Track me but not really:

Tracking undercoverage in metered data collection

Oriol J. Bosch| Department of Methodology, LSE

Patrick Sturgis| Department of Methodology, LSE

Jouni Kuha| Department of Methodology, LSE

o.bosch-jover@lse.ac.uk

orioljbosch

https://orioljbosch.com/

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Universitat Pompeu Fabra Barcelona

RECSM Research and Expertise Centre for Survey Methodology

Acknowledgements: I would like to thank Melanie Revilla for her always insightful comments

Funding: This project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No 849165; PI: Melanie Revilla); the Spanish Ministry of Science and Innovation under the "R+D+i projects" programme (grant number PID2019-106867RB-I00 /AEI/10.13039/501100011033 (2020-2024), PI: Mariano Torcal); and the BBVA foundation under their grant scheme to scientific research teams in economy and digital society, 2019 (PI: Mariano Torcal).

The rise of metered data

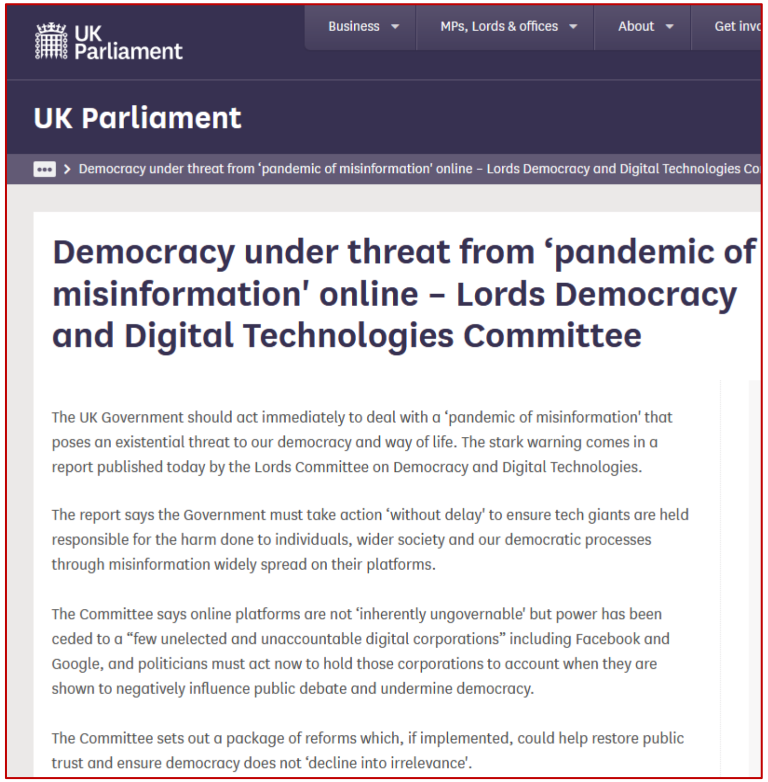
- It is becoming vital to better understand what people do online and what impact this has on online and offline phenomena.


The New York Times

OPINION
GUEST ESSAY

Does Instagram Harm Girls? No One Actually Knows.

Oct. 10, 2021


UK Parliament

Business MPs, Lords & offices About Get invo

UK Parliament

Democracy under threat from 'pandemic of misinformation' online – Lords Democracy and Digital Technologies Co

Democracy under threat from 'pandemic of misinformation' online – Lords Democracy and Digital Technologies Committee

The UK Government should act immediately to deal with a 'pandemic of misinformation' that poses an existential threat to our democracy and way of life. The stark warning comes in a report published today by the Lords Committee on Democracy and Digital Technologies.

The report says the Government must take action 'without delay' to ensure tech giants are held responsible for the harm done to individuals, wider society and our democratic processes through misinformation widely spread on their platforms.

The Committee says online platforms are not 'inherently ungovernable' but power has been ceded to a "few unelected and unaccountable digital corporations" including Facebook and Google, and politicians must act now to hold those corporations to account when they are shown to negatively influence public debate and undermine democracy.

The Committee sets out a package of reforms which, if implemented, could help restore public trust and ensure democracy does not 'decline into irrelevance'.

The rise of metered data

- It is becoming vital to better understand what people do online and what impact this has on online and offline phenomena.

- Self-reports might not be best suited for this

The Immensely Inflated News Audience: Assessing Bias in Self-Reported News Exposure Get access >

Markus Prior

Public Opinion Quarterly, Volume 73, Issue 1, Spring 2009, Pages 130–143, https://doi.org/10.1093/poq/nfp002

Published: 18 March 2009

Cite Permissions Share ▾

Abstract

Many studies of media effects use self-reported news exposure as their key independent variable without establishing its validity. Motivated by anecdotal evidence that people's reports of their own media use can differ considerably from independent assessments, this study examines systematically the accuracy of survey-based self-reports of news exposure. I compare survey estimates to Nielsen estimates, which do not rely on self-reports. Results show severe overreporting of news exposure. Survey estimates of network news exposure follow trends in Nielsen ratings relatively well, but exaggerate

The Journal of Politics > Volume 71, Number 3

< PREVIOUS ARTICLE NEXT ARTICLE >

Improving Media Effects Research through Better Measurement of News Exposure

Markus Prior

PDF PDF PLUS Abstract Full Text

Abstract

Survey research is necessary to understand media effects, but seriously impeded by considerable overreporting of news exposure, the extent of which differs across respondents. Consequently, apparent media effects may arise not because of differences in exposure, but because of differences in the accuracy of reporting exposure. Drawing on experiments embedded in two representative surveys, this study examines why many people overstate their exposure to television news. Analysis indicates that overreporting results from unrealistic demands on respondents' memory, not their motivation to misrepresent or provide superficial answers. Satisficing and social desirability bias do not explain overreporting. Instead, imperfect recall coupled with the

The rise of metered data

- It is becoming vital to better understand what people do online and what impact this has on online and offline phenomena.
- Self-reports might not be best suited for this

- Alternative: directly observe what people do online using digital tracking solutions, or *meters.*
  - **Group of tracking technologies**
  - **Installed on participants devices**.
  - **Collect traces** left by participants when **interacting with their devices online: e.g. URLs or apps visited**
- We call the resulting data: **metered data**.

The rise of metered data

webdataopp

- Since 2016, more than 60 papers published using metered data

Article

The International Journal of Press/Politics
2020, Vol. 25(3) 426–446

DOI: 10.1177/1940161220907018
journals.sagepub.com/home/hij
SAGE

Populist Attitudes and Selective Exposure to Online News: A Cross-Country Analysis Combining Web Tracking and Surveys

Sebastian Stier[1], Nora Kirkizh[1], Caterina Froio[2], and Ralph Schroeder[3]

Abstract

Research has shown that citizens with populist attitudes evaluate the news media more negatively, and there is also suggestive evidence that they rely less on established news sources like the legacy press. However, due to data limitations, there is still no solid evidence whether populist citizens have skewed news diets in the contemporary high-choice digital media environment. In this paper, we rely on the selective exposure framework and investigate the relationship between populist attitudes and the consumption of various types of online news. To test our theoretical assumptions, we link 150 million Web site visits by 7,729 Internet users in France, Germany, Italy, Spain, the United Kingdom, and the United States to their responses in an online survey. This design allows us to measure media exposure more precisely than previous studies while linking these data to demographic attributes and political attitudes of participants. The results show that populist attitudes leave pronounced

AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

ARTICLE

(Almost) Everything in Moderation: New Evidence on Americans' Online Media Diets

Andrew M. Guess

First published: 19 February 2021 | https://doi.org/10.1111/ajps.12589 | Citations: 13

This study was approved by the New York University Institutional Review Board (IRB-FY2016-1342). I would like to thank the editors and three anonymous reviewers for their detailed guidance and feedback on this article. I am grateful to Pablo Barberá, Neal Beck, Noah Buckley, Alex Coppock, Pat Egan, Albert Fang, Don Green, Trish Kirkland, Jeff Lax, Lucas Leemann, Yph Lelkes, Jonathan Nagler, Brendan Nyhan, Markus Prior, Jason Reifler, Robert Shapiro, Gaurav Sood, Lauren Young, and seminar participants at the Columbia University Department of Political Science, the Annenberg School for Communication at the University of Pennsylvania, the NYU Center for Data Science, and the Yale ISPS Experiments Workshop for extremely helpful comments and suggestions. Thanks also to those who provided valuable feedback during seminars at Brown University, Princeton University, Rutgers, Penn State, and NYU Abu Dhabi. I additionally benefited from comments by discussants and attendees at the 2016 Southern Political Science Association and Midwest Political Science Association annual meetings and the 2016 APSA Political Communication Pre-conference at Temple University. I am indebted to Doug Rivers, Brian Law, and Joe Williams at YouGov for facilitating access to the 2015 Pulse data, and to Ashley Grosse for making possible the survey on privacy attitudes. The 2016 data collection was generously supported by the American Press Institute. Some of the analysis was made possible by High Performance Computing (HPC) clusters at New York University.

The rise of metered data

- Since 2016, more than 60 papers published using metered data

- The benefits seem clear…but should we assume that metered data is unbiased?

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

Oriol J. Bosch[1] | **Melanie Revilla**[2]

[1]The London School of Economics and Political Science

[2]Research and Expertise Centre for Survey Methodology (RECSM), Universitat Pompeu Fabra

Correspondence
Oriol J. Bosch, Department of Methodology, The London School of Economics and Political Science, London, WC2B 4RR, United Kingdom
Email: o.bosch-jover@lse.ac.uk

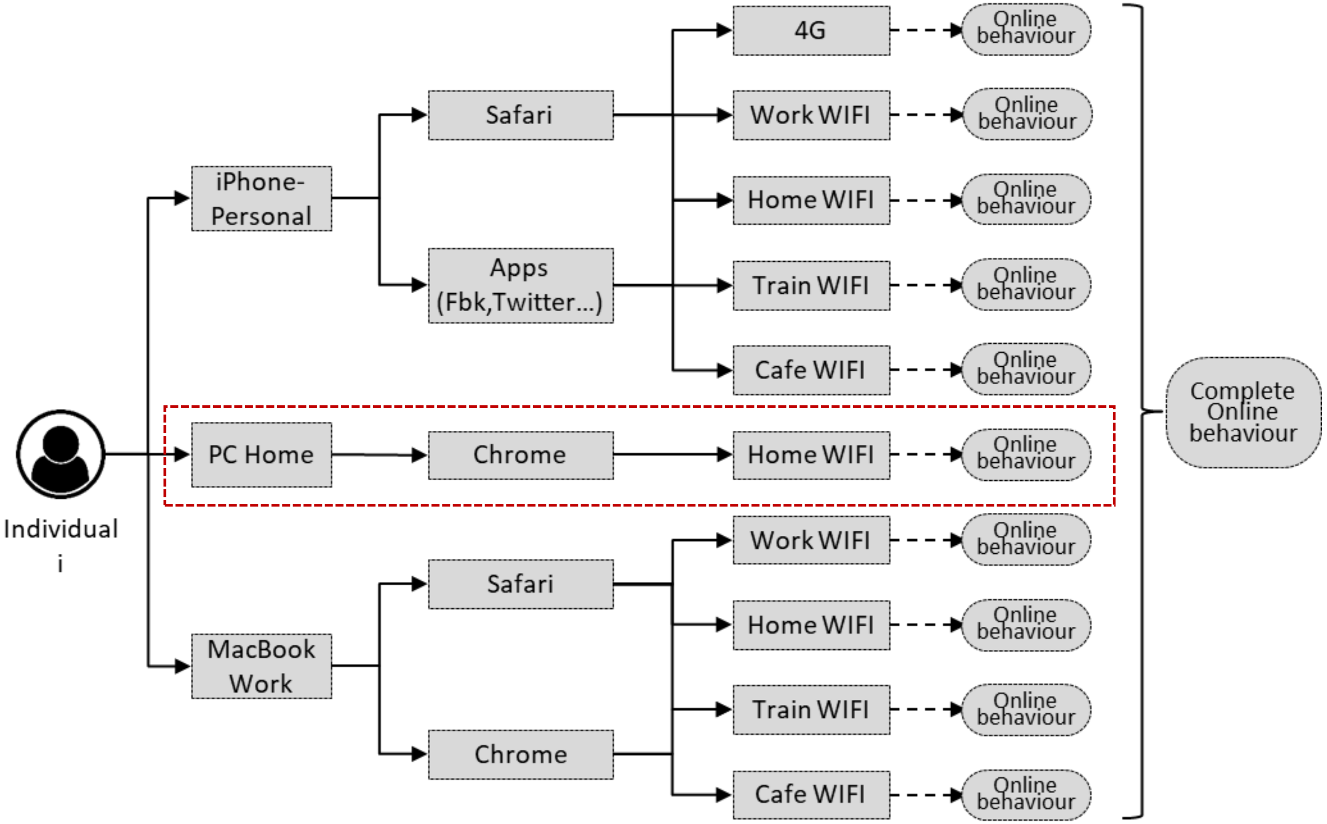
Funding information
European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No 849165)

Metered data, also called "web-tracking data", are generally collected from a sample of participants who willingly install or configure, onto their devices, technologies that track digital traces left when people go online (e.g., URLs visited). Since metered data allow for the observation of online behaviours unobtrusively, it has been proposed as a useful tool to understand what people do online and what impacts this might have on online and offline phenomena. It is crucial, nevertheless, to understand its limitations. Although some research has explored the potential errors of metered data, a systematic categorisation and conceptualisation of these errors are missing. Inspired by the Total Survey Error, we present a Total Error framework for digital traces collected with Meters (TEM). The TEM framework (1) describes the data generation and the analysis process for metered data and (2) documents the sources of bias and variance that may arise in each step of this process. Furthermore, using a case study, we show how the TEM can be applied in real life to identify, quantify and reduce metered data errors. This framework can help improve the quality of both stand-alone metered data research projects, as well as foster the understanding of how and when survey and metered data can be combined.

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

TRACKING UNDERCOVERAGE

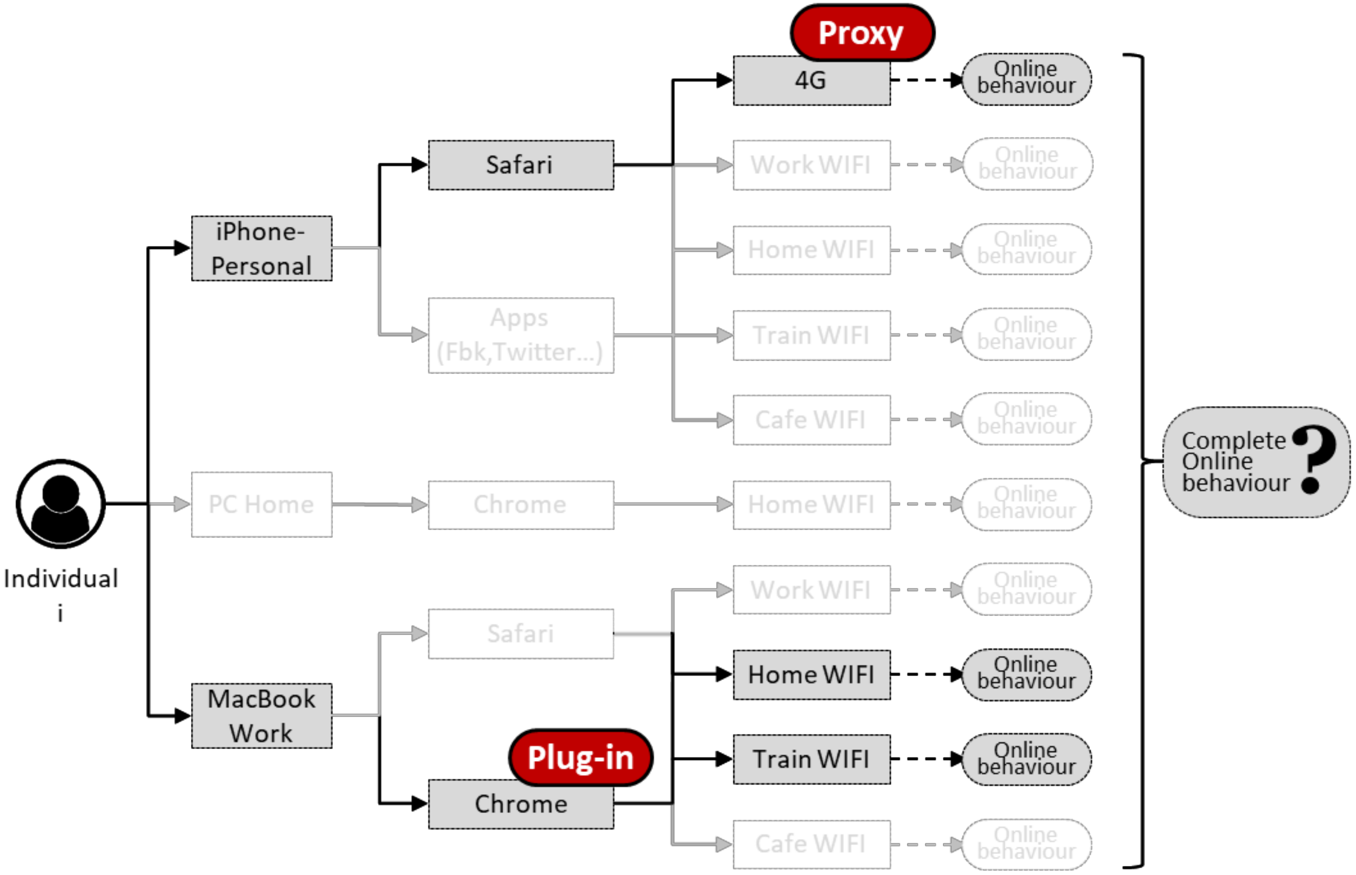
What do we mean by tracking undercoverage?



Objective: measuring individuals' behaviours

Reality: vector of those behaviours that individuals' do through all their *targets*

What do we mean by tracking undercoverage?

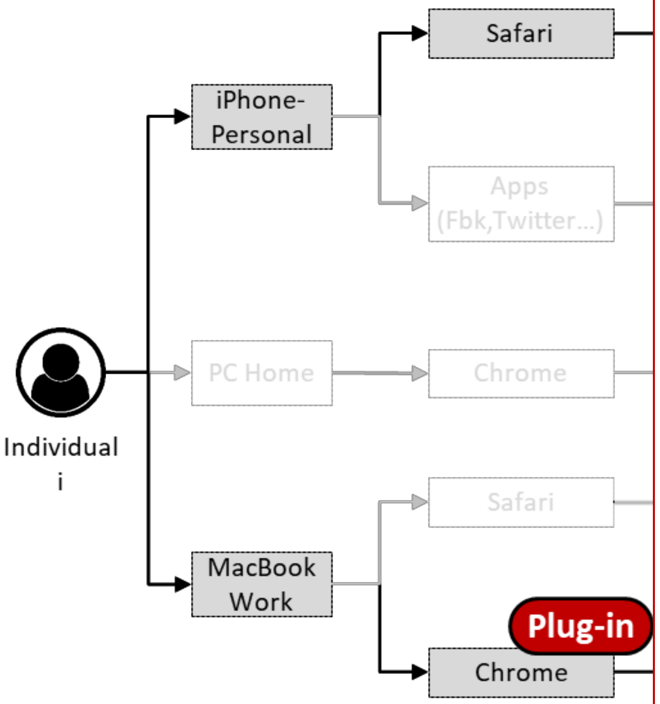


Undercoverage can prevent tracking a participant's complete online behaviour.

Different reasons:

- **Non-trackable targets**
- **Meter not installed**
- **Meter uninstalled**
- **New non-tracked target**

What do we mean by tracking undercoverage?




Pew Research Center

RESEARCH TOPICS ▾ | ALL PUBLICATIONS | METHODS | SHORT READS

Home > Research Topics > Methodological Research > Data Science

PEW RESEARCH CENTER | DECEMBER 8, 2020

Measuring News Consumption in a Digital Era

As news outlets morph and multiply, both surveys and passive data collection tools face challenges

BY MICHAEL BARTHEL, AMY MITCHELL, DORENE ASARE-MARFO, COURTNEY KENNEDY AND KIRSTEN WORDEN

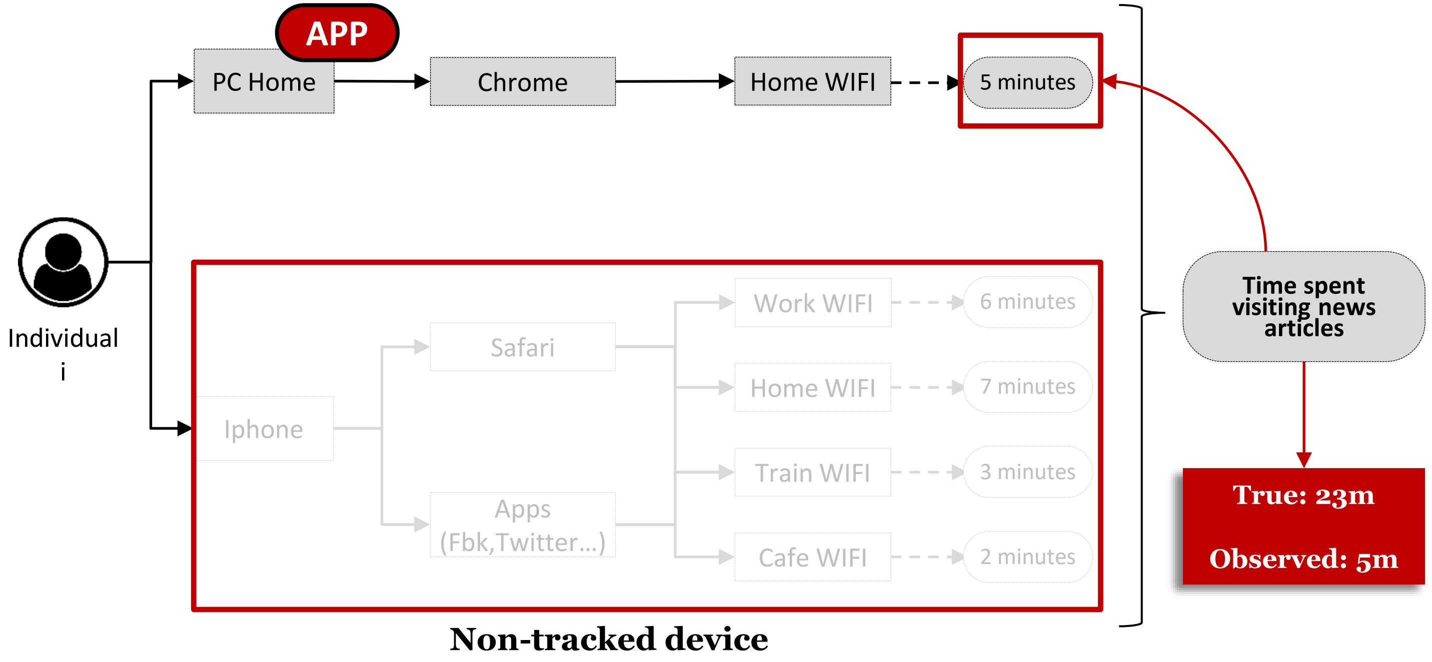
ndercoverage can prevent tracking a
articipant's complete online behaviour.

ifferent reasons:

- **Non-trackable targets**
- **Meter not installed**
- **Meter uninstalled**
- **New non-tracked target**

The consequences of tracking undercoverage

Partial observation



Partial observations can introduce **measurement errors**

- Can lead to underestimation of univariate estimates
- Biased multivariate estimates

OUR STUDY

Research questions

- What is the percentage of participants being undercovered in general (**RQ 1.1**) and in terms of their devices and browsers? (**RQ 1.2**)

- Which types of devices are not covered? (**RQ 2**)

- To what extent does undercoverage introduce bias to univariate (**RQ 3.1**) and multivariate estimates based on metered data? (**RQ 3.2**)

Data

TRI-POL project - Overview

- Three wave survey combined with metered data at the individual level
- Spain, Portugal, Italy + Argentina and Chile
- Netquest metered panels – Cross-quotas about gender, age, education and region

Data

TRI-POL project - Overview

- Three wave survey combined with metered data at the individual level
- Spain, Portugal, Italy + Argentina and Chile
- Netquest metered panels – Cross-quotas about gender, age, education and region

Survey part	Metered part
Questions: polarization, political trust, political communication...	Devices: Windows PC, MAC, iOS & Android mobile devices
Time: ≈30 minutes	Technologies: plug-in, apps and proxies
Fieldwork: September 21 – April 22	Time frame: 15 days before participants started the survey, 16 after starting
Sample Size: 1,289 (Spain), 1,231 (Italy), 1,028 (Portugal)	Sample size: 993 (Spain), 842 (Italy), 818 (Portugal)*

* Inverse probability weights computed using the random forest relative frequency method by Buskirk and Kolenikov (2015)

To measure undercoverage, we need to identify it

Our approach: combining survey and 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)?

	Yes
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	○	○

*Compare this information with device **paradata:** Information about **all** the devices and browsers in which they are tracked .*

OUR STUDY

To measure undercoverage, we need to identify it

Our approach: combining survey and 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)?

	Yes
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	○	○

RQ 1 & RQ2

This approach can be used to compute the proportion of participants undercovered, in general and for each kind of device / browser

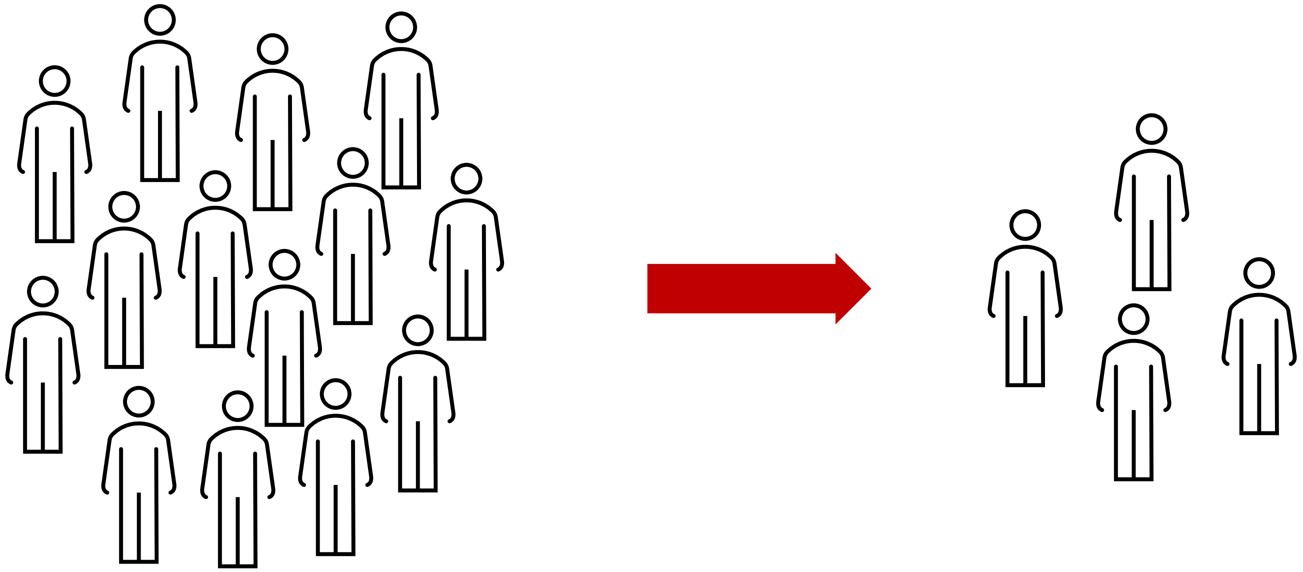
Simulating undercoverage bias (RQ3)

Knowing who is fully covered allows also to simulate bias for them

Simulating undercoverage bias (RQ3)

Knowing who is fully covered allows also to simulate bias for them

- We can treat those subsamples as our “population” of fully covered participants*



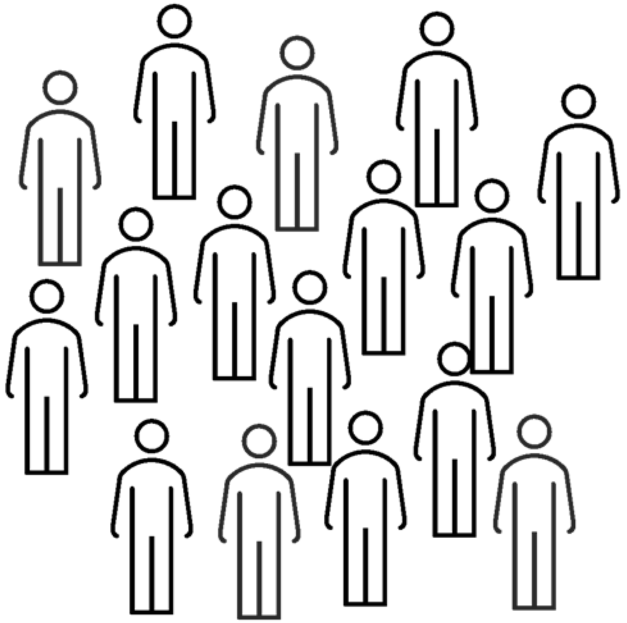
Full sample **Fully covered sample**

* Inverse probability weights computed using the random forest relative frequency method by Buskirk and Kolenikov (2015)

Simulating undercoverage bias (RQ3)

Simulation approach

We can estimate the true estimates of this fully covered subsamples…



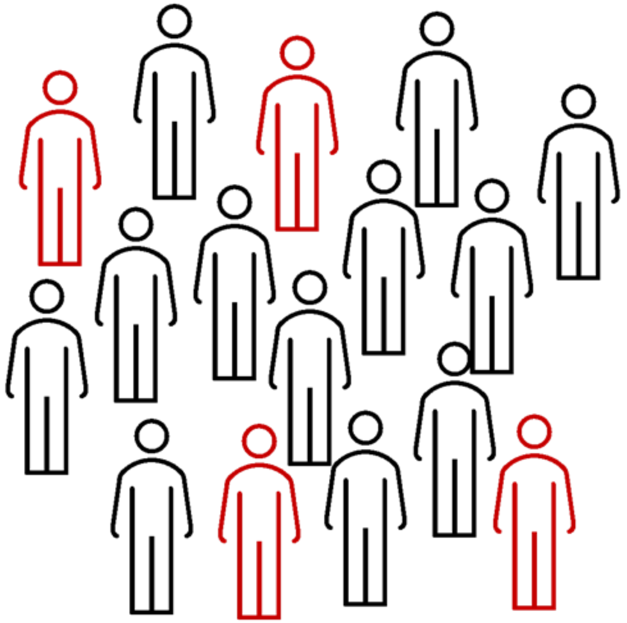
Under	Minutes mobile	Minutes PC	Total
Yes	20	4	24
No	10	6	16
Yes	5	14	19
Yes	26	9	35
No	3	32	35
Yes	14	3	17
No	17	6	23

Complete coverage → **True value:** 40 minutes

Simulating undercoverage bias (RQ3)

Simulation approach

…to then simulate how their estimates would change if some of their information was lost



Under	Minutes mobile	Minutes PC	Total
Yes	0	4	4
No	10	6	16
Yes	0	14	14
Yes	0	9	9
No	3	32	35
Yes	0	3	3
No	17	6	23

Simulated undercoverage → **Biased value:** 18 minutes

→ Difference: 18 minutes = *bias*

Simulating undercoverage bias (RQ3)

Simulating scenarios

- 3 different computer undercoverage scenarios:
  - 25%
  - 50%
  - 75%

With no **computer** covered

- 3 different mobile undercoverage scenarios:
  - 25%
  - 50%
  - 75%

With no **mobile** covered

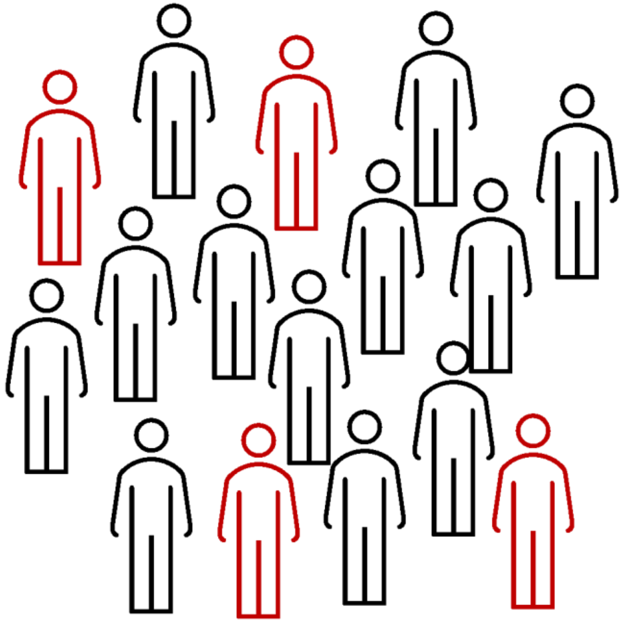
* **In our samples** ➡ % with no PC covered: 34.1 (Spain), 38.1 (Italy) , 27 (Portugal) | % with no mobile: 23.8 (Spain), 28.3 (Italy), 37.7 (Portugal)

Simulating undercoverage bias (RQ3)

Montecarlo simulations

For each scenario, we ran 1,000 random simulations.

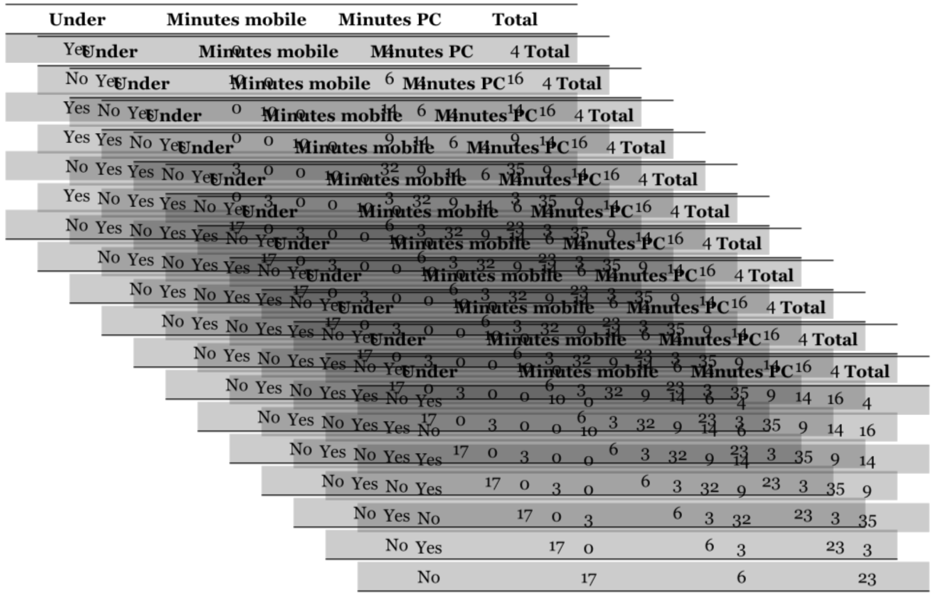
e.g. 25% with no ***computer*** covered ➡ 0.25 probability of being undercovered



Simulating undercoverage bias (RQ3)

Computing the bias

We then computed the *average estimate* of all 1,000 simulations.



Avg. undercovered estimate: 22 minutes
True estimate: 40 minutes
Difference: 18 minutes → *bias*

Simulating undercoverage bias (RQ3)

Simulation approach

We ran simulations for a variety of estimates

Univariate estimates:

- Average time spent on the Internet
- Average time spent on Social Network Sites (SNS)
- Proportion of participants visiting online news media outlets

Multivariate estimates

- Correlation between average time spent on SNS and trust in SNS
- Association between average number of visits to online news media outlets and political knowledge (OLS regression with controls*)

* Age, gender and education

PREVALENCE (RQ 1 & 2)

Proportion undercovered (RQ1)

	Spain	Italy	Portugal
Overall	80.5	83.1	85.7
Device*	69.7	76.1	77.5
Browser	35.1	26.8	39.3

Very high prevalence, with differences between device and browser

* 68% in the Pew Research Centre report, in the USA, using a probability-based panel and a different tracking provider

Is undercoverage evenly distributed across devices? (RQ2)

webdataopp

Proportion of users who use a specific type of device and not all of them are tracked

	Spain	Italy	Portugal
Windows PC	50.5	54.0	49.2
MAC	69.3	78.2	67.2
Android	44.7	47.8	53.1
iOS	93.4	80.9	95.4

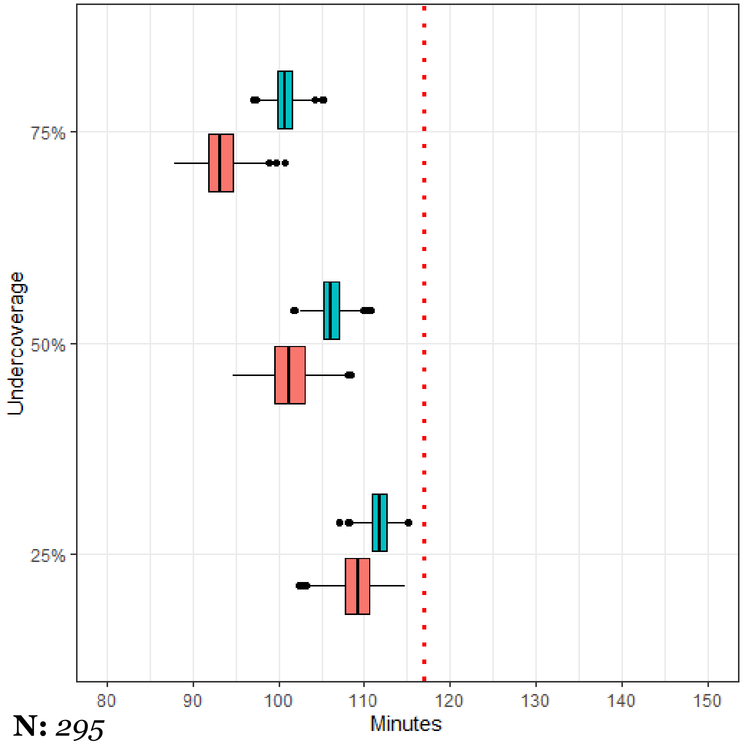
Apple devices present a substantially higher prevalence

SIMULATING BIAS (RQ3)

SIMULATING BIAS

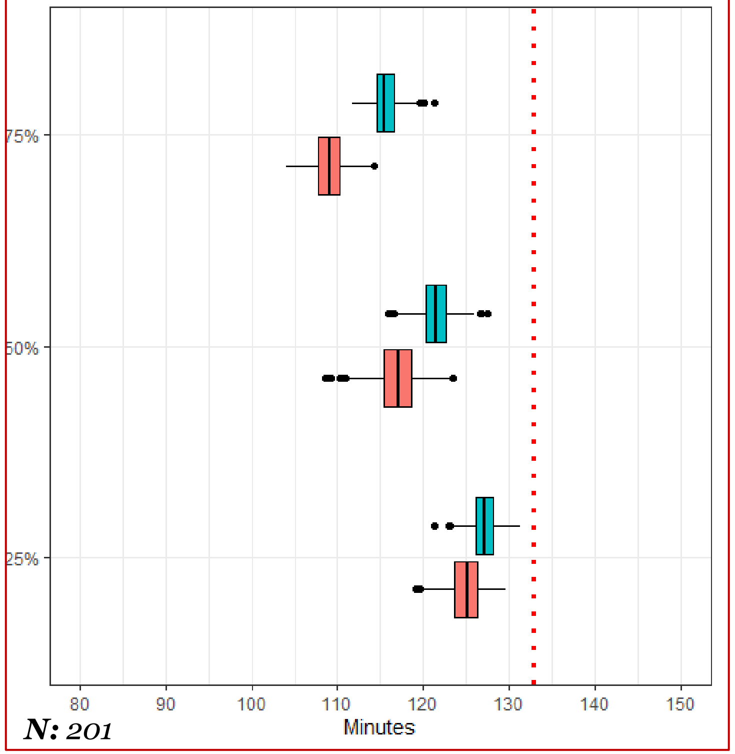
Average time spent on the Internet

SPAIN



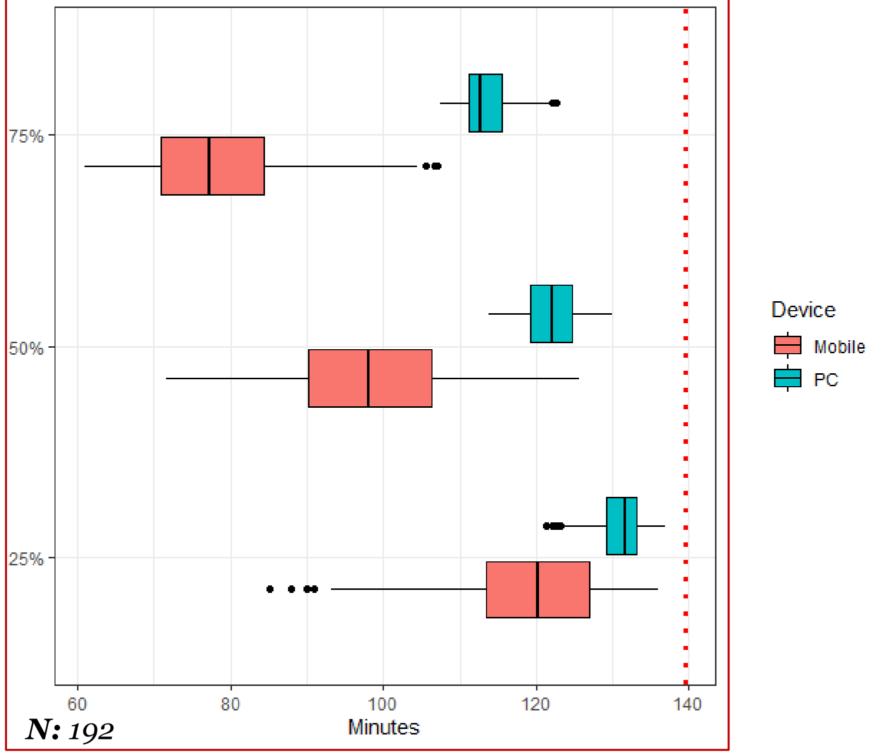
N: *295*

ITALY



N: *201*

PORTUGAL

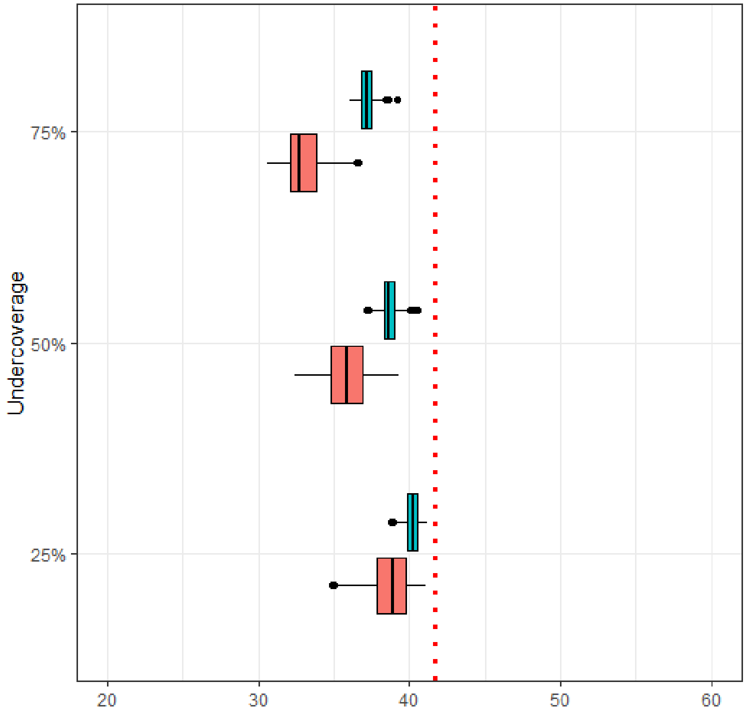


N: *192*

Avg. bias: *5 – 38 minutes* *5 – 23 minutes* *5 – 24 minutes*

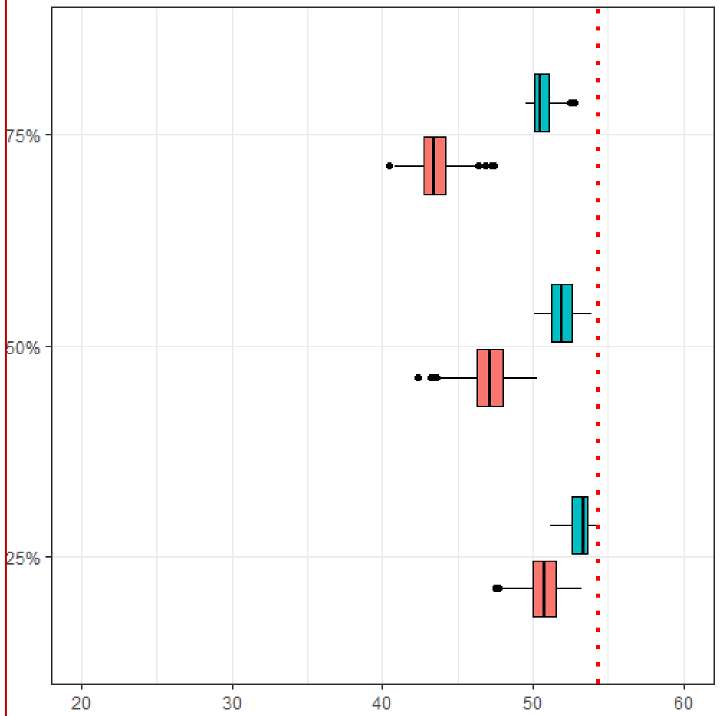
Average time spent on social network sites

SPAIN



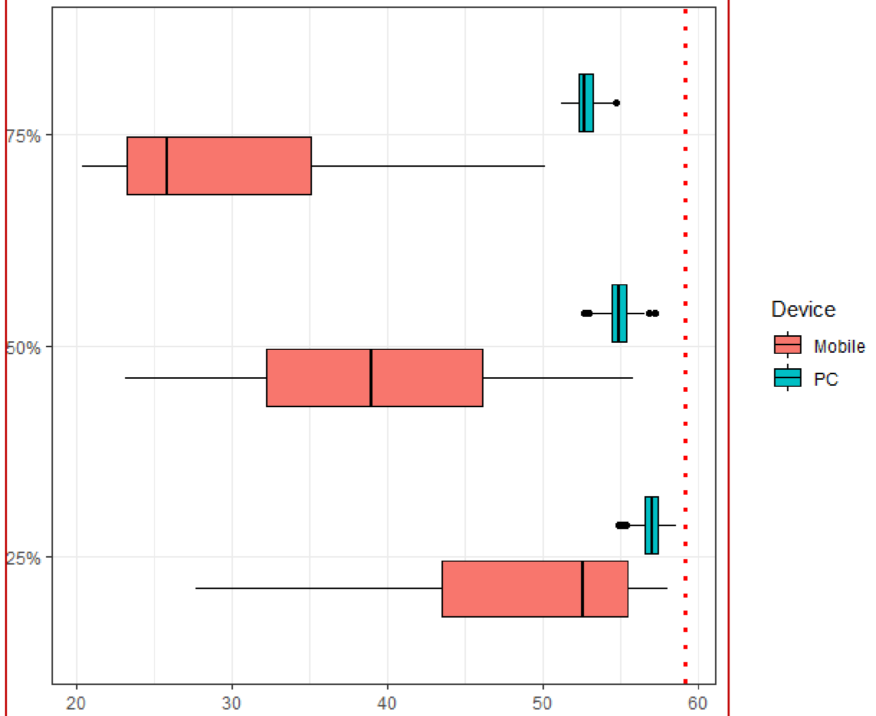
N: *295*

ITALY



N: *201*

PORTUGAL



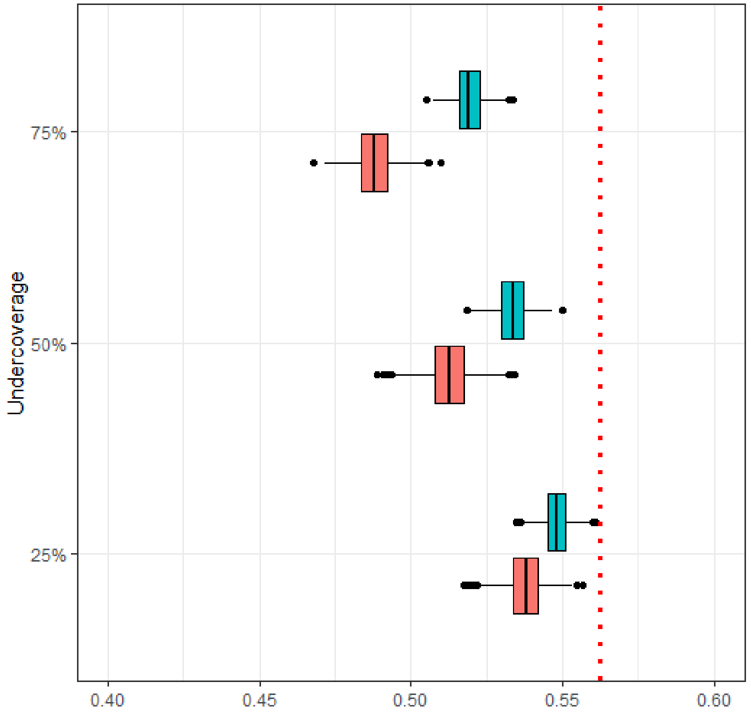
N: *192*

Avg. bias: *1 – 15 minutes* *1 – 11 minutes* *1 – 8 minutes*

SIMULATING BIAS

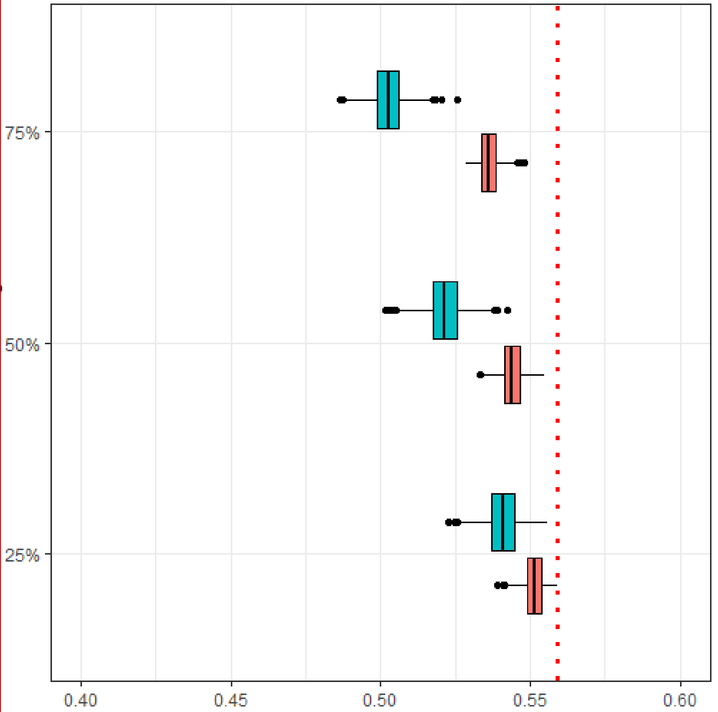
Proportion visiting online news media

SPAIN



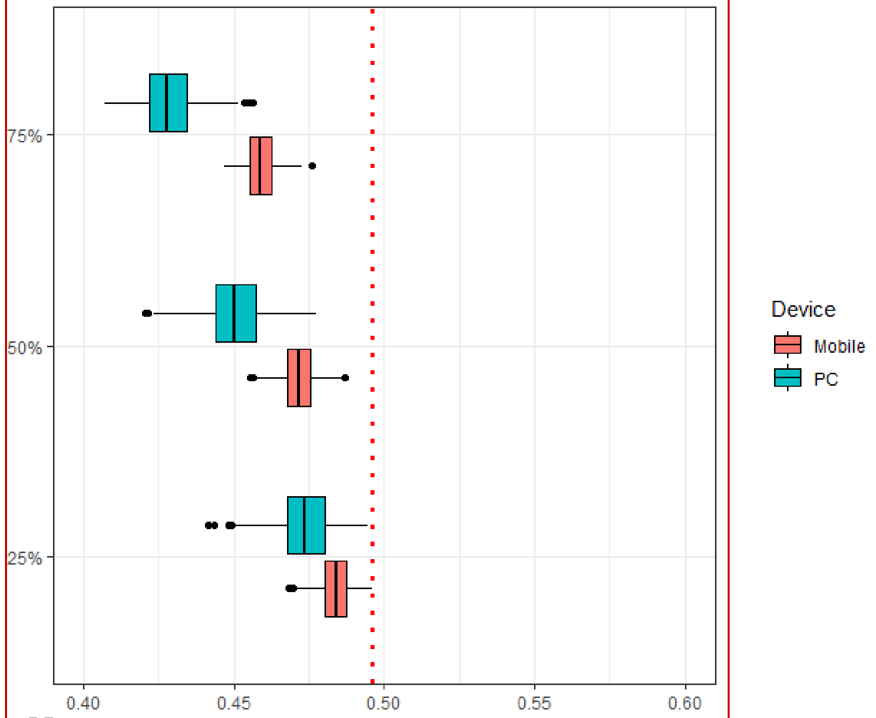
N: *295*

ITALY



N: *201*

PORTUGAL

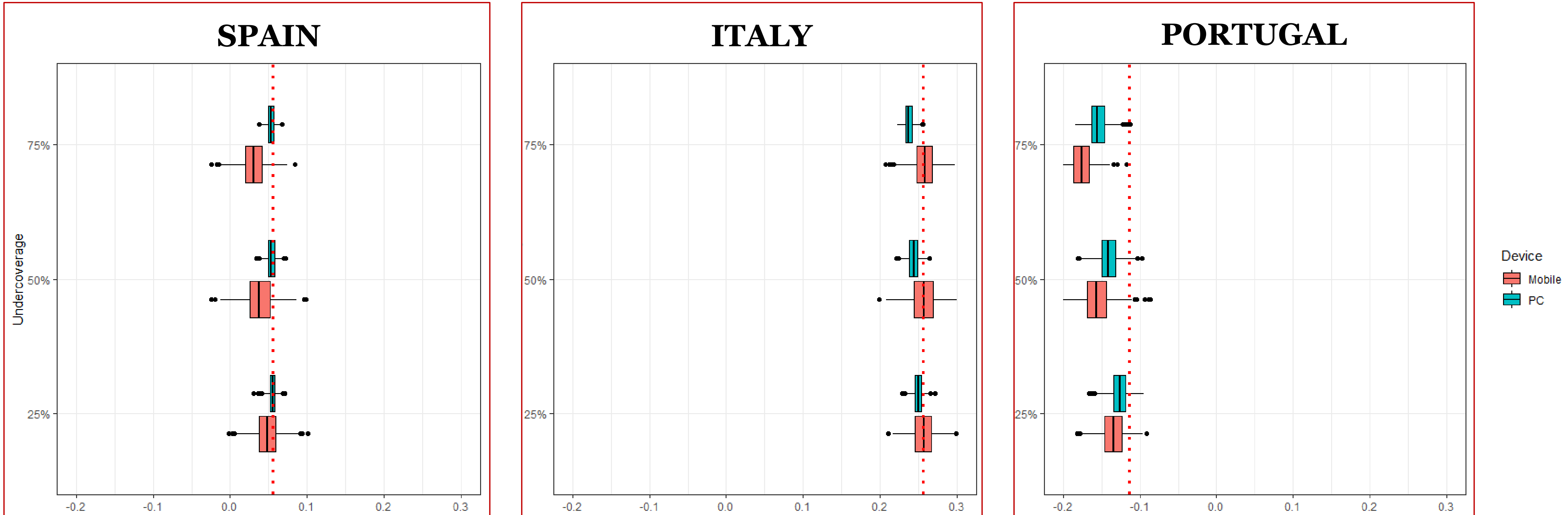


N: *192*

Avg. bias: *1 – 8 % point* *1 – 6 % point* *1 – 7 % point*

SIMULATING BIAS

Correlation between time spent on SNS and trust in SNS



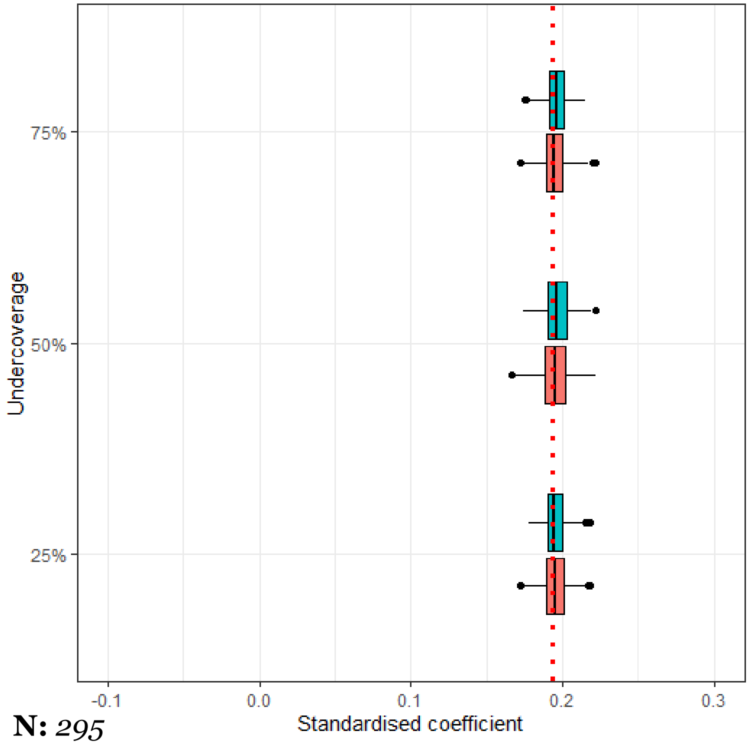
N: *295* **N:** *201* **N:** *192*

Avg. bias: *0.0 - 0.02* *0.0 - 0.02* *0.02 - 0.06*

SIMULATING BIAS

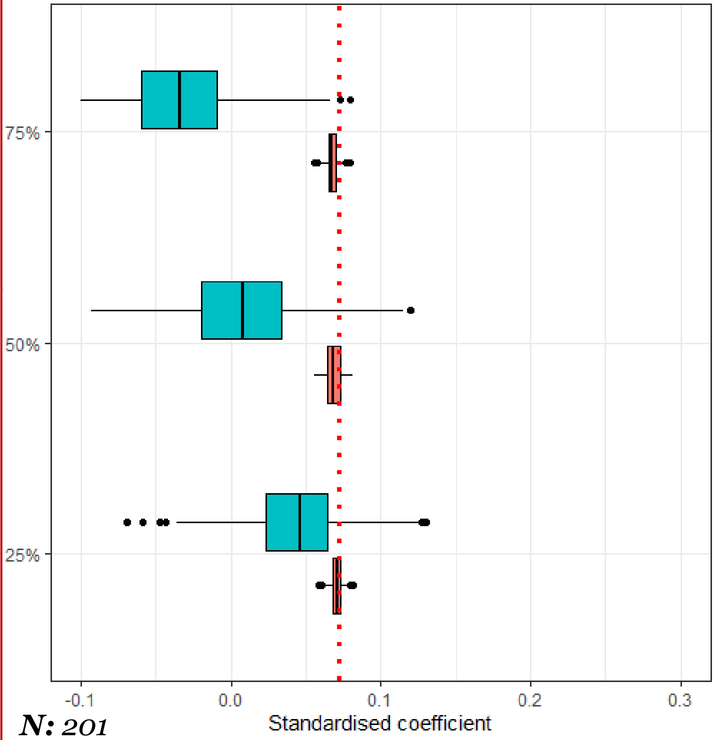
OLS coefficient: Political Knowledge ~ Nº visits to online news

SPAIN



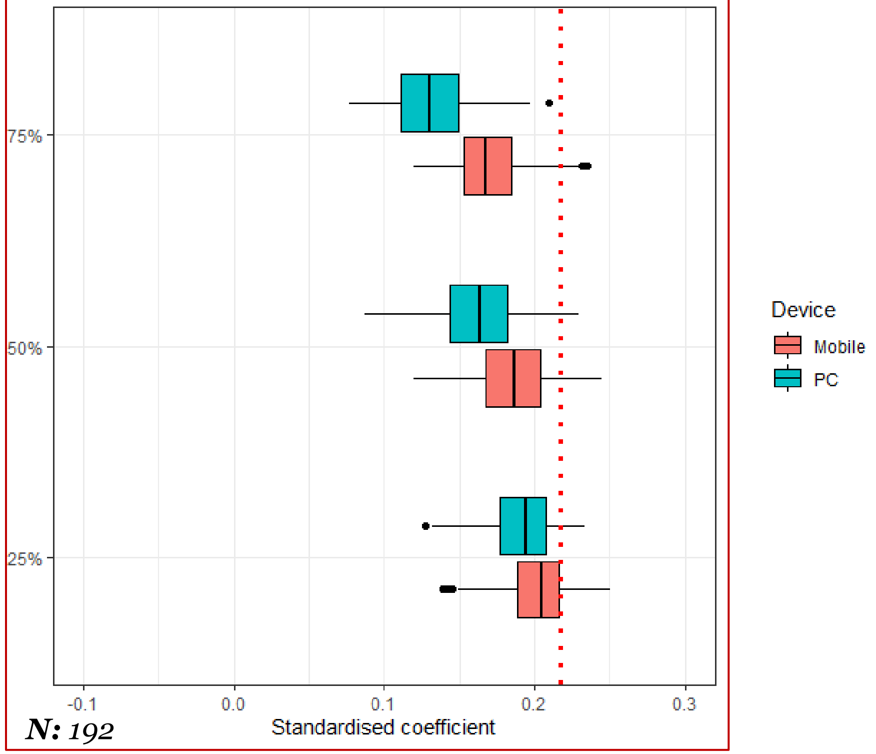
N: *295*

ITALY



N: *201*

PORTUGAL



N: *192*

Avg. bias: *0.002 -0.003* *0.00 – 0.11* *0.01 – 0.09*

* Control variables: age, gender, tertiary education

CONCLUSIONS

Take-home messages

- The prevalence of tracking undercoverage is high, mostly driven by device undercoverage (**RQ1**)
- Apple devices are more likely to be undercovered, specially iPhones and iPads (**RQ2**)
- Tracking undercoverage can bias *both* univariate and multivariate estimates (**RQ3**)
    - Higher undercoverage leads to higher bias
    - The extent varies across topics, as well as devices undercovered.

Take-home messages

- The prevalence of tracking undercoverage is high, mostly driven by device undercoverage (RQ1)
- Apple devices are more likely to be undercovered, specially iPhones and iPads (RQ2)
- Tracking undercoverage can bias *both* univariate and multivariate estimates (RQ3)
  - Higher undercoverage leads to higher bias
  - The extent varies across topics, as well as devices undercovered.

This can be extrapolated to other device-dependant digital trace data

Thanks!

Questions?

Oriol J. Bosch | Department of Methodology, LSE

o.bosch-jover@lse.ac.uk

orioljbosch

https://orioljbosch.com/