When survey science met online tracking: presenting an error framework for metered data

Oriol J. Bosch
The London School of Economics and Political Science, Department of Methodology
Research and Expertise Centre for Survey Methodology (RECSM), UPF
O.Bosch-Jover@lse.ac.uk

Melanie Revilla
Research and Expertise Centre for Survey Methodology (RECSM), UPF
melanie.revilla@upf.edu

RECSM Working Paper Number 62

February 2021

When survey science met online tracking: presenting an error framework for metered data

Abstract: Metered data (also called “web log data” or “web-tracking data”) is a type of data obtained from a meter willingly installed by participants on their devices. A meter refers to a heterogeneous group of technologies that allow tracking, at least, information about the URLs of the web pages visited. Metered data has the potential to replace part of survey data or to be combined with survey data to obtain higher quality data. It is crucial, nevertheless, to understand its limitations to mitigate potential errors. Although some research has explored some potential error causes a systematic categorization and conceptualization of these errors is missing.

We present a framework of all errors that can occur when using metered data. We adapt the Total Survey Error framework to accommodate it to the specific error generating processes and error causes of metered data. The adapted error framework shows 1) the data collection and analysis process of metered data and 2) how the unique characteristics of metered data can affect data quality. This framework can be useful to choose the best design options for metered data, but also to make better informed decisions while planning when and how to supplement or replace survey data.

Keywords — Metered data, digital trace data, passive data, web-tracking, error framework, total survey error

1. Introduction

Although surveys are one of the most common methods used for collecting data about various phenomena in social sciences and adjacent fields, measuring certain concepts using survey data remains challenging (e.g. online behaviours, travel and mobility). Cognitive operations involved in producing an answer might generate errors in the responses. For example, participants might only recall parts of some events or recall them inaccurately and self-reports might be based on guesses and assumptions of normality (Groves et al., 2009). Participants might also knowingly misreport their answers for sensitive questions. Besides, the granularity in which survey data can be collected is limited. This can be specifically problematic to measure online behaviours. Remembering online behaviours has become more difficult (Niederdeppe, 2016), with behaviours being increasingly fragmented across situations, devices and platforms (de Vreese and Neijens, 2016). Nonetheless, as (Araujo et al., 2017) note, measures of online behaviours, like online media consumption, remain crucial for political sciences , entertainment, marketing, or health communication. These challenges pushed researchers to supplement or substitute survey data with other data types. In particular, metered data, also called "web log data" (Dvir-Gvirsman et al., 2014), "digital trace data" (Bach et al., 2019), "online behavioural data" (Cid, 2018) or "web-tracking data" (Cardenal et al., 2018), has been used. This type of data is obtained from a meter willingly installed or configured by a sample of participants on their devices (PCs, tablets and/or smartphones). A meter refers to a heterogeneous group of tracking technologies that allow sharing with the researchers, at least, information about the URLs of the web pages visited by the participants. Depending on the technology used, HTML content, search terms, app usage, time or device information can also be collected. Metered data has the potential to bypass the challenges of self-reports by directly capturing the digital traces created by participants when interacting with their devices online, which could prove especially helpful when measuring online behaviours, with a granularity not achievable by surveys. This could allow researchers to capture objective data free of recall errors and memory limitations, in real time. Metered data has already been used to explore a variety of topics, including predicting voter turnout with online behaviours (Bach et al., 2019), studying the effect of Facebook on public agenda (Cardenal et al., 2018) or assessing the quantity and type of vaccine-related information Americans consume online and its relationship to social media use and attitudes toward vaccines (Guess et al., 2020).

Metered data, nevertheless, need to be used properly when aiming to make inferences about a theoretical concept for finite populations (e.g. average time spent visiting online news outlets for the adult population using Internet living in the UK). As any new data type, metered data comes with challenges and limits. Although limited attention has been put to metered data errors when used to draw statistical inferences for finite populations, some research has warned about potential errors (Jürgens et al., 2019; Revilla et al., 2017), which can specially affect the measurement quality of metered data. However, a systematic categorization and conceptualization of these errors has not been developed yet. Better understanding the causes and nature of errors affecting metered data when drawing statistical inferences for finite populations can help researchers make better informed decisions of when and how to supplement or replace survey questions. An approach to do so is developing a Total Error (TE) framework for metered data. TE is a paradigm used to refer to all the sources of bias and variance that may affect the accuracy and efficiency of data (Lavrakas, 2008). When operationalized as a framework, the TE conceptualizes and categorizes the different sources of errors, allowing to understand the data collection and analysis process, as well as to identify and estimate potential errors, the effects of those on estimates and how to minimize them (Biemer, 2010; Groves and Lyberg, 2010). TE frameworks can be used as a planning criterion: among a set of alternative design choices, the one with the smallest total error - considering the budget available - should be chosen. Although the TE paradigm has been mostly used to understand survey errors (e.g. Total Survey Error (TSE) framework), it can be applied to other types of data. In particular, during the last years, several TE frameworks have been developed for Big Data sources (Hsieh and Murphy, 2017).

Our main goal, therefore, is to provide a framework of all errors that can occur when using metered data. Following (Amaya et al., 2020)'s approach for Big Data, we consider that the TSE framework can be adapted to comprehensively identify and classify errors found in metered data, assuming that the error components presented in the TSE framework can also be found in metered data, making metered data errors directly comparable to those of surveys. Hence, instead of creating a completely new framework for metered data, we start from the TSE framework and accommodate it to the specific error generating processes and error causes of metered data. The adapted framework shows how the unique characteristics of metered data can affect data quality, but also allows comparing metered data errors with survey errors. Hence, the adapted framework can be useful to choose the best design options for metered data, but also to make better

informed decisions while planning when and how to supplement or replace survey data with metered data.

The rest of this paper is organized as follows. First, we present some background about the characteristics of metered data and the previous research on error frameworks. Second, we present the approach used to adapt the TSE framework to metered data errors. Third, we describe the data collection and analysis process of metered data. Fourth, we present in a systemized way the categorisation of the different error components and the specific error causes that might arise during the mentioned process. Finally, we discuss the practical implications of this adapted framework, and provide recommendations about how to use it.

2. Background

2.1. Distinctive aspects of metered data

Metered data is the data obtained from tracking the digital traces created by a sample of willing participants when interacting with their devices online (e.g. URLs, time stamps, HTML content). Therefore, two key design aspects of metered data can be identified. First, as surveys, metered data is collected from a designed sample of individuals. One approach is to draw a fresh sample of participants from a given frame and ask them if they consent to install the meter technology into their device (Guess, 2015; Haim and Nienierza, 2019). An alternative approach is to use already available pools of individuals with the meter installed to obtain the data. Different companies offer commercial opt-in online metered panels, which allow to obtain metered data from part or all the panellists (Revilla et al., 2017). This is the most used approach to date.

Second, metered data is nonreactive, meaning that data is not collected by soliciting a response by individuals (Sen et al., 2019) but by passively observing them. Metered data is collected by tracking the traces left by individuals when interacting with their devices online. To do so, it is required to set up a system that runs the tracking technology (e.g. app) and facilitates the collection, storage and extraction of data during and after the study (Harari et al., 2016). Hence, researchers must beforehand create, adapt or select the tracking technologies that they will ask

participants to install or configure into their devices. This technology can be either created from scratch (Guess, 2015) or obtained from a third party in the form of an open-source software (Haim and Nienierza, 2019) or of a commercial tracker (Guess et al., 2018; Revilla et al., 2017; Cardenal et al., 2018). The control over the configuration and capabilities of the technology varies depending on the approach used. The decision on which to use depends on researchers' will, but also on their skills and resources.

Regardless of who designs the meter, different tracking technologies can be used. These technologies change rapidly. However, so far available tracking technologies can be divided into four categories:

1) Apps that passively and continuously track information from the device and the device's browser(s).

2) Plug-ins that passively and continuously collect web browsing history and other device and browsing information.

3) Plug-ins that collect the available web browsing history at a given point in time, but without continuously tracking the device/browser activity.

4) Proxies which can be configured for the networks used to connect to the internet (e.g. WIFI at home, WIFI at work, 3/4/5G network, etc.) through a given device. Each internet connection made by the device through any of the configured networks passes through a set server. This information is automatically stored. However, both inputs and outputs are stored, and connections are made by the user (e.g. visiting a webpage) as well as the device (e.g. device checking for Facebook notifications). Thus, the raw information needs to be processed to determine which information qualifies as a user visit to a webpage. Proxies can be remote or local.

Deciding which tracking technology(ies) to use has important consequences. Thus, it is essential for researchers to think about these beforehand. Two main differences between categories should be considered. First, not all the different tracking technologies can collcet the same type of information nor with the same frequency, granularity and precision. Categories 1, 2 and 4 allow to collect information continuously, while category 3 collects information at one point in time.

Categories 1, 2 and 4 also allow to collect the length of the visit while only category 2 and, although with a much more complex process, category 4 can collect HTML information (e.g. textual and visual data), the actions of respondents in some webpages (i.e. functions they have interacted with) and information from incognito sessions. The capabilities of tracking technologies are determined, in part, by the companies producing and adapting the operating systems (OSs) of the devices (e.g. Google, Apple, Microsoft), which can allow or block different features.

Second, categories differ on the devices (PC or mobile), OSs (e.g. Android or iOS for mobile devices) and browsers (e.g. Chrome, Firefox or Explorer/Edge) in which they can be used. For instance, information from an Edge browser in a Windows computer might be trackable with a different technology than the information from a Chrome browser in the same device and OS. From now on, we will refer to the combinations of these three elements (device/OS/browser) as *targets*. Category 4 (only remote proxies) can be used in all types of targets whereas categories 1, 2 and 3 can only be used in some targets (e.g. tracking apps cannot be installed into iOS devices without breaching Apple's terms of service). In practice, different technologies are used in the same study. As an example, Appendix A shows the different technologies that the company Wakoopa (currently one of the main providers of meters) provides, as well as the type of information that these collect and for which devices they are used.

2.2. Classification of error sources

Classifying error sources is a good way of thinking about the quality of our data. Although data quality can be conceptualized in broad ways (e.g. accuracy, credibility, comparability, usability, relevance, accessibility, timeliness, completeness, and coherence), most error classifications have only focused on statistically computable quality indicators (accuracy). For surveys, classifications of errors have been developed for almost 80 years (Groves and Lyberg, 2010). Since (Deming, 1944)'s first classification of factors affecting the usefulness of a survey, the field's efforts moved from primarily focusing on sampling errors (Deming, 1950; Cochran, 1953) to a broader understanding of error sources. In 1979, the term *Total Survey Error* was first coined by (Anderson et al., 1979). The TSE framework was intended to list all potential error

sources affecting surveys. Besides, error sources were decomposed by variance and bias, sampling and nonsampling, and observation and nonobservation. Further variations of the TSE framework were developed during the following years, for instance, linking it to error notions of psychometrics and econometrics (Groves, 1989; Biemer and Lyberg, 2003). (Groves et al., 2009; Groves et al., 2004) proposed what is probably the most well-known framework for cross-sectional probability-based surveys, which links the steps of survey design, collection and estimation into the error sources (see Figure 1). This framework separates errors of representation and errors of measurement. Errors of representation refer to failures to measure eligible members of the population of interest. They include coverage errors, sampling errors, nonresponse errors and adjustment errors. Errors of measurement refer to deviations between the concept of interest for researchers and the processed measure collected, and include validity, measurement errors and processing errors. All these errors can affect the variance or bias of estimates, contributing to the overall mean square error (MSE) of a statistic. Although the TSE framework was initially conceived for probability-based cross-sectional surveys, it has already been expanded to understand error sources for longitudinal surveys (Lynn and Lugtig, 2017), nonprobability online panels (Unangst et al., 2019) and for cross-national comparisons in international surveys (Smith, 2009).

In the last years, the emergence of new types of data (e.g. web and sensors), normally englobed in the not-so-well defined term Big Data (see (Kitchin and McArdle, 2016) for an ontological discussion of the term), has offered new prospects for measurement, which can be used to substitute or enhance surveys. If we assume that these new types of data also measure an underlying true value, and that errors can deviate measures from the true value, the TSE can be used as a reference to develop TE frameworks for these types of data (Japec et al., 2015). Some researchers have used the TSE as a reference to develop new frameworks for specific types of data.

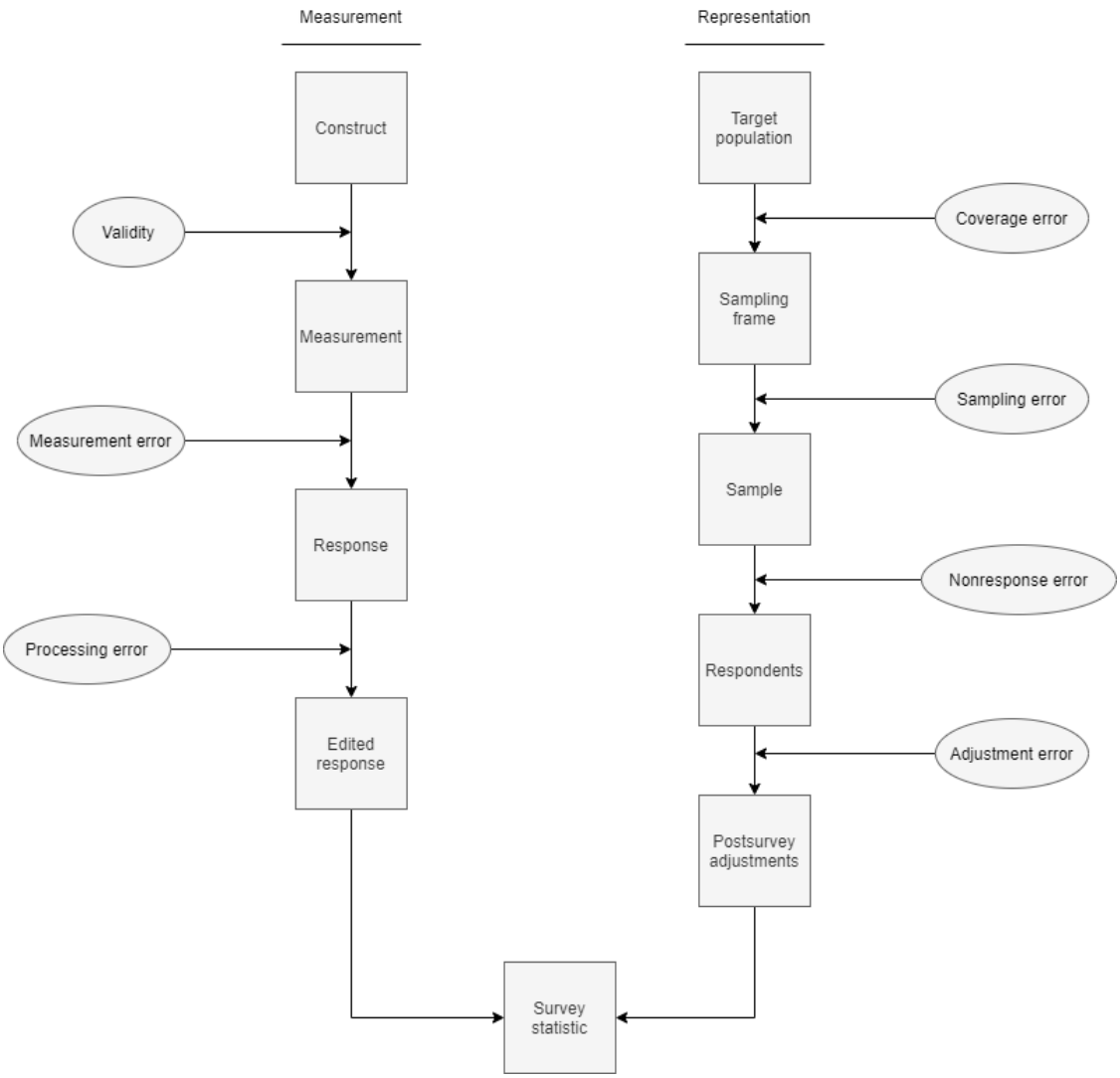

Figure1. Reproduction of the TSE framework by Groves et al. (2009: 48)

These approaches consider that, because of the distinctive data generation and collection processes (e.g. organic, nonreactive and found data) and the unstructured nature of data, new error components affect these types of data. For instance, (Hsieh and Murphy, 2017) presented the Total Twitter Error framework. It identifies three types of errors for Twitter data: coverage errors, query errors (i.e. errors associated with the keywords used to scrape data) and interpretation errors (i.e. errors introduced when coding tweets to create variables). This framework, hence, does not differentiate between errors of representation and measurement and does not use the same error components as the TSE. Similarly, (Sen et al., 2019) have developed the Total Error Framework for Digital Traces of Humans (TDE), mainly focused on Big Data

sources coming from web platforms, especially those from social media sites. The TDE is inspired by the TSE framework and separates representation (e.g. platform coverage errors, entity selection errors or adjustment errors) and measurement errors (e.g. platform affordances errors or signal selection errors). However, errors components are different and mostly focus on the problems related to data extraction from specific web platforms (Sen et al., 2019)(Sen et al. 2019: 1). Contrary to the idea of developing different frameworks for each new type of Big Data source, (Japec et al., 2015) proposed a framework that can be used for all Big Data sources: the Big Data Total Error (BDTE). The BDTE is closely modelled after the TSE framework but includes additional error components. The BDTE was not strictly conceived as a framework, but as a proposal of how Big Data frameworks could be developed. Building from the BDTE, (Biemer, 2020) presented a restructuration of the TSE which considers surveys and Big Data sources as matrices. Rows represent sample or population elements. Columns represent characteristics of the row elements. Cells are the values of the characteristics for each element. All errors can be categorized either at the row, column or cell level. For instance, unit nonresponse would be a type of row error (a missing unit) while specification errors would be a column error and measurement error would be a cell error. The authors argue that this restructuration should make the TSE applicable to Big Data sources. However, it neglects the errors produced before Big Data sources are transformed into rectangular matrices.

Although the previously presented frameworks are based on the TSE, they restructure it in different ways or consider new/different error components. A different approach, although closely connected to the BDTE, is the one presented by (Amaya et al., 2020) in the Total Error Framework (TEF). The authors extend the TSE framework to comprehensively identify and classify errors found in most Big Data sources, assuming that the error components presented in the TSE framework can also be found in Big Data, making Big Data errors directly comparable to those of surveys. The main difference between both types of data, hence, is not in the error components but in the error-generating processes and the specific error causes within each error component. In their framework, (Amaya et al., 2020) identify six steps in the data collection and analysis process, which are linked to different error components: define target population, generate and identify data sources, extract/transform/load, model, create estimates and draw inference. The TEF is mostly conceived for organic and found data. Thus, the data collection and analysis process, as well as the error causes presented in TEF, do not apply well to a design-

based type of data such as metered data. However, TEF proves that the TSE can be expanded/adapted to other new types of data that are or resemble Big Data. In this paper, hence, we follow a similar approach than (Amaya et al., 2020) for metered data.

3. A TE framework for metered data

In order to propose a TE framework for metered data, we use the seven error components of the TSE presented by Groves et al. (2009) as starting point (coverage errors, sampling errors, nonresponse errors, adjustment errors, validity, measurement errors and processing errors). However, since Groves et al. (2009)'s framework was conceived for probability-based cross-sectional surveys, some considerations must be made before adapting it to metered data.

First, metered data is longitudinal by nature. However, metered data can be used in a cross-sectional way, aggregating data points to create a measure for a given period (e.g. the time spent visiting online news outlets from 26 January to 27 April 2015, (Cardenal et al., 2019)). Several aspects of the survey errors and the interaction between different types of errors are different in longitudinal survey contexts (Lynn and Lugtig, 2017). Similar differences might exist between cross-sectional and longitudinal uses of metered data. However, considering that most past research used metered data in a cross-sectional way, for the sake of simplicity, we develop the framework for cross-sectional applications of metered data. We highlight, nonetheless, processes and error causes which could differ when researchers use metered data in a longitudinal way.

Second, as for surveys, probability and nonprobability-based methods can be used to select the sample of metered individuals. The data generation process and the ability to conceptualize and quantify errors varies depending on the method used (Unangst et al., 2019). For metered data, the use of metered opt-in online panels allows directly selecting individuals who have installed the meter in at least one target. The distinction between probability and non-probability sampling approaches is not meter specific and has been discussed in previous research already (Unangst et al., 2019; Pew Research Center, 2016). Thus, for the sake of simplicity, we present the data collection and analysis process when using a probabilistic approach and consider the error causes which would happen for a probability-based approach. However, since most research to date

used metered online opt-in panels, we also highlight the steps and errors which are different or non-existent in that case, even if we should note that large variations can exist within online opt-in panels (e.g. because of the methods used to recruit participants or select the samples).

Finally, metered data being a Big Data source, we borrow the terminology used by (Amaya et al., 2020) to refer to some error components, when it is better suited that the one presented by (Groves et al., 2009). Hence we refer to "validity" as "specification errors" and "nonresponse errors" as "missing data errors."

4. Metered data from a process perspective

Although metered and survey data share the same error components, differences exist in their error generating processes and the causes of errors. Hence, Figure 2 presents the data collection and analysis process of metered data in chronological order, linked with the error components. Figure 2 shows that researchers need to make decisions related to two main aspects: the sample and the measurement. On the measurement side (left set of boxes), the first decision is to *define the concept(s) of interest*, i.e. define what the researchers want to measure (elements of information that researchers want to collect): for example, the total time spent on social media websites. Next, researchers must *design the measurement(s)*, i.e. the specific instrument(s) to be used to gather information about the concept(s) of interest: for instance, the number of hours visiting a defined set of web pages corresponding to all social media websites. Next, researchers need to develop or choose the tracking technology(ies) that will be used to obtain the information needed to create the measurement(s). When using an opt-in online metered panel, panellists already have the meter installed. Thus, researchers only have the possibility to choose the panel with the best suited technology(ies) for their project.

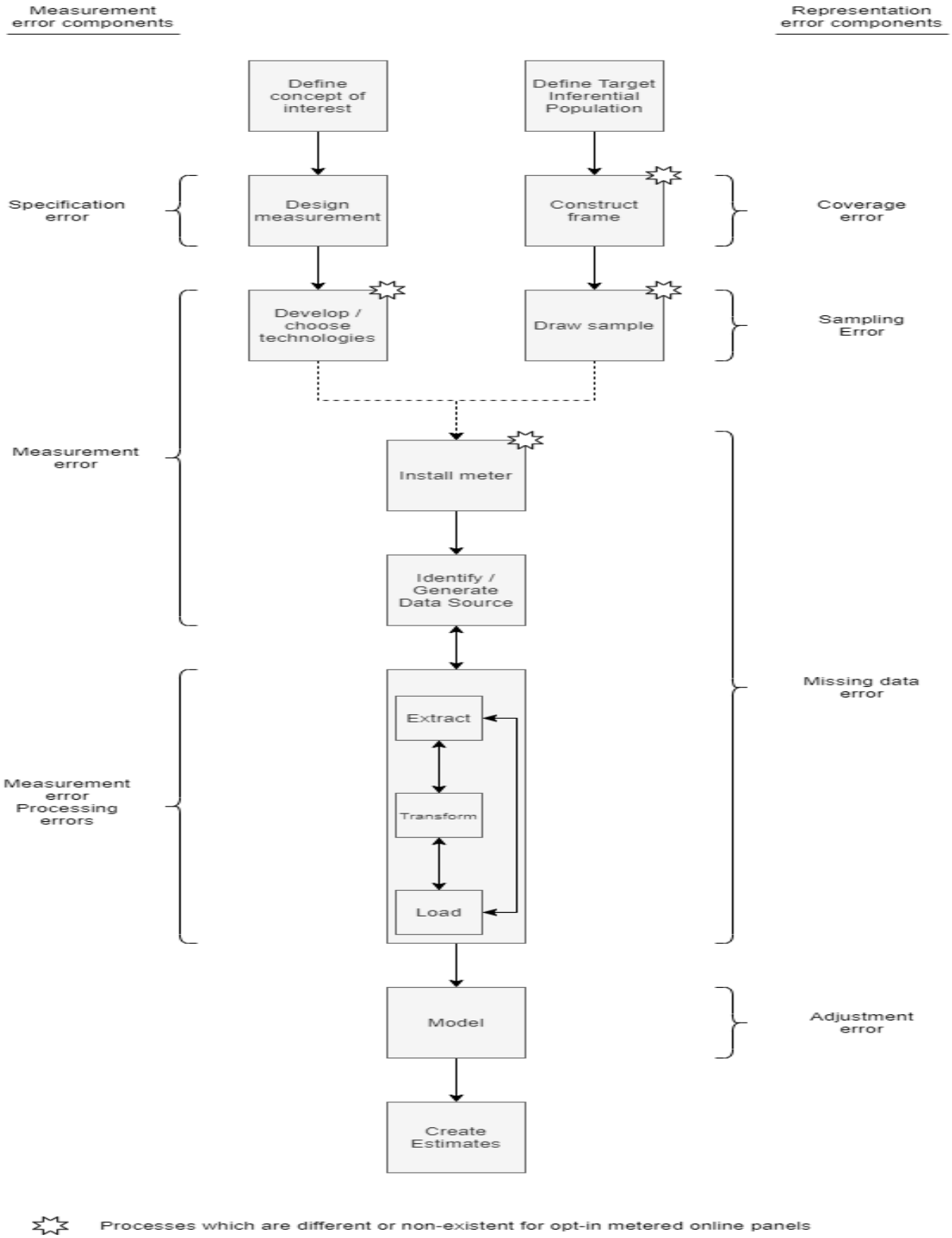

Figure 2. Data collection and analysis process for metered data.

On the representation side (right set of boxes), the first step is to *Define the Target Inferential Population*, i.e. to whom the researchers aim to conclude about. The second step is to *Construct the Frame*. A frame is a list (e.g. emails of university students), or a procedure (e.g. a map of houses), intended to identify the elements of the target population. When using a metered online panel, the panel acts as the frame (Unangst et al., 2019), essentially being a list of individuals with a meter installed or configured on at least one of their devices with an e-mail associated. Panel companies can follow different approaches to create such metered online panel. Currently, a common approach is to invite members of opt-in online panels who have already agreed to participate in surveys to additionally share metered data in exchange of an extra incentive (see (Revilla et al., 2021)). Such a frame does not aim to provide full coverage of the population but to include individuals with enough diversity to cover the needs of the projects conducted by the panel (Groves et al., 2009). The next step is to *Draw the Sample*, which means selecting a fraction of the frame from which measurement will be obtained. Ideally, this should be done using a probability-based sampling approach. In practice, for metered online panels, normally non-probability sampling approaches are used (e.g. quota sampling, see (Ochoa and Porcar, 2018)) because the full panel is not representative of the target populations.

Once the sample has been drawn and the technology chosen, sampled individuals can be asked to *Install the Meter* into their devices. Although researchers are interested in the sampled individuals, data of interest is produced by the individuals' interaction with a specific target (e.g. when visiting a webpage using a Chrome browser in a Windows PC). Therefore, to obtain data from the sampled individuals, they must install a given piece of software or configure their targets in a certain way (e.g. download an app into an Android smartphone to track the behaviour in the device's browsers or download a plug-in into a Chrome browser in a Windows PC). The process of inviting participants can include various phases, and there is no standard way of doing it yet. We illustrate this step with a process which could maximize the information available for researchers and allow understanding when data is missing. First, sampled individuals are contacted (e.g. push-to-web approach in which participants are asked to go online to answer some questions), introduced to the study and asked if they would consent to participate by installing a meter in at least one of their targets. Second, those which consent are asked about the targets they have and use. Hence, researchers know: a) which sampled individuals have and use the targets of interest (device/OS/browser) in which to install the meter and b) if targets used

by individuals can be tracked with the technology(ies) available or not. Third, those who accepted to install the meter and have trackable targets are provided with instructions about how to do it. This process can vary across targets and can be complex for some approaches (e.g. configuring proxies). Thus, individuals might decide not to install the meter or fail to successfully install it in some or all of the targets (e.g. an individual might download a plug-in into his/her PC's Chrome browser, but fail to configure a proxy in his/her iOS device). Once correctly installed, the meter starts collecting data from device and browser logs. When using an opt-in metered online panel, the process of installation of the technology is out of researchers' control. In the next step, the information collected by the meter is uploaded to a server which *Generates the Data Source*. Systems to collect and store data can be set in different ways depending on the technology. For example, for smartphone apps, (Harari et al., 2016) propose to do the following: a portal server receives the data collected by the meter and checks it against the participant manager, which provides the unique user ID. The portal server, then, stores the data collected in the data storage, which is normally a database that can handle large datasets (e.g. MySQL). This dataset allows to query the data to extract it and, when necessary, apply transformations to construct the final dataset for the analyses. When using a metered online panel, apart from the information generated after individuals have been sampled, the panel can also provide data already collected since participants joined the panel.

Once the dataset of interest has been identified and/or generated, come the steps of *Extraction, Transformation and Loading* of the metered data. These steps follow a similar process as the one described by (Amaya et al., 2020) for found Big Data sources. First, the data source can be extracted completely or partially. For instance, the researcher can specify to extract only information of certain domains (news sites from UK) and/or during a specific period (three months before and after the 2019 UK's general election). Then, the extracted data may be transformed to fit the researchers' objectives. As an example, if the information extracted from the news sites domains before and after the elections was HTML information from the content of UK politics articles, this information could be coded and processed (e.g. with a supervised machine learning algorithm) to generate an indicator of the degree of pro-Brexit or pro-Remain of the article. Finally, the data, once in the desired format, are loaded and stored on the researchers' devices or servers. These steps can be done simultaneously or iteratively.

Once data have been extracted, transformed and loaded, researchers can proceed to *Model*. This step involves adjusting the data to better reflect the target inferential population. Hence, it can include weighting for missing data, nonresponse or coverage deficiencies and/or imputation for missing data. Finally, with the adjusted and modelled data, an estimate can be created (e.g. the mean hours of media consumption).

5. Metered data from a quality perspective

Each step of the process from constructing the frame to creating the estimates contains some risk of errors. In the following subsections, we conceptualize the different error components for metered data. Since metered data and surveys share a similar process when it comes to drawing the sample from the frame, contacting sampled units and adjusting the estimates, some error causes are similar or shared with surveys. Those error causes have been explored extensively (Biemer, 2010; Groves et al., 2009). We mainly discuss, hence, the error causes specific to metered data. Table 1 summarizes those, by error component.

5.1. Specification errors

A specification error (also known as (in)validity) arises when the concept being measured differs from the concept of interest (Biemer, 2010). For surveys, researchers should first define the concept that they want to study and then design the survey question(s) to properly measure this concept. When constructing a measurement for metered data, researchers are constrained by what the meter can track and the form in which it can be tracked (e.g. URLs, time). Nonetheless, the logic behind specification errors is the same: considering the type of data that can be collected, researchers create the instrument(s) to measure the concept of interest, and if deviations from this concept of interest occur, specification errors appear.

Table 1. Specific Error Causes for Metered Data by Error Component

Error components	Specific error causes
Specification error	– Measuring concepts from which not enough data is available – Inferring attitudes – Defining valid information
Measurement error	– Non-trackable target – Meter not installed – Uninstalling the meter – New non-tracked device – Technology limitations – Technology errors – Hidden behaviours – Shared device – Social desirability – Extraction error
Processing error	– Coding error – Aggregation at the domain level – Data anonymization
Coverage error	– Non-trackable individuals
Sampling error	– Same error causes than for surveys
Missing data error	– Noncontact – Non-consent – Non-trackable target – Meter not installed – Uninstalling the meter – New non-tracked device – Technology limitations – Technology error – Hidden behaviour – Social desirability – Extraction error
Adjustment error	– Same error causes than for surveys

5.1.1. Defining what qualifies as valid information

When constructing a measurement for metered data, decisions must be taken whether to consider some piece of tracked information as part of the behaviour or attitude wanted to be measured or not. For instance, let us consider that Researcher 1 wants to measure the concept "average hours of consumption of online political news". To measure this concept, Researcher 1 considers the following measurement: "average time recorded of the visits to online political outlets' URLs." This measurement needs to be further developed, to discern which information will be part of it. First, Researcher 1 must define what is considered as a visit, for instance, setting a norm of how many seconds a participant must have spent on the site to qualify as a visit. Second, he/she needs to decide which online outlets to consider, which might publish political articles . Third, Researcher 1 must establish which URLs within the different outlets should be considered. For instance, The Guardian can be considered overall as an outlet which published political articles, but some URLs might not be political. Thus, Researcher 1 might decide to exclude all URLs starting with "theguardian.com/uk/sport". Finally, to compute an average which, represent the normal behaviour of an individual, a time frame of data needed must be stablished (e.g. one week). If due to these specifications, the defined measurement instrument deviates from the concept of interest, specification errors are introduced.

5.1.2. Measuring concepts with by-design missing data

Researchers might decide to measure a concept even being aware that part of the data will be missing-by-design. For instance, (Guess et al., 2018) intended to measure the total fake news consumption of a sample of Americans during the 2016 presidential election. However, they collected data only from metered PCs. Thus, the authors knew since the design stage that they will not be measuring the total fake news consumption, but only the fake news consumption from PCs. However, they used the collected data to make inferences about the total fake news consumption. This means that they make a strong assumption: that total fake news consumption can be inferred from PCs fake news consumption. If this is not the case, specification errors occur.

5.1.3. Inferring attitudes and opinions from behaviour

Metered data collects behavioural information which could serve as a good proxy to measure online behaviours. Other types of digital trace data have been used in the past to measure attitudes and opinions. For instance, (Barberá, 2015) inferred individuals left-right position based on the Twitter accounts that they followed. If a behavioural indicator (e.g. URLs visited) is used to infer about attitudes and opinions (e.g. left-right position) without a solid theory behind, it might produce weaker relationships, affecting the validity of the measurement.

5.2. Measurement errors / Missing data errors

When using metered data, measurement and missing data errors can be confounded. Thus, we discuss them together. On the one hand, *measurement errors* are produced when the value obtained from a sampled unit deviates from the true value that the measurement should have if no errors happened when collecting the data. For surveys, the measurement consists in (at least) one question and the values obtained are the provided answers. However, because of human memory limitations, interviewers' influence, deliberate falsification, or comprehension errors the answers might deviate from the true value of the measurement. For metered data, the measurement consists in the defined data that should be tracked from the sampled units (see example in section 5.1.1) and the values obtained are the tracked data. On the other hand, *missing data errors* are produced when information of some sampled units cannot be collected. This can happen at the unit level (i.e., no information is available for any measure for a given unit) or at the item level (i.e., information is not available for an item for a given unit). When data is missing, estimates are drawn based only on a subset of the sample. A bias is introduced when individuals with systematically missing data differ from individuals with available data,.

For surveys, participants can either answer or not answer (for whatever reason). Those not providing an answer are considered as missing, and since no information is available from them, they are excluded from the specific analyses. For metered data, instead of asking a question, participants' behaviours are recorded using a meter. Thus, a lack of data in the dataset (e.g. no adult website URLs recorded) might mean a true absence of behaviour (the individual has not

visited any URL) or a failure to capture data (e.g. the participant deactivated the meter to visit such URLs). If the absence of behaviour is provoked by a failure to capture data, the participant should be excluded from the analyses (the lack of behaviour cannot be considered as real, so the real value is unknown). If the lack of behaviour is real, then it should be considered as so (e.g. 0 minutes visiting theguardian.com during the last 15 days). Deciding whether the lack of information is considered as a missing is not straightforward. It requires additional information and often depends on the researchers' judgement. If an absence of behaviour which is due to missing data is considered as a true absence of behaviour, this will produce a measurement error equivalent to an underreporting. This might not be the case, nonetheless, when measuring non-behavioural concepts which require observations of specific behaviours. For instance, to compute the left-right orientation of participants using the visits to political news media website as a proxy, for those participants with no visit to any news media website, no left-right value will be computable. Thus, in this case, it cannot lead to an underreporting but only to a missing value.

5.2. 1. Noncontacts

In order to collect metered data, sampled individuals need to be contacted and asked to install the meter. As for surveys, the researchers, however, might fail to contact with some of the sampled units. For instance, the mail or e-mail invitation might never arrive or be seen by the sampled unit. In this scenario, the sampled individual will not become a participant, producing a missing data error equivalent to a unit nonresponse. No measurement errors are introduced by noncontacts.

5.2.2. Non-consents

Once contacted, individuals are asked to consent to install the meter into at least one of their targets. Individuals might not be willing to participate and install the meter. Hence, no information will be collected for that sampled unit. (Revilla et al., 2021) found, exploring the online metered opt-in panel Netquest in 9 countries, that acceptance rates (i.e. the proportion accepting to install the meter from those invited) ranged from 53% in Colombia to 28.1% in the United States. As for noncontact, the sampled unit will not become a participant. No measurement errors are introduced by non-consents.

5.2.3. Non-trackable targets

Some of the targets used by sampled units might not be trackable with the chosen technology(ies). Hence, information will not be collected for those targets. For those units with some non-trackable devices, if researchers know that 1) a participant cannot be tracked in one or some targets and 2) all the information of interest is produced by this or these targets, that information should be considered as missing. For instance, assume that some researchers want to measure the time spent browsing the internet with smartphones. Also, these researchers know that Participant 1 only uses non-trackable targets to access the internet with a smartphone. This lack of information will be considered as missing, meaning that the estimates will be computed excluding Participant 1. However, if researchers did not know that all the information was produced by non-trackable targets, this data could not be identifiable as missing. In this scenario, a missing data error would only happen if the loss of information provoked that all the information needed to compute a nonbehavioral measurement (e.g. left-right position) was missing.

Non-trackable devices can also provoke measurement errors, similar to underreporting in surveys, when the loss of information is partial (e.g. researchers only record 1 hour of internet consumption done with a tracked target but the true behaviour is 3 hours done by one tracked target and one non-trackable target) or is complete but cannot be identified as missing.

5.2.4. Meter not installed

Individuals who consent to be tracked still have to install or configure the meter into their targets. Several reasons might prevent them for installing the meter. They might not self-report the use of a device, provoking that researchers do not offer the option of tracking that device. They might decide not to do it even if they agreed (e.g. reading the instructions they realize it is too much burden). Besides, individuals might fail to successfully install or configure the meter in some or all devices, networks and/or browsers for various reasons (e.g. low IT skills, technical problems). For those targets, the missing data and measurement errors are the same as for non-trackable targets (see 5.2.3). Evidence suggests that not having the meter installed, either because devices were non-trackable or participants did not install them, might be an important problem. (Revilla et al., 2017), surveying panellists from an online metered opt-in panel in

Spain, found that almost 57% of the respondents have the meter installed in only one device whereas only 4% of them use only one device to go online. Similarly, a report from the (Pew Research Center, 2020) found that for those Ipsos' Knowledge panellists who accepted to install a meter in at least one of their devices, only 28% stated having all the devices that they use to access the internet metered, whereas 68% reported going online with non-tracked devices.

5.2.5. Uninstalling the meter

Participants accepting to be tracked at the beginning of the study can change their mind over time (e.g. lack of memory in the device, change on privacy concerns) and decide to uninstall the meter. Moreover, some participants may uninstall the meter accidentally. Both can happen for some or all of the tracked targets. For those targets with uninstalled meters, data will not be available from that point on. (Revilla et al., 2021), using data from the online metered opt-in panel Netquest in 9 countries, found that the proportion of invited panellists sending data after three months was around 16 percentage lower than the proportion who participated in the first place.

This can provoke missing data errors. If researchers know that the meter has been uninstalled (if it has been designed to give this information), and also know that all the information for a given variable was to be produced by that target during the unobserved period of time, this is considered as a missing data error similar to item nonresponse. However, if researchers do not have information about when the meter has been uninstalled, the uninstalling will only produce a missing data error when the loss of information provokes that all the information needed to compute a nonbehavioral measurement is missing. For longitudinal uses, uninstalling the meter for all the targets could be considered as attrition and, hence, equivalent to unit nonresponse for subsequent waves. When the loss of information is partial or cannot be identified as missing, measurement errors similar to the ones explained before can be produced.

5.2.6. New non-tracked targets

During the course of the study, participants might purchase new devices or substitute old ones, switch to new browsers or start using new networks. If these new targets are not tracked, their information will be lost. If the researchers know that a new target is being used, that the target is not tracked and that all the information of interest is produced by that or those targets, the loss of

information is considered a missing data error. However, if researchers do not have information about the new non-tracked target, it will only produce a missing data error when the loss of information provokes that all the information needed to compute a nonbehavioral measurement is missing. For longitudinal uses, if participants substitute all their tracked targets for new non-tracked ones, it could be considered as attrition and, as for uninstalling the meter, equivalent to unit nonresponse for subsequent waves. New non-tracked targets can also lead to measurement errors similar to the ones explained before.

5.2.7. Technology limitations

Tracking technologies are not perfect and can be subject to limitations that prevent them from capturing some types of data. Some of the current limitations are the following: 1) not all tracking technologies available allow to capture behaviours happening in incognito modes. 2) Although most technologies can capture domain-level information (i.e. theguardian.com) for all types of webpages, some approaches cannot capture subdomain-level information (i.e. tehguardian.com/sport/…) for https sites. 3) Behaviours happening inside apps, to the authors knowledge, cannot be capture with any technology at this day. For instance, it is not possible to know the news read inside a news outlet app or the profiles visited when using the Twitter app. 4) HTML content cannot be obtained from all tracking technologies. Therefore, depending on the technologies used to track the different targets of interest, some information might not be trackable. If the researchers know that all the information lost is due to technology limitations, the loss of information is considered a missing data error. As for other causes of error, for nonbehavioural measurements a complete lack of behaviour will produce a missing data error. Technology limitations can also lead to measurement errors similar to the ones explained before. For longitudinal uses, if technology limitations vary across time (e.g. new version solving some of the limitations or introducing news), measures of change will be affected by the changes in measurement errors' size.

5.2.8. Technology errors

The meter, as any technology, can suffer from technological errors. If the meter stops recording or fails to correctly record information, information will be lost. Several reasons can lead the meter to fail in terms of technology: 1) the device or a third-party app might shut down the

ability to collect data when the device is low of battery, in order to reduce the device's energy consumption. 2) If the meter is working through a proxy, the proxy generates raw data that must be processed to identify which part of the traffic came or was received, and which part of the traffic was done passively by the device (e.g. downloading Facebook information) or actively by the participant. This is normally done by trained algorithms. However, this is not purely accurate. 3) Since tracking technologies are built on top of OSs and browsers, when new versions of these software are released these can prevent the technologies from working (properly), provoking a loss of information until the technology is adapted to the new version. These errors can provoke an incorrect collection or a loss of information.

When technology errors occur, not enough information is available to identify missing data i.e. there is no information of either data being lost or which data was lost. Missing data errors will only happen when the loss of information provokes that all the information needed to compute a nonbehavioural measurement is missing. Nonetheless, technology errors can produce measurement errors. First, for proxies, if there is an incorrect collection of information (e.g. the algorithm incorrectly categorizes a passive behaviour done by the device as an active behaviour done by the participant), this will produce a measurement error similar to overreporting. Second, a loss of information will produce a measurement error similar to underreporting (for instance, if the participant visited some of the domains of interest while the meter was not recording). For longitudinal uses, technology errors will have a similar impact as technology limitations.

5.2.9. Hidden behaviours

Some technological approaches allow participants to disconnect the meter or to configure blacklists of domains not to be tracked. This allows participants to avoid the meter to track behaviours that they are not willing to share. For instance, when dealing with online banking or visiting sensitive web pages (e.g. adult websites).

As for technology errors, not enough information is available to determine if hidden behaviours provoke missing data errors. Only if a complete loss of information prevents computing a nonbehavioral measurement, it will be considered a missing data error. Nonetheless, if information from a given measurement is completely or partially lost, it produces a measurement error similar to an underreporting in surveys.

5.2.10. Social desirability / Hawthorne effect

Participants might change their behaviour if they know that they are being observed (Jürgens et al., 2019). Consequently, their observed behaviours could deviate from their habitual behaviours (when they are not observed). This can specially affect sensitive behaviours, with participants behaving in a more socially desirable way once they are observed. For instance, a participant visiting adult websites several times a week could avoid visiting such websites once he/she has installed the meter. Participants could also change their behaviours to visit more often socially desirable pages, like news sites, although it seems less likely. Although no experimental research has been conducted yet, preliminary evidence using non-experimental data suggests that individuals might not change they behaviour when observed (see (Toth and Trifonova, 2020)).

These changes of behaviours will produce measurement errors, unless they produce a complete loss of information needed to compute a nonbehavioral measurement, in which case it should be considered as a missing data error. For longitudinal uses, if participants start to behave differently, measures of change will be biased.

5.2.11. Extraction errors

Often researchers do not extract all the data, but select specific domains, periods of times or individuals for which/who to extract information. When specifying the domains or the period of time, incomplete or erroneous specifications can generate measurement errors. For instance, in the case of URLs, if a fake news domain is not specified in the query to extract data this would underestimate the total fake news consumption. This is not a specification error, since the error is produced not from the conceptualization phase, but as a mistake when creating and executing the queries which to extract the specified data. Extraction errors can also produce missing data errors. Indeed, problems with the query can leave sampled participants out of the final database, if their information is not extracted.

5.2.12. Shared devices

Metered data is produced by devices. Devices can be shared between different individuals. For instance, although a PC can be linked to a participant, the same PC might be shared by different members of the family. (Revilla et al., 2017) found, for a metered opt-in online panel, that more

than 60% of desktops, 40% of laptops and tablets, and 9% of smartphones used to go online by the participants were shared to some degree. Let us assume that some researchers want to measure the partisan news consumption. Participant 1 shares a metered PC with his/her father. During the metered period, Participant 1 does not visit any news media website. However, Participant's 1 father consumes an average of 1 hour of liberal news media outlets from the shared PC. Participant 1 will be considered to present a liberal consumption pattern, although he/she did not visit any news media website. Now assume that Participant 2 visits an average of 1 hour of liberal media outlets from a shared metered PC. Participant 2 shares his/her PC with his/her partner, who visits an average of 1 hour of conservative media outlets. Participant 2 will be considered to engage with both conservative and liberal media outlets equally, not being polarized. However, Participant's 2 true behaviour would be exclusively liberal.

Approaches to differentiate between the participant's behaviour and third-person's behaviours should be used (e.g. algorithms that differentiate between individuals' behaviours, see (Ochoa et al., 2017)). However, these approaches can still not perfectly discriminate between behaviours, so such errors are not perfectly accounted for. Moreover, these approaches are often not used. Shared devices, hence, introduce measurement errors. No missing data errors are produced by shared device. For longitudinal uses, if the patterns of sharing the devices vary across time, measures of change will be affected by variations in the sizes of measurement errors.

5.3. Processing errors

Processing errors are the errors introduced after the data are extracted and prior to estimation. They create deviations between the variables used for estimation and the observed ones. For survey data, processing errors can be produced during data entry, coding, editing, disclosure limitation, and variable conversions or transformations (Amaya et al., 2020). Considering that metered data can be unstructured, prior to estimation, extracted data might need to undergo a series of transformations before being converted into variables suited for estimation. These transformations happen during the *transformation* subprocess of *Extraction, Transformation and Loading* (see Figure 2). More complex transformations increase the risk of introducing processing errors.

5.3.1. Coding/categorization errors

Metered data can take an unstructured form like URLs, text, images or videos. Unstructured data often need to be processed and transformed to be useful for most researchers. This process might involve coding or categorizing the unstructured data into classes, labels, sentiments and so on. Categorization can be done manually (e.g. using MTurk coders, (Peterson et al., 2018)), using ad hoc machine learning algorithms (e.g. supervised machine learning to categorise the topic of news articles, (Peterson and Damm, 2019)), or using already available third party machine learning algorithms (e.g. a comparison of manual coders and Google Vision API to code images, (Bosch et al., 2018)). Manual coding can prompt the same errors as for survey data, i.e. that different persons coding the same raw data have different judgments or that coders systematically misinterpret and misclassify some information. Supervised learning algorithms present a similar problem, with errors linked to the accuracy of the machine learning model. The less accurate the model, the higher the chances of misclassifying. Finally, with third party algorithms, an extra problem is the lack of information of how the model is created. Third party models might be better or worse, but black boxes can prevent researchers from identifying the errors. Misclassifications can happen and these can be systematic, but the extent and the processes behind might be difficult to assess and correct.

5.3.2. Data aggregation

In some cases, the final analyses cannot be done using the data at the URL level due to vendors, privacy regulations or researchers' decisions. Then, data is aggregated at the domain level before being analysable (e.g. the domain for theguardian.com/uk/sport/ is The Guardian.com). By aggregating the data at the domain level, information is lost: 1) some concepts might not be measurable. For instance, if a researcher wants to measure the time spent visiting the sports section of The Guardian, this will not be possible since all the URLs with theguardian.com/uk/sport/ will be converted into theguardian.com. 2) Some concepts might be measured with less accuracy. Using a slightly modified version of the previous example, if a researcher wants to measure the time spent visiting sports related news, it would be possible to accurately measure the time spent for sports outlets (Eurosport). However, information from

generalist outlets (e.g. theguardian.com) would be lost. The final measure would underestimate the total time spent visiting sports news outlets.

5.3.3. Data anonymization

Data can be anonymized i.e. all the pieces of information that could lead to identifying participants are obscured. This can be done manually or using machine learning algorithms. Both approaches, however, can provoke errors. Thus, relevant information which was not intended to be hidden can be lost (see Ochoa and Paura, 2018).

5.4. Coverage errors

Coverage errors occur when the sampling frame from which the sample is drawn differs from the target population. Frames can suffer from undercoverage (i.e., units part of the target population are not in the frame), overcoverage (i.e., units in the frame are not part of the target population) or duplicate elements (i.e., units are listed more than once). Whether these problems introduce bias or variance depends on whether the duplicates and the over or undercovered are significantly different in terms of the statistic of interest. If researchers use an opt-in metered online panel, coverage errors occur when the full panel differs from the target population (Groves et al., 2009). Although coverage errors are unquantifiable per se when using an opt-in online panel, errors are linked to one or more panel practices (Unangst et al., 2019): for instance, their refreshment strategies or if they blend samples from different sources. Researchers can qualitatively assess panels beforehand to potentially reduce these errors.

5.4.1. Non-trackable individuals

Although participants might appear in a sampling frame, they might be non-trackable. Similarly as for online surveys, those individuals who do not use the Internet cannot provide information. Besides, for metered data, Internet users might only use non-trackable targets to access the Internet. Although these individuals might appear in the sampling frame, once contacted, they do not have the possibility to participate. Specific coverage errors related to Internet access/trackable devices are no evident and often cannot be assessed until sampled units are

contacted and the access to, or use of, the internet/trackable devices is assessed (Couper et al., 2007). If significant differences exist between those trackable and those non-trackable, coverage errors are introduced. This type of coverage error could be solved if sampled units which use the Internet but with non-trackable devices were provided with trackable devices by the researchers.

5.5. Sampling errors

Sampling errors are defined as the errors that arise because of analysing a subset of the population of interest rather than the entire population. When units in the sampling frame are given a zero chance of selection, in every potential sample drawn that unit will be systematically excluded. If the excluded units differ from the non-excluded ones in the frame, bias is introduced. For instance, in Address Based Sampling, when invitations are sent by mail, the selection of individuals within addresses is done by the address residents, following quasi-random protocols. This can lead to sampling errors.

Sampling also introduces variance into estimates. For a given sampling design, many different samples could be drawn. Each sample, by chance, would produce different values for the statistic of interest (e.g. average time spent visiting online political media outlets). The hypothetical dispersion of the values for all the different drawable samples measures the sampling variance, with small dispersions leading to low variance. Several factors can increase sampling variance, for instance, small sample sizes or the use of clustering.

If researchers use an opt-in metered online panel, nonprobability sampling is used (e.g. quota sampling). In this case, units are included with unknown probabilities. Therefore, the size of sampling errors is unknown. Regardless of the probability or non-probability nature of the approach used, the causes behind sampling errors do not necessarily differ between survey and metered data.

5.6. Adjustment error

When modelling and creating estimates, researchers can make use of weighting or imputation strategies with the objective of improving the representativeness of statistical estimates. Since metered data can be based on probability and nonprobability samples of participants drawn in a similar fashion as for surveys, similar weighting and imputation strategies can be used, with similar risks of producing errors. Hence, deficiencies in missing data and coverage error weighting adjustments, and imputation for item missing data, can introduce adjustment errors (see (Mercer et al., 2017): 256-257 for an example).

If metered data, however, is not supplemented with survey data or profiling information, accurately weighting for traditional sociodemographic and/or specific attitudinal variables might not be possible. Only using behavioural data to compute the weights might prove difficult since gold standard estimates for these are not normally available. Besides, even if metered data is supplemented with survey data or profiling information, a proportion of those being metered might not answer the questions, having missing information for those variables needed to compute the weights.

6. Discussion

In this paper, our main goal was to provide a framework to understand the errors that can occur when using metered data. To do so, we adapted the TSE framework by expanding (Groves et al., 2009)'s version to accommodate it to the specific error generating and error causes of metered data. The adapted framework serves as a guide to understand how the specific characteristics of metered data can affect data quality and to compare metered data errors with survey errors. This framework, first, shows that the data generation and analysis process of metered data is different than the ones of surveys and found Big Data sources presented in the TSE framework (Groves et al., 2009) and the TEF (Amaya et al., 2020), respectively. By presenting a detailed description of the process, researchers can understand all the steps that they should consider when planning to use metered data, as well as the errors associated with those. Second, this framework shows that metered data, as any data sources, is imperfect. Many errors, some already known and other

specific for metered data, can potentially affect data quality. Some causes of these errors, moreover, are new and specific for metered data. On the one hand, the measurement process of metered data, which is not based on developing and asking questions but on tracking and observing behaviours using specific technologies, provokes important differences. Technological errors, partial observation of behaviours or inaccurate observations due to shared devices are new problems. On the other hand, although the representation aspect of metered data is more similar to the one of survey data, sampled units are not contacted to answer a questionnaire but to install or configure a tracking technology into their targets. The different causes of nonresponse for metered data, as for instance not being able to install a tracking technology into an individual's target, substantially differ of those of surveys.

6.1. Main limits

This framework, nevertheless, has limits. First of all, although it tries to be as general as possible, highlighting differences when using metered opt-in online panels and for longitudinal uses, applying it to all the different ways in which metered data can be collected and analysed might not be possible. Moreover, as for TSE and TEF, this framework only considers a definition of data quality. Other factors should be considered when deciding whether to supplement or replace survey data with metered data (e.g. cost, timeliness, risks). For instance, metered data is currently an expensive alternative. In addition, privacy and ethical issues must be considered when planning to collect metered data. How data is collected, stored, processed, and shared can have important ethical implications. Especially important consideration must be placed in the tension between reproducibility and data protection. To protect participants privacy, metered data should not be shared nor made publicly available in its raw form. This can make results non-reproducible. Another limit lies in the lack of previous methodological research available. This framework is mostly conceptual. Empirical research is needed to understand how large these errors can be in different studies, and to discover new errors which might not be considered in this framework. Nonetheless, our framework can serve as a basis to empirically explore the extent to which error causes affect metered data. Apart from this, metered data is continuously evolving. Errors which appear here might be solved as technology evolves (e.g. shared devices

might be solved by implementing reliable algorithms). Besides, as technology evolves, new functionalities might allow to collect new types of data, which might introduce new challenges. Our framework, however, can serve as the basis for future adaptations if new changes outdate the current ones. Finally, this framework only considers the errors of metered data independently. Research must evaluate the error when combining metered data with surveys, which is common practice. For instance, how different harmonization methods affect the resulting estimates and models. Our framework can serve as a reference to explore these errors, but extra considerations might be needed.

6.2. Practical recommendations

Based on this framework, we propose some preliminary practical recommendations for researchers and practitioners using metered data. First, a clear definition of what concept the tracked information should measure is needed. This requires an exhaustive description of, for instance, what is considered as a visit and what domains and specific URLs should be used to measure the given concept. Besides, if measuring attitudes with what is essentially behavioural data, a clear theoretical argument must be made to justify that behaviours can be used to infer attitudes. For instance, (Barberá, 2015) inferred individuals left-right position using the Twitter accounts that they followed. To do so, the author clearly stated his assumptions, based on previous research: 1) individuals prefer to follow accounts whose ideological positions are similar to theirs, 2) following an account is a costly signal about individual's perception of their position and of the accounts followed, 3) offline network homophily can be extrapolated to online social networks and 4) individuals on Twitter behave similarly as when exposing themselves to political news, seeking opinion-reinforcing political information instead of opinion challenges. Second, researchers must consider the potential consequences that the different technologies can have on data quality before deciding which one(s) to use. The technologies used can have an impact on: which targets can be tracked (e.g. no iOS devices if using an app/SDK), the type of information trackable (e.g. no HTTPs subdomains if using a proxy), whether data is tracked continuously or at a given point in time (e.g. plug-ins that collect the available web browsing history at a given point in time), the accuracy of the data (e.g. proxies

might collect connections made by the device and not by the unit of interest) and potentially the willingness of individuals to install or configure the tracking technology (e.g. individuals need to manually configure the proxy following a guide which can reduce the willingness). All this can introduce different measurement and missing data errors. Besides, if different technologies are used to track different targets, differences in the propensity of installation and their measurement quality can introduce technology/mode effects, not only between participants, but also within. Given the seemingly impossible option of tracking all the targets with a unique technology that provides a high accuracy and low burden for individuals, researchers must consider the best trade-offs between cost, target coverage, burden, tracking accuracy and harmonization between different alternative designs, and clearly justify the final decisions. Third, considering that past research has found that willingness levels to install tracking technologies is low (e.g. 16.6% in Spain, (Revilla et al., 2019)), researchers should explore strategies to increase the willingness of individuals to install them. However, our paper shows that 1) different tracking technologies can present quite different installation and/or configuration processes; 2) in many cases multiple tracking technologies might need to be installed for the same individual; and 3) targets are normally unknown, so researchers might need to rely on individual's self-reports when deciding in which targets to ask for the meter to be installed. Hence, different individuals might present substantially different installation/configuration processes. All this need to be considered beforehand, in order to create specific instructions and encouragement strategies to maximize participation and complete installation. Fourth, missing data and measurement errors can be confounded. It is the responsibility of researchers to define strategies to maximise the information available to identify missing data. For instance, creating strategies that help identifying those which are not tracked in all the targets of interest. Fifth, even if no technological nor installation/willingness problems occur, it cannot be assumed that what is observed is a perfect representation of individual's real online behaviours. Targets might be shared with non-sampled individuals. In addition, individual's might change their behaviours when knowing they are observed. Finally, considering that design decisions can impact data quality, these decisions need to be properly reported. Most research done to date does not disclose the technology used, the specification of the measurements used, the specific targets metered (more specifically than reporting "desktops"). This is not an adequate practice since it does not give enough information to judge the implications for results, and it is not comparable

to the current practices when using survey research. The supplementary information provided by (Guess et al., 2018) represent a good example of how to transparently describe how the measured concepts have been defined and operationalized.

6.3. Conclusion

This framework shows that caution should be taken when using metered data to make inferences about a theoretical concept for finite populations. So far, metered data has been used without questioning its quality, assuming it should be higher than the one of surveys. Nevertheless, many errors can affect metered data. However, while survey errors have been studied for a long time and are generally well known, with (well) developed approaches to assess and correct them, metered data errors are still unclear, specific to the companies and technologies used and with very little research on how to assess and correct them. Although the size of these errors is unknown, metered data can be expected to be biased.

This does not imply that metered data should not be used. Some of the advantages of metered data are still relevant. Metered data allows to collect data in real time, with a high granularity and in an unobtrusive way. Besides, for behavioural concepts, although specification errors can happen, a strong relationship between measures and behaviours can be expected. These benefits might offset the potential errors presented in this framework. However, more work is needed to understand the consequences of metered data errors and the trade-offs of using metered data instead of survey data. Besides, researchers using metered data should at least acknowledge the potential errors and discuss the consequences they may have on their results.

Funding information: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 849165)

Acknowledgements: We are very thankful to Patrick Sturgis, Jouni Kuha, Mariano Torcal and Carlos Ochoa for their insightful comments during the preparation of this paper.

References

Amaya, A., Biemer, P.P., Kinyon, D., 2020. Total Error in a Big Data World: Adapting the TSE Framework to Big Data. Journal of Survey Statistics and Methodology 8, 89–119. https://doi.org/10.1093/jssam/smz056

Anderson, R., Kasper, J., Frankel, M., 1979. Total Survey Error: Applications toImprove Health Surveys.. Jossey-Bass Publishers, San Francisco.

Araujo, T., Wonneberger, A., Neijens, P., de Vreese, C., 2017. How Much Time Do You Spend Online? Understanding and Improving the Accuracy of Self-Reported Measures of Internet Use. Communication Methods and Measures 11, 173–190. https://doi.org/10.1080/19312458.2017.1317337

Bach, R.L., Kern, C., Amaya, A., Keusch, F., Kreuter, F., Hecht, J., Heinemann, J., 2019. Predicting Voting Behavior Using Digital Trace Data. Social Science Computer Review 089443931988289. https://doi.org/10.1177/0894439319882896

Barberá, P., 2015. Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. Political Analysis 23, 76–91. https://doi.org/10.1093/pan/mpu011

Biemer, P., Lyberg, L., 2003. Introduction to Survey Quality. Wiley, New York.

Biemer, P.P., 2010. Total Survey Error: Design Implementation, and Evaluation. Public Opinion Quarterly 74, 817–848. https://doi.org/10.1093/poq/nfq058

Biemer, P.P., 2020. Data Quality and Inference Errors, in: Foster, I., Ghani, R., Jarmin, R.S., Kreuter, F., Lane, J. (Eds.), Big Data and Social Science: Data Science Methods and Tools for Research and Practice. CRC Press.

Bosch, O.J., Revilla, M., Paura, E., 2018. Answering Mobile Surveys With Images: An Exploration Using a Computer Vision API. Social Science Computer Review 37, 669–683. https://doi.org/10.1177/0894439318791515

Cardenal, A.S., Aguilar-Paredes, C., Cristancho, C., Majó-Vázquez, S., 2019. Echo-chambers in online news consumption: Evidence from survey and navigation data in Spain. European Journal of Communication 34, 360–376. https://doi.org/10.1177/0267323119844409

Cardenal, A.S., Galais, C., Majó-Vázquez, S., 2018. Is Facebook Eroding the Public Agenda? Evidence From Survey and Web-Tracking Data. International Journal of Public Opinion Research. https://doi.org/10.1093/ijpor/edy025

Cid, E., 2018. 3 steps to adopt online behavioral data.

Cochran, W., 1953. Sampling Techniques. Willey.

Couper, M.P., Kapteyn, A., Schonlau, M., Winter, J., 2007. Noncoverage and nonresponse in an Internet survey. Social Science Research 36, 131–148. https://doi.org/10.1016/j.ssresearch.2005.10.002

Deming, E., 1950. Some Theory of Sampling. Wiley.

Deming, W.E., 1944. On Errors in Surveys. American Sociological Review 9, 359. https://doi.org/10.2307/2085979

Dvir-Gvirsman, S., Tsfati, Y., Menchen-Trevino, E., 2014. The extent and nature of ideological selective exposure online: Combining survey responses with actual web log data from the 2013 Israeli Elections. New Media & Society 18, 857–877. https://doi.org/10.1177/1461444814549041

Groves, R., 1989. Survey Errors and Survey Costs. Wiley, New York.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., Tourangeau, R., 2009. Survey Methodology, 2nd Edition, Wiley series in survey methodology. Wiley.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., Tourangeau, R., 2004. Survey Methodology. Wiley.

Groves, R.M., Lyberg, L., 2010. Total Survey Error: Past Present, and Future. Public Opinion Quarterly 74, 849–879. https://doi.org/10.1093/poq/nfq065

Guess, A., Nyhan, B., Reifler, J., 2018. Selective Exposure to Misinformation: Evidence from the consumption of fake news during the 2016 U. S. presidential campaign. European Research Council.

Guess, A.M., 2015. Measure for Measure: An Experimental Test of Online Political Media Exposure. Political Analysis 23, 59–75. https://doi.org/10.1093/pan/mpu010

Guess, A.M., Nyhan, B., O'Keeffe, Z., Reifler, J., 2020. The sources and correlates of exposure to vaccine-related (mis)information online. Vaccine 38, 7799–7805. https://doi.org/10.1016/j.vaccine.2020.10.018

Haim, M., Nienierza, A., 2019. Computational observation : Challenges and opportunities of automated observation within algorithmically curated media environments using a browser plug-in. Computational Communication Research. https://doi.org/10.5117/ccr2019.1.004.haim

Harari, G.M., Lane, N.D., Wang, R., Crosier, B.S., Campbell, A.T., Gosling, S.D., 2016. Using Smartphones to Collect Behavioral Data in Psychological Science. Perspectives on Psychological Science 11, 838–854. https://doi.org/10.1177/1745691616650285

Hsieh, Y.P., Murphy, J., 2017. Total Twitter Error, in: Total Survey Error in Practice. John Wiley & Sons Inc., pp. 23–46. https://doi.org/10.1002/9781119041702.ch2

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C., Usher, A., 2015. Big Data in Survey Research. Public Opinion Quarterly 79, 839–880. https://doi.org/10.1093/poq/nfv039

Jürgens, P., Stark, B., Magin, M., 2019. Two Half-Truths Make a Whole? On Bias in Self-Reports and Tracking Data. Social Science Computer Review 38, 600–615. https://doi.org/10.1177/0894439319831643

Kitchin, R., McArdle, G., 2016. What makes Big Data Big Data? Exploring the ontological characteristics of 26 datasets. Big Data & Society 3, 205395171663113. https://doi.org/10.1177/2053951716631130

Lavrakas, P., 2008. Total Survey Error, in: Encyclopedia of Survey Research Methods. Sage Publications. https://doi.org/https://dx.doi.org/10.4135/9781412963947.n585

Lynn, P., Lugtig, P.J., 2017. Total Survey Error for Longitudinal Surveys, in: Total Survey Error in Practice. John Wiley & Sons Inc., pp. 279–298. https://doi.org/10.1002/9781119041702.ch13

Mercer, A.W., Kreuter, F., Keeter, S., Stuart, E.A., 2017. Theory and Practice in Nonprobability Surveys. Public Opinion Quarterly 81, 250–271. https://doi.org/10.1093/poq/nfw060

Niederdeppe, J., 2016. Meeting the Challenge of Measuring Communication Exposure in the Digital Age. Communication Methods and Measures 10, 170–172. https://doi.org/10.1080/19312458.2016.1150970

Ochoa, C., Bort, C., Porcar, M., 2017. "Who is Who" with Behavioural Data: Metering data edge over survey data, in: ESOMAR BIG DATA WORLD 2017.

Ochoa, C., Paura, E., 2018. The Value of Machine Learning in Privacy: Results-oriented machine learning solution in securing PII data anonymisation, in: ESOMAR FUSION 2018: BIG DATA WORLD.

Ochoa, C., Porcar, J.M., 2018. Modeling the effect of quota sampling on online fieldwork efficiency: An analysis of the connection between uncertainty and sample usage. International Journal of Market Research 60, 484–501. https://doi.org/10.1177/1470785318779545

Peterson, E., Damm, E., 2019. A Window to the World: Americans Exposure to Political News From Foreign Media Outlets. https://doi.org/10.31235/osf.io/er48b

Peterson, E., Goel, S., Iyengar, S., 2018. Echo Chambers and Partisan Polarization: Evidence from the 2016 Presidential Campaign.

Pew Research Center, 2016. Evaluating Online Nonprobability Surveys.

Pew Research Center, 2020. Measuring News Consumption in a Digital Era.

Revilla, M., Couper, M.P., Ochoa, C., 2019. Willingness of online panelists to perform additional tasks. Methods, Data, Analyses. https://doi.org/10.12758/mda.2018.01

Revilla, M., Couper, M.P., Paura, E., Ochoa, C., 2021. Willingness to Participate in a Metered Online Panel. Field Methods 1525822X2098398. 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, 521–536. https://doi.org/10.1177/0894439316638457

Sen, I., Flock, F., Weller, K., Weiss, B., Wagner, C., 2019. A Total Error Framework for Digital Traces of Humans.

Toth, R., Trifonova, T., 2020. Somebody's Watching Me: Smartphone Use Tracking and Reactivity. https://doi.org/10.31235/osf.io/7aqdx

Unangst, J., Amaya, A.E., Sanders, H.L., Howard, J., Ferrell, A., Karon, S., Dever, J.A., 2019. A Process for Decomposing Total Survey Error in Probability and Nonprobability Surveys: A Case Study Comparing Health Statistics in US Internet Panels. Journal of Survey Statistics and Methodology 8, 62–88. https://doi.org/10.1093/jssam/smz040

de Vreese, C.H., Neijens, P., 2016. Measuring Media Exposure in a Changing Communications Environment. Communication Methods and Measures 10, 69–80. https://doi.org/10.1080/19312458.2016.1150441