report errors, which are to be expected and are well researched (e.g., Bartels, 1993; Chaffee & Schleuder, 1986), but also to errors on the side of web trackers. In addition, Jurgens et al. (2019) has shown that the size of self-reported errors of the time spent on the Internet, when compared with web tracking data, is different for individuals tracked only on PCs, mobile devices or in both types of devices, even when controlling for key sociodemographic variables. While this could be caused by a plethora of different issues, tracking undercoverage could be partially responsible for these results if the type of device participants are tracked on is associated with likelihood of being undercovered.

Nonetheless, much is still uncertain. Nothing is known about the characteristics of tracking undercoverage nor the mechanisms behind this phenomenon. Additionally, the size and direction of tracking undercoverage bias is still unknown. Considering this, the paper explores three research questions:

(1) What is the prevalence of tracking undercoverage? (**RQ. 1**)
(2) What characteristics are associated with tracking undercoverage? (**RQ. 2**)
(3) To what extent does device undercoverage bias estimates of media exposure? (**RQ. 3**)

Defining tracking undercoverage bias

Here we provide a formal definition of tracking undercoverage and undercoverage bias. Consider respondents $i = 1,\ldots,n$ in a sample from a target population. For these respondents, we measure variables using their web tracking data, which are denoted Y_{ijk} for individual i, variable $j = 1,\ldots,J$, and device $k = 1,\ldots,K$. We can, without loss of generality, assume that $Y_{ijk} \geq 0$, and take them to represent time/visits spent in a given period on specific internet activities on different devices. For example, the time that an individual has spent on social media, through their personal smartphone. If our interest lays at the level of the individual, hence, our variable of interest will be the sum of the time each participant has spent doing those activities, through all their devices,$Y_{ij} = \sum_{k=1}^{K} Y_{ijk}$.

Without considering other sources of error, the true value of any of these given variables will be the sum of all the behaviors done across the devices that an individual uses to go online. Said more formally, define $d_{ik} = 1$ if individual i uses device k at all, and $d_{ik} = 0$ if they do not. Similarly, let $d^{\star}{}_{ik} = 1$ if device k is recorded in web tracking data for individual i and $d^{\star}{}_{ik} = 0$ if it is not, and let $e_{ik} = 0$ if $d^{\star}{}_{ik} = d_{ik}$ and $e_{ik} = 1$ if $d_{ik}{}^{\star}\ d_{ik}$.

There is undercoverage of device k for individual i if $e_{ik} = 1$, i.e. $d_{ik} = 1$ but $d^{*}_{ik} = 0$ (we assume that the false recording case $\mathrm{d}_{ik} = 0, \mathrm{d}^{*}_{ik} = 1=$ does not occur). With this in mind, the true value of Y_{ij} can be written as $Y_{ij} = \sum_{k=1}^{K} d_{ik} Y_{ijk}$ and its measured values from web tracking data as $Y^{*}_{ij} = \sum_{k=1}^{K} d_{ik} Y_{ijk}$. Hence, undercoverage leads to measurement error in the measured values if $Y^{*}_{ij} - Y_{ij} = \sum_{k=1}^{K} e_{ik} Y_{ijk} \neq 0$. The expected size of this error depends on the probabilities of undercoverage $P(e_{ik} = 1)$, but also on the distribution of Y_{ijk} across the devices and on the correlations between e_{ik} and Y_{ijk}. Said otherwise: it is not only about missing one device, but about how much the person used that device for the activity of interest.

Variables computed with web tracking data can be used to estimate a plethora of parameters of interest across populations of interest. For example, the average daily time a person spends on the internet, the proportion of their internet time they spend on news sites, the average variance of political ideology scores of the news sites that they visit, or the correlation between the time a person spends on news sites and their self-reported voting behavior in elections. We can think about these parameters of interest as some function of (Y_i, Z_i, V_i) across the individuals in the population. Let Y_i = ($Y_{i1},\ldots,Y_{iJ}$) denote the vector of these variables for individual i. This could be, for example, the time people spend on the Internet in general, and on Instagram. A vector of variable Z_i that will be measured in the survey part of the data collection. Examples include the respondent's level of education and self-reported level of political interest. And a vector of variables V_i which characterize the internet sources and activities under consideration. These could be obtained from external sources (in which case they do not depend on respondent i) or from the survey. An example of the former is the identification of a website as a news site vs. not, and an example of the latter the respondent's self-report of which news site they have most trust in

Measurement errors introduced to Y_{ij} can lead to *undercoverage bias* in estimates of the parameters of interest that use Y^{*}_{ij} instead of the true values Y_{ij}. The size of this bias will depend not only on the magnitude of the measurement errors but also on the definition of the parameter and the joint distribution of all the variables involved in it. Here, it is worth noting that the bias does not need to be toward zero, i.e., undercoverage does not need to lead to underestimation. Downward bias is inevitable only if the parameter is a simple total or mean, such as the average time spent on the internet, since then every missed device can only reduce the estimate. For other types of parameters, however, the undercoverage bias can be in any direction. For example, the average proportion of time that individuals spend on news sites will be underestimated, overestimated, or estimated with little bias, depending on whether and how the individual uses the devices that are included in tracking tend to differ from how they use the devices that are not included.

Data and variables

The TRI-POL dataset

We use data from the first wave of the TRI-POL project (Torcal et al., 2023), the goal of which is to understand whether and how online behaviors are related to affective polarization across Southern European and Latin American countries (https://www.upf.edu/web/tri-pol).[1] TRI-POL conducted a three-wave survey between September 2021 and March 2022. Questionnaire responses were matched at the individual level with metered data. Data were collected through the Netquest opt-in metered panels (https://www.netquest.com), which consist of individuals who have meter(s) already installed in their devices and who can also be contacted to conduct surveys. We can thus link these respondents' online behavior with their questionnaire data. When the panelists join the metered panels, they must agree to install the meter on at least one device (PC, tablet, or smartphone), and they receive more incentives if they install it on more devices (up to a maximum of three). Here we use the data collected in Italy, Portugal, and Spain.

Cross quotas for age and gender, and quotas for educational level, and region were used in each country to ensure a sample matching on these variables to the general online populations. Survey questions were used to measure attitudinal and demographic variables, while metered data were used to measure variables related to the general Internet use as well as consumption of specific news media outlets, political news, and social media (see the TRI-POL data protocols in footnote 2 to check the specific URLs defined to measure these concepts). Metered data was collected for the 15 days prior to and following participants completing the questionnaire. The meter logged each URL accessed by the panelists, along with timestamps indicating the initial visit to the URL, and the duration in seconds during which the URL remained the active content within the browser, or in the case of mobile devices, on the smartphone screen. Netquest's technology only captures active behaviors, filtering out passive transmissions of data that could add noise to the data.[2] It is important to note that a URL or app was classified as 'active' when it was the foremost content displayed in the browser or on the device's screen. This definition excludes any other URLs or apps that might have been open in separate tabs or screens, as they were not considered active during this time frame. The duration of active engagement was computed as the elapsed time between the moment the URL or app first gained 'active' status within the browser or device and the point at which a different URL or app took over as the active content in the browser or device. A visit was defined as any opened URL/app lasting one second or more, and operationalization that has been shown to produce slightly more reliable media exposure measures than other stricter thresholds (Bosch, 2023). Participants were tracked on iOS and Android mobile devices, and Windows and MAC computers, using the tracking solutions provided by Wakoopa (https://www.wakoopa.com/). Windows and MAC devices were tracked with desktop apps and/or web browser plug-ins, Android devices through apps and iOS devices through manually configured proxies. Hence, the type of technologies used to track each panelist depend on the devices they use to go online. More information about the collectible data and the characteristics of each of the tracking technologies used can be found in Torcal et al. (2023).

Challenges were faced when filling some of the specific cross-quotas with participants from the metered panel. Hence, in some cases panelists were invited without a meter installed to fill some of the quotas. Thus, in total, for the first wave, 3,548 respondents completed the survey, but only 2,653 had the meter installed in at least one mobile (smartphone or tablet) or PC device: 993 in Spain, 818 in Portugal and 842 in Italy. For the analyses, we combine the samples from the three countries. No significant differences are observed between the full sample and the subsample of tracked participants,

[1] More information about the data collection strategy of both survey and digital trace data can be found in the TRI-POL data protocols: https://osf.io/3t7jz/

[2] For iOS devices, which are tracked with manually configured proxies, these proxies generate raw data that must be processed to identify which part of the tracked traffic was done passively by the device or actively by the participant. Netquest does this using their own proprietary algorithms before sending the data to their clients. This process is still susceptible to false positives or negatives, but the extent of these is unknown.

across a selection of demographic, political and technological variables (see Supplementary Material 1, i.e., SOM 1). Nonetheless, to correct for any unobserved difference in the sample characteristics between the full sample of respondents, and those with the meter installed, estimates from the subsample of metered participants were weighted with inverse probability weights, computed using the random forest relative frequency method (Buskirk et al., 2015).

Identifying undercoverage and its characteristics

We first had to identify when a participant was undercovered, and through which devices. To do so, two pieces of information were needed: which sources were tracked, and which sources panelists used to go online. The first piece of information was obtained using paradata about the technology with which participants were being tracked, the type of device, the operating system, whether it was a tablet or smartphone and, for plug-ins, the browser in which they were installed.[3] The second piece of information was obtained by asking participants questions about which devices and browsers they used to access the internet during the 15 days before the start of the survey. SOM 2 shows the wordings of these questions (English translations), as well as the paradata available. We were able to identify any mismatches between the self-reported devices used to go online, and the paradata of those that we tracked. With this, we were able to identify how many of devices were not covered, and which type of devices they were (e.g., Windows or MAC).

In terms of browser undercoverage, we did not have information about the number of browsers that participants used within each of their devices, but only the types of browsers (e.g., Chrome or Firefox) used on all their Windows PC, MAC, and Android devices (see SOM 2 to see the exact question used). Hence, for undercovered panelists, it would not be possible to discern whether a browser is not covered because the panelist did not install it in a tracked device, or because it was installed on an untracked device. Conversely, for fully covered panelists, we know that all their devices have a tracking technology installed; if a browser is not tracked, it is because the technology installed in the device is not tracking that browser. Hence, for fully covered panelists we were able to identify the types of browsers that they were not covered on, in general and within each type of device (e.g., Windows PC). No information was computed for undercovered panelists.

It should be noted, nonetheless, that this identification strategy is affected by errors, given that the self-reported measures of the number and types of devices and browsers used by participants can be affected by measurement errors. In SOM 4 we present an exhaustive assessment of whether measurement errors should be expected, and to what extent. Based on SOM 4, limited measurement errors are expected.

Predictors of tracking undercoverage

To assess which individual characteristics are associated with undercoverage, we conducted three logistic regressions predicting whether a participant was not fully tracked in terms of device (1= at least one device not tracked, 0 = fully covered). In the first model, we focus on the profile of the people more likely of being undercovered. Hence, as independent variables, we introduced sociodemographic information about participants' sex (male = 0, female = 1), education (0 = not completed high education, i.e., post-secondary education such as university or superior technical training, 1 = completed high education), age group (18–24 = 1, 25–34 = 2, 35–44 = 4, 45–54 = 5, +55 = 6), and country of residence. Also included were political variables measuring participants' political ideology (0 = left, 10 = right) and political interest (1 = not at all, 2 = a little, 3 = a fair amount, 4 = a lot), as well as a measure of panel loyalty (the number of years a participant had been part of the Netquest panels) a self-reported measure of participants' Internet use (as hours spent on the Internet on a typical day). The second model introduced variables relating to the devices that people use. Therefore, as

[3]No information about the networks in which proxies were configured was available.

predictors, we introduced variables for the self-reported number of Windows PCs, MACs, Android, and iOS devices that participants used to go online during the 15 days prior to the survey. We also included dummy variables for whether the participants used both PCs and mobile devices to go online, or only mobile or PCs. In Model 3, we combined all predictors.

Simulated estimates of undercoverage bias

The bias introduced by device undercoverage cannot be directly quantified, since data that would be collected from the uncovered devices is, by definition, not observed. Hence, to assess the likely extent of this bias, we developed a simulation approach to examine the magnitude of bias introduced by device undercoverage.

As Figure 2 exemplifies, the simulation worked in the following way. We selected the subsample of 688 participants identified as fully covered in terms of device (as explained in section 4.2). In the simulation, this represents the ground truth of a sample in a study where every respondent is completely covered. Hence, we assume that, for this subsample, we can observe participants' complete online behavior in terms of devices, allowing us to get estimates unaffected by device undercoverage. After selecting this subsample, for some participants, we removed information from some of their devices, to simulate a situation where there is undercoverage. We specifically defined 13 undercoverage scenarios, varying the proportion of participants affected by undercoverage (from 25% to

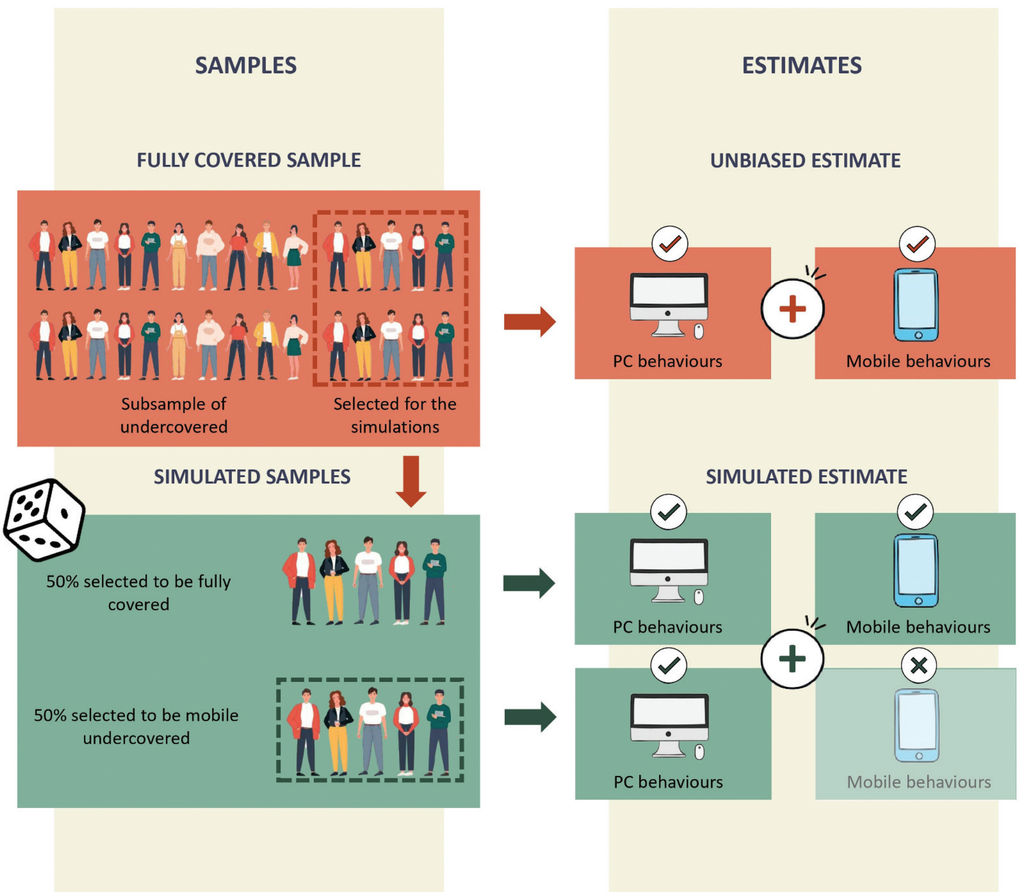

Figure 2. Diagram exemplifying the simulation approach.

75%), and the type of undercoverage (PC, mobile, or PC and mobile). Estimates of the statistics of interest from these reduced samples represent estimates that we would obtain from samples that were affected by undercoverage. For example, how would the original sample estimate change if 50% of the sample had all their mobile data deleted?

Instead of doing this only once, we created 1,000 random samples for each scenario, with participants randomly being selected as undercovered or not every time a sample was created. This allowed to get a distribution of possible biases, rather than just having a single bias estimate. We did this for a plethora of statistics of interest, both univariate and multivariate. For example, the average time people in the sample spent on the Internet. After averaging the results across samples, we compared them to the original estimate, to get an estimate of the bias introduced by the different types of undercoverage simulated.

Simulation specifications

We defined 13 simulation scenarios, each with a different probability for participants of having none of their PCs or mobile devices covered. Ideally, simulations would have been designed to explore undercoverage at the device level. However, web tracking data at the level of individual devices were not available, only aggregated at the level of PC and mobile. The probability of being selected to be undercovered was independent for different participants. Table 1 presents the details for each scenario. Scenario 13 represents the exact undercoverage observed in our sample: 33% of participants had none of their mobile devices covered, and 33% had none of their PCs covered. We expect this prevalence to represent the typical undercoverage in the Netquest panel, at the time we collected the data, which can be understood as the most realistic scenario out of all tested.

For scenarios simulating both PC and mobile undercoverage happening at the same time, two caveats must be considered. First, the previously defined independence was relaxed to constrain the simulations to not allow both PC and mobile undercoverage for the same participant. Additionally, scenarios were not fully crossed (e.g., there is no scenario with 75% of mobile undercoverage and 50% of PC undercoverage). Both constraints were introduced to avoid having scenarios with participants having no data for both their PCs and mobile devices, which would have meant they were not tracked in any device. This is an unrealistic scenario in web tracking studies, especially those using opt-in metered panels. For the sake of clarity and simplicity, we excluded them. Hence, for simulations dealing with both PC and mobile undercoverage, the selection process was twofold: in the first step, all participants had a random chance to be selected to have their PCs untracked. In the second step, a subsample of those fully tracked on PC were randomly selected to have their mobile devices untracked. Furthermore, and for the same reason, fully covered participants using only PCs or mobile

Table 1. Scenarios for the simulations.

Scenario	*P*(PC undercoverage)	*P*(Mobile undercoverage)
1	.25	.0
2	.50	.0
3	.75	.0
4	.0	.25
5	.0	.50
6	.0	.75
7	.25	.25
8	.25	.50
9	.25	.75
10	.50	.25
11	.50	.50
12	.75	.25
*13**	.33	.33

*Scenario 13 represents the actual undercoverage in the sample.

devices were kept in the sample, but were not affected by undercoverage, given that this would have resulted in them having no device covered whatsoever.

For each scenario, we created 1,000 random allocations. In each of them, devices were set to be uncovered with the probabilities shown in Table 1. Estimates of several parameters of interest, as described below, were calculated for each such dataset.[4] Monte Carlo standard errors were computed using the R "rsimsum" package (Gasparini, 2018; White, 2010).

Simulations were conducted for a range univariate and multivariate statistics selected based on the conceptualization of undercoverage bias presented in section 3. First, we selected a variety of univariate statistics to compute, covering simple counts, ratios and binary indicators. Specifically, the univariate statistics were the following:

- **Average time spent on the Internet**. This measure captures the time spent by each participant on any URL and online app (Araujo et al., 2017).
- **Average time spent on Social Network Sites (SNSs)**. This corresponds to the time spent on any URL or app identified as being an SNS (i.e., Facebook, Instagram, Snapchat, TikTok, Twitter, WhatsApp, Messenger, YouTube) (Scharkow et al., 2020).
- **Proportion of non-users of online news**. We replicate Reiss (2022) approach to measure news avoiders. Non-users were defined as those who during the period of 1 month never visited an URL or app defined as "news" (Palmer et al., 2020). In our case, we only consider the consumption of written "news."
- **Total number of media consumed**. We replicated Padr´o-Solanet and Balcells (2022) and computed a statistic showing the total number of media consumed by participants (during the first wave), to measure the variety of their media diet.
- **Proportion of Internet time spent on SNSs**. We computed a measure of importance of SNSs consumption over all the time participants spend on the Internet: ((Time on SNSs)/(Time on the Internet) × 100).
- **Proportion of Internet time spent on news media outlets**. We computed a measure of importance of news media consumption over all the time participants spend on the Internet: ((Time on news media outlets)/(Time on the Internet) × 100).

For multivariate statistics, we focused on a range of statistics, both bivariate and multivariate. Our expectation was that biases will be more substantial for stronger associations; hence, we computed associations expected to yield small (e.g., Time spent SNSs ˜ Trust in SNSs) to high (e.g., Age ˜ Instagram use) coefficients. Specifically, we computed the following five statistics:

- **Correlation between the average time spent on SNSs and trust in SNSs**. We computed the Spearman[5] correlation between trust in SNSs (0 to 10 scale, 0 being "I don't trust it at all," and 10 "Completely trust"), and the average time spent on SNSs during the month of tracking.
- **Association between the trust in news and news avoidance**. We ran a logit regression with news avoidance as the dependent variable, trust in news as the main independent variable (0 to 10 scale, 0 being "I don't trust it at all," and 10 "Completely trust"), and several common control variables (age, gender, higher education, left-right self-placement, and country).
- **Association between the total number of media consumed and ideological extremism**. Similarly as Padro-Solanet and Balcells (2022), we ran an OLS regression with a measure of ideological extremism as the dependent variable, the total number of media as the main

[4]Weights applied to compute these estimates. The weights bring closer the already very close sample of fully covered participants and the full sample, on a selection of sociodemographic, political, and technological variables.

[5]Given the zero-inflated nature of the web tracking measures used, we use Spearman instead of Pearson. Spearman correlation is non-parametric and does not rely on the distributional assumptions of the data that Pearson does.

independent variable, and several common control variables (age, gender, higher education, political interest, left-right self-placement, and country).

- **Correlation between age and Instagram use**. We computed the Spearman correlation between the age of the participant (continuous), and the average time spent on Instagram during the month of tracking.
- **Correlation between the average time spend on SNSs and on news sites**. We computed the Spearman correlation between the average time spent on SNSs during the month of tracking, and the average time spent on news media sites.

Results

The prevalence and characteristics of tracking undercoverage

Table 2 presents the proportion of participants that had all their devices covered, and the median number of untracked devices for those who were not fully tracked.

Results are additionally presented for the number and types of devices used. Only 26% had all reported devices tracked, which is very similar to the 28% of fully tracked participants that the Pew Research Center (2020) found in a probability-based sample in the United States. Of those who were not fully tracked, the median number of untracked devices was 2.

Table 2 also shows that the prevalence of device undercoverage differs depending on participant's number and types of devices. The more reported devices, the higher the proportion of participants not fully tracked and the median number of untracked devices. Specifically, while 34% of individuals using two devices were fully tracked, this number drops to 1% and 0% for those using four and five or more devices, respectively. There are also clear differences between those who use only PCs and mobile devices, and those who use both, with the latter group yielding a three-times lower proportion of fully covered participants.

These estimates also show notable differences in the prevalence of undercoverage depending on the types of devices that participants use. For PCs, while 49% of those using a Windows PC had all their

Table 2. Proportion of participants fully covered, and median number of untracked devices, in general and per specific sub-groups of participants.

	n	% fully covered	Median number of untracked devices
General coverage			
All participants	2653	26	2
Participants who reported using ...			
1 device	207	100	0
2 devices	1103	34	1
3 devices	611	13	2
4 devices	305	1	3
+5 devices	416	0	6
Only PC	66	77	1
Only mobile	264	66	1
Both PC and mobile	2312	20	2
Device specific coverage†			
Participants who reported using ...			
PC			
Any type	2379	47	1
Windows	2305	49	1
MAC	302	27	1
Mobile			
Any type	2577	41	1
Android	2340	52	1
iOS	782	10	1

† Values for the device specific undercoverage are computed over all members of those subsamples. Therefore, the 49% fully covered of "windows" means that, out of all windows users, 49% have all their self-reported windows devices covered.

Windows PCs tracked, only 27% of MAC users had all their devices tracked. Similarly for mobile, while 52% of Android users had all their devices tracked, only 10% of iOS users had all their devices tracked. Therefore, tracking undercoverage is more prevalent for Apple devices, especially iPads and iPhones. Indeed, 85% of participants with at least one iOS mobile device had none of them tracked. This means that, for 29% of the participants, almost everything that they did through their iOS devices was missed. This is likely to be because Netquest, the panel provider, tracks these devices using proxies, which require participants to manually configure them according to a complex and burdensome process.

Even if all devices are covered, we might still miss people's behaviors if we do not track the browsers that they use within those devices. Table 3 presents the proportion of participants that, having all their devices tracked, also have full browser coverage. Results are additionally presented for different subgroups, depending on the types of devices and browsers that they use to go online.

Only 38% of those who were fully covered in terms of device, also had all browsers tracked. This shows that even if we track people on all the devices that they use to go online, there is still a high rate of undercoverage. Focusing on the different subgroups by device used, we see that browser undercoverage is mainly a phenomenon affecting PCs. Of those with all Android and/or iOS devices covered, respectively 91% and 100% of them had all browsers within those devices covered. Conversely, these numbers go down to 51% and 48% for fully covered Windows and MAC users. This is to be expected; while mobile tracking technologies tend to track all browsers used within a device, most technologies used to track PCs have to be installed as plug-ins in each browser that participants use. Additionally, it is more common to have multiple browsers on a PC than a mobile device. This might result from participants not installing the plug-ins in all the browsers they use to go online. Another potential hypothesis might be that participants satisfice by installing trackers on browsers they do not use, to get the incentives without having their behaviors tracked.

Furthermore, Table 3 shows that the prevalence of browser undercoverage is highly dependent on the browsers that people use. While we observe very high coverage rates for Chrome (97%), all other browsers are substantially lower (0% to 26%). In particular, for Internet Explorer or Safari users, none of those participants had all those specific browsers covered. This points to technological limitations on the side of the panel provider: if most panelists are tracked with web browser plug-ins, and these are

Table 3. Browser coverage conditional on device full coverage.

	n	% fully covered browser
Device specific browser coverage		
Fully covered in terms of . . . *		
All devices	688	38
Windows PC	1109	51
MAC	83	48
Android	1224	91
iOS	79	100
Browser specific coverage		
Fully covered using . . . †		
Internet Explorer	130	0
Chrome	401	97
Firefox	124	26
Safari	7	0
Other	132	15

*Values for the device specific browser undercoverage are computed over those participants that are fully covered in terms of the specific devices listed. Therefore, the 91% fully covered of "Android" means that, out of all those participants that have all their Android devices covered, 91% have all their browsers covered. †Values for the specific types of browsers are computed over all those fully covered participants that self-reported using the listed browsers. Therefore, the 97% fully covered of "Chrome" means that, out of all fully covered (in terms of device) participants using Chrome, 97% have all their self-reported Chrome covered.

Table 4. Characteristics associated with tracking undercoverage.

	Model 1		Model 2		Model 3	
	Coefficient	SE	Coefficient	SE	Coefficient	SE
Type of user						
Mobile only			.64	.17	0.62	0.18
PC only			.17***	.07	.19***	.09
Self-reported number of . . .						
Windows PC			7.07***	1.45	7.37***	1.56
MAC			10.13***	2.92	11.93***	3.69
Android			4.72***	.63	4.31***	.60
iOS			6.88***	1.03	7.13***	1.16
Hours of internet consumption	1.02	.02			0.97	.02
Years in panel	.96**	.01			.93***	.00
Ideology	1.01	.02			1.02	.02
Political interest	1.06	.05			1.01	.06
Age						
25-34	.95	.19			1.08	.25
35-44	.85	.16			0.95	.22
45-54	.81	.15			0.87	.20
55+	.81	.15			1.04	.24
Female	.86	.07			1.25*	.13
Tertiary education	1.32**	.12			0.98	.11
Country						
Portugal	1.27*	.15			1.12	.16
Italy	1.33**	.14			1.26	.16
Constant	2.99**	.76	.05***	.01	.07***	
AIC	3307.1		2542.3		2263.7	
n	2374		2641		2366	

Coefficients reported in odds ratios. *p < .05, **p < .01, ***p <.001.

only available for Chrome and Firefox, almost all behaviors done through other types of browsers will be missed.

Table 4 presents the results of three logistic regression predicting whether an individual was not fully tracked in terms of device (1 = at least one device not tracked, 0 = fully covered). Results show that the more longstanding the panelists and people who completed tertiary education have a lower probability of being undercovered, Model 1 also shows significant differences in terms of the country of residence, with people residing in Italy and Portugal being more likely to not be fully covered than those from Spain. In Model 2, we observe that for each additional device, the odds of being undercovered increase by a factor of 4.7 (Android), 6.9 (iOS), 7.1 (Windows PC), and 10.1 (MAC). The odds of a participant who uses only PCs being undercovered are 83% lower than for those using both PCs and mobile devices.

Model 3 shows that education and country of residence of individuals are no longer significant (and the odds ratios are smaller), suggesting that the main driver behind those effects is that the number and types of devices used vary across educational levels and country groups. Additionally, we find that women are slightly more likely to be undercovered than men. All in all, these models seem to suggest that, although there might be some slight differences in the demographics of those being fully tacked and those not, the biggest driver behind someone being undercovered is the type and number of devices that they use to go online.

The bias introduced by tracking undercoverage

Table 5 presents the results of the simulations for estimating biases due to device undercoverage. The top row (in bold) shows the “true value” which is taken as the estimate from the subset of fully covered participants, and the following rows show the estimates under the different undercoverage scenarios.

We can see that tracking undercoverage results in biases for most of the univariate statistics considered. In many instances, these biases are of a substantial magnitude, both in absolute and

Table 5. Average estimates of the univariate statistics under different scenarios of simulated undercoverage.

Undercoverage scenarios		Univariate estimates					
P(PC undercoverage)	*P*(Mobile undercoverage)	Time Internet	Time SNSs	% News avoiders	Num. media exposed	% Internet time spent on SNS	% Internet time spent on news
.00	**.00**	**221′**	**64′**	**17%**	**4.3**	**29%**	**1.6%**
(Full coverage)		(*.18*)	(*.08*)	(*.01*)	(*.29*)	(*1.22*)	(*.23*)
.25	.00	206	61	22	3.4	30	1.4
		(*.06*)	(*.02*)	(*.02*)	(*.01*)	(*.01*)	(*.00*)
.50	.00	191	58	27	3.0	31	1.3
		(*.06*)	(*.02*)	(*.03*)	(*.01*)	(*.01*)	(*.00*)
.75	.00	176	55	32	2.7	32	1.1
		(*.06*)	(*.02*)	(*.02*)	(*.01*)	(*.01*)	(*.00*)
.00	.25	188	53	23	3.5	25	1.8
		(*.13*)	(*.05*)	(*.03*)	(*.01*)	(*.01*)	(*.00*)
.00	.50	153	42	28	3.2	23	2.0
		(*.16*)	(*.06*)	(*.03*)	(*.01*)	(*.01*)	(*.00*)
.00	.75	119	31	33	2.8	20	2.2
		(*.13*)	(*.05*)	(*.03*)	(*.01*)	(*.01*)	(*.00*)
.25	.25	172	50	27	2.4	27	1.6
		(*.12*)	(*.04*)	(*.04*)	(*.01*)	(*.02*)	(*.00*)
.25	.50	138	39	33	.9	24	1.9
		(*.13*)	(*.05*)	(*.04*)	(*.01*)	(*.02*)	(*.00*)
.25	.75	104	28	38	.2	21	2.1
		(*.13*)	(*.05*)	(*.04*)	(*.01*)	(*.02*)	(*.00*)
.50	.25	157	47	32	1.5	28	1.5
		(*.13*)	(*.05*)	(*.04*)	(*.01*)	(*.02*)	(*.00*)
.50	.50	123	36	37	1.0	25	1.7
		(*.12*)	(*.04*)	(*.04*)	(*.00*)	(*.02*)	(*.00*)
.75	.25	142	44	37	.5	29	1.3
		(*.16*)	(*.05*)	(*.04*)	(*.01*)	(*.02*)	(*.00*)
.33*	.33*	157	45	31	2.0	26	1.7
(Sample undercoverage)		(*.14*)	(*.02*)	(*.04*)	(*.01*)	(*.02*)	(*.00*)

Averages computed over the 1,000 simulated scenarios, with the exception of the full coverage scenario, which represent the sample estimate. Empirical Monte Carlo standard errors in brackets.

relative terms (see Figure 3). In absolute terms, the direction of the effects is as expected: while device undercoverage reduces the average time participants spend on the Internet (by between 15 and 117 minutes) and SNSs (by 3–36 minutes), it increases the estimated proportion identified as news avoiders (by 5% to 21% points). Furthermore, undercoverage reduces the average number of media outlets exposed to (by 0.9 to 4 fewer media), which implies that people do not have consistent media diets across devices.

The effect of tracking undercoverage is less pronounced for the percentage of time spent on SNSs and on news. In these cases, the bias would be introduced if tracking undercoverage has a different effect on the numerator (time on SNSs/news) than the denominator (time on the Internet). What we observe is that, while undercovered estimates deviate from the full coverage, the deviations are smaller and more varied. For instance, for percentage of time spent on SNSs, PC undercoverage alone inflates the estimates by around 1% to 3% points, while mobile undercoverage reduces it by 4% to 9% points.

If we focus on the relative size of these biases (see Figure 3), undercoverage is responsible for a relative overestimation of 29% to 123% of the participants identified as news avoiders, and an underestimation of 21% to 93% of the number of media that people consume on average. Figure 2 also shows that undercoverage leads to an underestimation of 5% to 53% of the average time spent on the Internet and SNSs. Although the relative bias for the percentage of time spent on SNSs and on news is smaller than for the other statistics, in some scenarios these are still substantial, with the estimated percentage of Internet time spent on SNSs underestimated by up to 31%, while the percentage of Internet time spent on news can be both under- and overestimated by up to 31%.

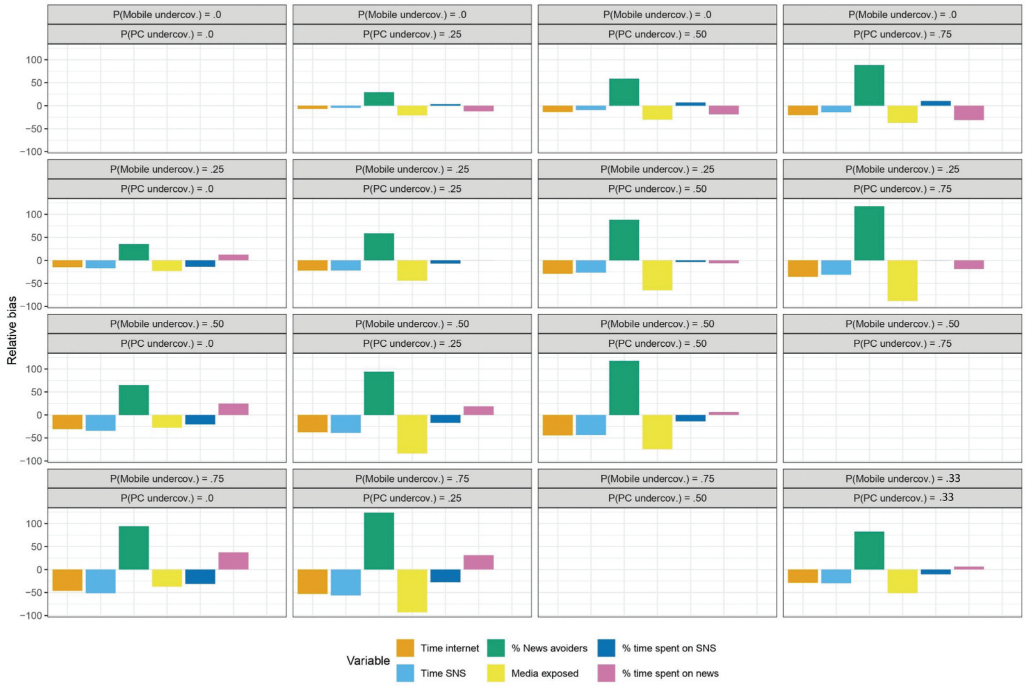

Figure 3. Estimates of relative bias for the univariate estimates. Blank panels represent unrealistic scenarios that we have not simulated. The last panel represents the actual undercoverage in the TRI-POL dataset.

Additionally, as we would expect, there is an association between the extent of undercoverage and the size of the bias. In the scenarios with only PC or mobile undercoverage, the scenarios with 50% and the 75% undercoverage reveal biases two and three times the size of the scenario with 25% of the sample, respectively. In general, mobile undercoverage introduces more bias than PC undercoverage. This can be observed both in the scenarios with mobile and PC undercoverage alone, and when combined.

Table 5 also shows the estimated bias that undercoverage would introduce at the observed undercoverage level of TRI-POL dataset (around 33% for both PCs and mobile devices), which is expected to be very similar to what other studies have experienced. This shows that the observed time spent on the Internet and on SNSs is underestimated by 29% and 30%, or between 64' and 19' lower than what we would observe without undercoverage, respectively (221' vs 157'/64' vs 45'). In addition, undercoverage is estimated to lead the observed proportion of news avoiders to be overestimated by 82%, or 14% points higher than what it would be without undercoverage (17% vs 31%). The number of media outlets that we observe people being exposed to be less than half of what we would observe with full coverage (4.3 vs 2.1). On the other hand, at this level of undercoverage both the estimated percentage of time spent on SNSs and on news would be very similar to full coverage (29% vs 26%/1.6% vs 1.7%).

Table 6 presents the simulation estimates of bias for the set of multivariate statistics, again the "true scores" are in bold in the first row of the table. This shows that, overall, the biases of tracking undercoverage are considerably smaller than for the univariate quantities but that there is wide variability between them. For three of the statistics the biases are close to zero. Two of them are highly affected by tracking undercoverage: the correlations between time spent on the Internet and age and time spent on SNSs and news media outlets.

Focusing on the association between the time spent on Instagram and age, device undercoverage reduces the estimated correlation by 24% (−.41 vs −.29), a difference with potentially important substantive implications. For the correlation between the time spent on SNSs and on news media

Table 6. Average estimates of the multivariate statistics under different scenarios of simulated undercoverage.

Undercoverage scenarios		Multivariate estimates				
P(PC undercoverage)	*P*(Mobile undercoverage)	Time SNSs ~ trust SNSs	News avoidance ~ trust news	Polarization ~ N° media consumed	Time Instagram ~ Age	Time SNSs ~ Time News
.00	**.00**	**.03**	**.89**	**−.01**	**−.41**	**.16**
(Full coverage)		(*.02*)	(*.02*)	(*.02*)	(*.02*)	(*.02*)
.25	.00	.03	.92	−.01	−.41	.19
		(*.00*)	(*.00*)	(*.00*)	(*.00*)	(*.00*)
.50	.00	.03	.93	−.01	−.40	.22
		(*.00*)	(*.00*)	(*.00*)	(*.00*)	(*.00*)
.75	.00	.02	.95	−.01	−.39	.24
		(*.00*)	(*.00*)	(*.00*)	(*.00*)	(*.00*)
.00	.25	.02	.90	−.01	−.32	.27
		(*.00*)	(*.00*)	(*.00*)	(*.00*)	(*.00*)
.00	.50	.00	.91	−.01	−.24	.35
		(*.00*)	(*.00*)	(*.00*)	(*.00*)	(*.00*)
.00	.75	−.02	.91	−.01	−.17	.40
		(*.00*)	(*.00*)	(*.00*)	(*.00*)	(*.00*)
.25	.25	.01	.92	−.01	−.31	.28
		(*.00*)	(*.00*)	(*.00*)	(*.00*)	(*.00*)
.25	.50	−.01	.92	.02	−.23	.33
		(*.00*)	(*.00*)	(*.00*)	(*.00*)	(*.00*)
.25	.75	−.03	.92	.03	−.17	.35
		(*.00*)	(*.00*)	(*.00*)	(*.00*)	(*.00*)
.50	.25	.01	.93	.00	−.31	.27
		(*.00*)	(*.00*)	(*.00*)	(*.00*)	(*.00*)
.50	.50	−.01	.93	.00	−.23	.29
		(*.00*)	(*.00*)	(*.00*)	(*.00*)	(*.00*)
.75	.25	.00	.94	.01	−.30	.26
		(*.00*)	(*.00*)	(*.00*)	(*.00*)	(*.00*)
.33	.33	.01	.93	−.01	−.29	.29
(Sample undercoverage)		(*.00*)	(*.00*)	(*.00*)	(*.00*)	(*.00*)

Averages computed over the 1,000 simulated scenarios, with the exception of the full coverage scenario, which represent the sample estimate. Empirical Monte Carlo standard errors in brackets. "Time SNSs ~ trust SNSs," "Time Instagram ~ Age" and "Time SNSs ~ Time News" expressed as correlation coefficient, "News avoidance ~ trust news" as odds ratios, and "Polarization ~ N° media consumed" as non-standardized regression coefficients.

outlets, we observe the opposite effect: device undercoverage now inflates the correlation by between 3% and 24%; the estimated correlation increases from .16 to .29. This might be due to the difference in the strength of the associations in the full coverage scenario: while the first three statistics show small and weak associations under full coverage, the associations showed for the other two statistics are substantially higher.

Discussion

Overall, we found that tracking undercoverage is highly prevalent in a commercial panel in the three studied countries: 74% of participants had at least one of the devices they used to go online not tracked. These results are in line with the 68% of device undercoverage found by the Pew Research Center (2020) in a US probability-panel. Additionally, of those fully tracked, 62% had at least one web browser uncovered, showing that undercoverage is multi-layered: even when we track all participant's devices, we can still fail to observe online behaviors within those devices. For most participants in this study, we could not track at least some of what they did online during the period of observation.

Besides describing the prevalence of tracking undercoverage, our results also identify the main issues faced when trying to track online behaviors, and potential approaches to reduce them. Our data shows that, at least in the context of our panel company and period studied, there are special difficulties when tracking devices and browsers other than Windows, Android, Chrome and Firefox. Indeed, between 90% of those participants that reported browsing online with iPhones and/or iPads had at least one of those not tracked (a large majority had all of them untracked), and none of the self-reported Internet Explorer and Safari browsers were tracked. Considering that our tracking approach is based on current standard research practice (provided by Wakoopa), these results highlight the technological limitations that the field still faces when tracking anything apart from Android and Windows devices, and mainstream browsers. Our analysis also shows that the main determinant of tracking undercoverage is the number, type, and combination of devices that participants use, pointing to the importance of creating tailored recruitment approaches based on participants' self-reported information on what devices and browsers they use to go online. This has several practical implications. First, if the behaviors that people do online vary across devices and browsers, web tracking data will systematically miss some behaviors more than others. Hence, the size of the biases will differ across statistics of interest. Second, if the probability of using these devices and browsers varies across key demographics, tracking undercoverage will introduce differential errors.

The results of our simulations show that most statistics derived with web tracking data are highly likely to be biased due to tracking undercoverage. Specifically, when simulating the bias at the level found in the TRI-POL dataset, tracking undercoverage led to very large biases across a broad range of quantities of substantive interest. For example, the estimated proportion of participants identified as not consuming news almost doubled from 17% to 31% and the number of media outlets exposed dropped by 50% under this scenario. Our results confirm that the higher the level of undercoverage, the larger the bias introduced. The type of devices missed is also important; mobile undercoverage leads to higher biases than PC undercoverage. There are two potential and complementary explanations for this. On the one hand, internet consumption is more common through mobile devices. On the other hand, most of the online behaviors that we have tested are done more often through mobile devices (Festic et al., 2021). The bias due to device undercoverage also varies depending on the statistics of interest. For univariate statistics, undercoverage underestimates count variables (e.g., counting number of visits, time, or media), but will under- or overestimate proportions engaging in specific behaviors when the underlying variables used are simple counts (e.g., people avoiding news). Remarkably, the biases were considerably larger for count and binary statistics, such as number of media exposed and percentage of news avoiders, over time-based ones. Researchers interested in these types of statistics should be particularly aware of the biases of undercoverage. In terms of multivariate statistics, we observe big variations depending on the associations tested. Although just a speculation, results suggest that when the true association is small, undercoverage might be irrelevant. Nonetheless, for associations that are expected to be more substantial, our results propose that tracking undercoverage might heavily deviate the estimated correlation coefficients from their true value.

There are some limitations in our research design that should be acknowledged. First, participants were recruited using an opt-in online panel of already tracked participants. Although this is the most common approach in the literature, it is unclear whether these results would replicate when conducted on another opt-in panel using a different recruiting and tracking approach. Second, tracking undercoverage has been identified combining paradata and self-reports, the latter being sensitive to measurement errors. Since participants might have trouble properly recalling the devices/browsers used to go online, the estimates of undercoverage cannot be themselves expected to be free of errors, although SOM 4 suggests these errors might not be substantial. Furthermore, the results from the simulations must be understood together with several caveats: simulations are based on a small subsample of fully covered participants. Although no relevant differences are observed between the subsample used and the full sample of participants in key sociodemographic variables, and the weighting approach should reduce some of the potential problems introduced by this, it is to expect that these results to be biased on the side of representation, as well as noisier than desired. For example, our results clearly show that those fully tracked use a significantly lower number of devices. If there is an association between the number of devices and the size of the errors, using a sample of fully covered participants might have introduced biases to our own simulated estimates. Additionally, measurement errors might have affected the identification of the sample of fully covered participants, both introducing false negatives and positives. While this sample is only used as a realistic ground truth to exemplify how errors might deviate estimates from their "fully covered" values, these potential false positives and negatives could introduce noise to our estimates.

Although we would ideally have simulated undercoverage at the level of individual devices and browsers, we were limited by the granularity of our dataset, forcing us to focus only on the effect of full mobile/PC undercoverage (i.e., having all mobile or PC devices undercoverage). Considering that partial undercoverage (i.e., some but not all mobile/PC devices undercovered) is more common, the simulated bias is expected to be lower than realistically expected. Furthermore, the mechanism leading to undercoverage in our simulation is at random (everyone has the same probability). This might not be realistic in real life given that some people being more prone to be undercovered because of some of their demographics, as our multivariate analyses suggest. Additionally, undercoverage could also be associated with how individuals use specific devices or browsers, leading to missingness not at random. This could exacerbate the biases that might be affecting real web tracking studies. In addition, our results cannot address whether there is an association between the type of device undercovered – mobile or PC, brand, sharing status, and technology used -, and the size of the errors introduced. Future research should explore this, to fully understand what devices should be prioritized in tracking studies and to be able to predict undercoverage errors even in the absence of simulations. Finally, our simulations focused on undercoverage bias in cross-sectional settings. Some interesting new combinations of issues could also arise in longitudinal studies, when we consider bias in genuinely longitudinal estimands or when the undercoverage itself changes from wave to wave for some respondents, which are worth exploring in future research.

While these limitations show that our own error estimates must be understood with many caveats, our results have relevant implications for both private companies and scholars. First, fieldwork companies offering panels of already tracked individuals should increase efforts to minimize device undercoverage. This can be achieved by increasing the resources allocated to making sure that participants install tracking technologies in all their devices/browsers, and they keep them installed and updated. Device coverage can also be increased by improving the capabilities of tracking beyond Windows, Android, Chrome, and Firefox. In parallel, they should be more transparent to their clients, disclosing the approaches used to assure full coverage, and up to date information of the level of undercoverage of each of their panelists. As an example, in the context of media research and ratings, established vendors such as ComScore or Nielsen undergo third-party audits, in order to be accredited by the Media Ratings Council. These audits guarantee that companies providing this data follow clear guidelines on how to measure, for example, audiences or ad impressions. Researchers should be aware of the implications of using currency and non-currency providers, and the industry could benefit from clearer more homogeneous standards.

Tech companies can also improve their practices. Currently, some of the challenges faced when tracking specific devices (e.g., iOS) can be linked to tech companies' terms of services, which limit the information that apps can track from users' devices. Although this can be beneficial in many instances, these companies need to acknowledge that if their products and services might lead to negative effects to their users, it should be possible for individuals to willingly access and share this information with academics for research purposes. Hence, research-based tracking technologies should not be treated in the same way than those that exploit personal data for commercial purposes.

Our results also have direct implications when revisiting the results from previous studies using web tracking data. If tracking undercoverage has affected previously published research in similar ways as we find in this paper, it is to be expected that their results might be to some extent biased. For example, research exploring binary outcomes might have underestimated the prevalence of behaviors such as misinformation exposure. Nonetheless, it is important to remark that we have no evidence to suggest that biases found in our study can be extrapolated to other samples and statistics of interest. Hence, this paper only suggests that results coming from studies that do not guarantee full coverage should not be taken as unbiased. This should not discredit the usefulness, relevance and, in many cases, pioneering nature of most past research published using web tracking data.

When it comes to scholars, those using web tracking data – and other types of digital trace data by extension – should not assume that this data is without error. In recent years, there has been a great deal of excitement about the "zero measurement error" of meters and how they can solve longstanding problems of self-reports in areas of research such as media consumption. Nonetheless, following the example of previous media scholars dealing with people meters (Milavsky, 1992; Webster, 2005), it would be recommendable to embrace a more nuanced approach, one that asks scholars to understand not only the benefits but also the caveats of digital data. Therefore, instead of assuming that web tracking data can be complete or unbiased, scholars should focus on understanding how to work with these limitations. This can mean finding ways of reducing or adjusting for errors such as tracking undercoverage, or identifying context in which these errors might be negligible. For example, our results already show that undercoverage is more problematic for some statistics of interest (count and binary variables), than for others (ratios or multivariate statistics). Additionally, web tracking data might be more suitable for specific populations than others, especially those who use fewer devices or mainly use easily trackable devices. For example, in global south countries such as India or Indonesia, around 70% of Internet users only use mobile (Comscore, 2017), with Apple having a marginal market share (4% in India, 12.4% in Indonesia, Statcounter (2024)).

All in all, researchers using web tracking data should reflect these limitations in their analysis plans and reporting practices. Best practice when using metered data must involve:

(1) Identify what participants are affected by undercoverage, and to what extent. Our approach can be used as a template to replicate or build on. If this is the case, researchers should be cautious about measurement errors, with further research needed to test the size and prevalence of these and reduce them if necessary.
(2) Report the proportion of people affected by tracking undercoverage, and some information about the characteristics. This should be similar as to what is done with nonresponse or drop-out rates in survey research, allowing readers and secondary data users to understand the quality of the data. The TRI-POL data protocols can be used as a good example of transparency (Torcal et al., 2023).
(3) When possible, researchers should try to simulate the extent and ways in which tracking undercoverage might bias their results, in a similar way as robustness checks are conducted when using survey data.

Even though our findings raise serious concerns about the quality of meter data, there are also reasons to be optimistic. While device and browser undercoverage are highly prevalent and result in large biases, there is clear room for improvement in the future. Our findings point toward some of the areas where low-hanging fruit can lead to big improvements. Specifically, much of this undercoverage seems to be linked to the current limitations faced by the tracking technologies that we use. With extra investment, most of these limitations could be addressed. Our results show that it is possible to develop approaches to identify, estimate, and report these errors. Doing so in a similar way as we did should be easy and mostly inexpensive to any other researcher dealing with digital trace data and, especially, web tracking data.

Disclosure statement

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

Funding

The work was supported by the H2020 European Research Council [849165]; H2020 European Research Council Ministerio de Ciencia e Innovación Leverhulme Trust Large Centre Grant LCDS Fundacion BBVA. Oriol J. Bosch is supported by an ERC Advanced Grant (835079, PI M.C Mills), Leverhulme Trust Large Centre Grant LCDS (RC-2018-003, PI M.C.Mills).

References

Amaya, A., Biemer, P. P., & Kinyon, D. (2020, February). Total error in a big data world: Adapting the TSE framework to big data. *Journal of Survey Statistics and Methodology*, *8*(1), 89–119. https://doi.org/10.1093/jssam/smz056

Araujo, T., Wonneberger, A., Neijens, P., & de Vreese, C. (2017, July). How much time do you spend online? Understanding and improving the accuracy of self-reported measures of internet use. *Communication Methods and Measures*, *11*(3), 173–190. https://doi.org/10.1080/19312458.2017.1317337

Bartels, L. M. (1993). Messages received: The political impact of media exposure. *The American Political Science Review*, *87*(2), 267–285. https://doi.org/10.2307/2939040

Bosch, O. J. (2023). Validity and reliability of digital trace data in media exposure measures: A multiverse of measurements analysis. https://doi.org/10.31235/osf.io/fyr2z

Bosch, O. J., Asensio, M., & Roberts, C. (2024). What is the best way of collecting data donations in an online survey? An experiment assessing the feasibility of different data donation approaches to measure mobile and app usage. https://doi.org/10.31235/osf.io/q8v35

Bosch, O. J., & Revilla, M. (2022a). The challenges of using digital trace data to measure online behaviors: Lessons from a study combining surveys and metered data to investigate affective polarization. *SAGE Research Methods Cases*. https://doi.org/10.4135/9781529603644

Bosch, O. J., & Revilla, M. (2022b). When survey science met web tracking: Presenting an error framework for metered data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185*(Suppl._2), S408–S436. https://doi.org/10.1111/rssa.12956

Buskirk, T. D., Saunders, T., & Michaud, J. (2015). Are sliders too slick for surveys? An experiment comparing slider and radio button scales for smartphone, tablet and computer based surveys. *Methods, Data, Analyses: A Journal for Quantitative Methods and Survey Methodology*, *9*(2), 229–260.

Cardenal, A. S., Aguilar-Paredes, C., Galais, C., & Perez-Montoro, M. (2019, October). Digital technologies and selective exposure: How choice and filter bubbles shape news media exposure. *The International Journal of Press/Politics*, *24*(4), 465–486. https://doi.org/10.1177/1940161219862988

Chaffee, S. H., & Schleuder, J. (1986, September). Measurement and effects of attention to media news. *Human Communication Research*, *13*(1), 76–107. https://doi.org/10.1111/j.1468-2958.1986.tb00096.x

Chang, L. C., & Krosnick, J. A. (2003). Measuring the frequency of regular behaviors: Comparing the “typical week” to the “past week”. *Sociological Methodology*, *33*(1), 55–80. https://doi.org/10.1111/j.0081-1750.2003.t01-1-00127.x

Comscore. (2017). The global mobile report.

Ernala, S. K., Burke, M., Leavitt, A., & Ellison, N. B. (2020). How well do people report time spent on Facebook? *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, Hawai‘i, USA.

Festic, N., Bu¨chi, M., & Latzer, M. (2021). How long and what for? tracking a nationally representative sample to quantify internet use. *Journal of Quantitative Description: Digital Media*, *1*. https://doi.org/10.51685/jqd.2021.018

Gasparini, A. (2018). Rsimsum: Summarise results from monte carlo simulation studies. *Journal of Open Source Software*, *3*(26), 739. https://doi.org/10.21105/joss.00739

Gonzalez-Bailon, S., & Xenos, M. A. (2022). The blind spots of measuring online news exposure: A comparison of self-reported and observational data in nine countries. *Information, Communication Amp;amp; Society*, *26*(10), 2088–2106. https://doi.org/10.1080/1369118X.2022.2072750

Guess, A. M. (2021, February). (Almost) everything in moderation: New evidence on Americans’ online media diets. *American Journal of Political Science*. https://doi.org/10.1111/ajps.12589

Guess, A. M., Nyhan, B., & Reifler, J. (2020, March). Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, *4*(5), 472–480. https://doi.org/10.1038/s41562-020-0833-x

He, Y., & Tsvetkova, M. (2023). A method for estimating individual socioeconomic status of twitter users. *Sociological Methods Amp*; *Research*, 004912412311686. 004912412311686. https://doi.org/10.1177/00491241231168665

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems*, *12*(12), 767–797. https://doi.org/10.17705/1jais.00282

Jurgens, P., & Stark, B. (2022). Mapping exposure diversity: The divergent effects of algorithmic curation on news consumption. *Journal of Communication*, *72*(3), 322–344. https://doi.org/10.1093/joc/jqac009

Jurgens, P., Stark, B., & Magin, M. (2019, February). Two half-truths make a whole? On bias in self-reports and tracking data. *Social Science Computer Review*, *089443931983164*(5), 600–615. https://doi.org/10.1177/0894439319831643

Kashyap, R., Weber, I., Al Tamime, R., & Fatehkia, M. (2020). Monitoring global digital gender inequality using the online populations of Facebook and Google. *Demographic Research*, *43*(27), 779–816. https://doi.org/10.4054/DemRes.2020.43.27

Konitzer, T., Allen, J., Eckman, S., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2021). Comparing estimates of news consumption from survey and passively collected behavioral data. *Public Opinion Quarterly*, *85*(S1), 347–370. https://doi.org/10.1093/poq/nfab023

Mangold, F., Stier, S., Breuer, J., & Scharkow, M. (2021). The overstated generational gap in online news use? a consolidated infrastructural perspective. *New Media Amp;Amp; Society*, *24*(10), 2207–2226. https://doi.org/10.1177/1461444821989972

Milavsky, J. R. (1992). How good is the a.c. Nielsen people-meter system? A review of the report by the committee on nationwide television audience measurement. *Public Opinion Quarterly*, *56*(1), 102–115. https://doi.org/10.1086/269299

Napoli, P. M. (2005, October). Audience measurement and media policy: Audience economics, the diversity principle, and the local people meter. *Communication Law and Policy*, *10*(4), 349–382. https://doi.org/10.1207/s15326926clp1004_1

Ohme, J., Araujo, T., Boeschoten, L., Freelon, D., Ram, N., Reeves, B. B., & Robinson, T. N. (2023). Digital trace data collection for social media effects research: APIs, data donation, and (screen) tracking. *Communication Methods and Measures*, *18*(2), 124–141. https://doi.org/10.1080/19312458.2023.2181319

Padro-Solanet, A., & Balcells, J. (2022). Media diet and polarisation: Evidence from spain. *South European Society & Politics*, *27*(1), 75–95. https://doi.org/10.1080/13608746.2022.2046400

Palmer, R., Toff, B., & Nielsen, R. K. (2020). "The media covers up a lot of things": Watchdog ideals meet folk theories of journalism. *Journalism Studies*, *21*(14), 1973–1989. https://doi.org/10.1080/1461670X.2020.1808516

Parry, D. A., Davidson, B. I., Sewall, C. J., Fisher, J. T., Mieczkowski, H., & Quintana, D. S. (2021). A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nature Human Behaviour*, *5*(11), 1535–1547. https://doi.org/10.1038/s41562-021-01117-5

Pew Research Center. (2020). Measuring news consumption in a digital era. *Technical report*.

Price, V., & Zaller, J. (1993). Who gets the news? Alternative measures of news reception and their implications for research. *Public Opinion Quarterly*, *57*(2), 133. https://doi.org/10.1086/269363

Prior, M. (2009, July). Improving media effects research through better measurement of news exposure. *The Journal of Politics*, *71*(3), 893–908. https://doi.org/10.1017/S0022381609090781

Rampazzo, F., Raybould, A., Rampazzo, P., Barker, R., & Leasure, D. R. (2022). "Update: I'm pregnant!": Inferring global use of fertility tracking apps. https://doi.org/10.31235/osf.io/2va3n

Reiss, M. V. (2022). Dissecting non-use of online news – systematic evidence from combining tracking and automated text classification. *Digital Journalism*, *11*(2), 363–383. https://doi.org/10.1080/21670811.2022.2105243

Revilla, M. (2022, April). How to enhance web survey data using metered, geolocation, visual and voice data? *Survey Research Methods*, *16*(1), 1–12.

Revilla, M., Couper, M. P., Paura, E., & Ochoa, C. (2021). Willingness to participate in a metered online panel. *Field Methods*, *33*(2), 202–216. https://doi.org/10.1177/1525822X20983986

Revilla, M., Ochoa, C., & Loewe, G. (2017). Using passive data from a meter to complement survey data in order to study online behavior. *Social Science Computer Review*, *35*(4), 521–536. https://doi.org/10.1177/0894439316638457

Scharkow, M. (2016, January). The accuracy of self-reported internet use—A validation study using client log data. *Communication Methods and Measures*, *10*(1), 13–27. https://doi.org/10.1080/19312458.2015.1118446

Scharkow, M., Mangold, F., Stier, S., & Breuer, J. (2020). How social network sites and other online intermediaries increase exposure to news. *Proceedings of the National Academy of Sciences*, *117*(6), 2761–2763. https://doi.org/10.1073/pnas.1918279117

Sen, I., Fl¨ock, F., Weller, K., Weiß, B., & Wagner, C. (2021, 08). A total error framework for digital traces of human behavior on online platforms. *Public Opinion Quarterly*, *85*(S1), 399–422. https://doi.org/10.1093/poq/nfab018

Statcounter. (2024). *Mobile vendor market share India*. https://gs.statcounter.com/vendor-market-share/mobile/india

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2019, April). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*, *089443931984366*(5), 503–516. https://doi.org/10.1177/0894439319843669

Stier, S., Mangold, F., Scharkow, M., & Breuer, J. (2021). Post post-broadcast democracy? News exposure in the age of online intermediaries. *The American Political Science Review*, *116*(2), 768–774. https://doi.org/10.1017/S0003055421001222

Taneja, H. (2016, April). Using commercial audience measurement data in academic research. *Communication Methods and Measures*, *10*(2–3), 176–178. https://doi.org/10.1080/19312458.2016.1150971

Taneja, H., Wu, A. X., & Edgerly, S. (2017, May). Rethinking the generational gap in online news use: An infrastructural perspective. *New Media & Society*, *20*(5), 1792–1812. https://doi.org/10.1177/1461444817707348

Torcal, M., Carty, E., Comellas, J. M., Bosch, O. J., Thomson, Z., & Serani, D. (2023). The dynamics of political and affective polarisation: Datasets for Spain, Portugal, Italy, Argentina, and Chile (2019-2022). *Data in Brief*, *48*, 109219. https://doi.org/10.1016/j.dib.2023.109219

Webster, J. (2005, October). *Ratings analysis* (3 ed.). Lawrence Erlbaum Associates.

White, I. R. (2010). Simsum: Analyses of simulation studies including monte carlo error. *The Stata Journal: Promoting Communications on Statistics & Stata*, *10*(3), 369–385. https://doi.org/10.1177/1536867X1001000305