When survey science met online tracking:

Presenting an error framework for metered data

Oriol J. Bosch | THE LONDON SCHOOL OF ECONOMICS / RECSM-UPF

Melanie Revilla | RECSM-UPF

o.bosch-jover@lse.ac.uk

orioljbosch

https://orioljbosch.com/

upf. Universitat Pompeu Fabra *Barcelona*

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

RECSM Research and Expertise Centre for Survey Methodology

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No849165), PI: Melanie Revilla

When survey science met online tracking:

Presenting an error framework for metered data

+

Track me but not really:

Device undercoverage and its consequences when tracking online behaviours

Oriol J. Bosch | THE LONDON SCHOOL OF ECONOMICS / RECSM-UPF

Melanie Revilla | RECSM-UPF

o.bosch-jover@lse.ac.uk

orioljbosch

https://orioljbosch.com/

Universitat Pompeu Fabra Barcelona

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

RECSM Research and Expertise Centre for Survey Methodology

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No849165), PI: Melanie Revilla

INTRODUCTION

Tracking online behaviours using a meter

Definition

Metered data is obtained from a meter willingly installed or configured by a sample of participants on their devices (PCs, tablets and/or smartphones).

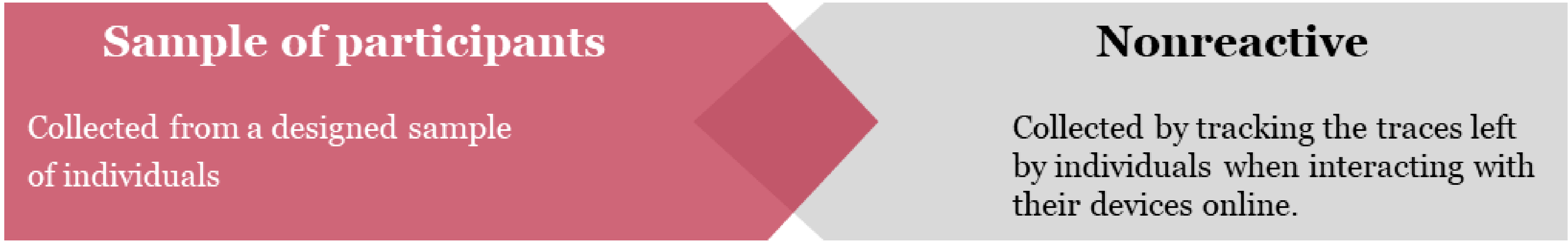
A ***meter*** refers to a heterogeneous group of tracking technologies that allow sharing with the researchers, at least, ***information about the URLs of the web pages visited by the participants***.

Tracking online behaviours using a meter

Definition

Metered 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***.



Sample of participants

Collected from a designed sample of individuals

Nonreactive

Collected by tracking the traces left by individuals when interacting with their devices online.

Tracking online behaviours using a meter

Benefits of metered data

- Objective and free of recall errors
- Continuously collected in real time
- Pre-designed sample of participants

BACKGROUND

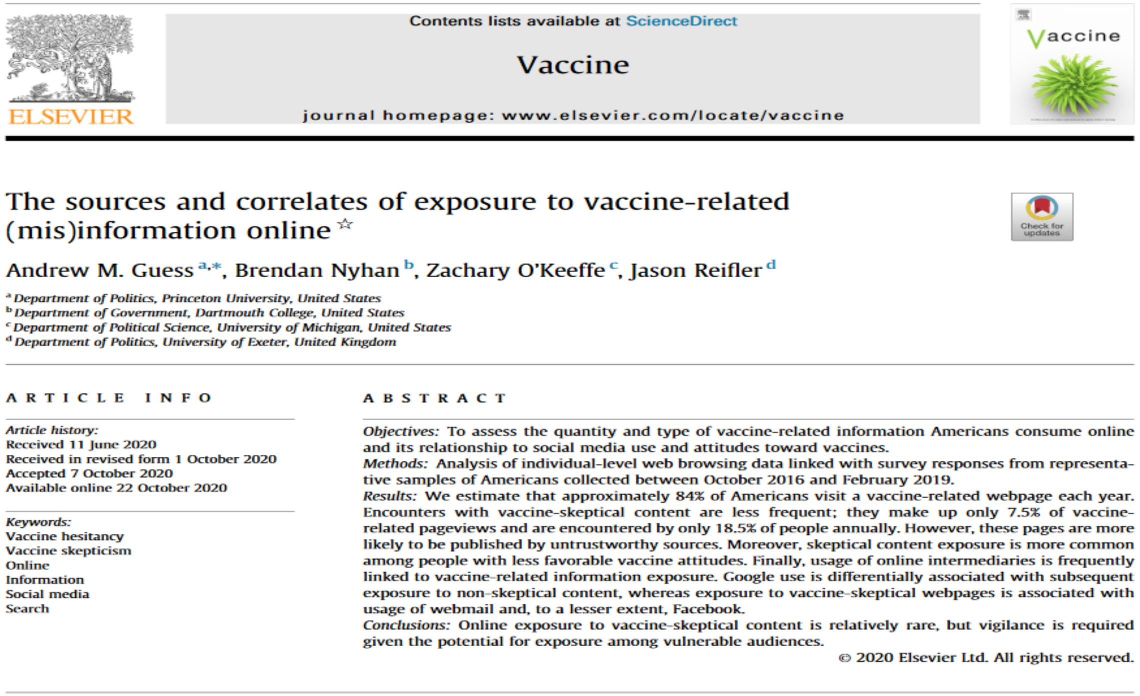
Metered data in past research

33 papers identified, 26 since 2019

BACKGROUND

Metered data in past research

33 papers identified, 26 since 2019



Contents lists available at ScienceDirect

Vaccine

journal homepage: www.elsevier.com/locate/vaccine

The sources and correlates of exposure to vaccine-related (mis)information online ☆

Andrew M. Guess[a,*], Brendan Nyhan[b], Zachary O'Keeffe[c], Jason Reifler[d]

[a] *Department of Politics, Princeton University, United States*
[b] *Department of Government, Dartmouth College, United States*
[c] *Department of Political Science, University of Michigan, United States*
[d] *Department of Politics, University of Exeter, United Kingdom*

ARTICLE INFO

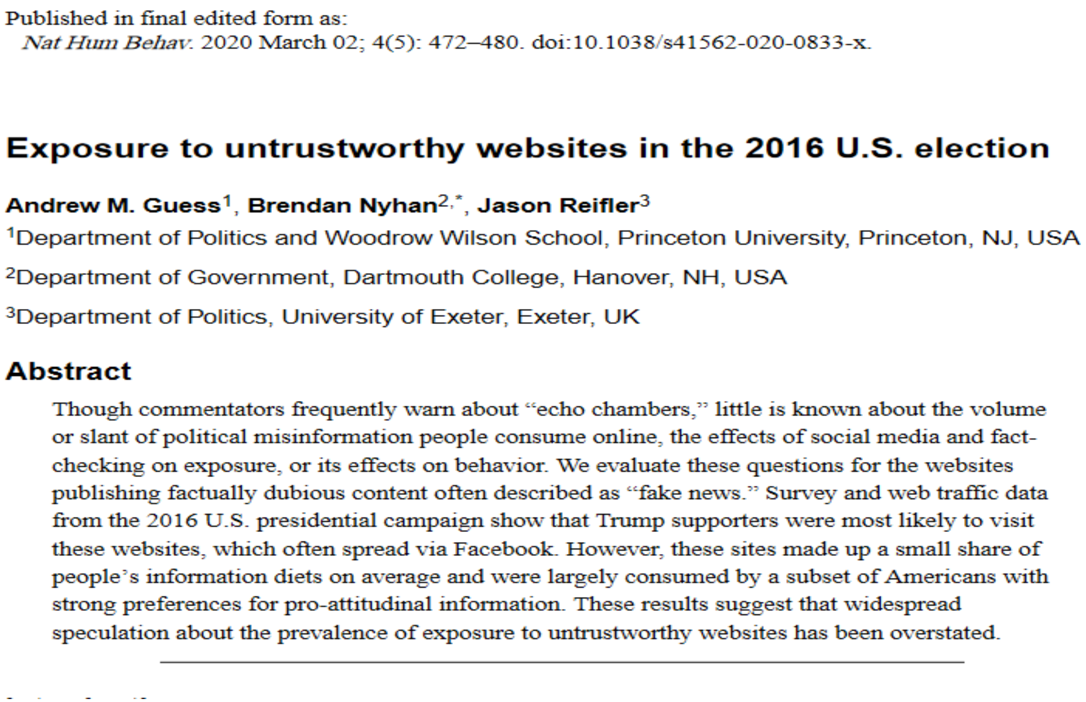
Article history:
Received 11 June 2020
Received in revised form 1 October 2020
Accepted 7 October 2020
Available online 22 October 2020

Keywords:
Vaccine hesitancy
Vaccine skepticism
Online
Information
Social media
Search

ABSTRACT

Objectives: To assess the quantity and type of vaccine-related information Americans consume online and its relationship to social media use and attitudes toward vaccines.
Methods: Analysis of individual-level web browsing data linked with survey responses from representative samples of Americans collected between October 2016 and February 2019.
Results: We estimate that approximately 84% of Americans visit a vaccine-related webpage each year. Encounters with vaccine-skeptical content are less frequent; they make up only 7.5% of vaccine-related pageviews and are encountered by only 18.5% of people annually. However, these pages are more likely to be published by untrustworthy sources. Moreover, skeptical content exposure is more common among people with less favorable vaccine attitudes. Finally, usage of online intermediaries is frequently linked to vaccine-related information exposure. Google use is differentially associated with subsequent exposure to non-skeptical content, whereas exposure to vaccine-skeptical webpages is associated with usage of webmail and, to a lesser extent, Facebook.
Conclusions: Online exposure to vaccine-skeptical content is relatively rare, but vigilance is required given the potential for exposure among vulnerable audiences.

© 2020 Elsevier Ltd. All rights reserved.



Published in final edited form as:
Nat Hum Behav. 2020 March 02; 4(5): 472–480. doi:10.1038/s41562-020-0833-x.

Exposure to untrustworthy websites in the 2016 U.S. election

Andrew M. Guess[1], Brendan Nyhan[2,*], Jason Reifler[3]

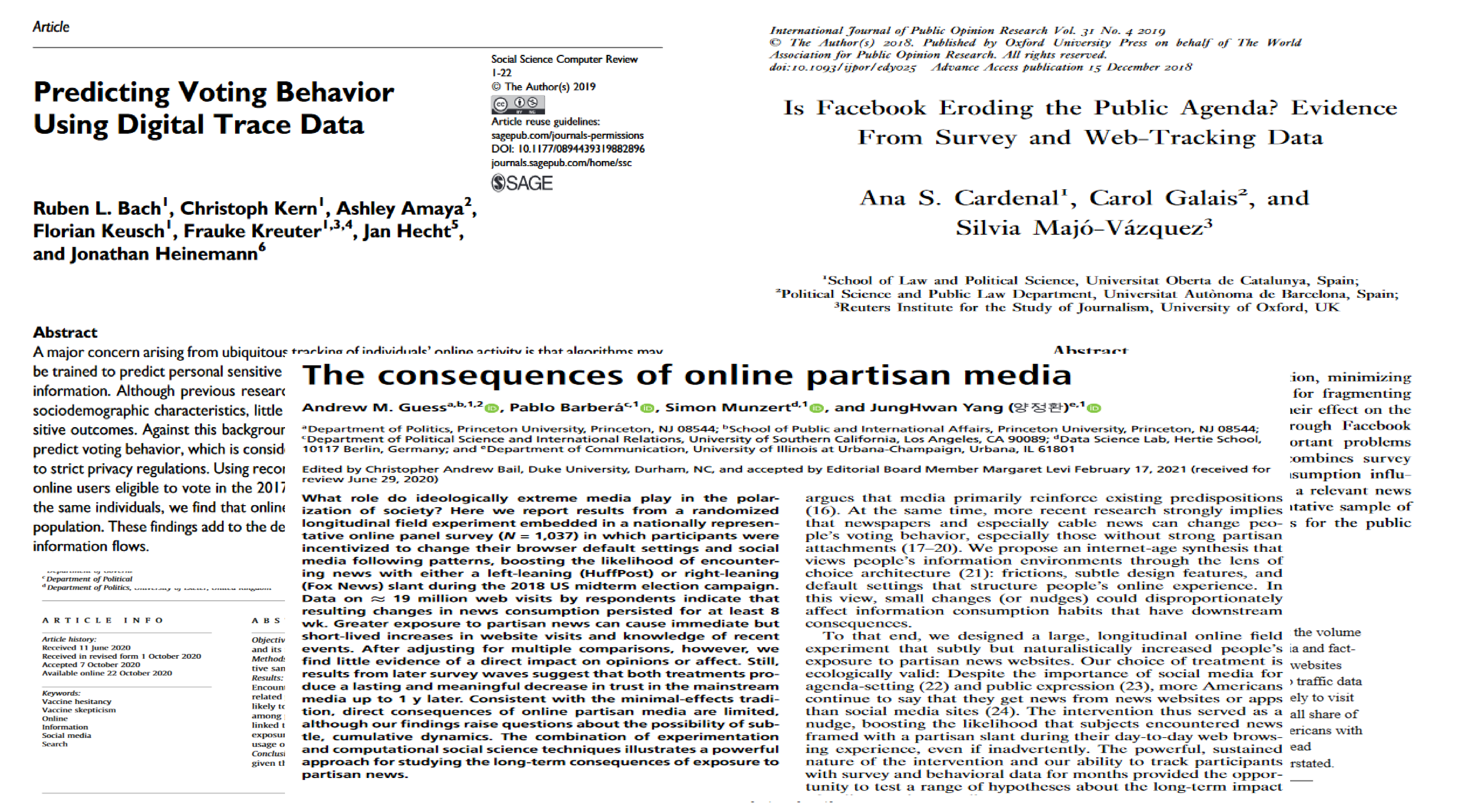
[1]Department of Politics and Woodrow Wilson School, Princeton University, Princeton, NJ, USA

[2]Department of Government, Dartmouth College, Hanover, NH, USA

[3]Department of Politics, University of Exeter, Exeter, UK

Abstract

Though commentators frequently warn about "echo chambers," little is known about the volume or slant of political misinformation people consume online, the effects of social media and fact-checking on exposure, or its effects on behavior. We evaluate these questions for the websites publishing factually dubious content often described as "fake news." Survey and web traffic data from the 2016 U.S. presidential campaign show that Trump supporters were most likely to visit these websites, which often spread via Facebook. However, these sites made up a small share of people's information diets on average and were largely consumed by a subset of Americans with strong preferences for pro-attitudinal information. These results suggest that widespread speculation about the prevalence of exposure to untrustworthy websites has been overstated.



Article

Social Science Computer Review
1-22
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0894439319882896
journals.sagepub.com/home/ssc
SAGE

Predicting Voting Behavior Using Digital Trace Data

Ruben L. Bach[1], Christoph Kern[1], Ashley Amaya[2], Florian Keusch[1], Frauke Kreuter[1,3,4], Jan Hecht[5], and Jonathan Heinemann[6]

Abstract

A major concern arising from ubiquitous tracking of individuals' online activity is that algorithms may
be trained to predict personal sensitive
information. Although previous researc
sociodemographic characteristics, little
sitive outcomes. Against this backgrour
predict voting behavior, which is consid
to strict privacy regulations. Using recor
online users eligible to vote in the 2017
the same individuals, we find that online
population. These findings add to the de
information flows.

International Journal of Public Opinion Research Vol. 31 No. 4 2019
© The Author(s) 2018. Published by Oxford University Press on behalf of The World Association for Public Opinion Research. All rights reserved.
doi:10.1093/ijpor/edy025 Advance Access publication 15 December 2018

Is Facebook Eroding the Public Agenda? Evidence From Survey and Web-Tracking Data

Ana S. Cardenal[1], Carol Galais[2], and Silvia Majó-Vázquez[3]

[1]School of Law and Political Science, Universitat Oberta de Catalunya, Spain;
[2]Political Science and Public Law Department, Universitat Autònoma de Barcelona, Spain;
[3]Reuters Institute for the Study of Journalism, University of Oxford, UK

Abstract

ion, minimizing
for fragmenting
neir effect on the
rough Facebook
ortant problems
combines survey
nsumption influ-
a relevant news
ntative sample of
s for the public

the volume
ia and fact-
websites
traffic data
ely to visit
all share of
ericans with
ead
rstated.

The consequences of online partisan media

Andrew M. Guess[a,b,1,2], Pablo Barberá[c,1], Simon Munzert[d,1], and JungHwan Yang (양정환)[e,1]

[a]Department of Politics, Princeton University, Princeton, NJ 08544; [b]School of Public and International Affairs, Princeton University, Princeton, NJ 08544; [c]Department of Political Science and International Relations, University of Southern California, Los Angeles, CA 90089; [d]Data Science Lab, Hertie School, 10117 Berlin, Germany; and [e]Department of Communication, University of Illinois at Urbana-Champaign, Urbana, IL 61801

Edited by Christopher Andrew Bail, Duke University, Durham, NC, and accepted by Editorial Board Member Margaret Levi February 17, 2021 (received for review June 29, 2020)

What role do ideologically extreme media play in the polarization of society? Here we report results from a randomized longitudinal field experiment embedded in a nationally representative online panel survey ($N = 1{,}037$) in which participants were incentivized to change their browser default settings and social media following patterns, boosting the likelihood of encountering news with either a left-leaning (HuffPost) or right-leaning (Fox News) slant during the 2018 US midterm election campaign. Data on ≈ 19 million web visits by respondents indicate that resulting changes in news consumption persisted for at least 8 wk. Greater exposure to partisan news can cause immediate but short-lived increases in website visits and knowledge of recent events. After adjusting for multiple comparisons, however, we find little evidence of a direct impact on opinions or affect. Still, results from later survey waves suggest that both treatments produce a lasting and meaningful decrease in trust in the mainstream media up to 1 y later. Consistent with the minimal-effects tradition, direct consequences of online partisan media are limited, although our findings raise questions about the possibility of subtle, cumulative dynamics. The combination of experimentation and computational social science techniques illustrates a powerful approach for studying the long-term consequences of exposure to partisan news.

argues that media primarily reinforce existing predispositions (16). At the same time, more recent research strongly implies that newspapers and especially cable news can change people's voting behavior, especially those without strong partisan attachments (17–20). We propose an internet-age synthesis that views people's information environments through the lens of choice architecture (21): frictions, subtle design features, and default settings that structure people's online experience. In this view, small changes (or nudges) could disproportionately affect information consumption habits that have downstream consequences.

To that end, we designed a large, longitudinal online field experiment that subtly but naturalistically increased people's exposure to partisan news websites. Our choice of treatment is ecologically valid: Despite the importance of social media for agenda-setting (22) and public expression (23), more Americans continue to say that they get news from news websites or apps than social media sites (24). The intervention thus served as a nudge, boosting the likelihood that subjects encountered news framed with a partisan slant during their day-to-day web browsing experience, even if inadvertently. The powerful, sustained nature of the intervention and our ability to track participants with survey and behavioral data for months provided the opportunity to test a range of hypotheses about the long-term impact

[c] Department of Political
[d] Department of Politics,

ARTICLE INFO

Article history:
Received 11 June 2020
Received in revised form 1 October 2020
Accepted 7 October 2020
Available online 22 October 2020

Keywords:
Vaccine hesitancy
Vaccine skepticism
Online
Information
Social media
Search

ABS

Objectiv
and its
Method:
tive san
Results:
Encoun
related
likely to
among
linked t
exposur
usage o
Conclusi
given th

COMMUNICATION METHODS AND MEASURES
2017, VOL. 11, NO. 3, 173–190
http://dx.doi.org/10.1080/19312458.2017.1317337

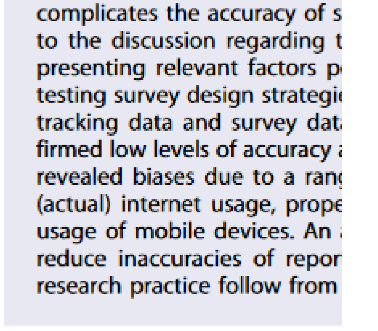
Routledge
Taylor & Francis Group

OPEN ACCESS

How Much Time Do You Spend Online? Understanding and Improving the Accuracy of Self-Reported Measures of Internet Use

Theo Araujo, Anke Wonneberger, Peter Neijens, and Claes de Vreese

Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Amsterdam, The Netherlands



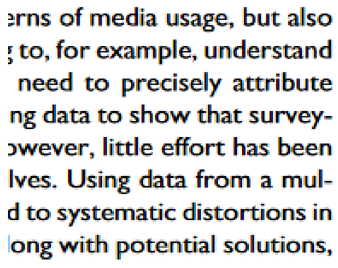
ABSTRACT
Given the importance of survey measures of online media use for communication research, it is crucial to assess and improve their quality, in particular because the increasingly fragmented and ubiquitous usage of internet
complicates the accuracy of s
to the discussion regarding t
presenting relevant factors p
testing survey design strategi
tracking data and survey dat
firmed low levels of accuracy a
revealed biases due to a rang
(actual) internet usage, prope
usage of mobile devices. An a
reduce inaccuracies of repor
research practice follow from

Article

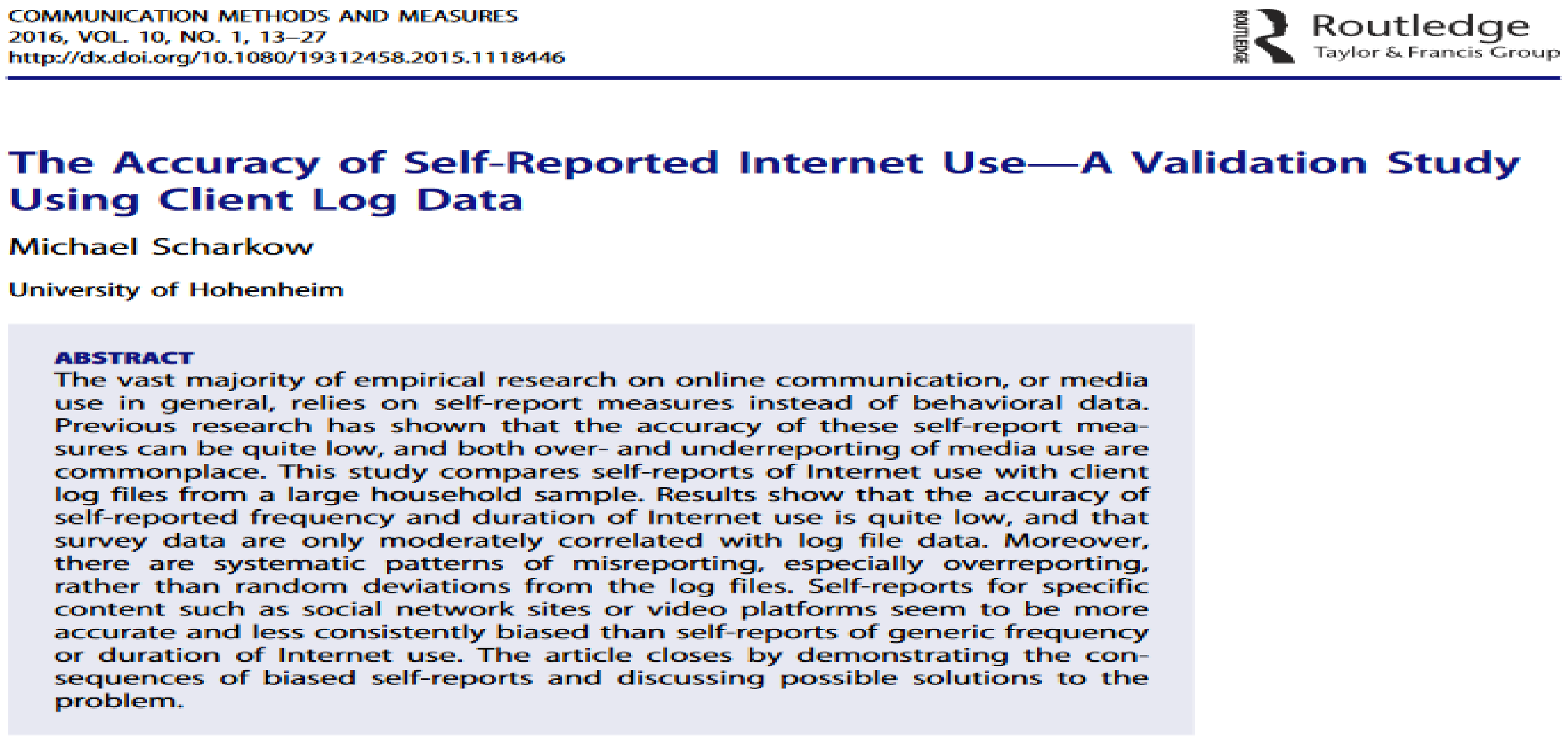
Two Half-Truths Make a Whole? On Bias in Self-Reports and Tracking Data

Social Science Computer Review
2020, Vol. 38(5) 600-615
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0894439319831643
journals.sagepub.com/home/ssc
SAGE

Pascal Jürgens[1], Birgit Stark[1], and Melanie Magin[2]



erns of media usage, but also
; to, for example, understand
need to precisely attribute
ng data to show that survey-
owever, little effort has been
lves. Using data from a mul-
d to systematic distortions in
long with potential solutions,



COMMUNICATION METHODS AND MEASURES
2016, VOL. 10, NO. 1, 13–27
http://dx.doi.org/10.1080/19312458.2015.1118446

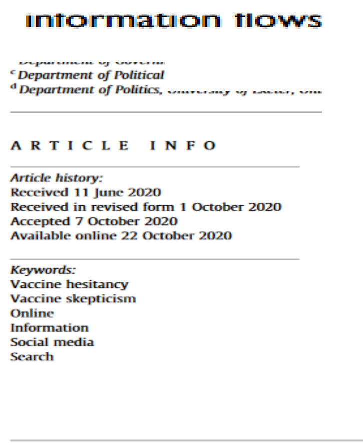
Routledge
Taylor & Francis Group

The Accuracy of Self-Reported Internet Use—A Validation Study Using Client Log Data

Michael Scharkow

University of Hohenheim

ABSTRACT
The vast majority of empirical research on online communication, or media use in general, relies on self-report measures instead of behavioral data. Previous research has shown that the accuracy of these self-report measures can be quite low, and both over- and underreporting of media use are commonplace. This study compares self-reports of Internet use with client log files from a large household sample. Results show that the accuracy of self-reported frequency and duration of Internet use is quite low, and that survey data are only moderately correlated with log file data. Moreover, there are systematic patterns of misreporting, especially overreporting, rather than random deviations from the log files. Self-reports for specific content such as social network sites or video platforms seem to be more accurate and less consistently biased than self-reports of generic frequency or duration of Internet use. The article closes by demonstrating the consequences of biased self-reports and discussing possible solutions to the problem.



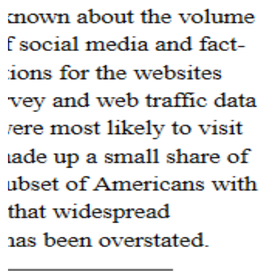
information flows

[c] *Department of Political*
[d] *Department of Politics,*

A R T I C L E I N F O

Article history:
Received 11 June 2020
Received in revised form 1 October 2020
Accepted 7 October 2020
Available online 22 October 2020

Keywords:
Vaccine hesitancy
Vaccine skepticism
Online
Information
Social media
Search



known about the volume
f social media and fact-
ions for the websites
vey and web traffic data
vere most likely to visit
nade up a small share of
ubset of Americans with
that widespread
has been overstated.

Inferences for finite populations

Metered data can potentially suffer from different types of errors

Shared devices and observation of only part of the activity

- 60% of desktops, 40% of laptops and tablets, and 9% of smartphones shared to some degree(Revilla et al., 2017)
- 28% with the meter installed in all devices (Pew Research Center, 2020)

Technical issues and reactivity / social desirability bias (Jurgens et al., 2020; Toth and Trifonova, 2020)

Substantive conclusions vary depending on what is considered as a visit (3 seconds / 30 seconds / 120 seconds) (Mangold et al., 2021)

BACKGROUND

Inferences for finite populations

Metered data can potentially suffer from different types of errors

Shared devices and observation of only part of the activity

- 60% of desktops, 40% of laptops and tablets, and 9% of smartphones shared to some degree(Revilla et al., 2017)
- 28% with the meter installed in all devices (Pew Research Center, 2020)

Technical issues and reactivity / social desirability bias (Jurgens et al., 2020; Toth and Trifonova, 2020)

Substantive conclusions vary depending on what is considered as a visit (3 seconds / 30 seconds / 120 seconds) (Mangold et al., 2021)

BACKGROUND

Inferences for finite populations

Metered data can potentially suffer from different types of errors

Shared devices and observation of only part of the activity

- 60% of desktops, 40% es shared to some degree(Revilla et al.,
- 28% with the meter i 2020)

Technical issues and reacti 020; Toth and Trifonova, 2020)

Substantive conclusions vary depending on what is considered as a visit (3 seconds / 30 seconds / 120 seconds) (Mangold et al., 2021)

A systematic **categorization** and **conceptualization** of metered data errors

Not available

BACKGROUND

Inferences for finite populations

Metered data can potentially suffer from different types of errors

Shared devices and observation of only part of the activity

- 60% of desktops, 40 degree(Revilla et al.,
- 28% with the meter i 2020)

Technical issues and reacti 20; Toth and Trifonova, 2020)

Substantive conclusions vary depending on what is considered as a visit (3 seconds / 30 seconds / 120 seconds) (Mangold et al., 2021)

A systematic **categorization** and **conceptualization** of metered data errors

Not available

Nor empirical demonstrations of (many of) those errors!

FIRST PAPER

Main goals and contribution

Total Error Framework for metered data

- #1 **Summarize** the data collection and analysis process for metered data.

- #2 **Conceptualize and categorize** all errors components (e.g. measurement errors) and causes (e.g. social desirability) that can occur when using metered data.

FIRST PAPER

Main goals and contribution

Total Error Framework for metered data

- #1 **Summarize** the data collection and analysis process for metered data.
- #2 **Conceptualize and categorize** all errors components (e.g. measurement errors) and causes (e.g. social desirability) that can occur when using metered data.

1) Choose the best design options for metered data.

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

3) Help assess research using metered data.

FIRST PAPER

Main goals and contribution

Total Error Framework for metered data

- #1 **Summarize** the data collection and analysis process for metered data.
- #2 **Conceptualize and categorize** all errors components (e.g. measurement errors) and causes (e.g. social desirability) that can occur when using metered data.

Bosch, O.J., and M. Revilla (2021). **“When survey science met online tracking: presenting an error framework for metered data.”** RECSM Working Papers Series, 62

FIRST PAPER

Approach

Adapting instead of reinventing

- Follow approach by Amaya et al (2020) with their **Total Error Framework for Big Data**

- 7 error components of the TSE (Groves et al., 2009) as starting point:
  - Coverage errors, sampling errors, *missing data errors*, adjustment errors, *specification errors*, measurement errors and processing errors

RESULTS

Data collection and analysis process

RESULTS

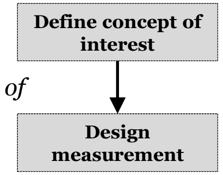
Data collection and analysis process

Define concept of interest

Average hours of consumption of online political news

RESULTS

Data collection and analysis process



Define concept of interest

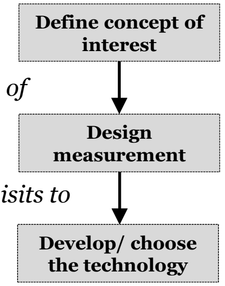
Average hours of consumption of online political news

Design measurement

Average time recorded of the visits to online political outlets' URLs.

RESULTS

Data collection and analysis process



Define concept of interest

Design measurement

Develop/ choose the technology

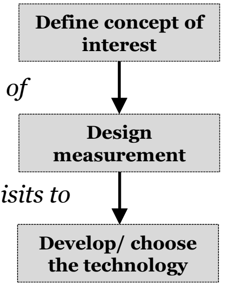
Average hours of consumption of online political news

Average time recorded of the visits to online political outlets' URLs.

Proxy for IOS/ App for others

RESULTS

Data collection and analysis process


Define concept of interest

Design measurement

Develop/ choose the technology

Average hours of consumption of online political news

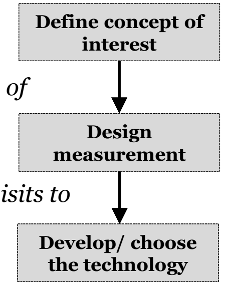
Average time recorded of the visits to online political outlets' URLs.

Proxy for IOS/ App for others

Define target inferential population

People living in the UK older than 18

Data collection and analysis process



Define concept of interest

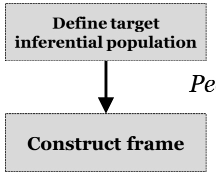
Average hours of consumption of online political news

Design measurement

Average time recorded of the visits to online political outlets' URLs.

Develop/ choose the technology

Proxy for IOS/ App for others



Define target inferential population

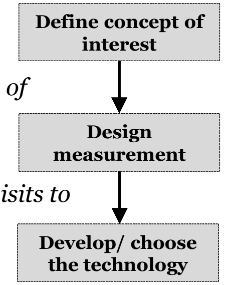
People living in the UK older than 18

Construct frame

Postal Address Frame

RESULTS

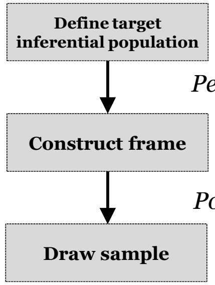
Data collection and analysis process



Average hours of consumption of online political news

Average time recorded of the visits to online political outlets' URLs.

Proxy for IOS/ App for others



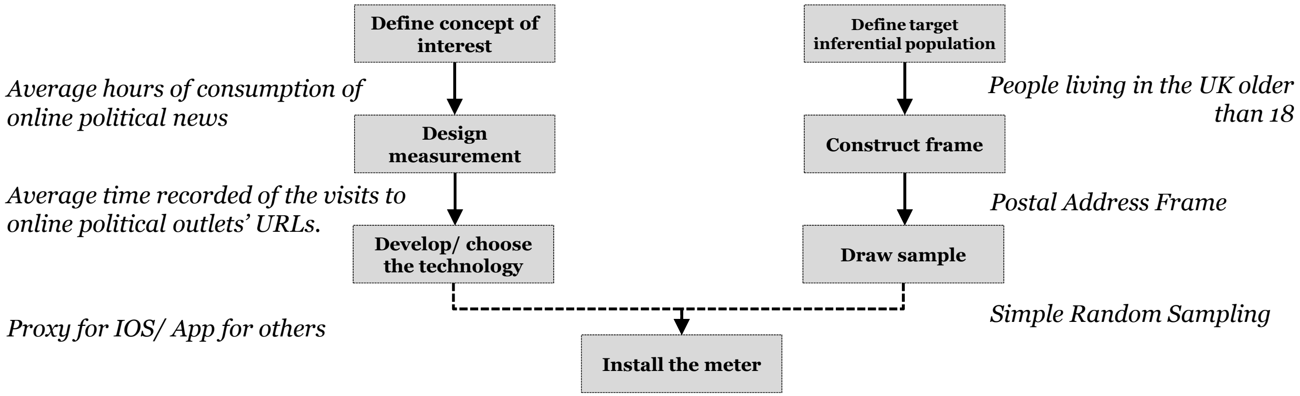
People living in the UK older than 18

Postal Address Frame

Simple Random Sampling

RESULTS

Data collection and analysis process



Define concept of interest

Average hours of consumption of online political news

Design measurement

Average time recorded of the visits to online political outlets' URLs.

Develop/ choose the technology

Proxy for IOS/ App for others

Define target inferential population

People living in the UK older than 18

Construct frame

Postal Address Frame

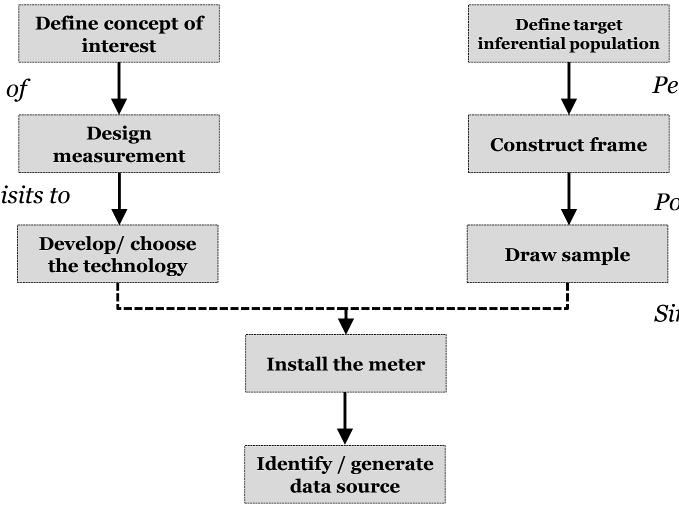
Draw sample

Simple Random Sampling

Install the meter

RESULTS

Data collection and analysis process


Define concept of interest
Design measurement
Develop/ choose the technology
Define target inferential population
Construct frame
Draw sample
Install the meter
Identify / generate data source

Average hours of consumption of online political news

Average time recorded of the visits to online political outlets' URLs.

Proxy for IOS/ App for others

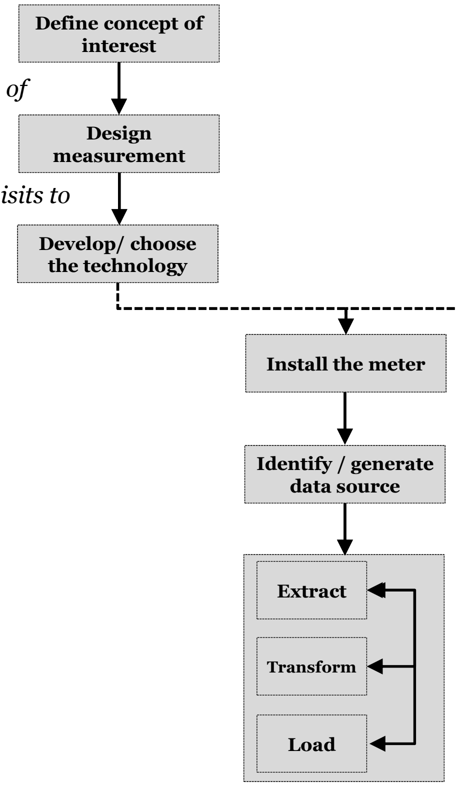
People living in the UK older than 18

Postal Address Frame

Simple Random Sampling

RESULTS

Data collection and analysis process


Define concept of interest
Design measurement
Develop/ choose the technology
Define target inferential population
Construct frame
Draw sample
Install the meter
Identify / generate data source
Extract
Transform
Load

Average hours of consumption of online political news

Average time recorded of the visits to online political outlets' URLs.

Proxy for IOS/ App for others

People living in the UK older than 18

Postal Address Frame

Simple Random Sampling

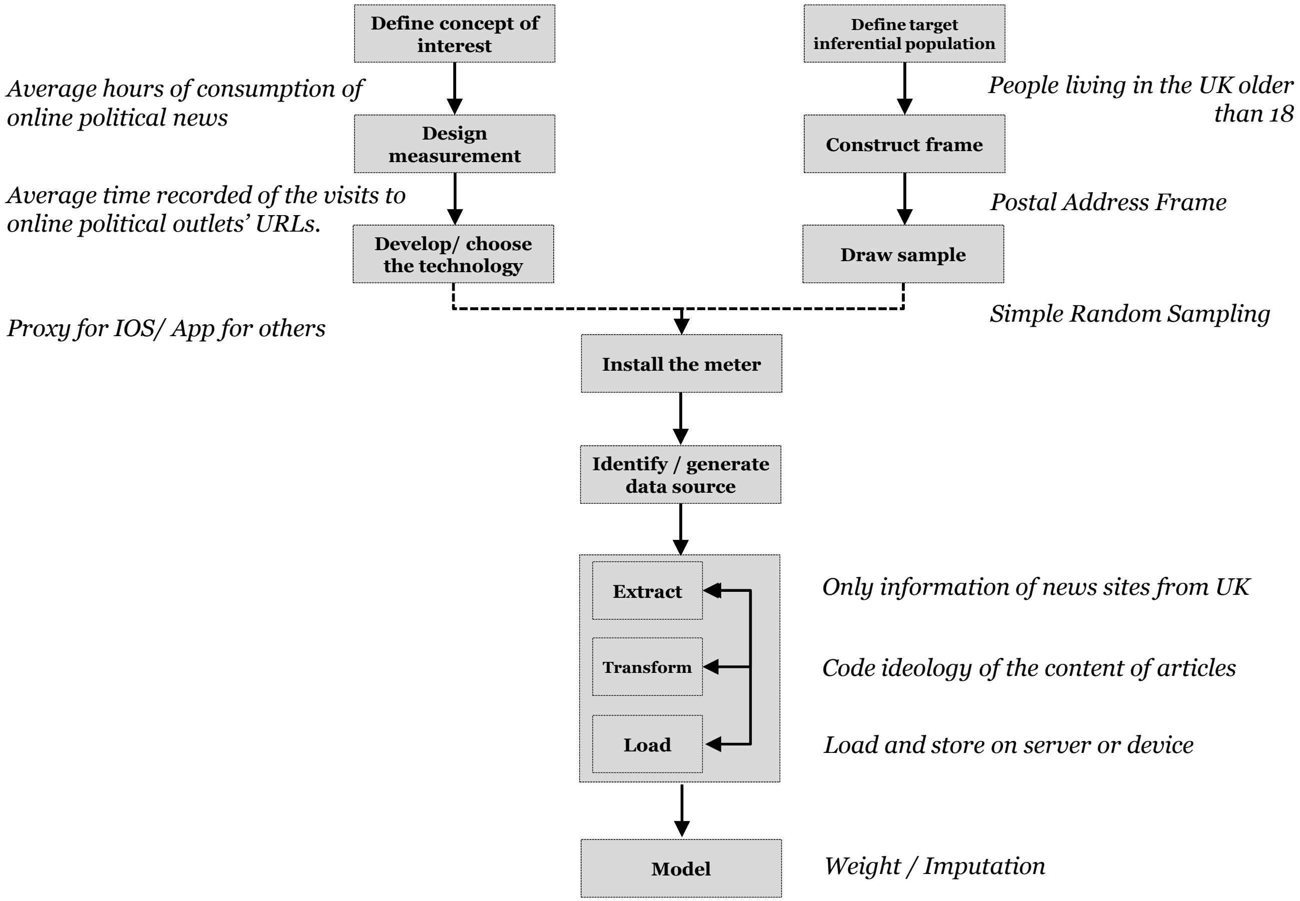
Only information of news sites from UK

Code ideology of the content of articles

Load and store on server or device

RESULTS

Data collection and analysis process



Define concept of interest

Average hours of consumption of online political news

Design measurement

Average time recorded of the visits to online political outlets' URLs.

Develop/ choose the technology

Proxy for IOS/ App for others

Define target inferential population

People living in the UK older than 18

Construct frame

Postal Address Frame

Draw sample

Simple Random Sampling

Install the meter

Identify / generate data source

Extract

Only information of news sites from UK

Transform

Code ideology of the content of articles

Load

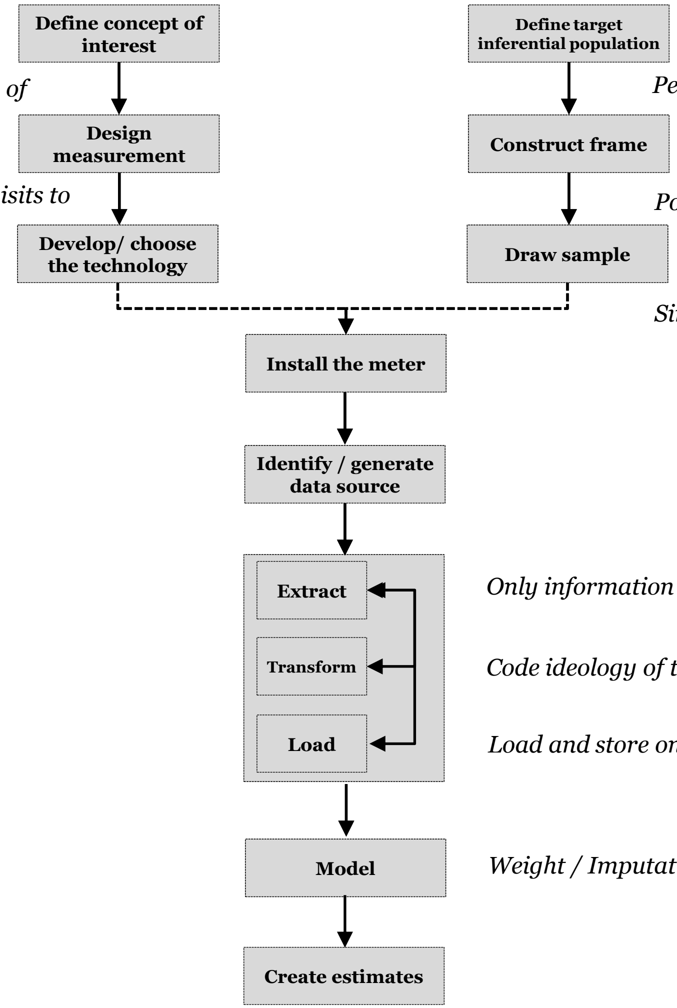
Load and store on server or device

Model

Weight / Imputation

RESULTS

Data collection and analysis process



Define concept of interest

Average hours of consumption of online political news

Design measurement

Average time recorded of the visits to online political outlets' URLs.

Develop/ choose the technology

Proxy for IOS/ App for others

Define target inferential population

People living in the UK older than 18

Construct frame

Postal Address Frame

Draw sample

Simple Random Sampling

Install the meter

Identify / generate data source

Extract

Only information of news sites from UK

Transform

Code ideology of the content of articles

Load

Load and store on server or device

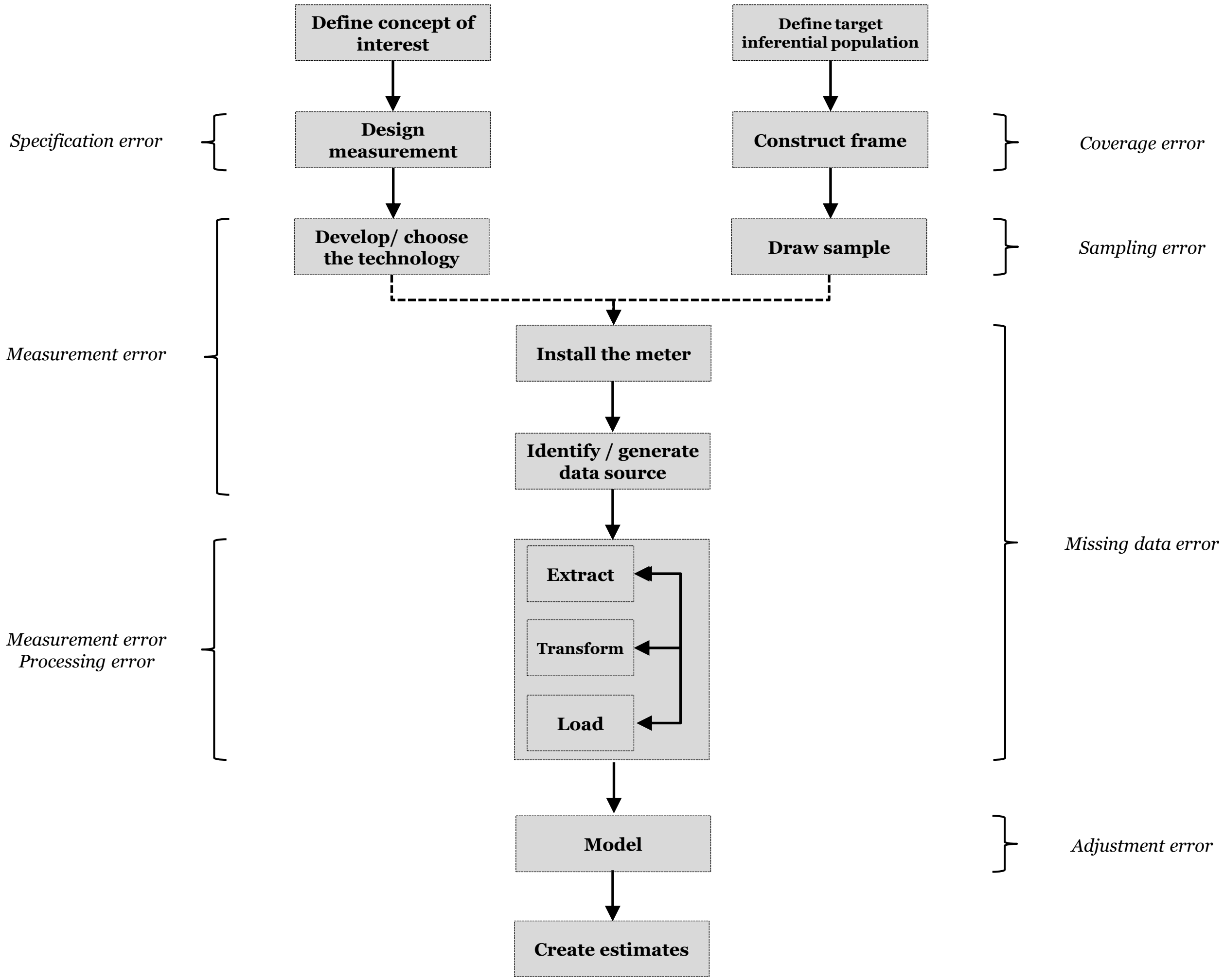
Model

Weight / Imputation

Create estimates

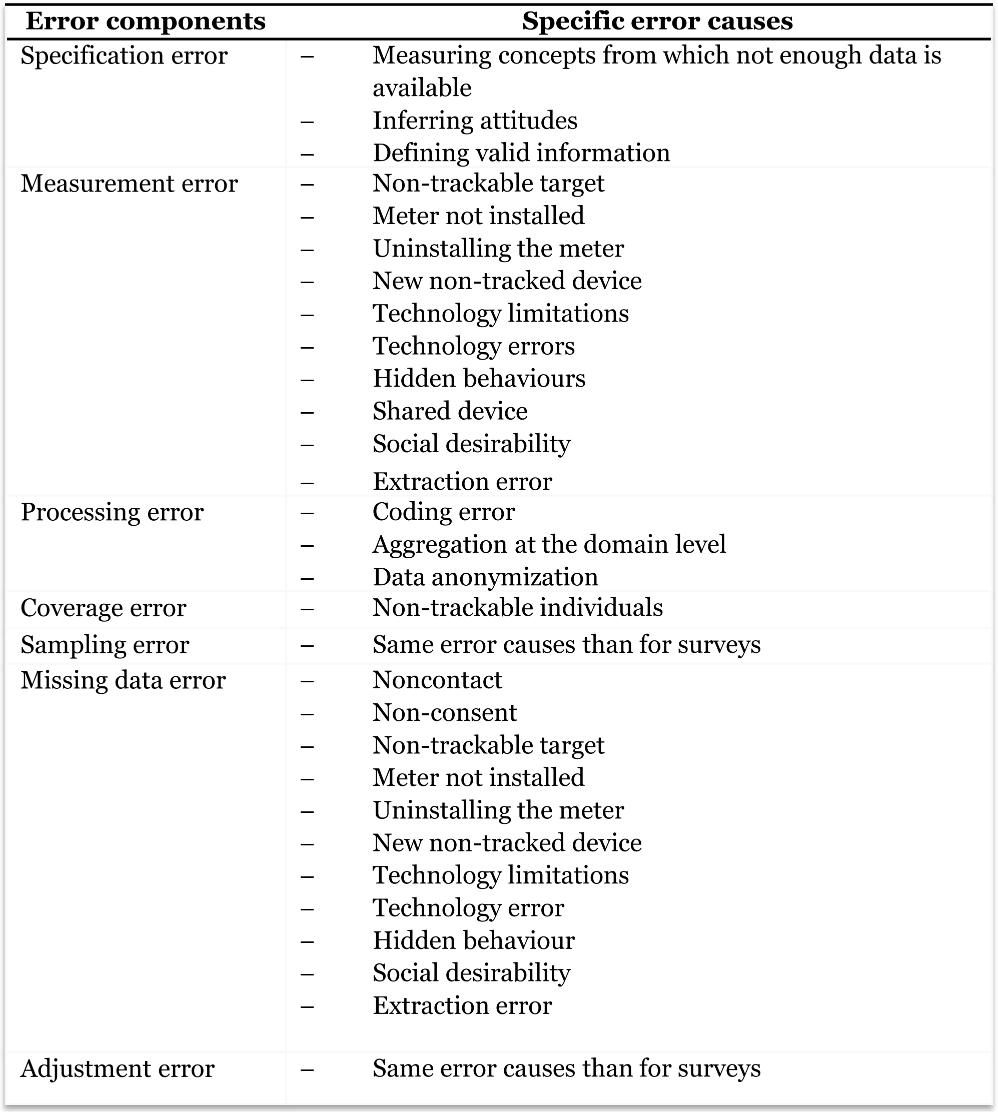
RESULTS

Data collection and analysis process



RESULTS

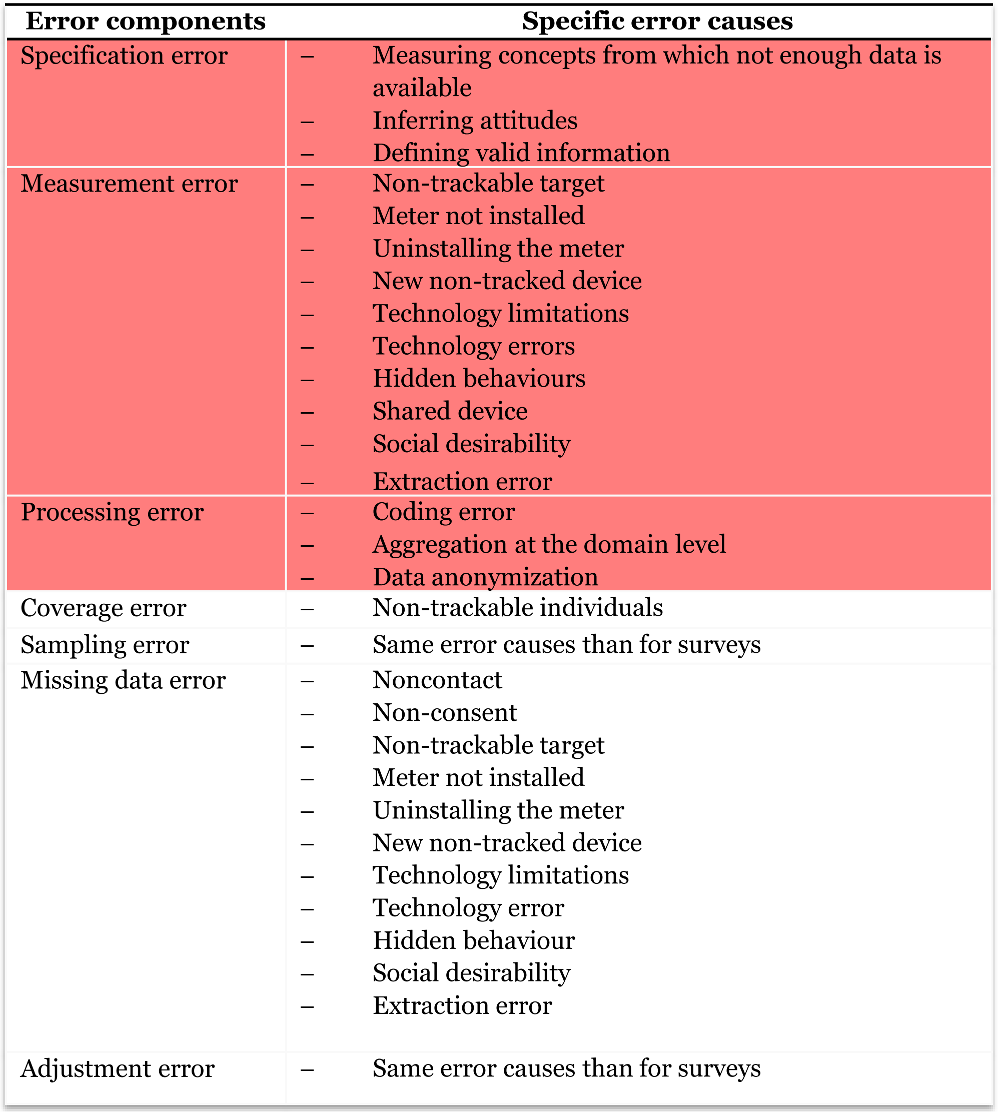
Error components and their causes



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

RESULTS

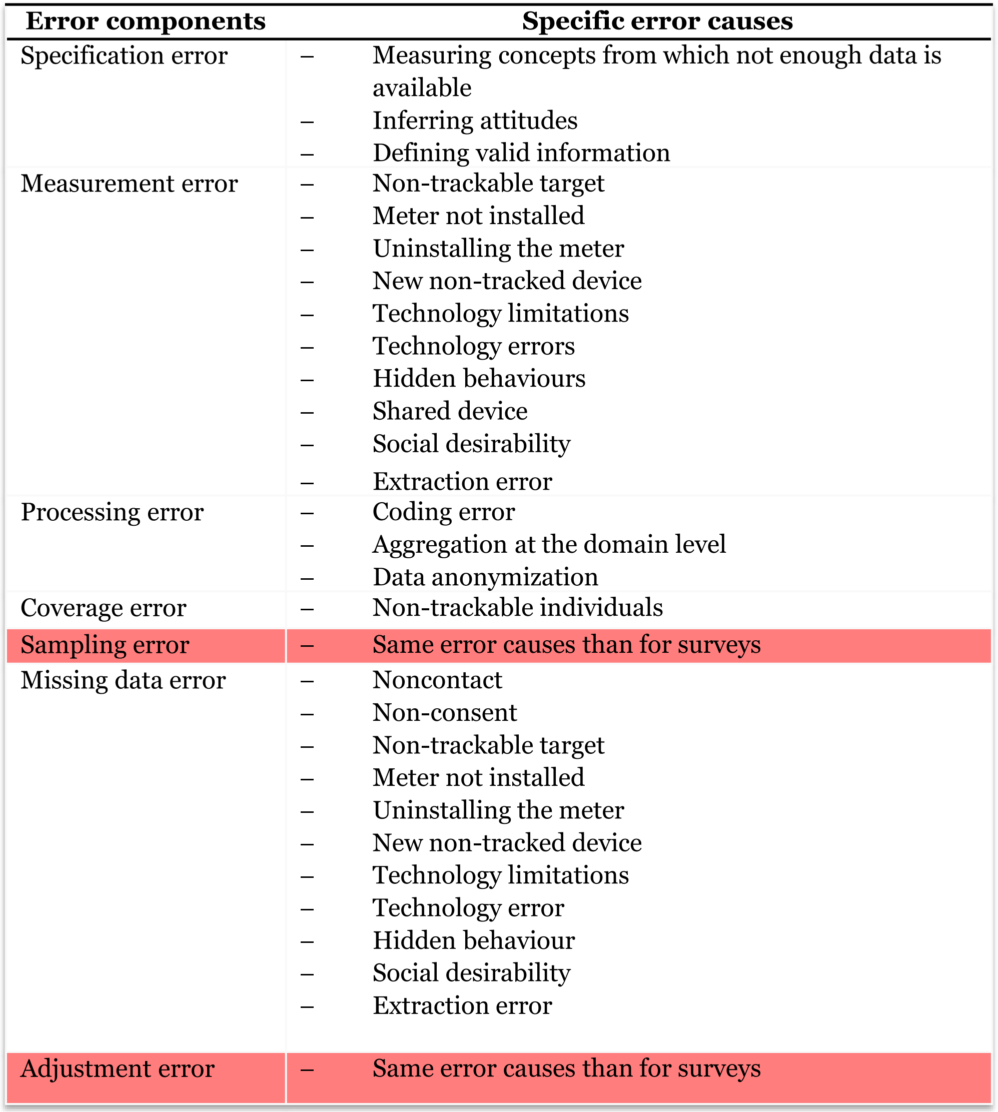
Error components and their causes



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

Most specific error causes on the side of measurement

Error components and their causes



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

Sampling and adjustment errors have no specific error causes

RESULTS

Practical recommendations

1. **Clearly define what your tracked data is measuring beforehand**

RESULTS

Practical recommendations

1. **Clearly define what your tracked data is measuring beforehand**

Concept: *average hours of consumption of online political news*
Measure: *average time recorded of the visits to online political outlets' URLs.*

Practical recommendations

1. **Clearly define what your tracked data is measuring beforehand**

Concept: *average hours of consumption of online political news*
Measure: *average time recorded of the visits to online political outlets' URLs.*

- What is considered a visit?

RESULTS

Practical recommendations

1. **Clearly define what your tracked data is measuring beforehand**

Concept: *average hours of consumption of online political news*
Measure: *average time recorded of the visits to online political outlets' URLs.*

- What is considered a visit?
- Which online outlets?

RESULTS

Practical recommendations

1. **Clearly define what your tracked data is measuring beforehand**

Concept: *average hours of consumption of online political news*

Measure: *average time recorded of the visits to online political outlets' URLs.*

- What is considered a visit?
- Which online outlets?
- Which URLs should be considered political?

Practical recommendations

1. **Clearly define what your tracked data is measuring beforehand**

Concept: *average hours of consumption of online political news*

Measure: *average time recorded of the visits to online political outlets' URLs.*

- What is considered a visit?
- Which online outlets?
- Which URLs should be considered political?
- What time frame to use to compute an average?

Practical recommendations

2. **Consider the impact of the chosen technologies on data quality**

RESULTS

Practical recommendations

2. Consider the impact of the chosen technologies on data quality

Apps

Where?
Device

Devices
Not iOS

Continuous?
Yes

Types of data
URLs, Time, Device, Search terms, Incognito

Plug-in A

Where?
Browser

Devices
Only PC & MAC

Continuous?
Yes

Types of data
URLs, Time, Device, Search terms, Incognito, HTML

Plug-in B

Where?
Browser

Devices
Only PC & MAC

Continuous?
No

Types of data
URLs, Time, Device

Proxy

Where?
Network

Devices
All

Continuous?
Yes

Types of data
URLs, Time, Device

RESULTS

Practical recommendations

2. Consider the impact of the chosen technologies on data quality

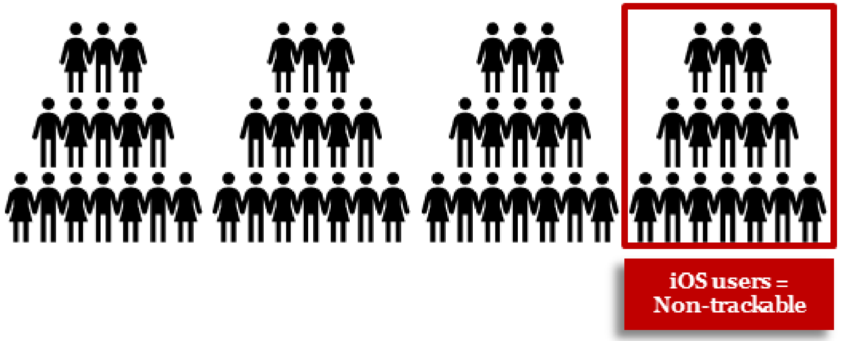
Apps

Where?
Device

Devices
Not iOS

Continuous?
Yes

Types of data
URLs, Time, Device, Search terms, Incognito



Practical recommendations

2. Consider the impact of the chosen technologies on data quality

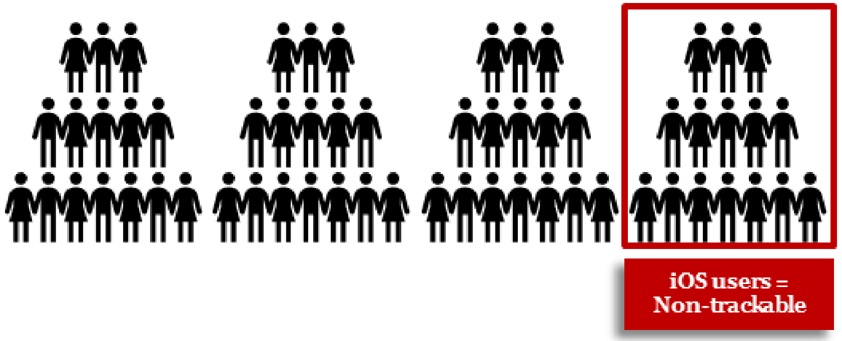
Apps

Where?
Device

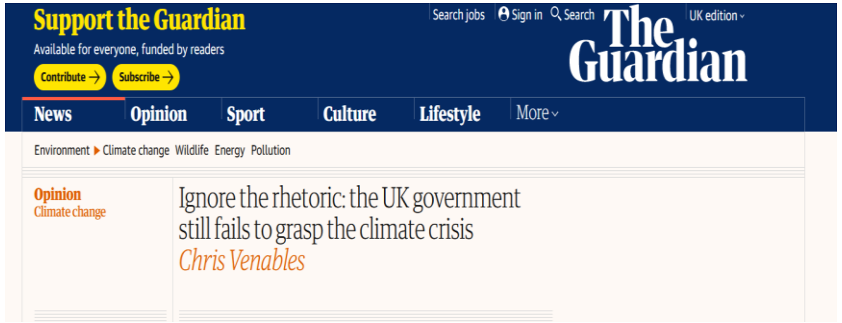
Devices
Not iOS

Continuous?
Yes

Types of data
URLs, Time, Device, Search terms, Incognito



iOS users = Non-trackable



Support the Guardian

Available for everyone, funded by readers

Contribute → Subscribe →

Search jobs | Sign in | Search | The Guardian | UK edition

News | Opinion | Sport | Culture | Lifestyle | More

Environment ▸ Climate change Wildlife Energy Pollution

Opinion
Climate change

Ignore the rhetoric: the UK government still fails to grasp the climate crisis

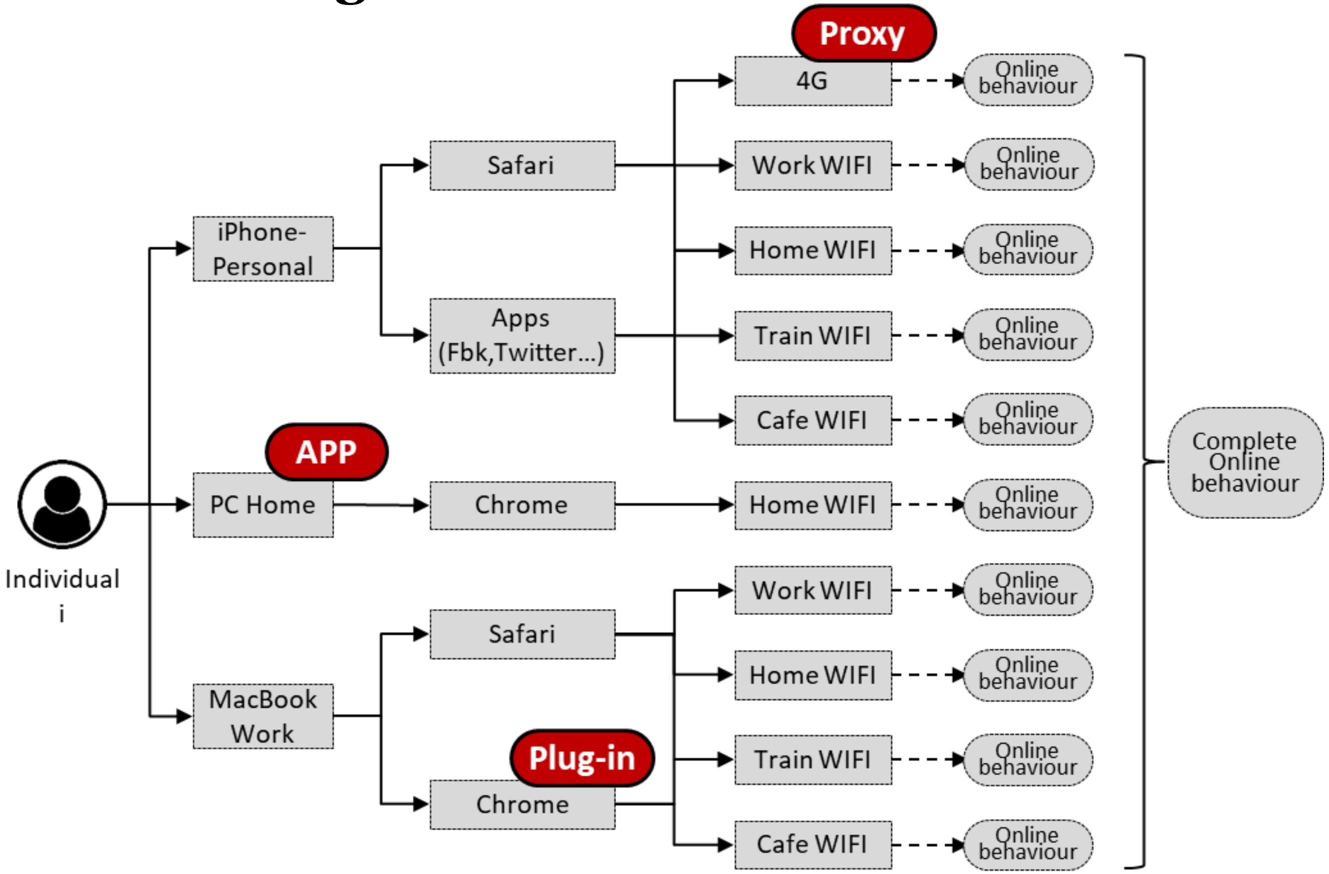
Chris Venables

Practical recommendations

3. **Explore strategies to increase the willingness of individuals to install the meter in all targets**

Practical recommendations

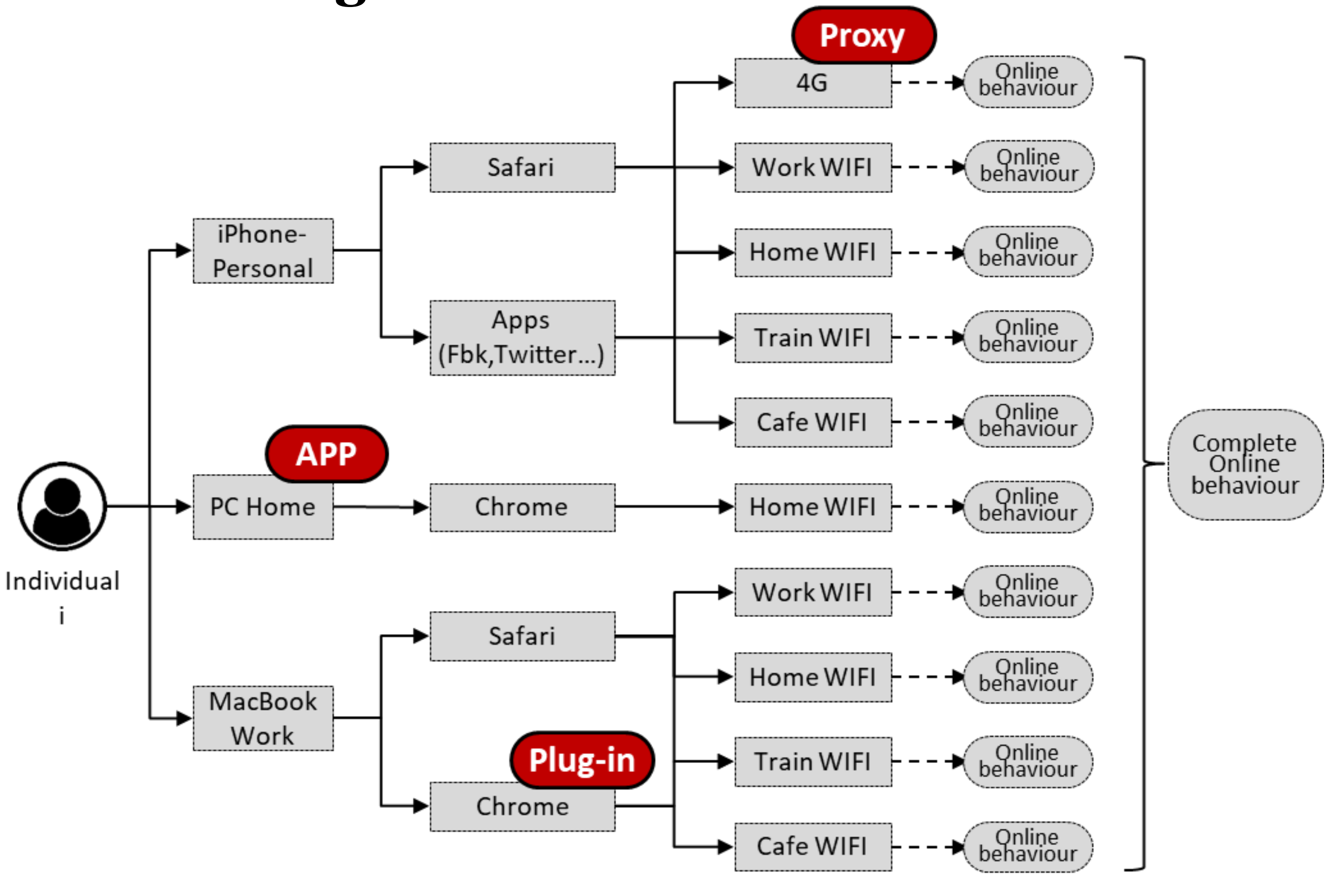
3. Explore strategies to increase the willingness of individuals to install the meter in all targets



RESULTS

Practical recommendations

3. Explore strategies to increase the willingness of individuals to install the meter in all targets



- Multiple tracking technologies might need to be installed for the same participant.
- Tracking technologies present different installations processes.
- Targets (devices / browsers / networks used) are unknown.

Practical recommendations

What if we fail to properly address recommendations 2 & 3?

Practical recommendations

What if we fail to properly address recommendations 2 & 3? → Undercoverage

Practical recommendations

What if we fail to properly address recommendations 2 & 3? → **Undercoverage**

Different levels of undercoverage.

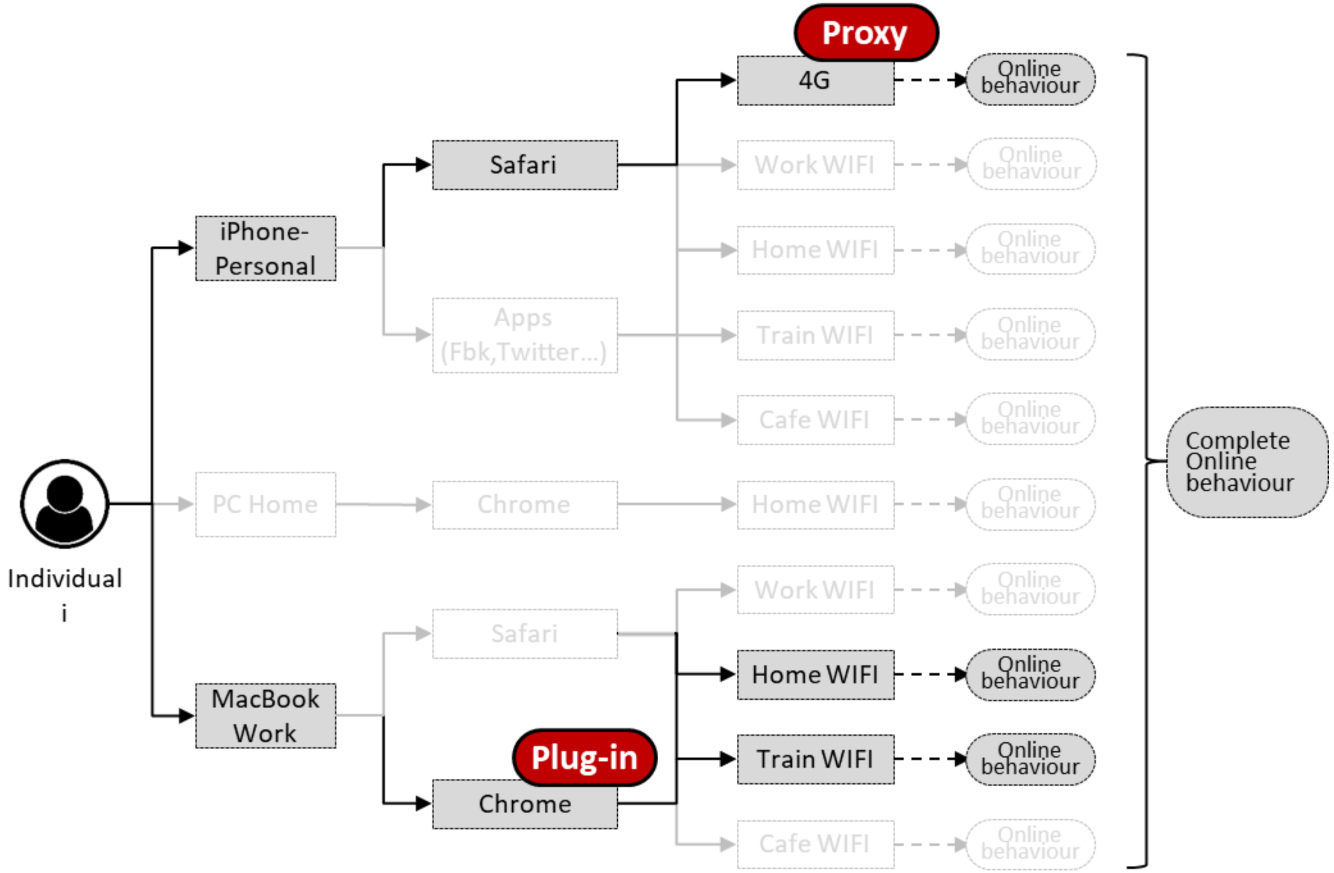
- **Device:** at least one device used by a participant is not tracked
- **Browser:** at least one web-browser used by a participant is not tracked
- **In-app:** the behaviours happening inside apps are not tracked.
- **Network:** at least one network from which a participant connect to the Internet is not tracked

Undercoverage can prevent tracking the complete online behavior

RESULTS

Practical recommendations

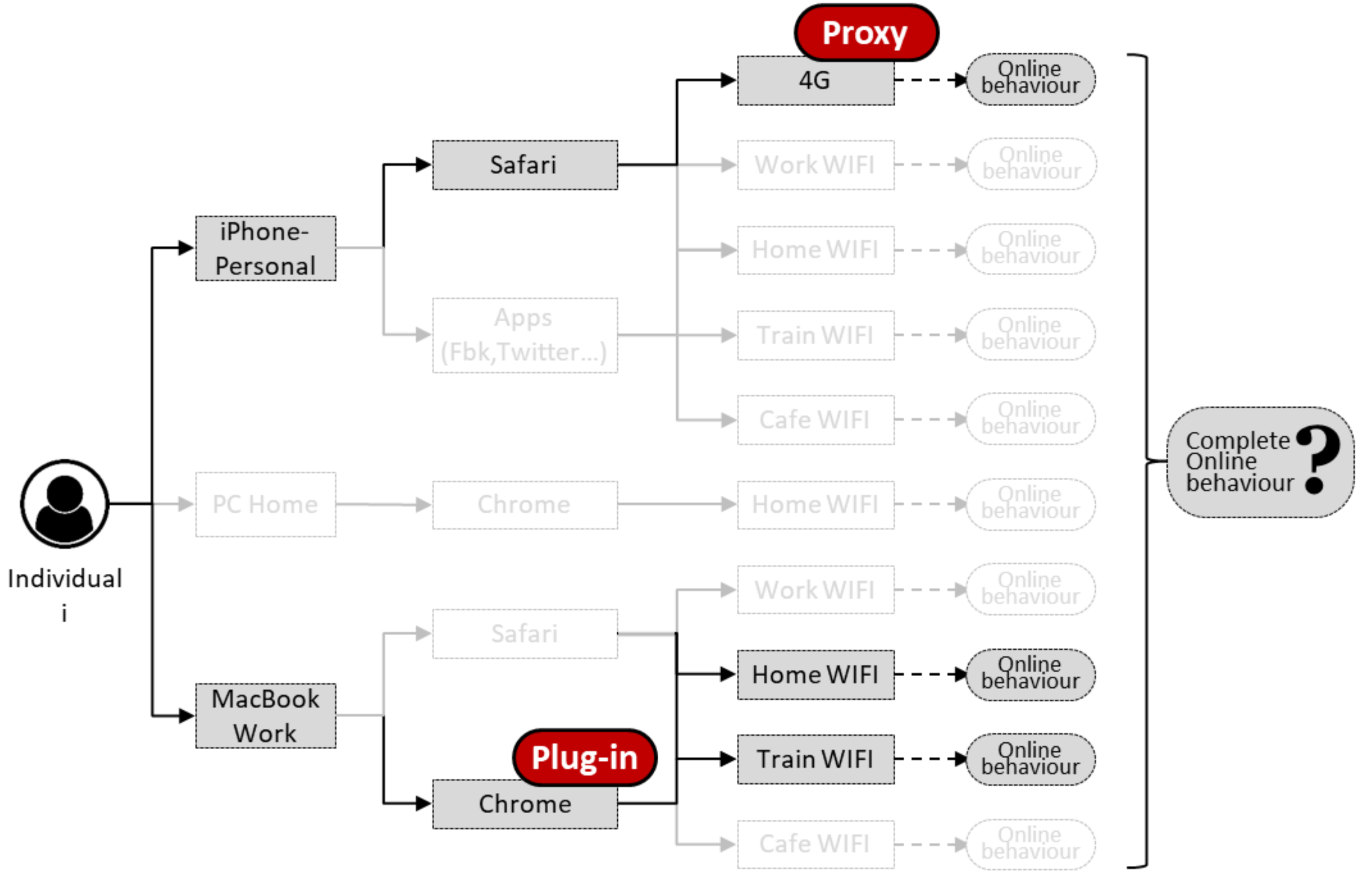
What if we fail to properly address recommendations 2 & 3? → Undercoverage



RESULTS

Practical recommendations

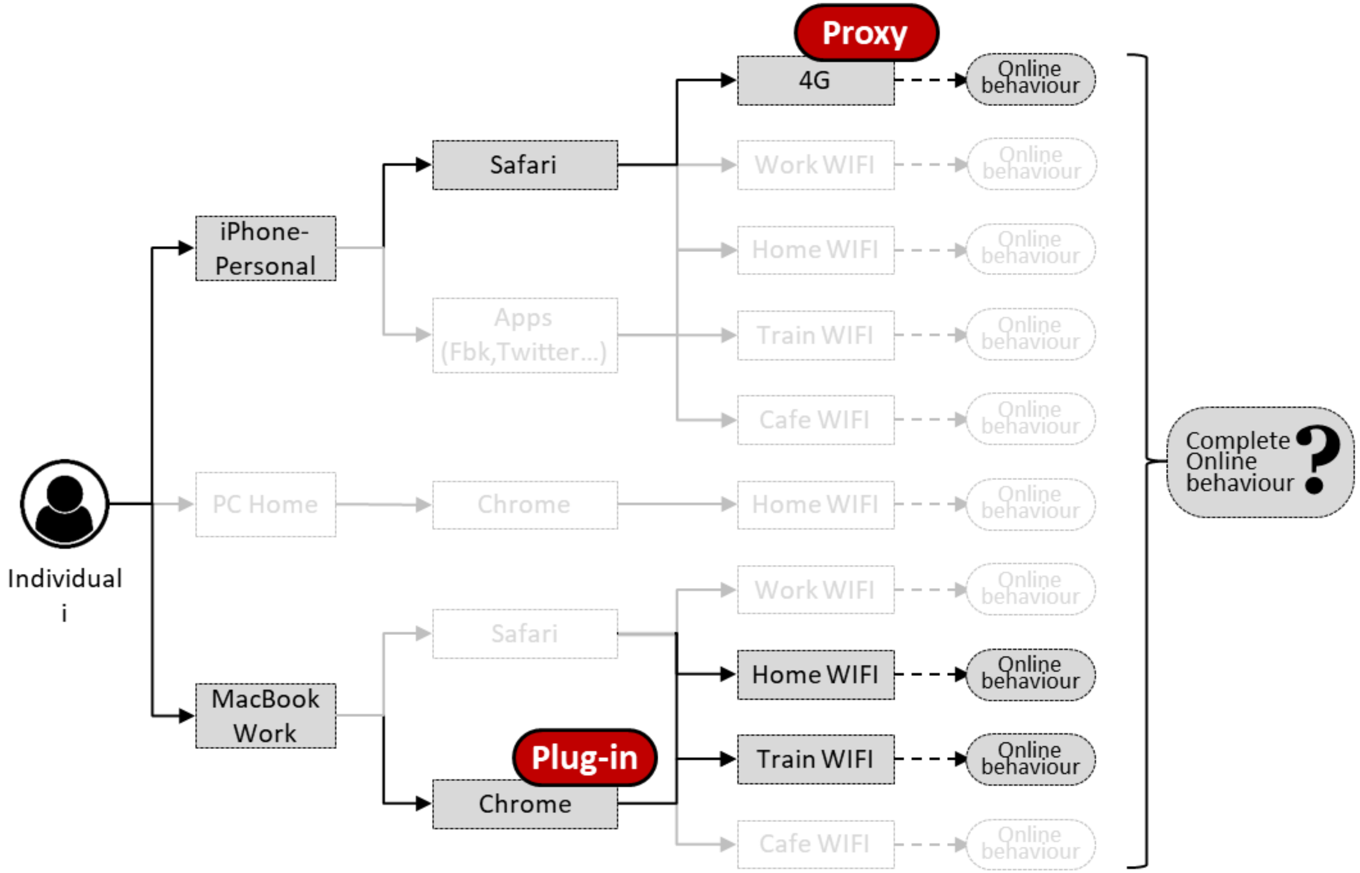
What if we fail to properly address recommendations 2 & 3? → Undercoverage



RESULTS

Practical recommendations

What if we fail to properly address recommendations 2 & 3? → Undercoverage



68%

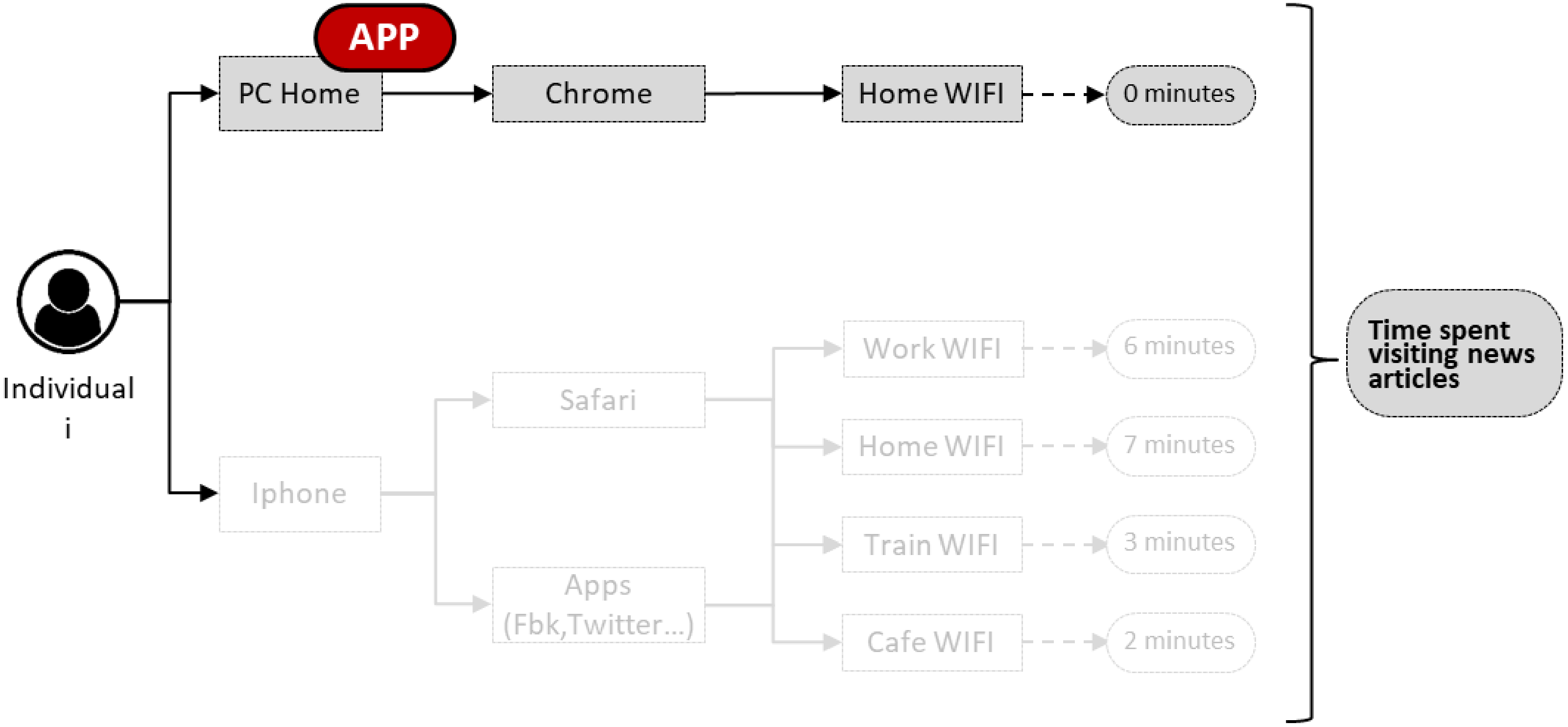
Pew Research Center, 2020. Measuring News Consumption in a Digital Era.

Practical recommendations

Undercoverage might be present, so what?

Practical recommendations

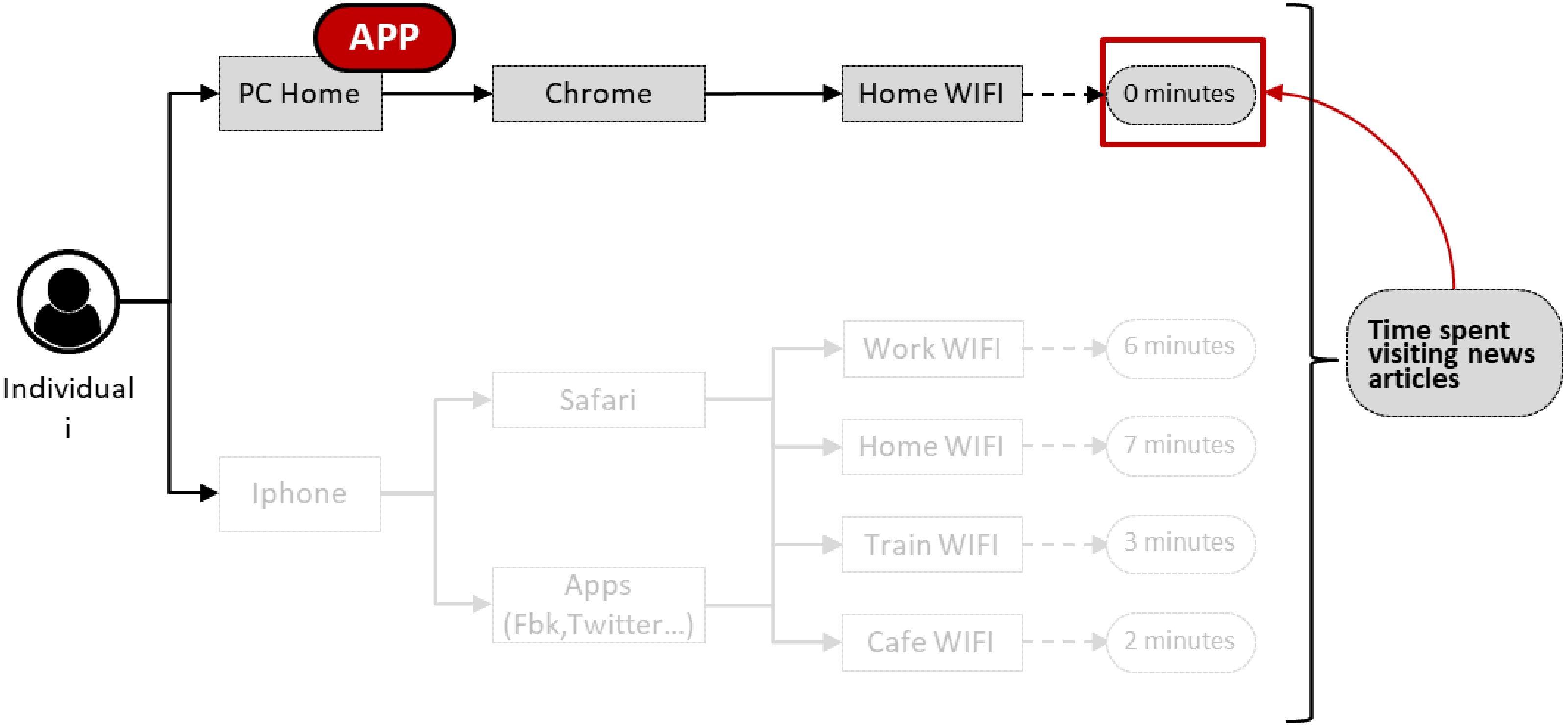
Undercoverage might be present, so what?



RESULTS

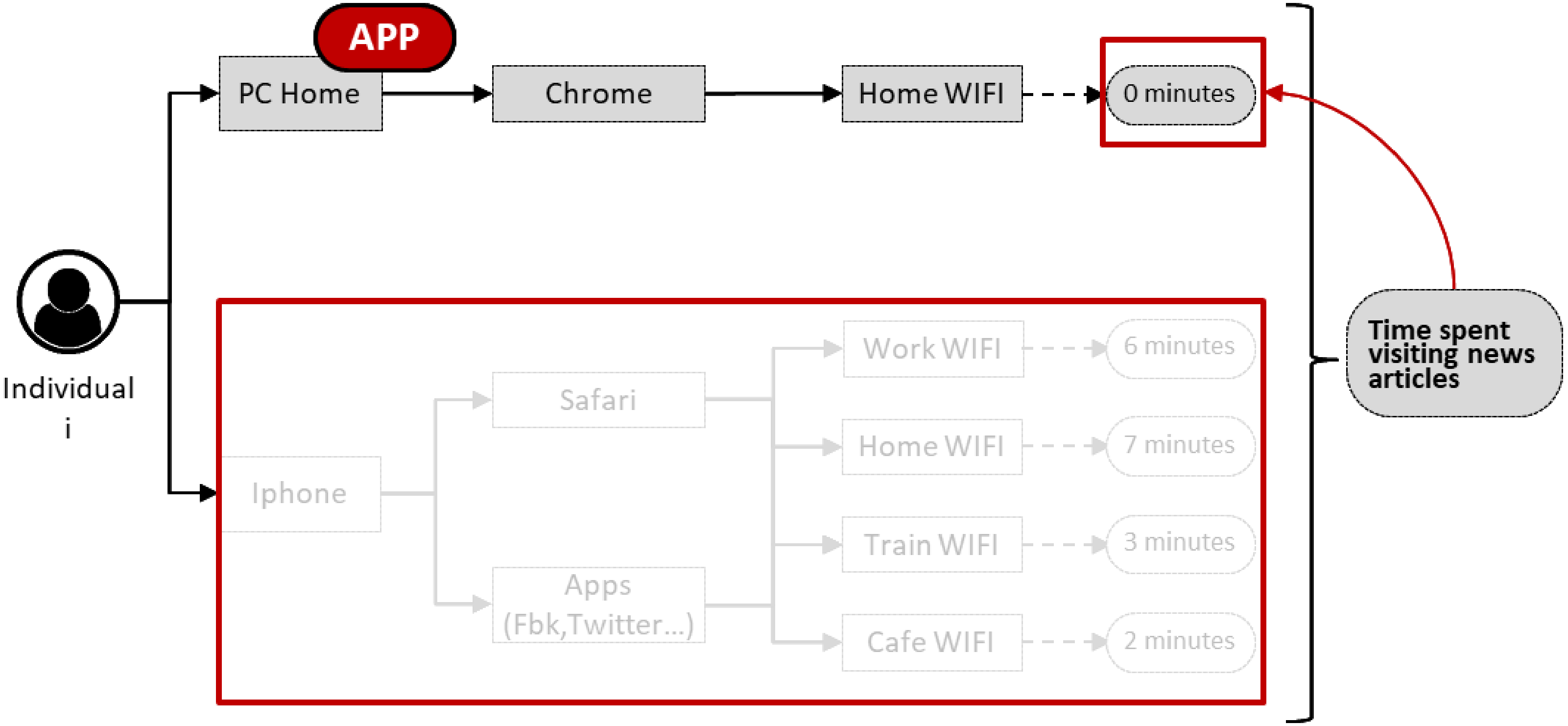
Practical recommendations

Undercoverage might be present, so what?



Practical recommendations

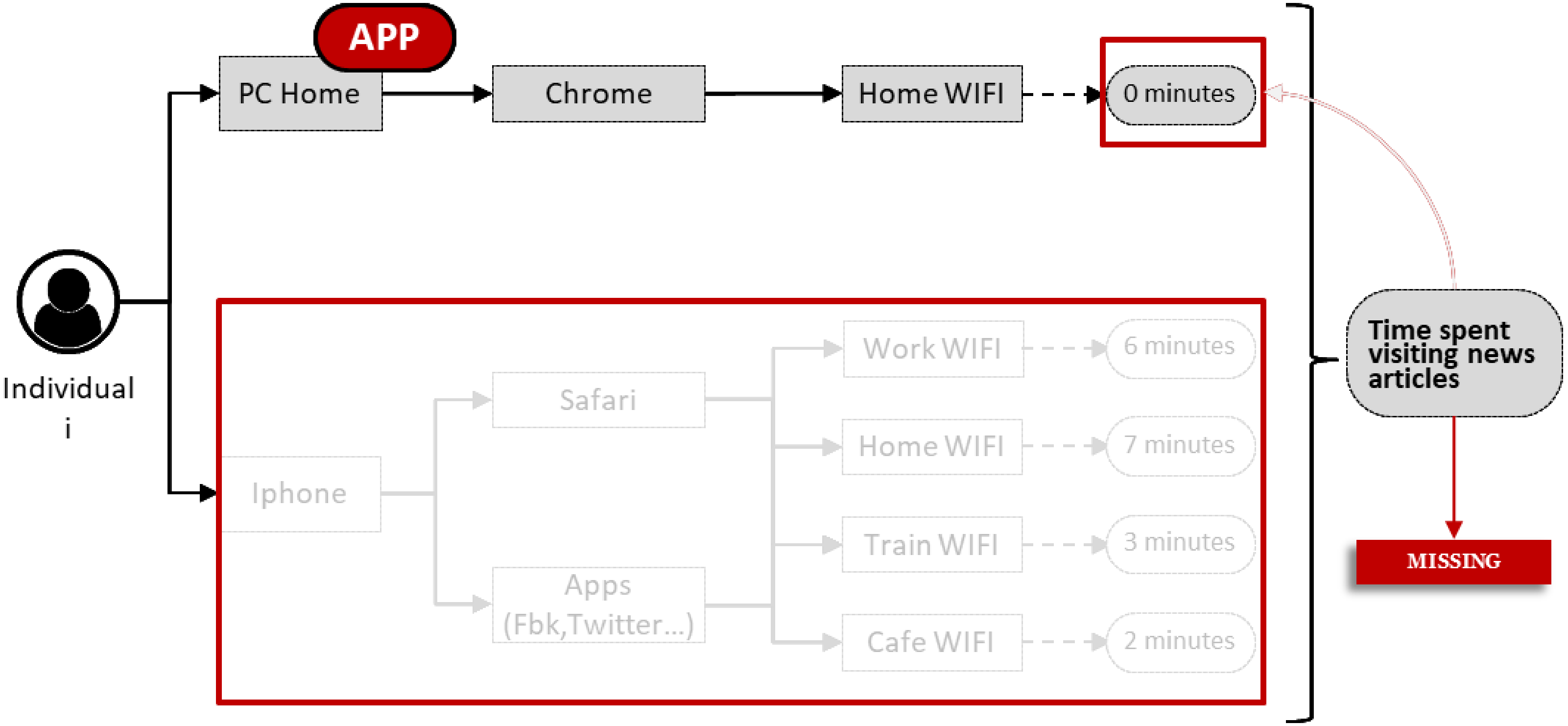
Undercoverage might be present, so what?



RESULTS

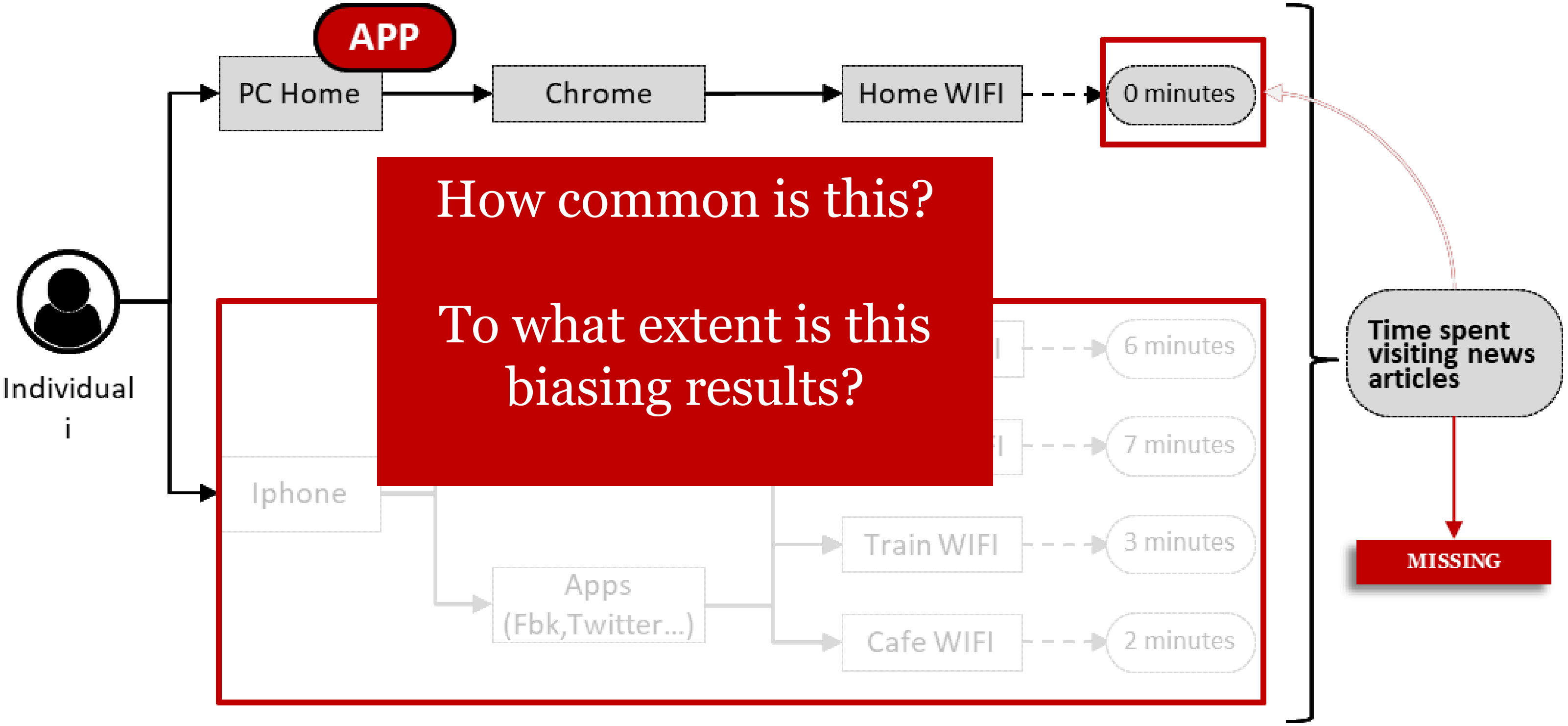
Practical recommendations

Undercoverage might be present, so what?



Practical recommendations

Undercoverage might be present, so what?



How common is this?

To what extent is this biasing results?

Practical recommendations

4. **Define strategies to maximise the information available to identify missing data**

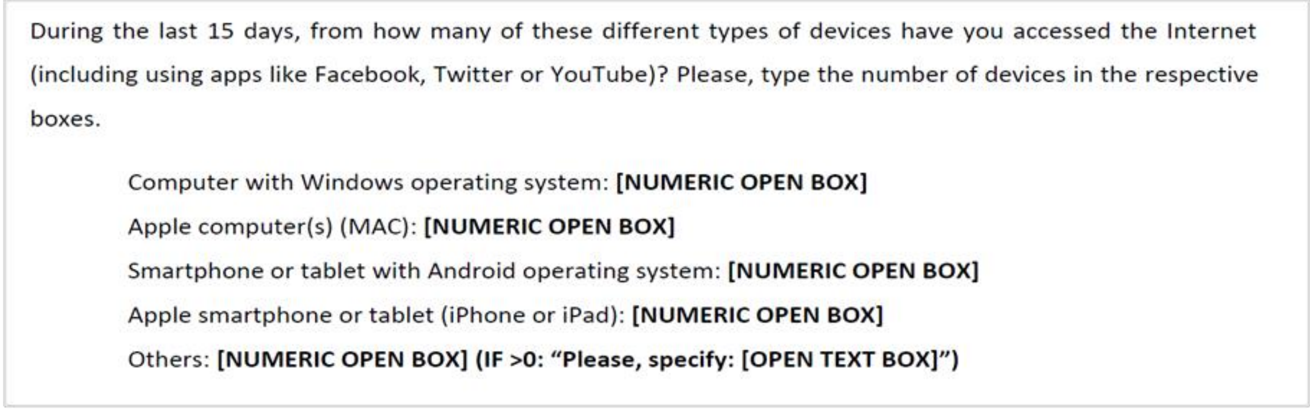
This is still not very clear at the moment. However…

RESULTS

Practical recommendations

4. Define strategies to maximise the information available to identify missing data

This is still not very clear at the moment. However… we can combine survey & paradata



During the last 15 days, from how many of these different types of devices have you accessed the Internet (including using apps like Facebook, Twitter or YouTube)? Please, type the number of devices in the respective boxes.

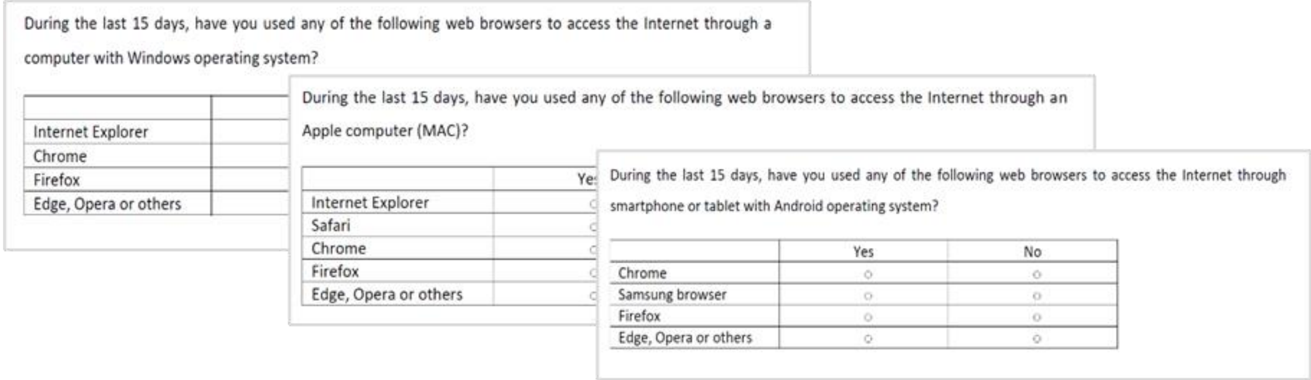
Computer with Windows operating system: **[NUMERIC OPEN BOX]**

Apple computer(s) (MAC): **[NUMERIC OPEN BOX]**

Smartphone or tablet with Android operating system: **[NUMERIC OPEN BOX]**

Apple smartphone or tablet (iPhone or iPad): **[NUMERIC OPEN BOX]**

Others: **[NUMERIC OPEN BOX] (IF >0: "Please, specify: [OPEN TEXT BOX]")**



During the last 15 days, have you used any of the following web browsers to access the Internet through a computer with Windows operating system?

Internet Explorer	
Chrome	
Firefox	
Edge, Opera or others	

During the last 15 days, have you used any of the following web browsers to access the Internet through an Apple computer (MAC)?

	Ye
Internet Explorer	○
Safari	○
Chrome	○
Firefox	○
Edge, Opera or others	○

During the last 15 days, have you used any of the following web browsers to access the Internet through smartphone or tablet with Android operating system?

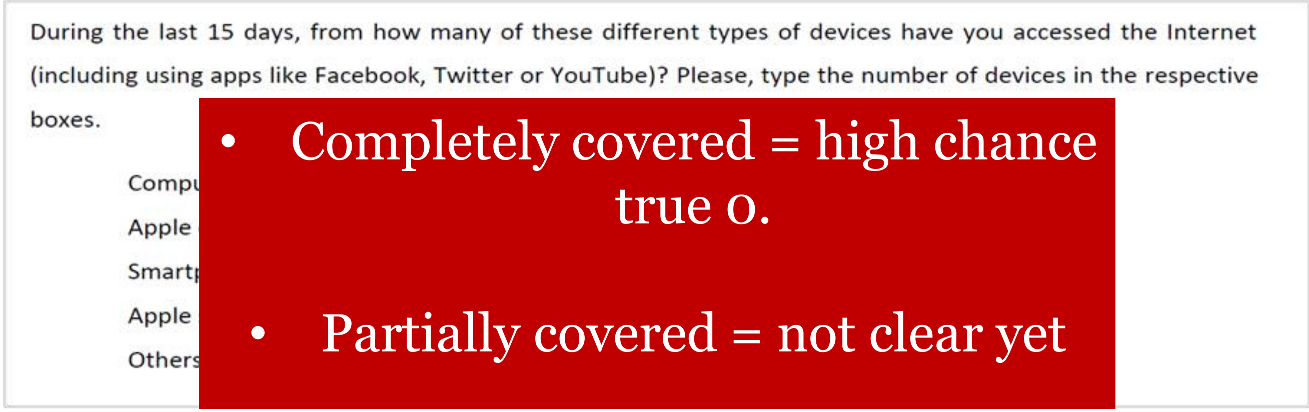
	Yes	No
Chrome	○	○
Samsung browser	○	○
Firefox	○	○
Edge, Opera or others	○	○

RESULTS

Practical recommendations

4. Define strategies to maximise the information available to identify missing data

This is still not very clear at the moment. However… we can combine survey & paradata



During the last 15 days, from how many of these different types of devices have you accessed the Internet (including using apps like Facebook, Twitter or YouTube)? Please, type the number of devices in the respective boxes.

Compu

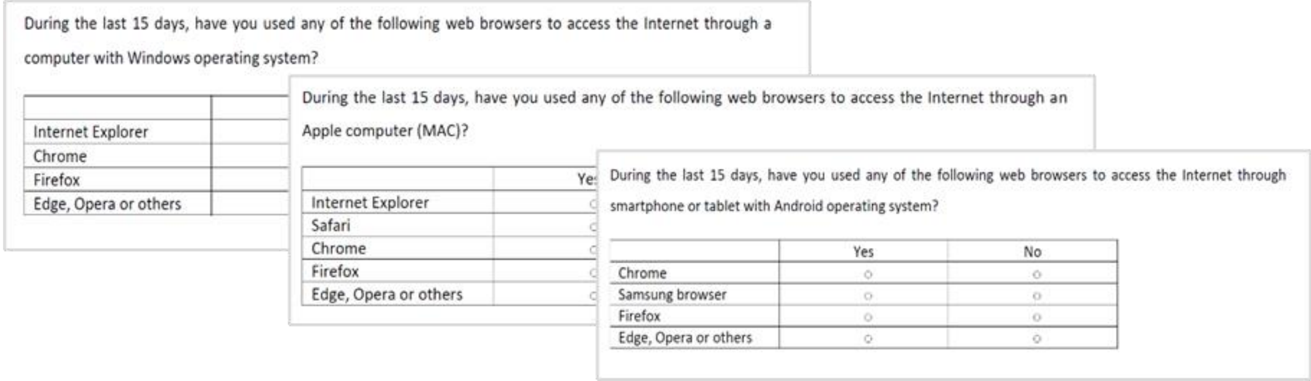
Apple

Smartp

Apple

Others

- Completely covered = high chance true 0.
- Partially covered = not clear yet



During the last 15 days, have you used any of the following web browsers to access the Internet through a computer with Windows operating system?

Internet Explorer	
Chrome	
Firefox	
Edge, Opera or others	

During the last 15 days, have you used any of the following web browsers to access the Internet through an Apple computer (MAC)?

	Ye
Internet Explorer	○
Safari	○
Chrome	○
Firefox	○
Edge, Opera or others	○

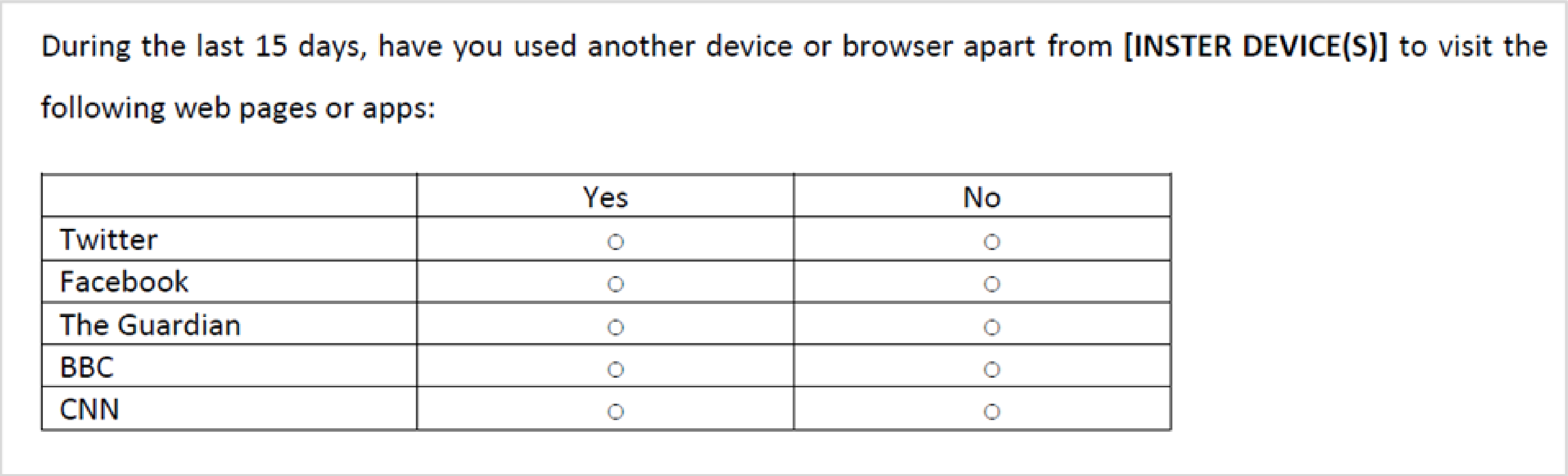
During the last 15 days, have you used any of the following web browsers to access the Internet through smartphone or tablet with Android operating system?

	Yes	No
Chrome	○	○
Samsung browser	○	○
Firefox	○	○
Edge, Opera or others	○	○

Practical recommendations

4. Define strategies to maximise the information available to identify missing data

This is still not very clear at the moment. However… we can combine survey & paradata



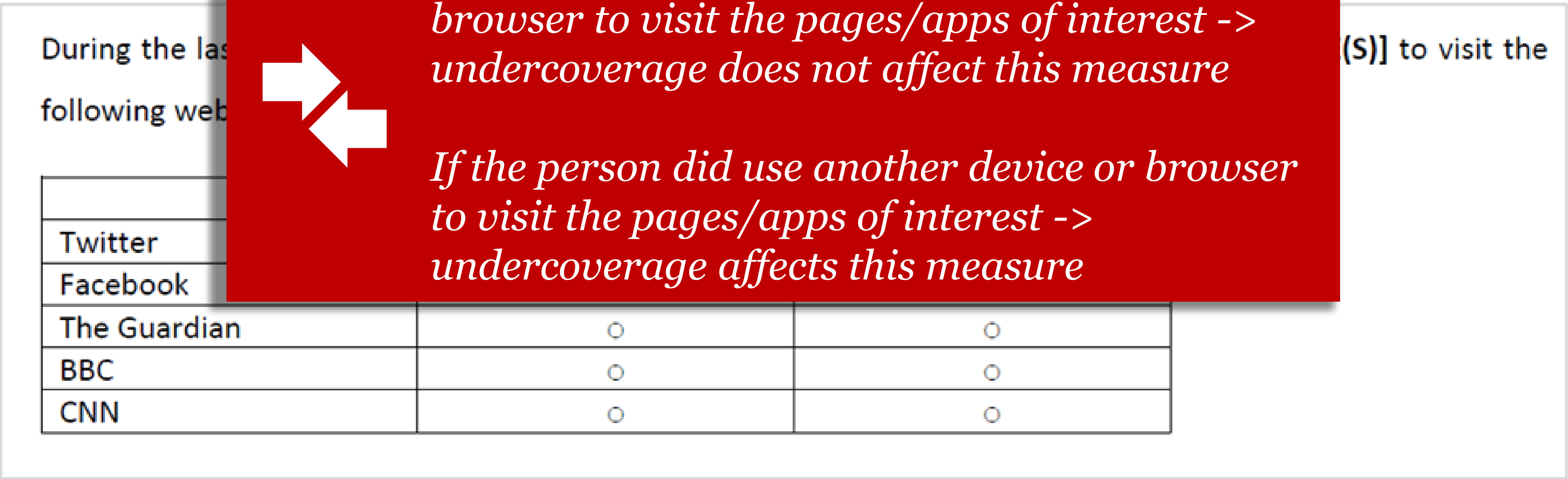
During the last 15 days, have you used another device or browser apart from **[INSTER DEVICE(S)]** to visit the following web pages or apps:

	Yes	No
Twitter	○	○
Facebook	○	○
The Guardian	○	○
BBC	○	○
CNN	○	○

Practical recommendations

4. Define strategies to maximise the information available to identify missing data

This is still not very clear at the moment. However… we can combine survey & paradata



During the las … (S)] to visit the

following web …

Twitter		
Facebook		
The Guardian	○	○
BBC	○	○
CNN	○	○

If the person did not use another device or browser to visit the pages/apps of interest -> undercoverage does not affect this measure

If the person did use another device or browser to visit the pages/apps of interest -> undercoverage affects this measure

Most likely cannot be done for every web page/app of interest

CONCLUSIONS

Limits

1. One specific definition of data quality.
2. Lack of previous empirical research.
3. Tracking technologies are constantly evolving.
4. Metered data errors are considered independently.

Take-home messages

1. Using metered data is complex and many decisions must be taken.
2. Reporting these decisions and conducting robustness checks is necessary.
3. More empirical research is needed.
4. This framework can help on all these aspects.

5. Identifying when a lack of behaviour is real or a product of undercoverage is key
6. Confounding both phenomena can inflate measurement and missing data errors.

Thanks!

Questions?

Oriol J. Bosch

o.bosch-jover@lse.ac.uk

orioljbosch

https://orioljbosch.com/

Bosch, O.J., and M. Revilla (2021). **“When survey science met online tracking: presenting an error framework for metered data.”** RECSM Working Papers Series, 62

upf. Universitat Pompeu Fabra *Barcelona*

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

RECSM Research and Expertise Centre for Survey Methodology

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No849165), PI: Melanie Revilla