New Opportunities to Enhance or Replace Conventional Web Survey Data

30 November 2022

Melanie Revilla | IBEI

***Acknowledgments**:*

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 849165); 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)); and the BBVA foundation under their grant scheme to scientific research teams in economy and digital society, 2019.

I want to thank Oriol Bosch, Patricia Iglesias, and Carlos Ochoa for their feedback on previous drafts of this presentation.

Which new opportunities?

WHICH NEW OPPORTUNITIES

Main idea

Smartphones are everywhere!

- More people have smartphones than toilets worldwide[1]

So there are also used to participate in **web surveys**

- Smartphones used in
  - 79% of surveys completed by Millennials
  - 36% of surveys completed by Boomers[2]

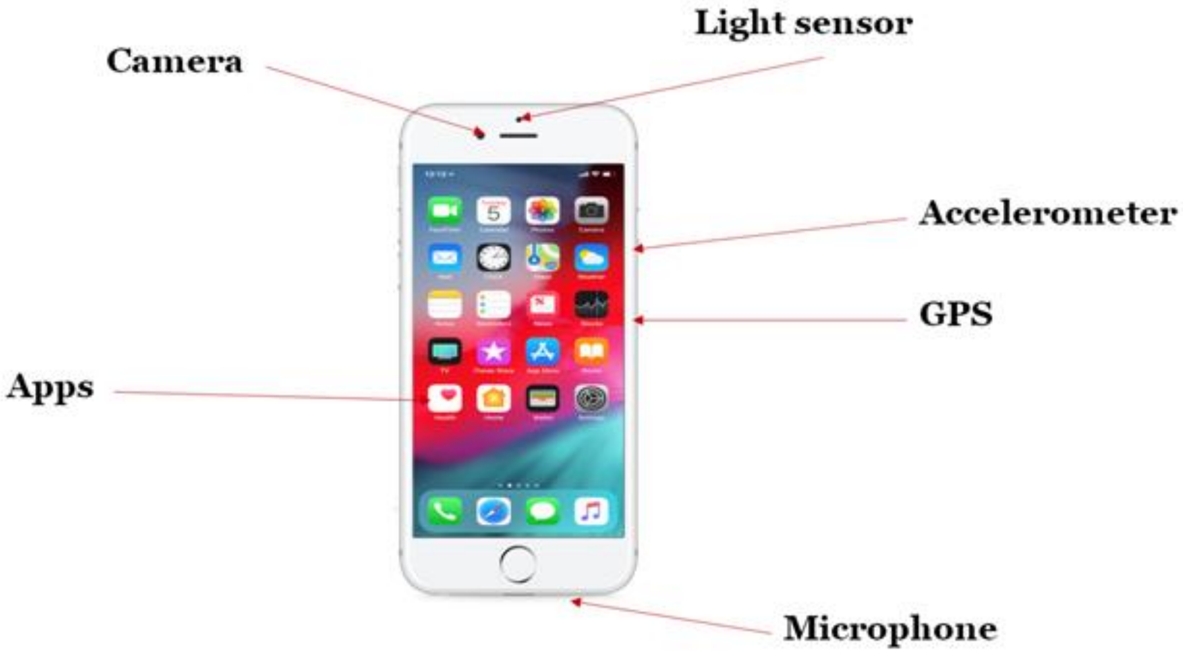
Creates both new challenges and new opportunities

[1] https://www.globalcitizen.org/en/content/access-denied-toilets-Harpic-Waterorg-RB/

[2] Average for the US Netquest panel in 2017/2018

Main idea

- Focus on possibility to **collect other data types**
  - Lot of different data types
  - Each one has its own potential benefits and risks
  - Important to study them separately
  - But also a lot in common



WHICH NEW OPPORTUNITIES

New data types considered

In-the-moment surveys triggered by such data

METERED DATA

Obtained through a tracking application (“meter”) installed by the participants on their devices to register at least the URLs of the webpages visited

GEOLOCATION DATA

Obtained through a tracking application installed on participants’ devices to register at least the GPS coordinates

Most of those data can also be collected for PCs

VISUAL DATA

Screenshots
Photos/videos taken during the survey
Visual files saved on (or accessible from) the device

VOICE DATA

Dictation
Voice recording

New data types considered

In-the-moment surveys triggered by such data

METERED DATA

Obtained through a tracking application ("meter") installed by the participants on their devices to register at least the URLs of the webpages visited

GEOLOCATION DATA

Obtained through a tracking application installed on participants' devices to register at least the GPS coordinates

Benefits not expected for all concepts but enough applications to make the investigation worth it

VISUAL DATA

Screenshots
Photos/videos taken during the survey
Visual files saved on (or accessible from) the device

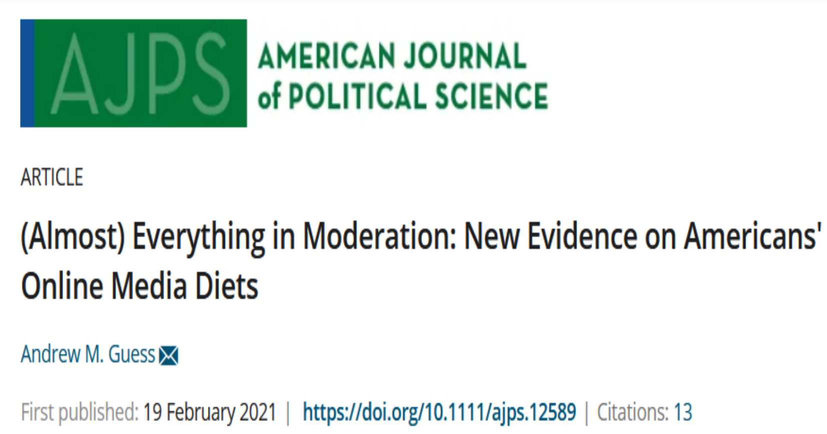
VOICE DATA

Dictation
Voice recording

METERED DATA

Metered data are already used in substantive research

More than **70 papers** published since 2016 using metered data



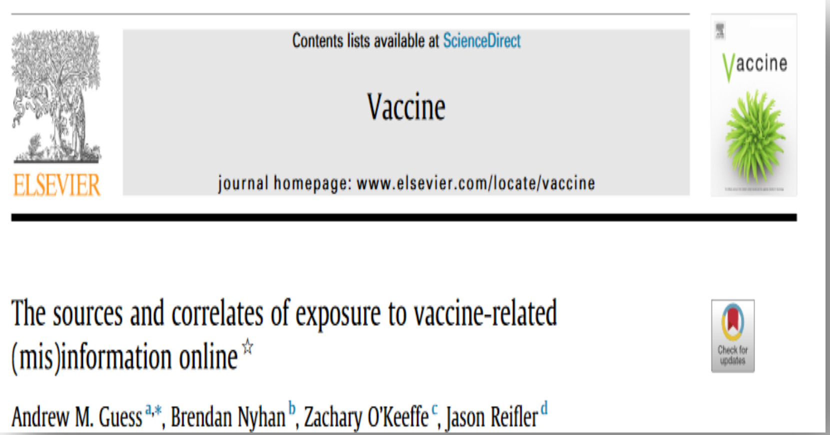
AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

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



Contents lists available at ScienceDirect

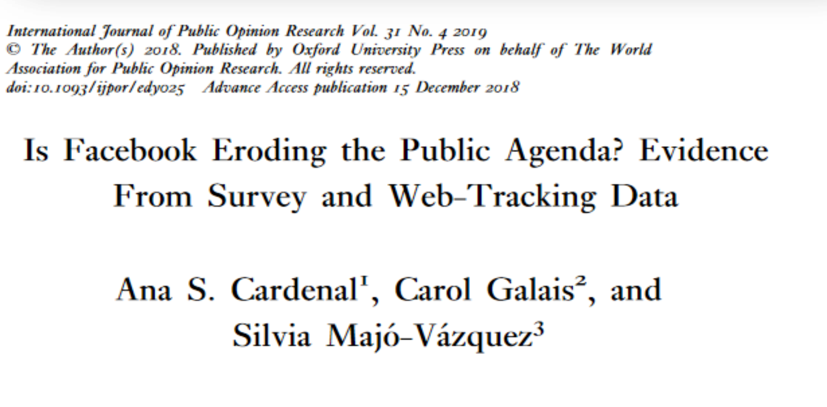
Vaccine

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

ELSEVIER

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

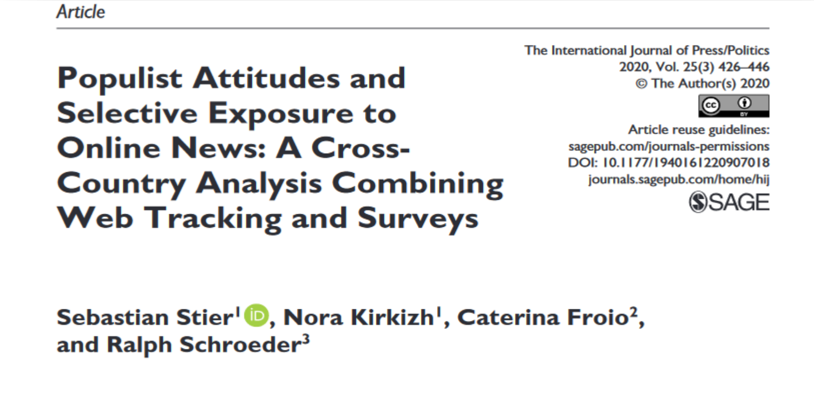
Andrew M. Guess[a,*], Brendan Nyhan[b], Zachary O'Keeffe[c], Jason Reifler[d]



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]



Article

The International Journal of Press/Politics
2020, Vol. 25(3) 426–446
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
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

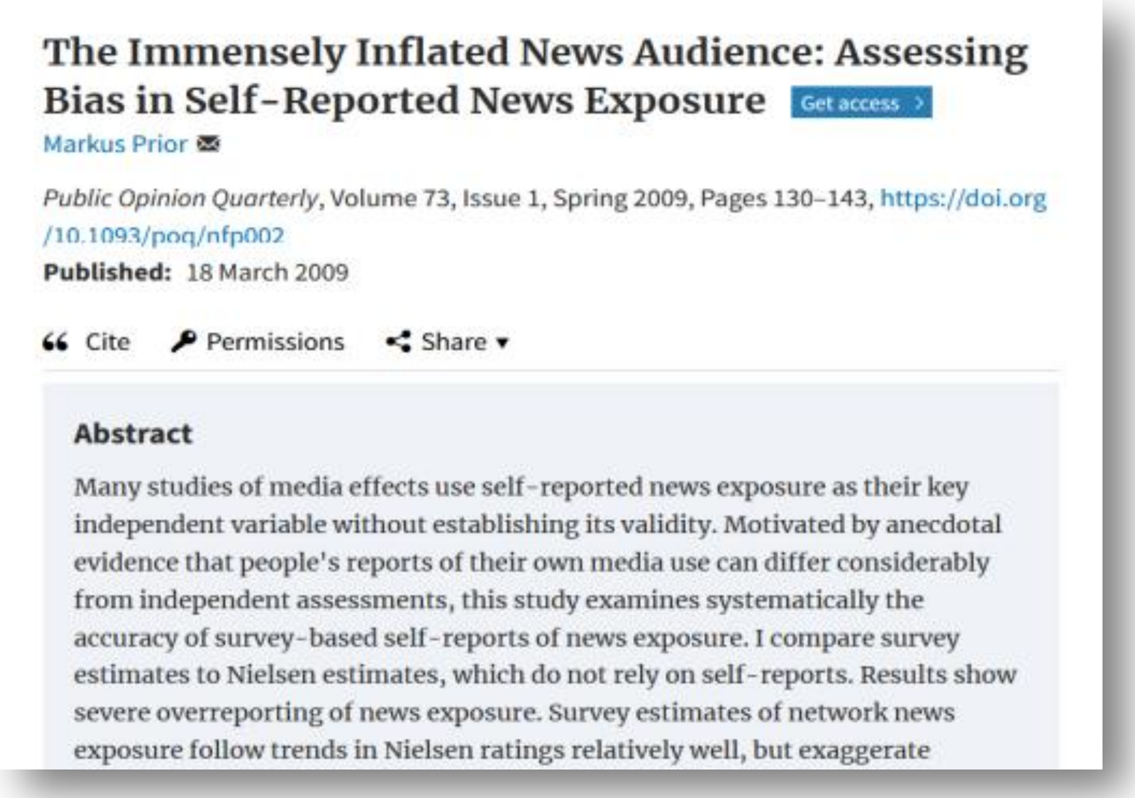
Sebastian Stier[1], Nora Kirkizh[1], Caterina Froio[2], and Ralph Schroeder[3]

METERED DATA

But almost no methodological research

- Researchers usually assume that measures based on metered data are **perfect**

- Many even use them as the **gold standard**, to which they compare self-reported measures to assess their bias



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

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

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

How could metered data help?

HOW COULD METERED DATA HELP?

Expected benefits

Researchers

Reduce some of the issues related to measurement errors

Massive amount of data
Continuous /real time

➡ New insights

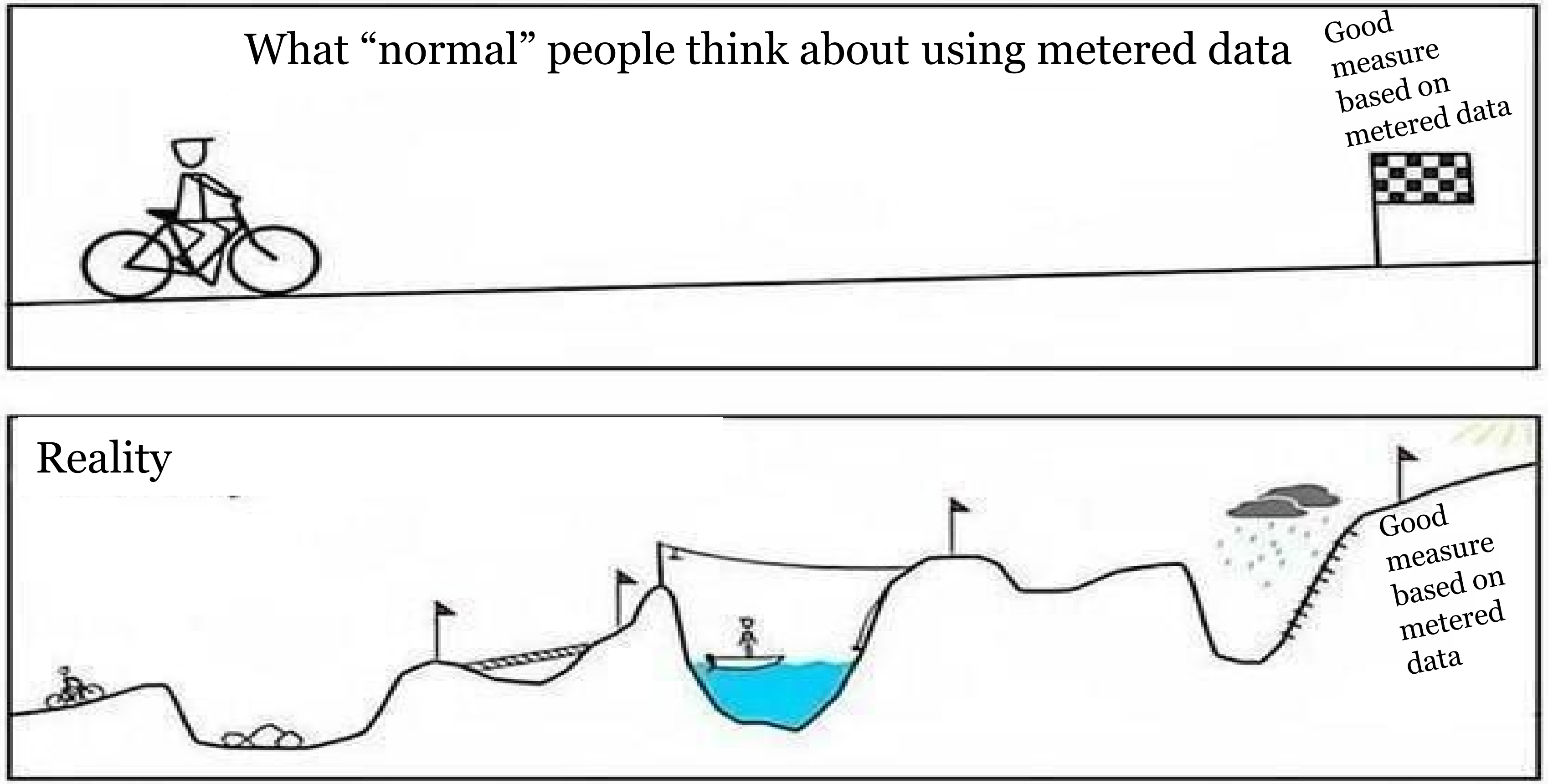
Benefits

Participants

Reduced time dedicated to provide information

Reduced effort

But this is not that easy…


What "normal" people think about using metered data
Good measure based on metered data
Reality
Good measure based on metered data

THIS IS NOT THAT EASY

Also new challenges

webdataopp

Disadvantages

Researchers

More expensive

Dependence on private companies

Selection bias?

Data protection/ethical issues?

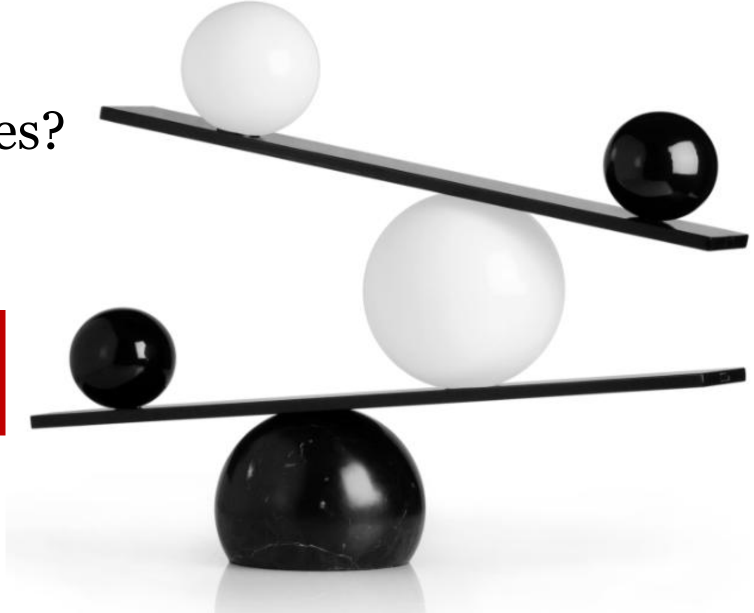
Different types of errors

Participants

Privacy issues?

Loss of control?

New skills needed?



Benefits

Researchers

Reduce some of the issues related to measurement errors

Massive amount of data
Continuous /real time

➡ New insights

Participants

Reduced time dedicated to provide information

Reduced effort

THIS IS NOT THAT EASY

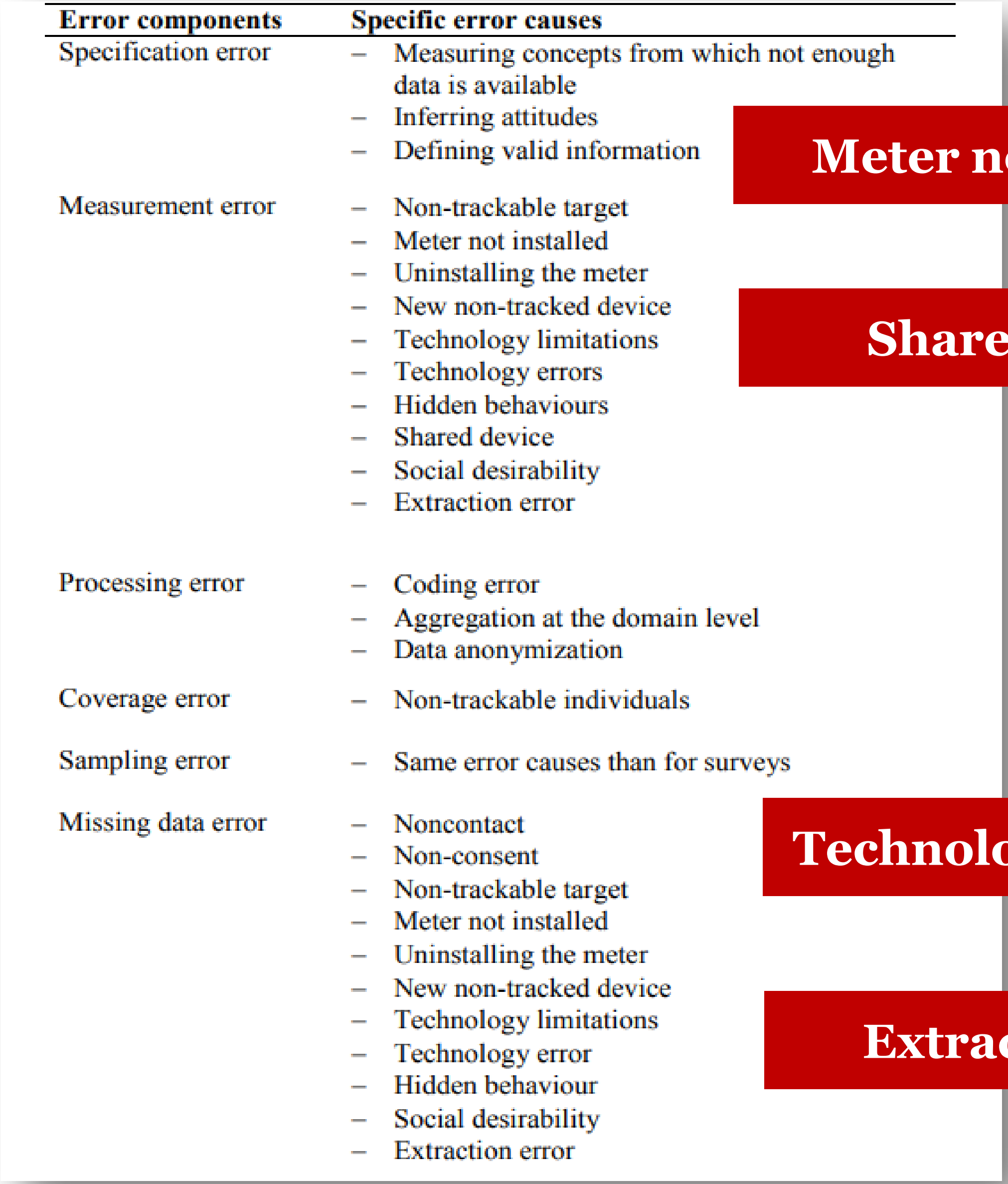
Different types of errors

- Many possible kinds of errors
  - We developed a **Total error framework for metered data** (TEM) = adaptation of the total survey error (TSE) framework to metered data[1]
  - Provides an overview of all possible errors and their causes

[1] https://doi.org/10.1111/rssa.12956

THIS IS NOT THAT EASY

Different types of errors



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

Meter not installed

Shared devices

Technology limitations

Extraction errors

THIS IS NOT THAT EASY

Size of the errors

- Next, we investigated how large some of these errors are and to what extent they may affect the final estimates[1]

- Focus on **tracking undercoverage**
  - ➢Participants do not install the meter in all devices/browsers

TRI-POL data[2]	Spain, Portugal, Italy 3 survey waves + metered data 2 weeks before/after each survey
Survey+meter	Comparing survey answers to information from the meter We found **80-85%** of undercovered
Simulations	Biased univariate and multivariate estimates

[1] https://www.upf.edu/documents/244683118/246905697/Undercoverage+-+AAPOR.pdf/b8a75290-0465-160f-670d-c9bddad468ce; [2] https://www.upf.edu/web/tri-pol

THIS IS NOT THAT EASY

Size of the errors

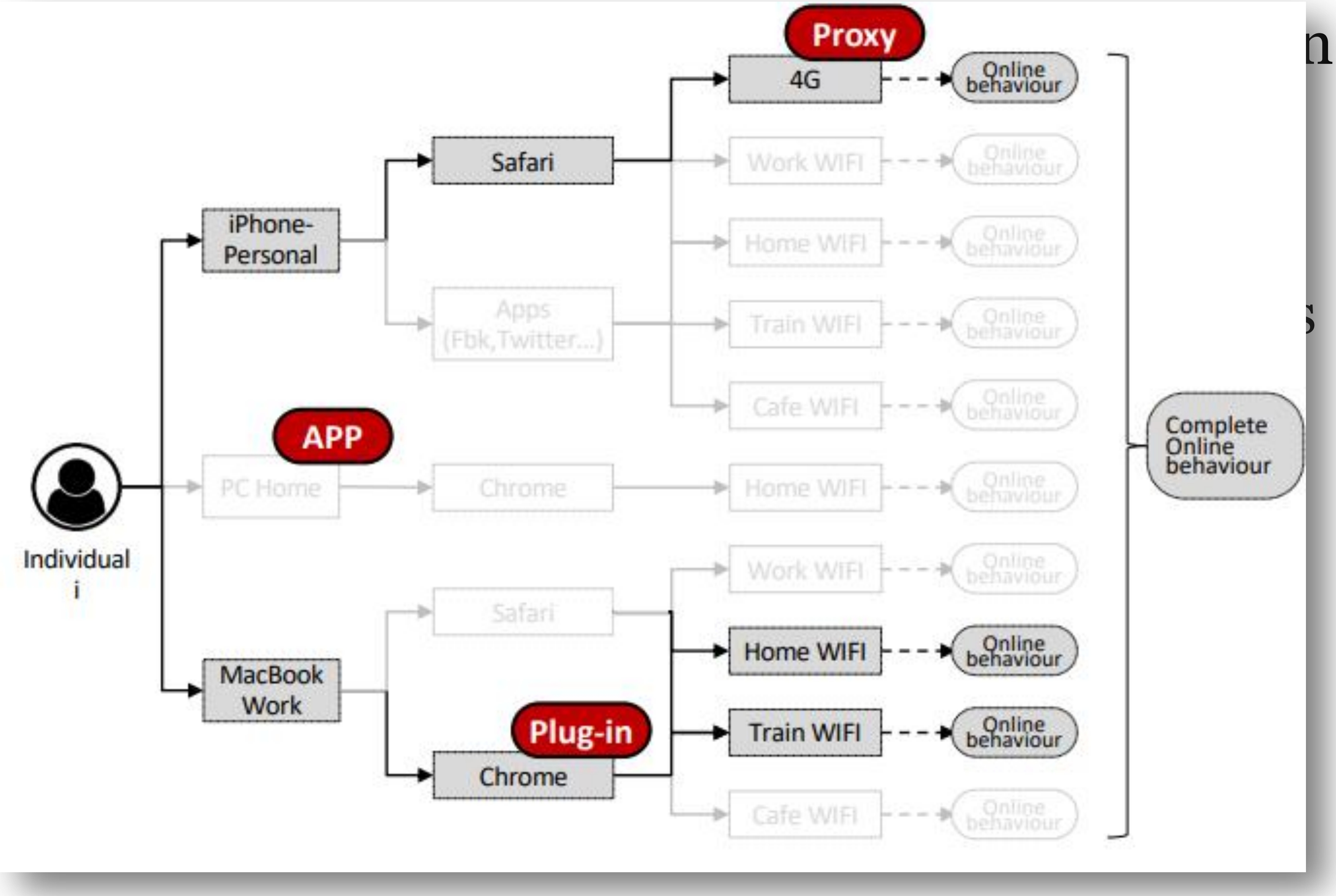
- Next, we ir nd to what exten

- Focus on t
  - ➢Particip s

TRI-POL data[2]

Survey+meter

Simulations

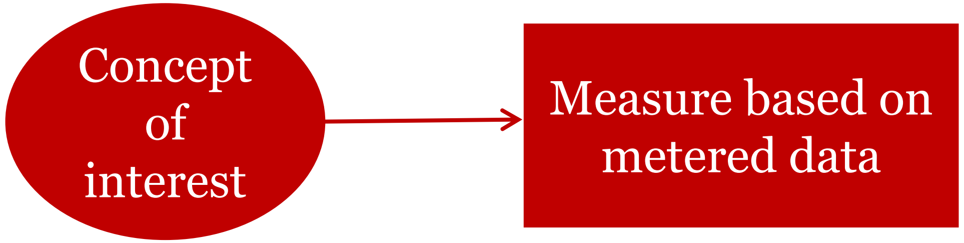


[1] https://www.upf.edu/documents/244683118/246905697/Undercoverage+-+AAPOR.pdf/b8a75290-0465-160f-670d-c9bddad468ce; [2] https://www.upf.edu/web/tri-pol

THIS IS NOT THAT EASY

Validity

- We studied the **validity of measures** based on metered data
- Focus on "**online (written) news media exposure**"
- How to create a measure of this concept using metered data?



- Many decisions
  - Which URLs are considered "online written **news media**"?
  - What is considered as **being "exposed"**?
  - How many **days of tracking** should be used?
  - Etc.

THIS IS NOT THAT EASY

Validity

- Combining all these decisions → theoretically we could create **>8,000** variables that should all measure the same concept of interest

Characteristics	Choices
Metric	Visits, Seconds, Days, Media
List of traces	
List of media	Own, Tranco, Alexa, Cisco, Majestic
Top media	10, 20, 50, 100, 200, All
Information	All domain level, subdomains defined as political
Exposure	
Time threshold	1 second, 30 seconds, 120 seconds
Devices	PC only, Mobile only, All, All without apps
Tracking period	2, 5, 10, 15, 31 days

THIS IS NOT THAT EASY

Validity

- How do these decisions affect the **convergent** and **predictive validity** of the measures?

Convergent validity	All variables measuring the same concept should highly correlate with each other
Predictive validity	Measures that correlate more with political knowledge assumed to be better

- TRI-POL data
  - Average to low convergent validity
  - High fluctuations in predictive validity depending on the choices

Conclusions

CONCLUSIONS

Still a lot to be done

➡ More research needed for all 4 types of data

- Learn more about the errors of those data
  - Types of errors, their size and how they affect the results in different contexts

- Better understand **when** to use those data
  - Need to identify when benefits > disadvantages, balancing those for researchers and participants
  - Need to understand better the mechanisms

CONCLUSIONS

Still a lot to be done

➡ More research needed for all 4 types of data

- Better understand **how** to use those data
  - To replace?
    - But errors will always be there → need to **acknowledge them** and think about **their consequences**
  - To combine?
    - Can provide **different but complementary information**

CONCLUSIONS

Look from different perspectives

THE BLIND MEN AND THE ELEPHANT

"And so these men of research
Disputed loud and long,
Each in his own opinion
Exceeding stiff and strong,
Though each was partly in the right,
And all were in the wrong!"

John Godfrey Saxe (1816-1887)

Thanks!

Questions?

Melanie Revilla | IBEI

✉ mrevilla@ibei.org

https://www.upf.edu/web/webdataopp