Challenging the Gold Standard: A Methodological Study of the Quality and Errors of Web Tracking Data

Oriol Bosch Jover
Oriol.Bosch-Jover@demography.ox.ac.uk

The importance of measuring what people do online

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


We must prevent a vaccine 'infodemic' from fuelling the Covid pandemic
Melinda Mills

Wise governments will take a leaf out of the anti-vaxxers' book by creating campaigns that persuade through engagement


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'.

Web tracking data to understand online behaviours

- Survey self-reports are still the **most common approach**

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

But they might be affected by many errors

Web tracking data to understand online behaviours

- Survey self-reports are still the **most common approach**

- More and more availability of **digital traces to directly observe media exposure**

Individual-level approach: web trackers

Direct observations of online behaviours using tracking solutions, or *meters.*

Group of tracking technologies (plug-ins, apps, proxies, etc)

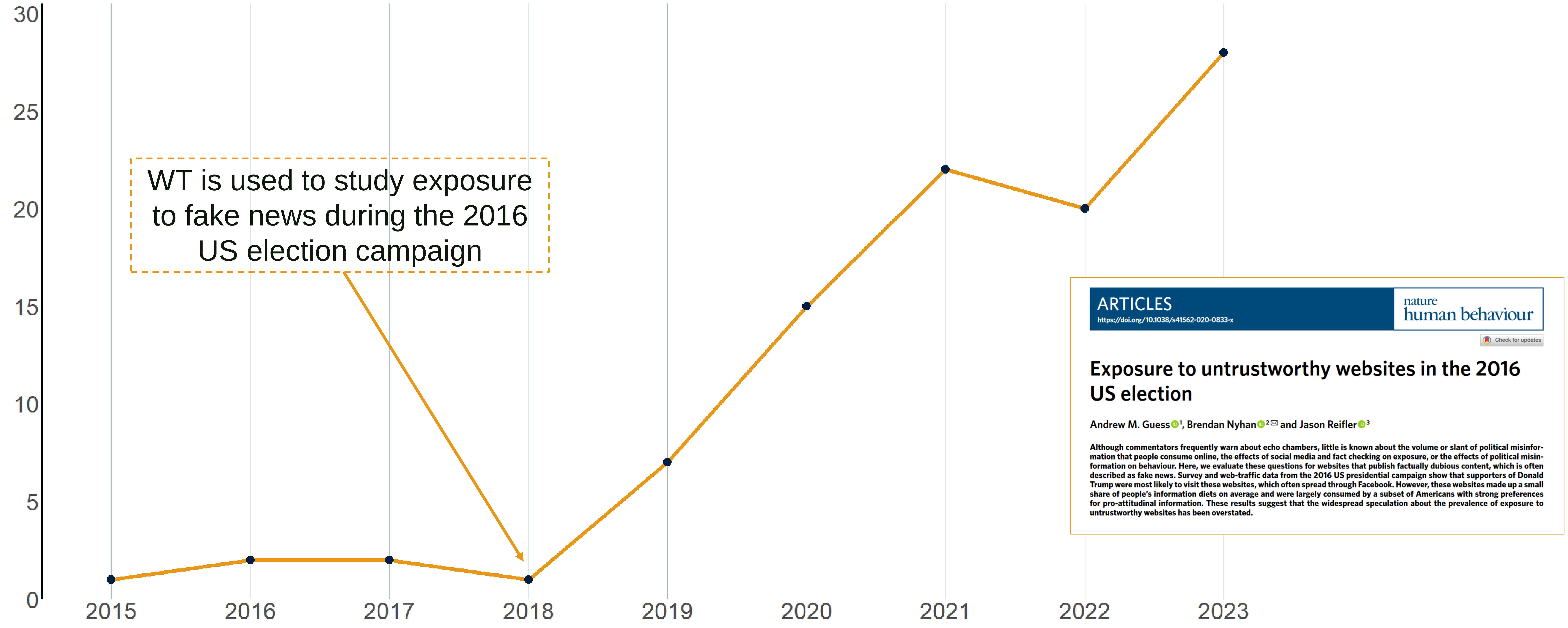
Installed on participants devices

Collect traces left by participants when **interacting with their devices online: URLs, apps visited, cookies…**

The rise of web tracking data

Web tracking data is becoming part of social scientists' toolkit

Number of papers using web tracking data



What is this thesis about?

A critical assessment of the quality of web tracking data

While web tracking data enjoys gold standard status, no evidence supports this

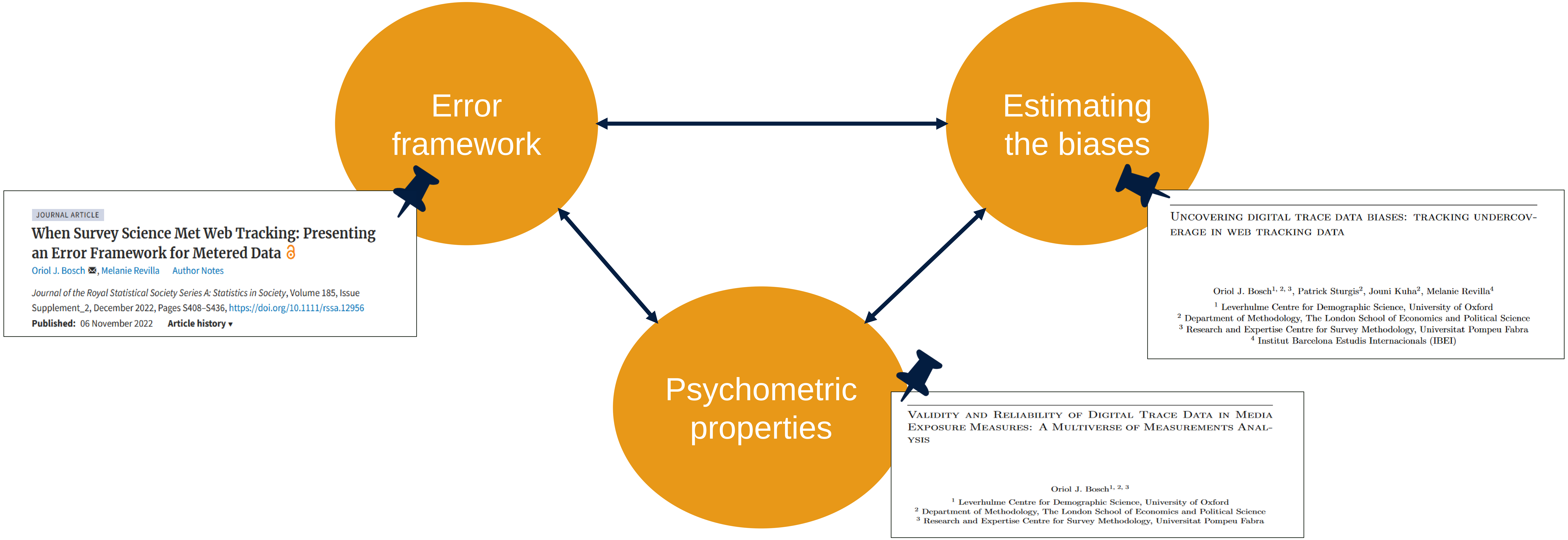
- What is the quality of web tracking data?
- Its errors?
- What are the best practices when using this data?

> "The development of a sophisticated measurement theory is a precondition for digital trace data to be meaningfully integrated into the social sciences."

Jungherr, A. (2018). Normalizing Digital Trace Data. In *Digital Discussions* (pp. 9-35). Routledge.

A multidimensional exploration of web tracking data

I combine **survey** and **computational** methods to understand how we can improve the use of **web tracking data** in the social sciences



TRI-POL: the triangle of polarization

Three wave survey combined with **web tracking data** at the individual level (both PC and mobile data)

Netquest metered panels

- **Cross-quotas:** gender, age, education and region
- **Sample size:** ≈1,200 per country

Spain, Portugal, **Italy, Argentina and Chile**

ELSEVIER

Data in Brief

Available online 9 May 2023, 109219

In Press, Journal Pre-proof What's this?

Data Article

The dynamics of political and affective polarisation: Datasets for Spain, Portugal, Italy, Argentina, and Chile (2019-2022)

Mariano Torcal [1], Emily Carty [2], Josep Maria Comellas [3], Oriol J. Bosch [4], Zoe Thomson [1], Danilo Serani [2]

An error framework for web tracking data

Received: 11 February 2021 | Accepted: 25 August 2022
DOI: 10.1111/rssa.12956

ORIGINAL ARTICLE

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

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

[1]Department of Methodology, The London School of Economics and Political Science, London, UK

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

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

Funding information
Fundación BBVA; H2020 European Research Council, Grant/Award Number: 849165; Ministerio de Ciencia e Innovación, Grant/Award Number: PID2019-106867RB- I00 /AEI/10.13039/501100011033

Abstract

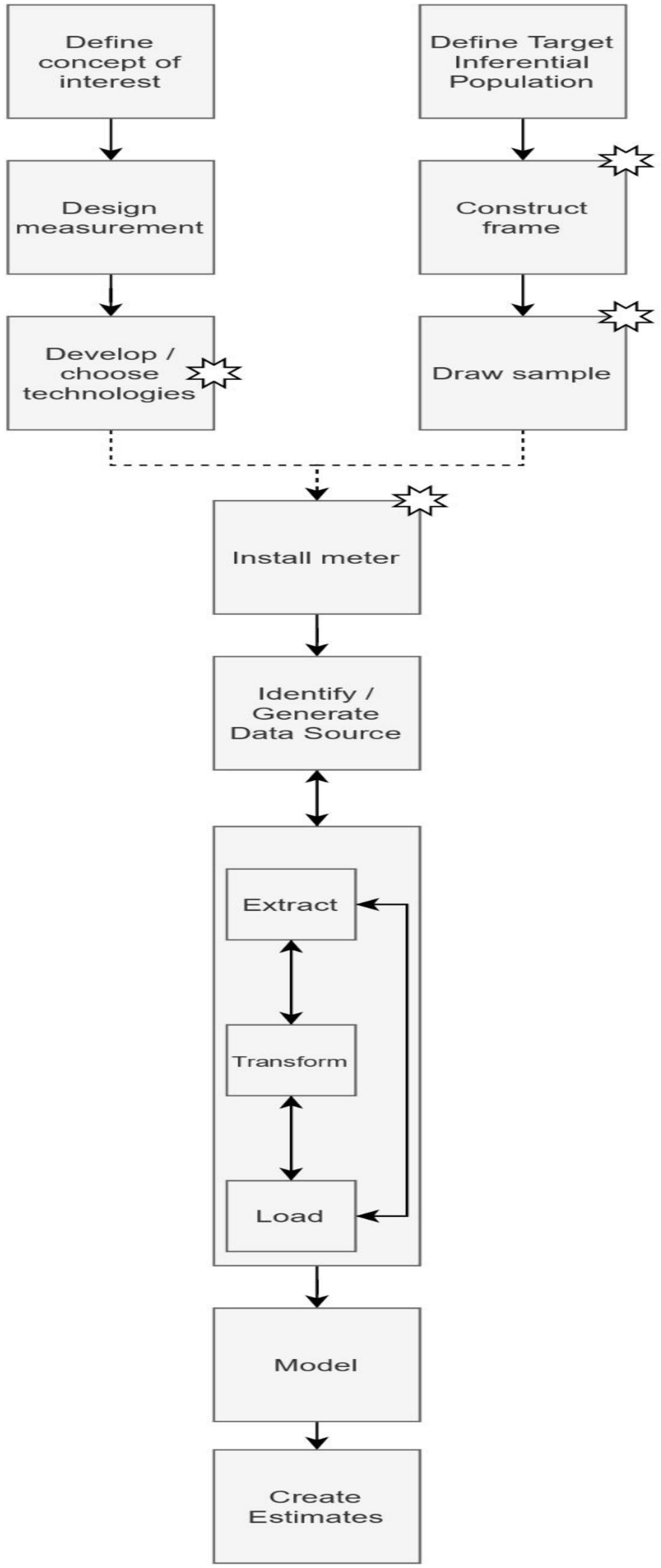
Metered data, also called web-tracking data, are generally collected from a sample of participants who willingly install or configure, onto their devices, technologies that track digital traces left when people go online (e.g., URLs visited). Since metered data allow for the observation of online behaviours 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 have 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. Using a case study we also show how the TEM can be applied in real life to identify, quantify and reduce metered data errors. Results suggest that metered data might indeed be affected by the error sources identified in our framework and, to some extent, biased. This framework can help improve the quality of both stand-alone metered data research projects, as well as

A complex and burdensome process

- In general, web tracking data is used to **make inferences** about a **concept of interest** for a given **population.**

- There are **many steps** to follow when collecting web tracking data.

- Many decisions can be made for each step, all with **potential impact on data quality.**



Error components	Specific error causes
Specification error	– Defining what qualifies as valid information – Measuring concepts with by-design missing data – Inferring attitudes and opinions from behaviours
Measurement error	– Tracking undercoverage – Technology limitations – Technology errors – Hidden behaviours – Social desirability – Extraction errors – Misclassifying non-observations – Shared devices
Processing error	– Coding error – Aggregation at the domain level – Data anonymization
Coverage error	– Non-trackable individuals
Sampling error	– Same error causes as for surveys
Missing data error	– Non-contact – Non-consent – Tracking undercoverage – Technology limitations – Technology errors – Hidden behaviours – Social desirability – Extraction errors – Misclassifying non-observations
Adjustment error	– Same error causes than for surveys

This can lead to many errors

- There are many specific problems that can introduce errors to web tracking data

Error components	Specific error causes
Specification error	– Defining what qualifies as valid information – Measuring concepts with by-design missing data – Inferring attitudes and opinions from behaviours
Measurement error	– Tracking undercoverage – Technology limitations – Technology errors – Hidden behaviours – Social desirability – Extraction errors – Misclassifying non-observations – Shared devices
Processing error	– Coding error – Aggregation at the domain level – Data anonymization
Coverage error	– Non-trackable individuals
Sampling error	– Same error causes as for surveys
Missing data error	– Non-contact – Non-consent – Tracking undercoverage – Technology limitations – Technology errors – Hidden behaviours – Social desirability – Extraction errors – Misclassifying non-observations
Adjustment error	– Same error causes than for surveys

This can lead to many errors

- There are many specific problems that can introduce errors to web tracking data
- Most are on the side of measurement
- The representation side is quite similar to surveys

Uncovering the biases of web tracking data

UNCOVERING DIGITAL TRACE DATA BIASES: TRACKING UNDERCOVERAGE IN WEB TRACKING DATA

Oriol J. Bosch[1, 2, 3], Patrick Sturgis[2], Jouni Kuha[2], Melanie Revilla[4]

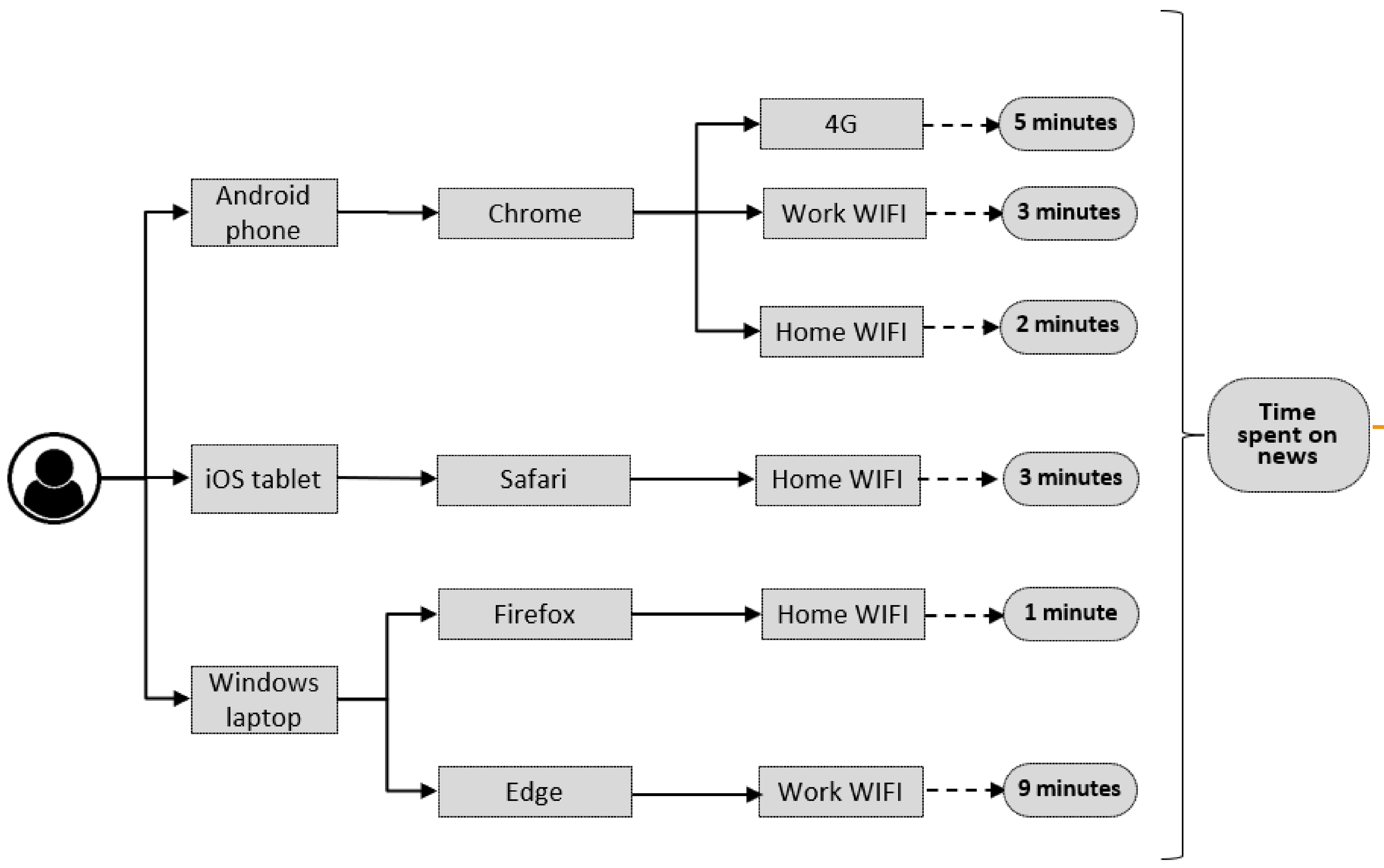
[1] Leverhulme Centre for Demographic Science, University of Oxford
[2] Department of Methodology, The London School of Economics and Political Science
[3] Research and Expertise Centre for Survey Methodology, Universitat Pompeu Fabra
[4] Institut Barcelona Estudis Internacionals (IBEI)

Abstract

In the digital age, understanding people's online behaviours is vital. Digital trace data has emerged as a popular alternative to surveys, many times hailed as the gold standard. This study critically assesses the use of web tracking data to study online media exposure. Specifically, we focus on a critical error source of this type of data, tracking undercoverage: researchers' failure to capture data from all the devices and browsers that individuals utilize to go online. Using data from Spain, Portugal, and Italy, we explore undercoverage in commercial online panels and simulate biases in online media exposure estimates. The paper shows that tracking undercoverage is highly prevalent when using commercial panels, with more than 70% of participants affected. In addition, the primary determinant of undercoverage is the type and number of devices employed for internet access, rather than individual characteristics and attitudes. Additionally, through a simulation study, it demonstrates that web tracking estimates, both univariate and multivariate, are often substantially biased due to tracking undercoverage. This represent the first empirical evidence demonstrating that web tracking data is, effectively, biased. Methodologically, the paper showcases how survey questions can be used as auxiliary information to identify and simulate web tracking errors.

Keywords:
Digital trace data · Web tracking data · Undercoverage · Bias · Media exposure · Monte Carlo simulation

Tracking undercoverage



Objective: measuring individuals' behaviours.

Reality: we only measure what we can manage to track.

True: 23m
Observed: 11m
Measurement error: -12m

Identifying who is not fully tracked

The devices they say they use through self-reports

−

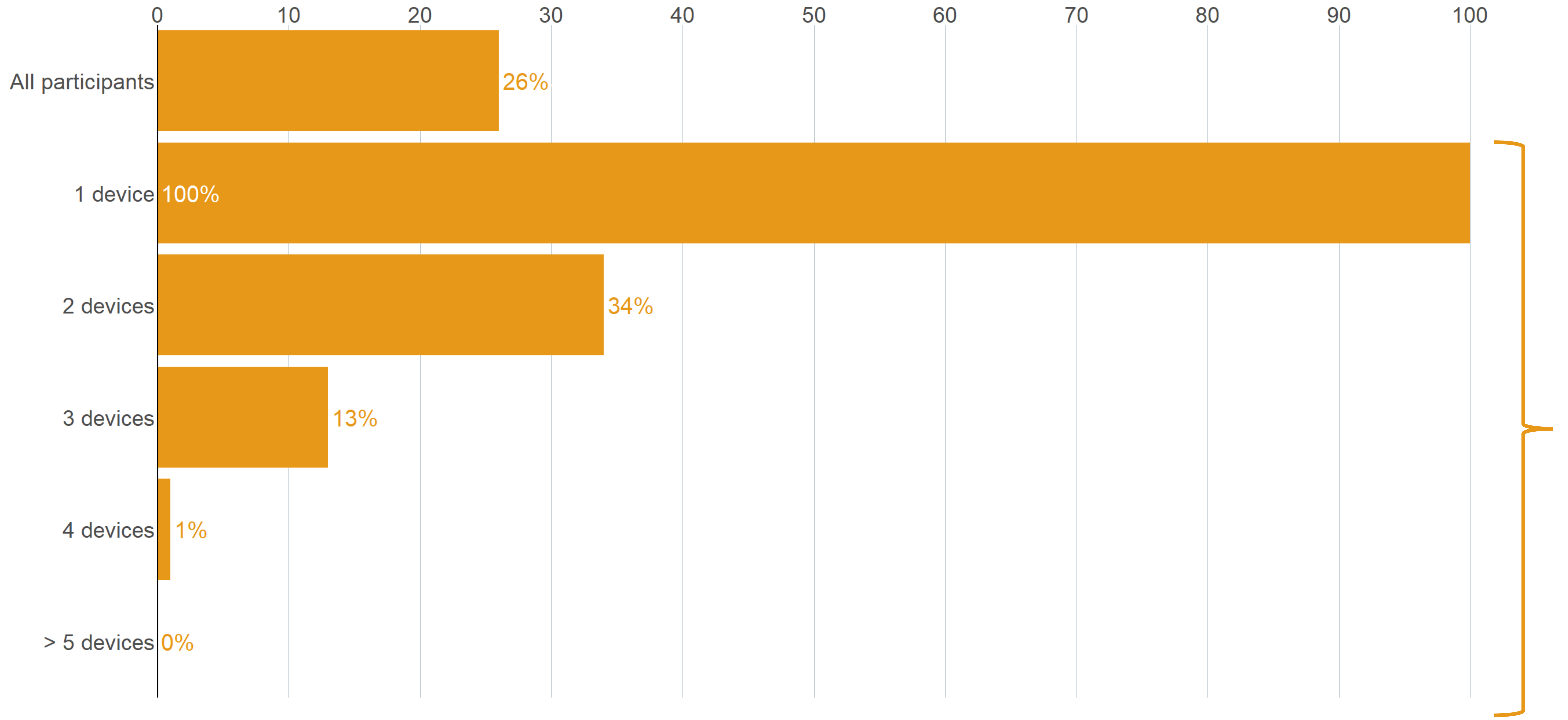
The devices we know we track through paradata

The devices we estimate they have undercovered

How big of a problem is this?

Most people do not have all their devices covered, especially those who use many

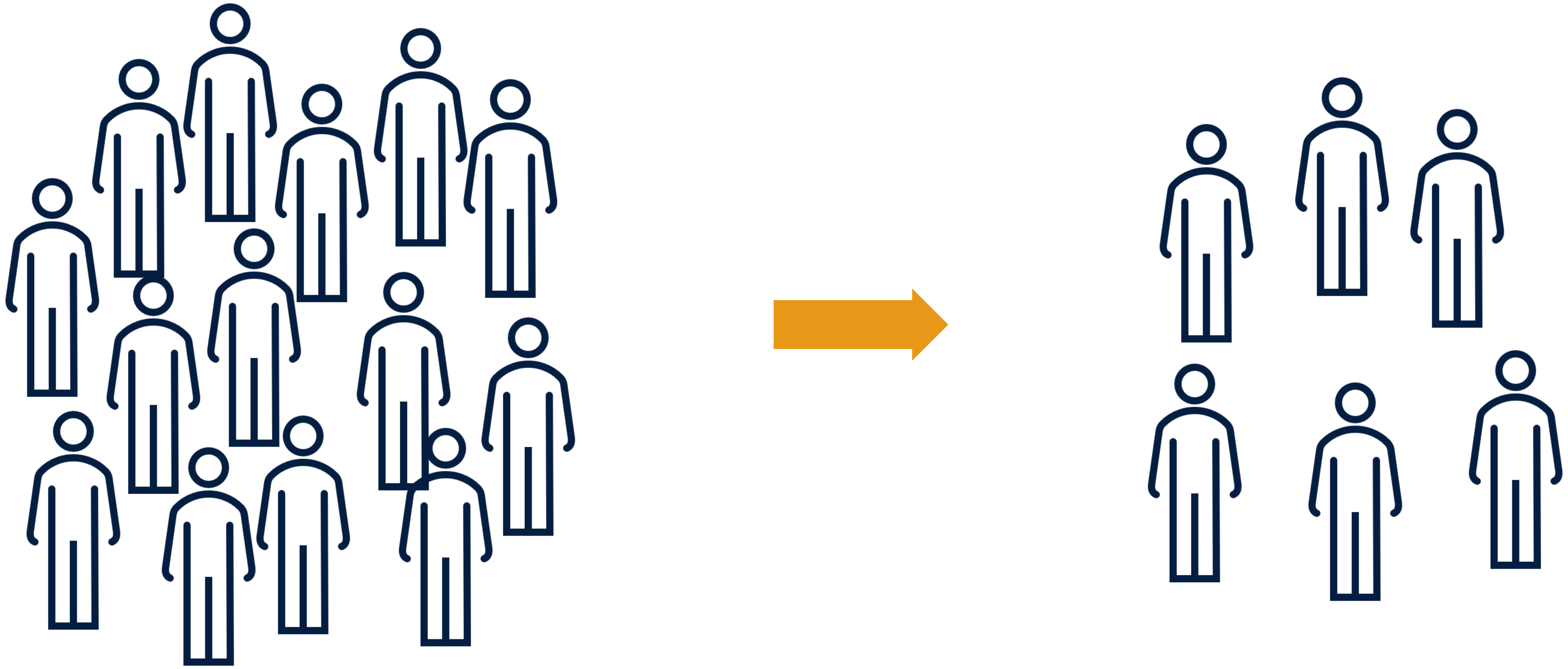
% of people with all their devcies covered



The more devices people report using, the less likely it is for them to be fully covered

Simulating undercoverage bias

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



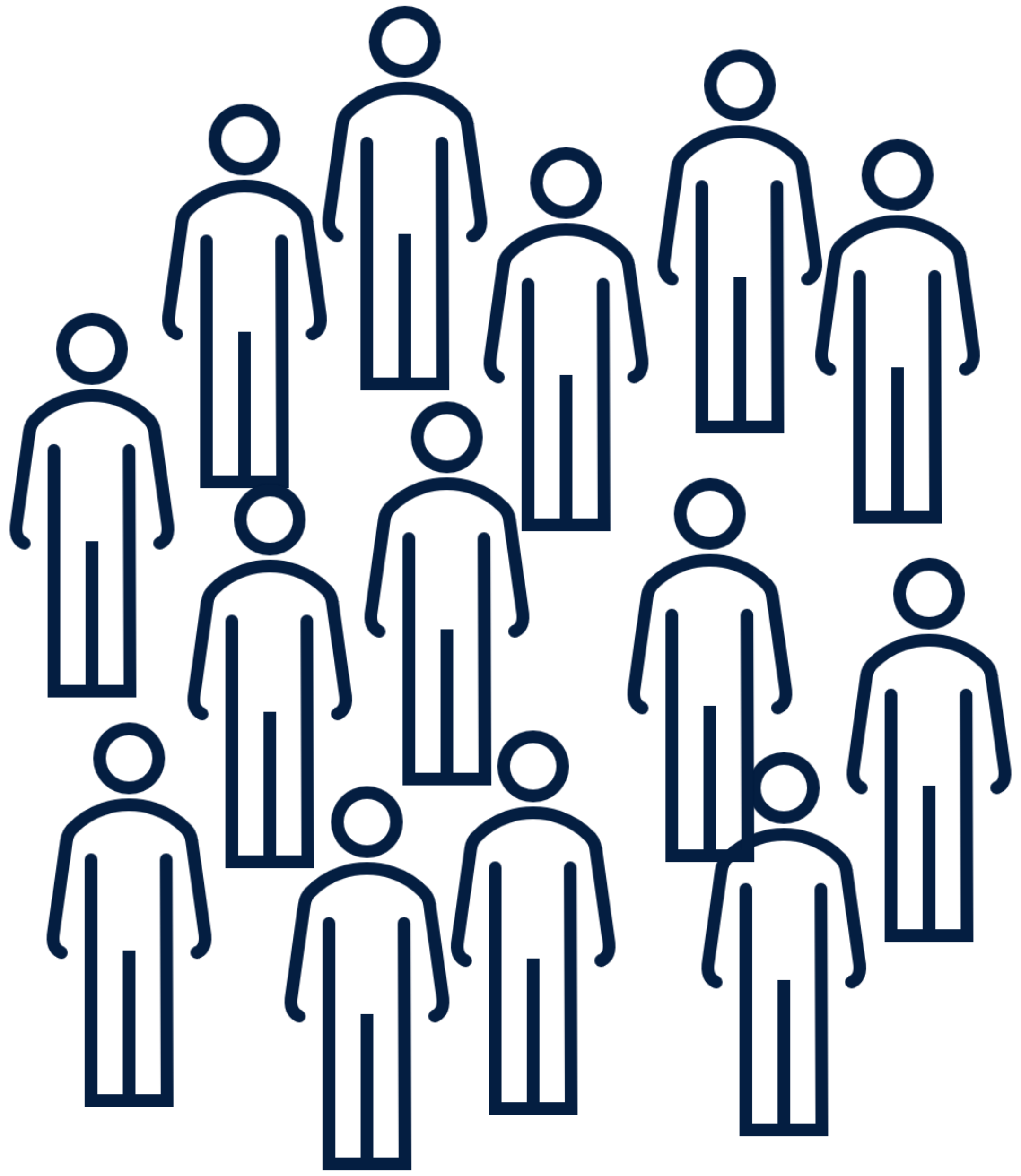
Full sample

Fully covered sample

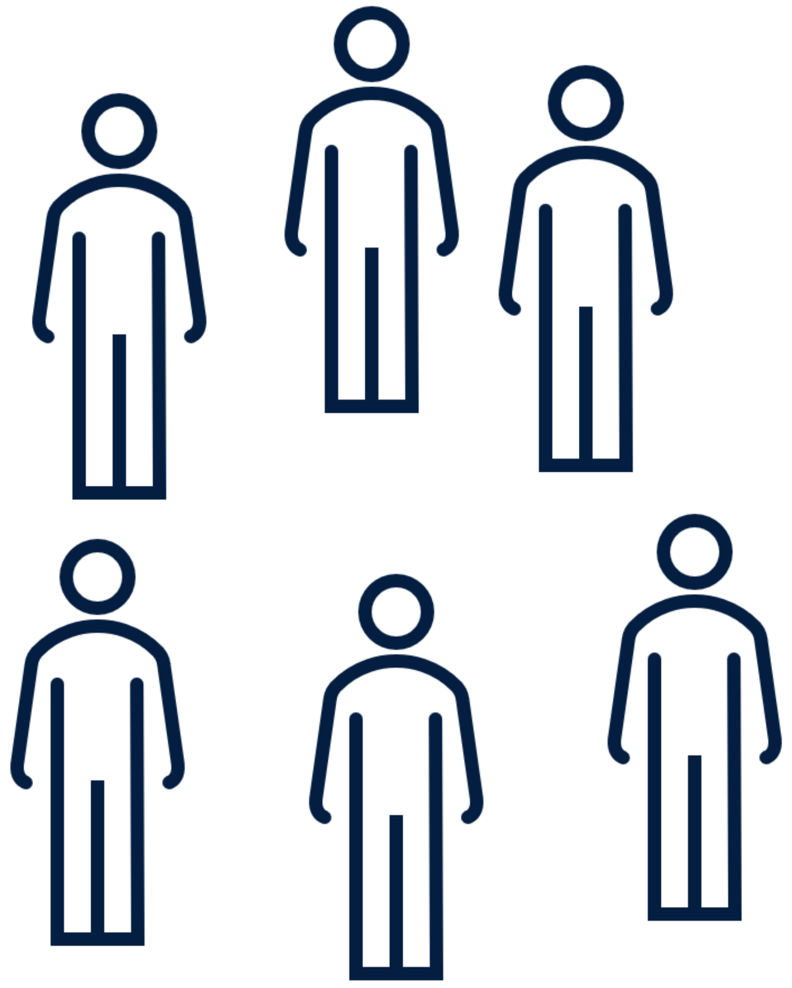
Simulating undercoverage bias

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

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



Full sample

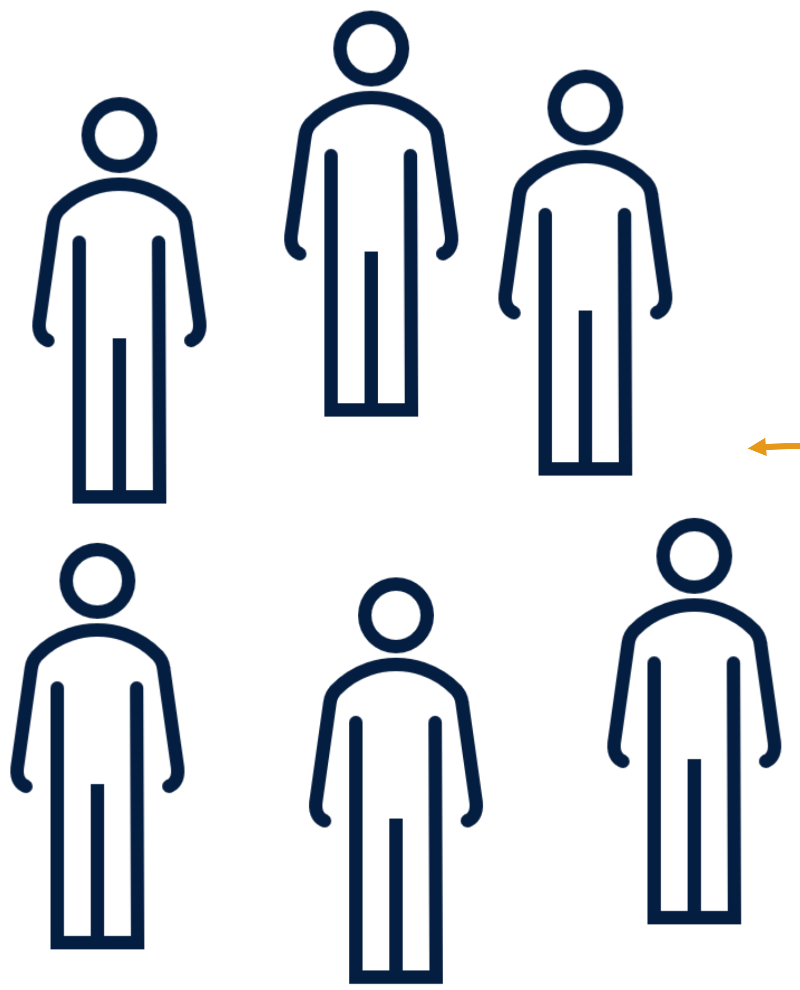


Fully covered sample

Simulating undercoverage bias

Simulation approach

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



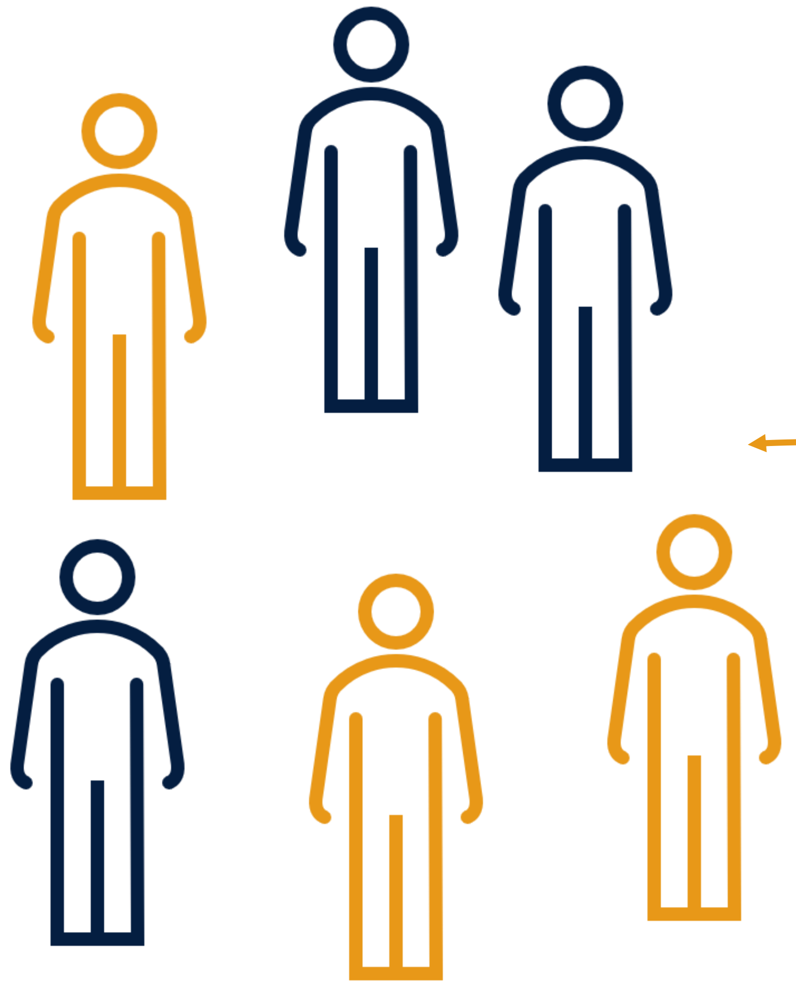
Under	Minutes mobile	Minutes PC	Total
Yes	20	4	24
No	10	6	16
Yes	5	14	19
Yes	26	9	35
No	3	32	35
Yes	14	3	17
No	17	6	23

Complete coverage → **True value:** 169 minutes

Simulating undercoverage bias

Simulation approach

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



…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

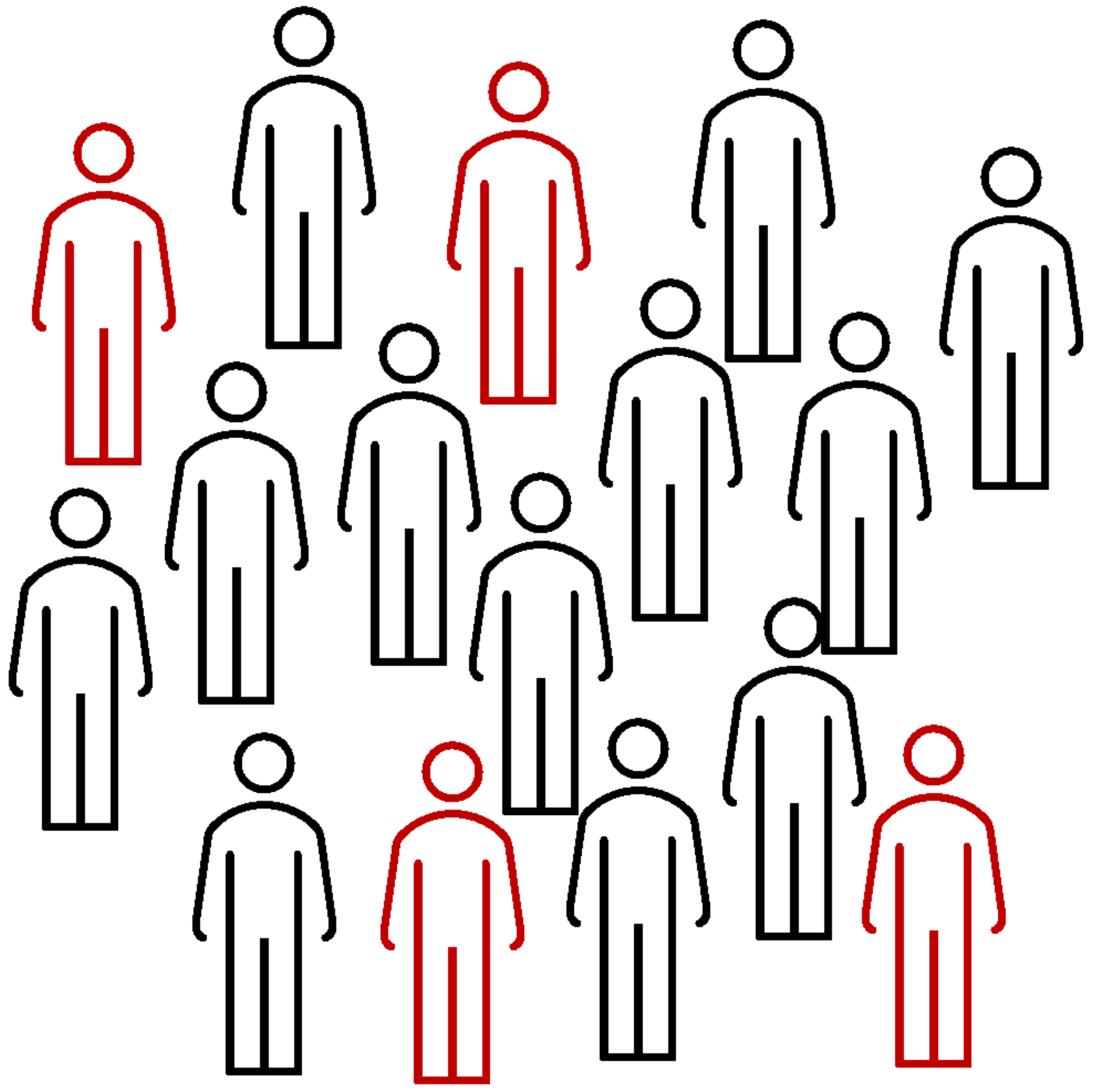
Complete coverage → **Biased value:** 104 minutes

Simulating undercoverage bias

Monte Carlo simulations

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

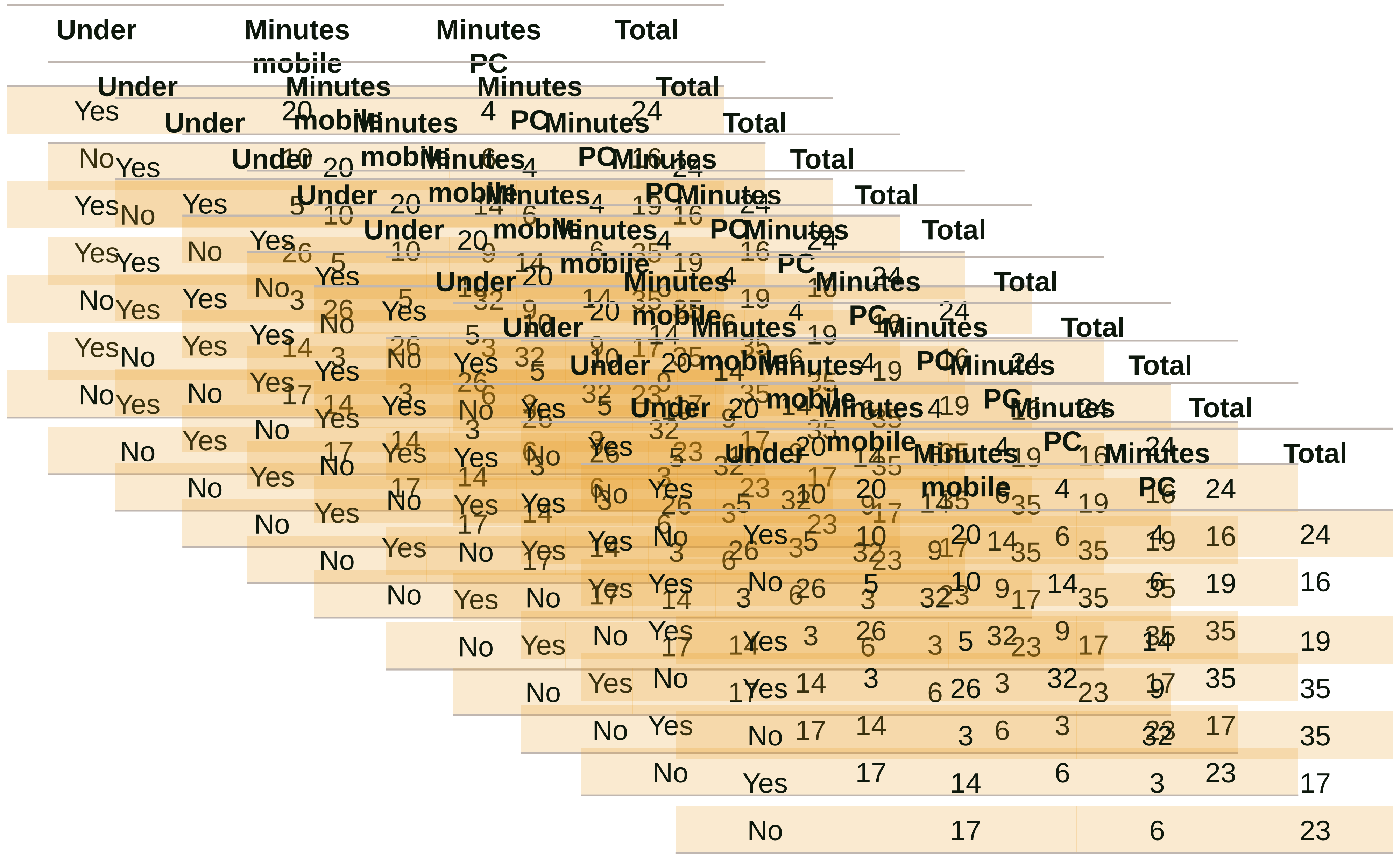
For example, in each simulation a participant has a 0.25 chance of losing all their mobile devices



Simulating undercoverage bias

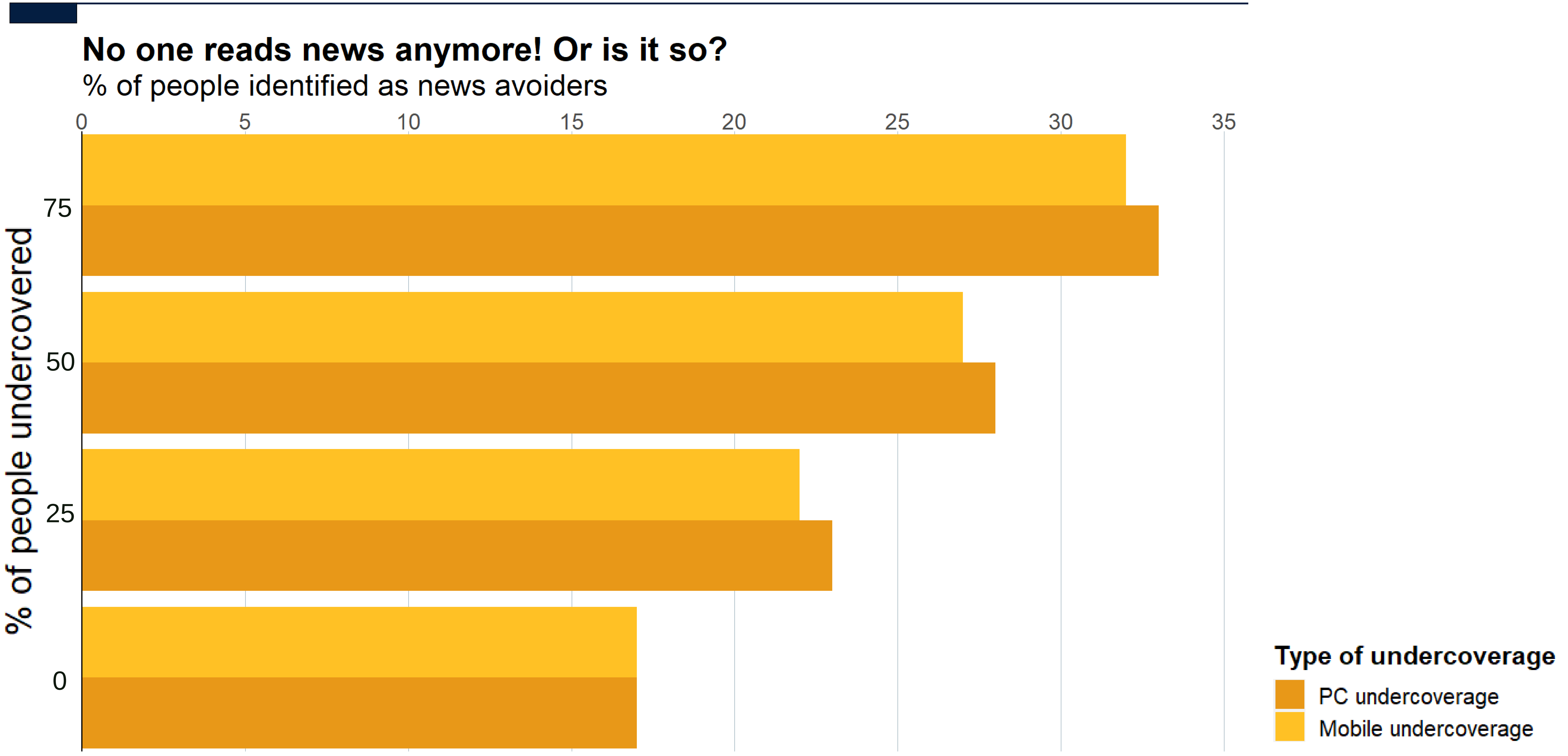
Computing the bias

We then compute the average estimate of all 1,000 simulations



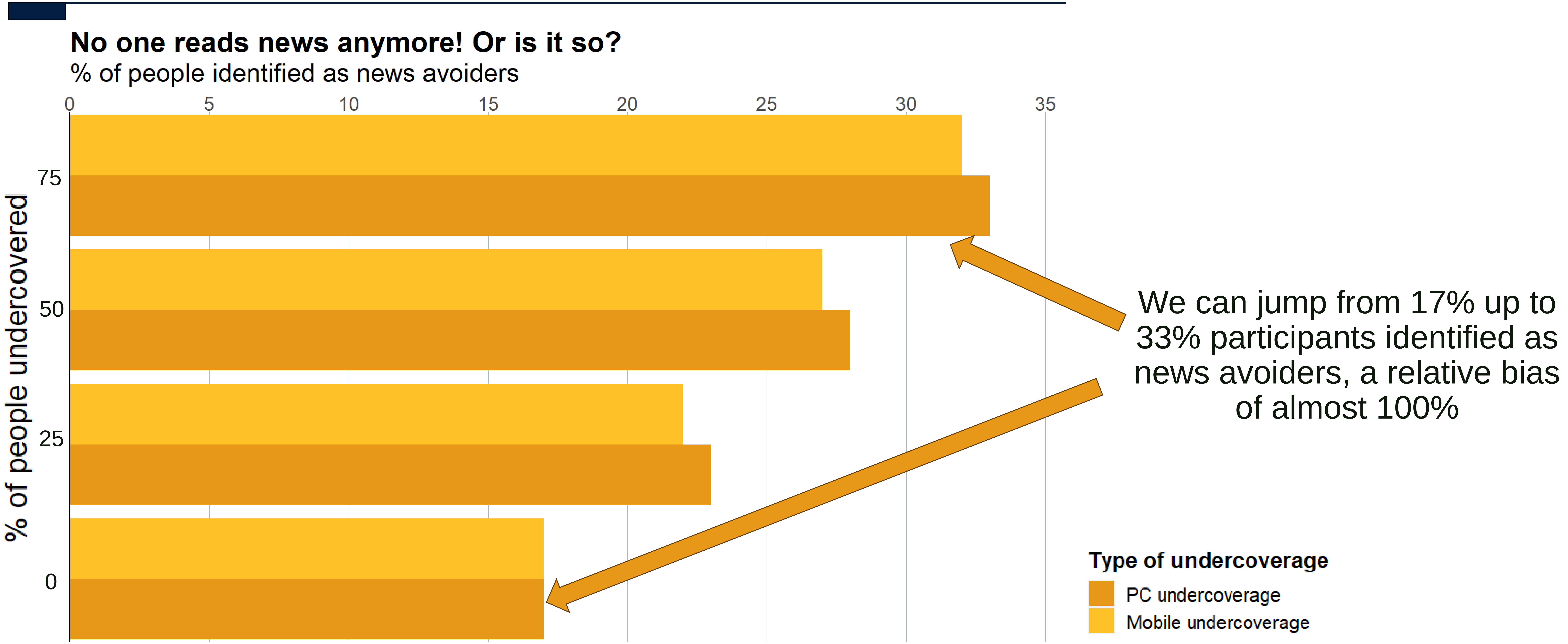
Avg. undercovered estimate: 22 minutes
True estimate: 40 minutes
Difference: 18 minutes *bias*

What are the sizes of the simulated biases?



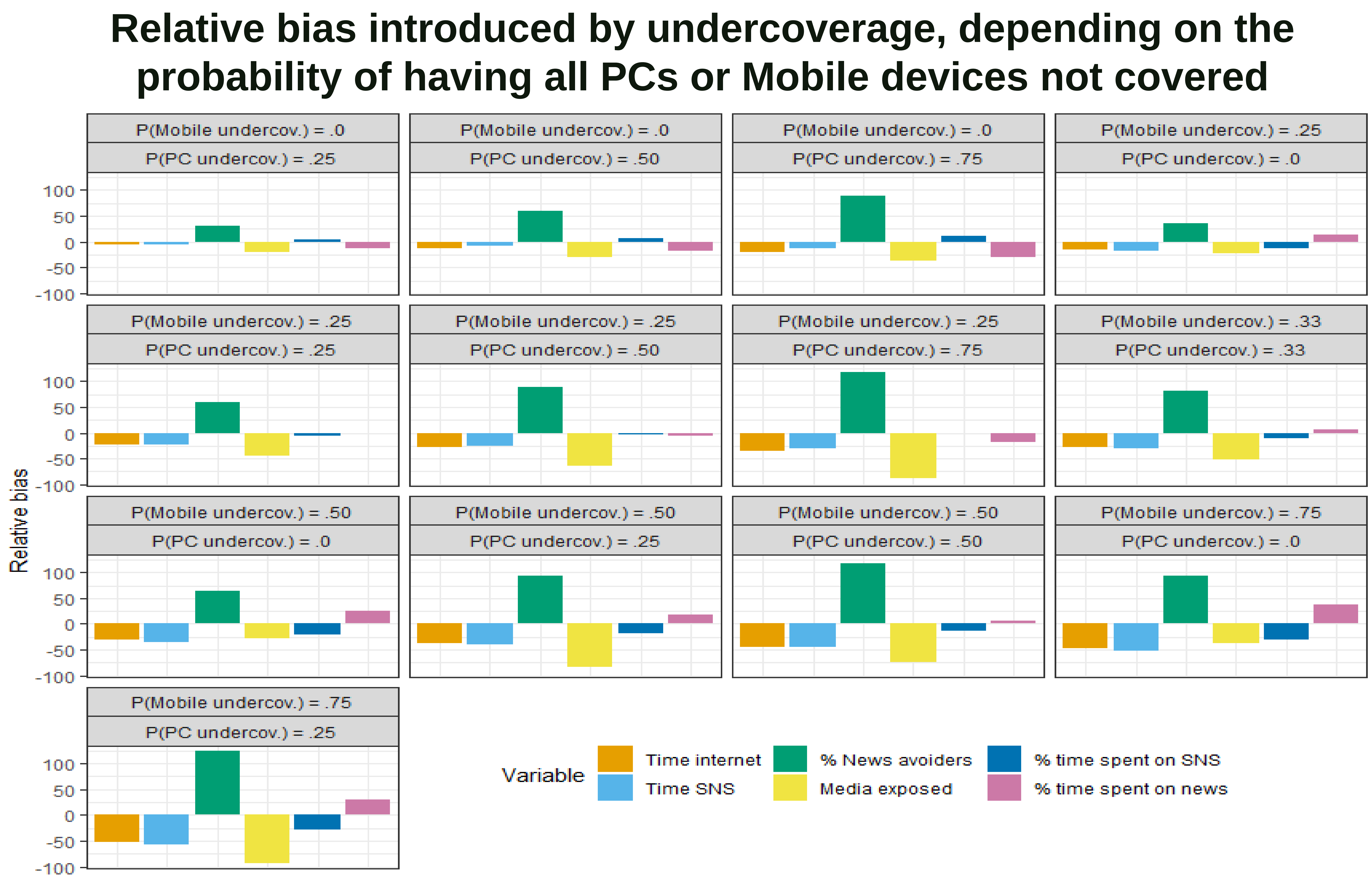
Sources: Uncovering digital trace data biases: tracking undercoverage in web tracking data

What are the sizes of the simulated biases?



Sources: Uncovering digital trace data biases: tracking undercoverage in web tracking data

This happens across statistics and scenarios



What are the psychometric properties of web tracking measures?

Validity and Reliability of Digital Trace Data in Media Exposure Measures: A Multiverse of Measurements Analysis

Oriol J. Bosch[1, 2, 3]

[1] Leverhulme Centre for Demographic Science, University of Oxford
[2] Department of Methodology, The London School of Economics and Political Science
[3] Research and Expertise Centre for Survey Methodology, Universitat Pompeu Fabra

Abstract

Given the doubts about survey self-reports, media exposure research has turned to web tracking data. However, web tracking data is also biased. To improve the understanding of the quality of web tracking measures of media exposure, this paper estimates their validity and reliability. It additionally identifies design choices to optimize these. Using data from a cross-national three-wave survey, combined with web tracking, this paper conducts a multiverse analysis to assess the validity and reliability of +2,500 measures of media exposure. Results show an overall high reliability (0.86). In terms of predictive validity, the association between media exposure measures and political knowledge appears weak. This raises questions about the predictive validity of web tracking measures, and previous critiques to surveys self-reports. Additionally, results suggest that design choices impact the quality of web tracking measures. Methodologically, the paper presents a multiverse of measurements approach, improving the transparency of web tracking research.

Keywords:
Digital trace data · Web tracking data · Media exposure · Reliability · Validity · Multiverse analysis

Web tracking measures are biased, but how valid and reliable are they?

- To decide when to use web tracking data or surveys to measure specific concepts, it is helpful to compare their psychometric properties

Exposure, Attention, or “Use” of News? Assessing Aspects of the Reliability and Validity of a Central Concept in Political Communication Research

William P. Eveland, Jr. and Myiah J. Hutchens
Ohio State University

Fei Shen
City University of Hong Kong

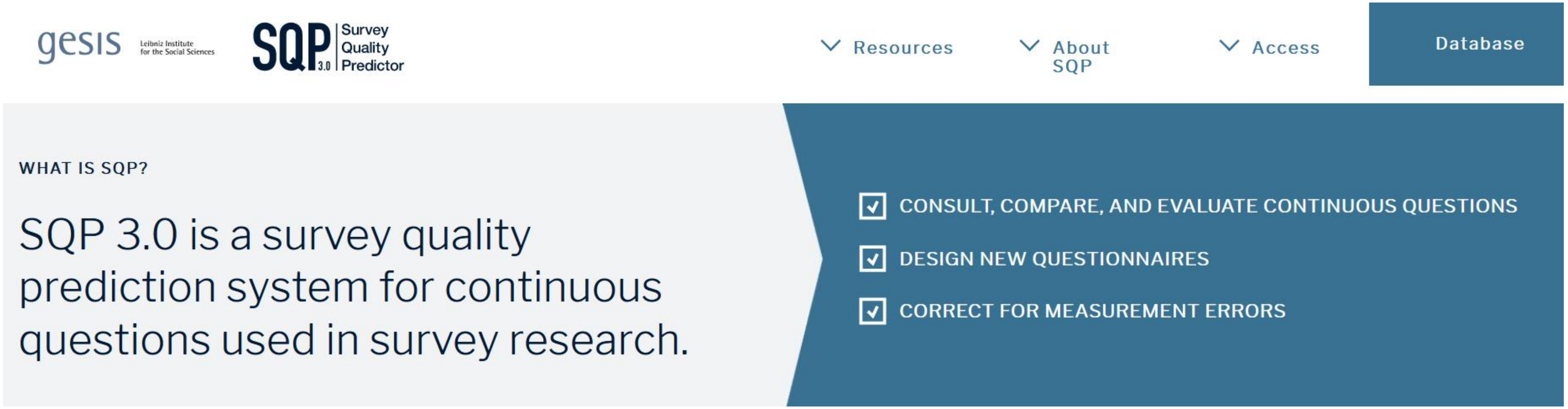
COMPARING ESTIMATES OF NEWS CONSUMPTION FROM SURVEY AND PASSIVELY COLLECTED BEHAVIORAL DATA

TOBIAS KONITZER
JENNIFER ALLEN
STEPHANIE ECKMAN
BAIRD HOWLAND
MARKUS MOBIUS
DAVID ROTHSCHILD*
DUNCAN J. WATTS

Web tracking measures are biased, but how valid and reliable are they?

- To decide when to use web tracking data or surveys to measure specific concepts, it is helpful to compare their psychometric properties

- This is also important when deciding which specific web tracking measure to use, if there are multiple options



The case of media exposure

- Most used measure in web tracking research
- The quality of surveys has been heavily criticised

The state of research on media effects is one of the most notable embarrassments of modern social science. The pervasiveness of the mass media and their virtual monopoly over the presentation of many kinds of information must suggest to reasonable observers that what these media say and how they say it has enormous social and political consequences. Nevertheless, the scholarly literature has been much better at refuting, qualifying, and circumscribing the thesis of media impact than at supporting it. (Bartels, 1993, p. 267)

Bartels, L. M. (1993). **Message received: The political impact of media exposure**. *American Political Science Review*, 87, 267-285

How can we measure media exposure?

Concept: The extent to which an individual encounters **written news media**

How can we measure media exposure?

Concept: The extent to which an individual encounters **written news media**

Characteristics	Potential choices
Metric	Visits, Seconds, Days, Media
List of traces	
List of media	Tranco, Alexa, Cisco, Majestic
Top media	10, 20, 50, 100, 200, All
Information	Broad definition of news, only those identified as “hard” news
Exposure	
Time threshold	1 second, 30 seconds, 120 seconds
App behaviour	Included, excluded
Tracking period	2, 5, 10, 15

Across the three countries, this concept can be measured with +7,500 **different measures**

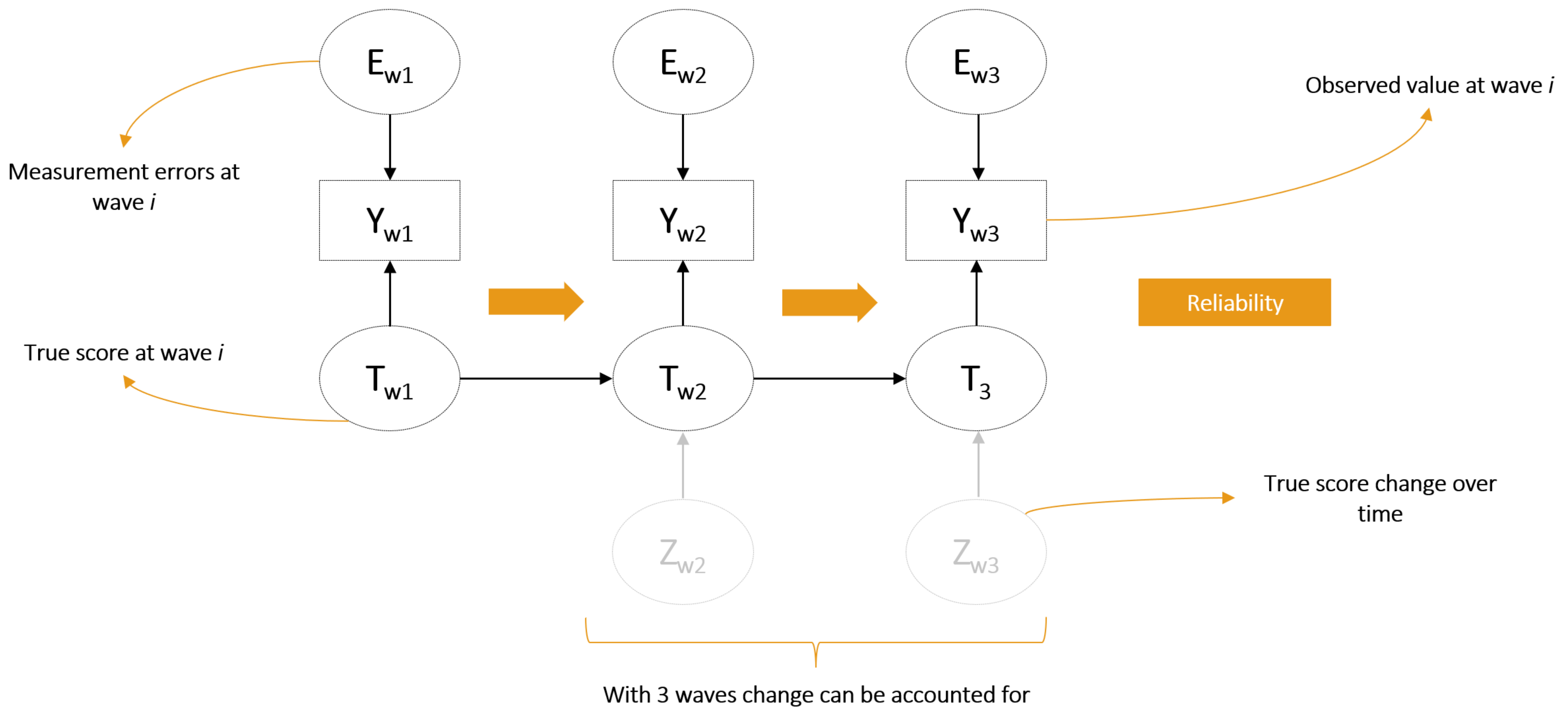
Estimating the validity and reliability of the multiverse

Estimating the validity and reliability of the multiverse

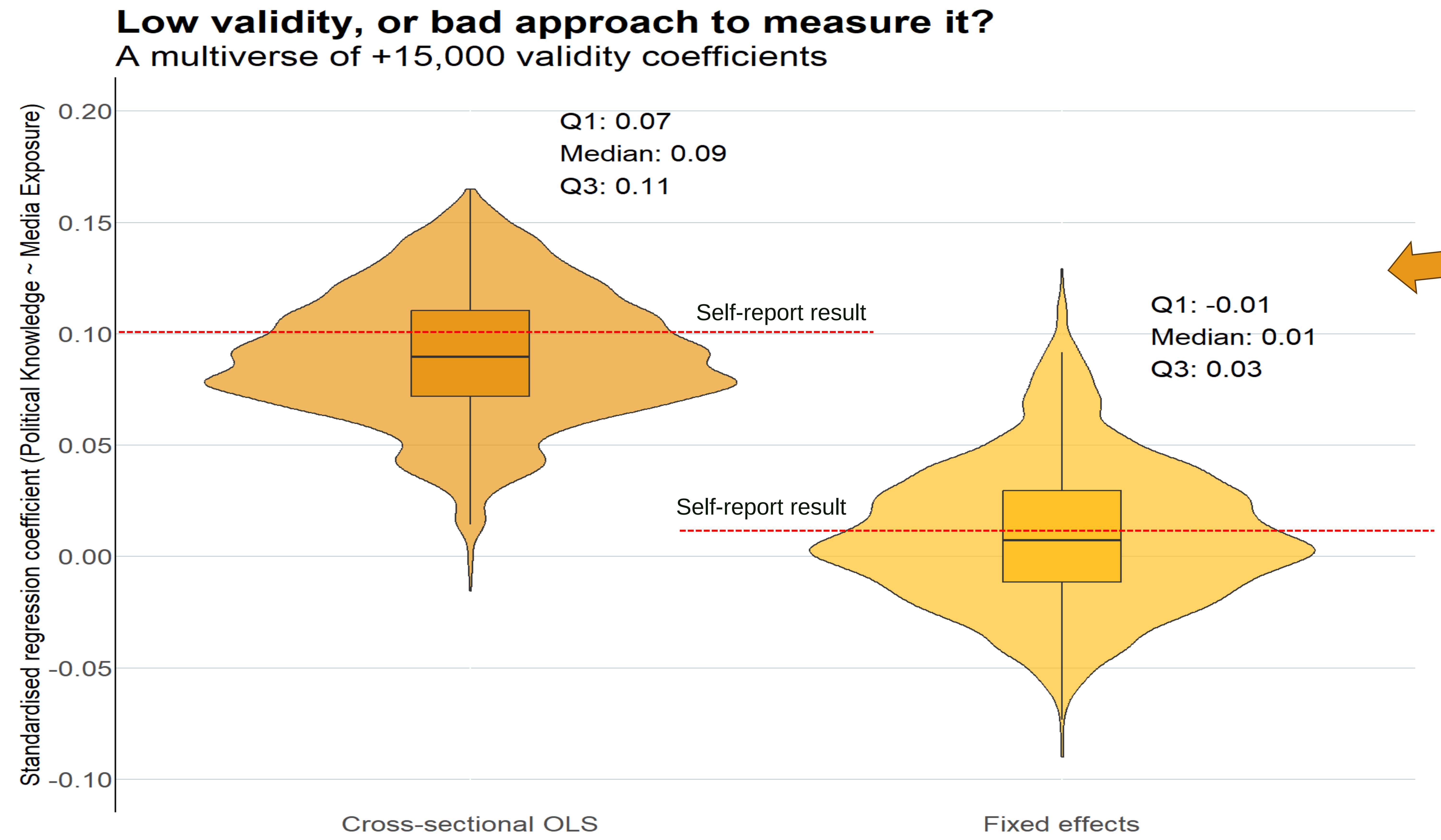
- **Predictive validity**: the association between media exposure and political knowledge

Estimating the validity and reliability of the multiverse

- **Predictive validity**: the association between media exposure and political knowledge
- **Reliability**: Is the measure consistent across multiple observations?



The validity of media exposure measures

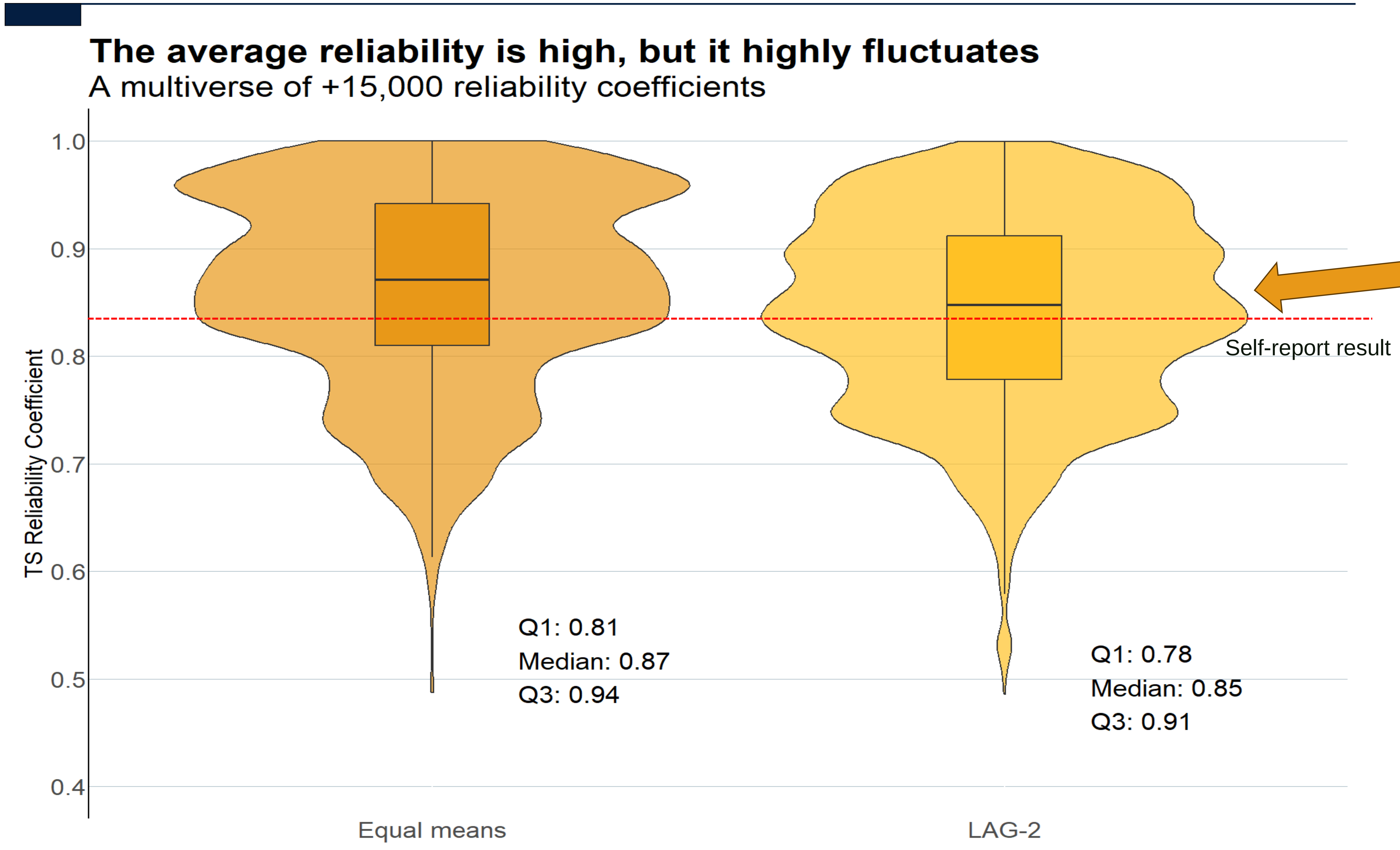


Source: Validity and Reliability of Digital Trace Data in Media Exposure Measures: A Multiverse of Measurements Analysis

Media exposure is a bad predictor of political knowledge (gains).

An indictment of the predictive validity of measures? Or of the method use in the literature?

The reliability of media exposure measures



The average is ≈ .86, but there are clear fluctuations.

Just like for surveys!

Conclusions

Take-home messages

- This thesis puts into **question the gold standard status** of web tracking data

- There are **many potential errors** affecting web tracking data, with some such as **tracking undercoverage clearly biasing results**

- The **reliability and validity** of web tracking measures of media exposure is **similar than those of surveys**, for better or for worse

The bigger picture

I am optimistic! Errors should always be expected, this does not discredit digital trace data

The thesis shows that we can (1) **understand these errors**, (2) **quantify them**, and (3) **identify** which **design decision** might produce the **highest validity and reliability**…

…in a faster and more efficient way than with surveys!

A world of unexplored opportunities, we can improve how we study:

- Digital inequalities
- Digital wellbeing
- Democratic processes
- The relationship between misinformation and health outcomes

Thanks! Questions?

Oriol Bosch Jover
Oriol.Bosch-Jover@demography.ox.ac.uk

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