Are media exposure measures created with digital trace data any good?

An approach to assess and predict the true-score reliability of web tracking data.

Oriol J. Bosch | Department of Methodology, LSE

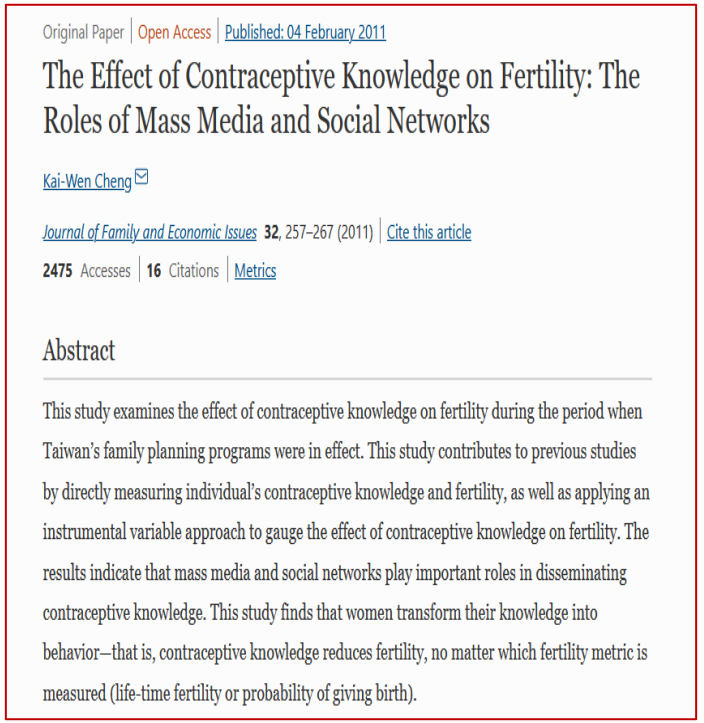
o.bosch-jover@lse.ac.uk
orioljbosch
https://orioljbosch.com/

Acknowledgements: I would like to thank Melanie Revilla, Mariano Torcal, Patrick Sturgis and Jouni Kuha

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).

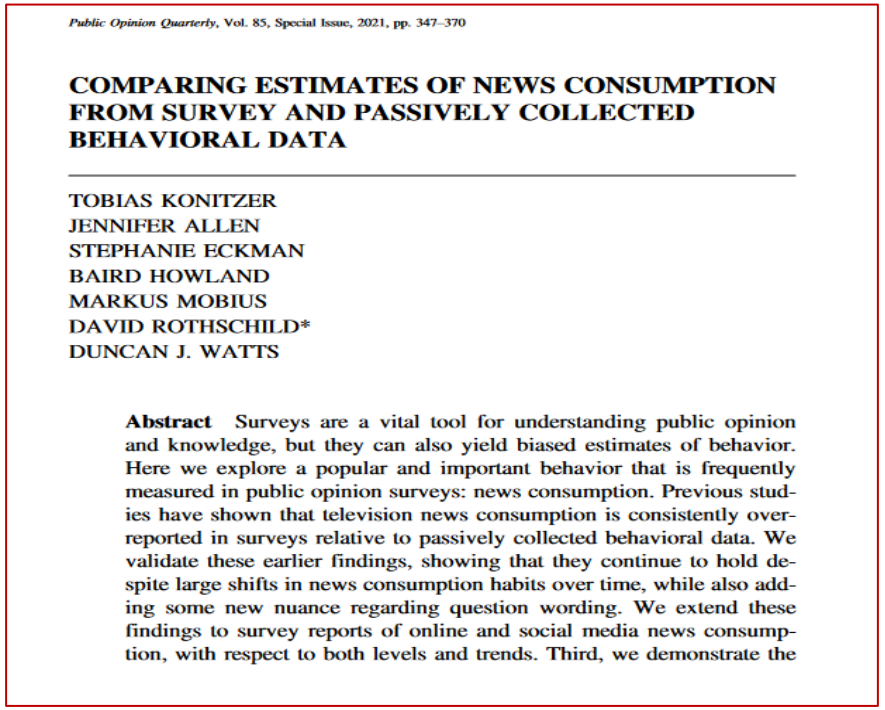
The importance of media exposure

- Increased importance of understanding **the extent** and **the type of media/content people are exposed** to
- As well as its **effect** on how people **think, feel,** and **behave**



The importance of media exposure

- Increased importance of understanding **the extent** and **the type of media/content people are exposed** to
- As well as its **effect** on how people **think, feel,** and **behave**
- We can now measure this with **digital trace data**



Public Opinion Quarterly, Vol. 85, Special Issue, 2021, pp. 347–370

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

Abstract Surveys are a vital tool for understanding public opinion and knowledge, but they can also yield biased estimates of behavior. Here we explore a popular and important behavior that is frequently measured in public opinion surveys: news consumption. Previous studies have shown that television news consumption is consistently overreported in surveys relative to passively collected behavioral data. We validate these earlier findings, showing that they continue to hold despite large shifts in news consumption habits over time, while also adding some new nuance regarding question wording. We extend these findings to survey reports of online and social media news consumption, with respect to both levels and trends. Third, we demonstrate the

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…**

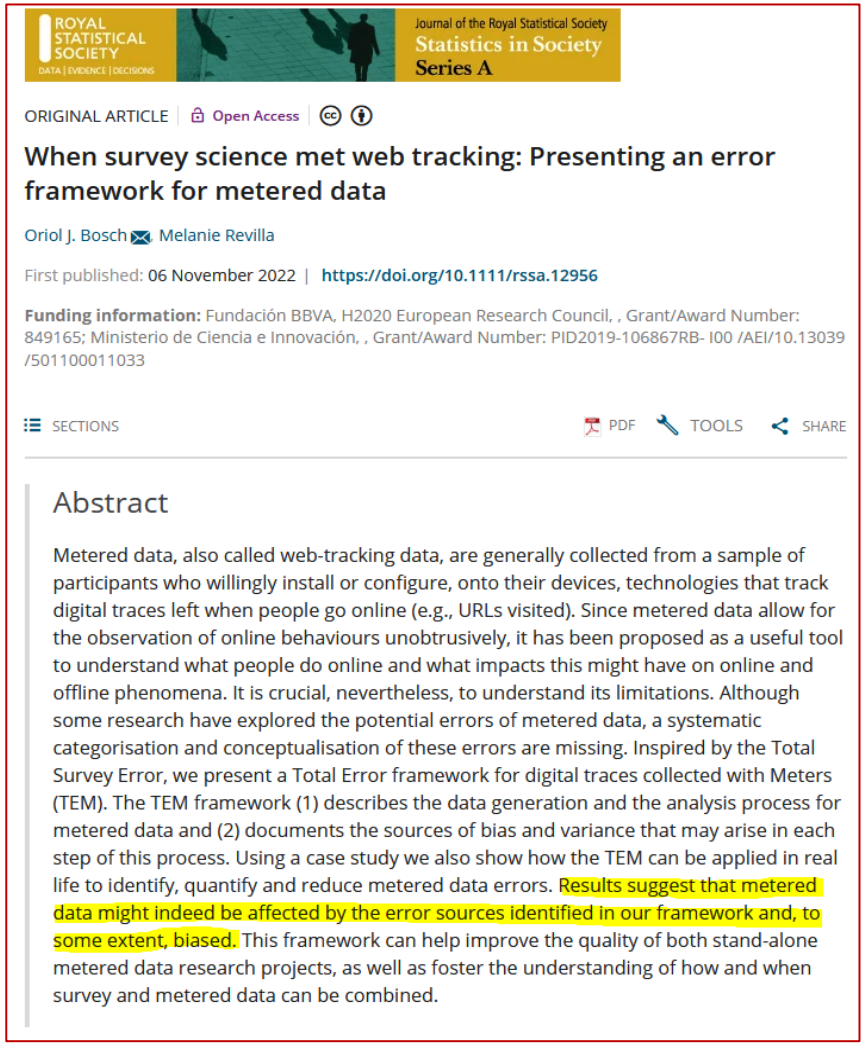
Great, we will get unbiased measures!

Is web tracking data actually unbiased?

Little but growing evidence that **web tracking data is affected by errors**

But still not near what we know about surveys!

My pitch: adapt decades of knowledge in psychometrics and survey methodology to improve how we use digital trace data



ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Royal Statistical Society
Statistics in Society
Series A

ORIGINAL ARTICLE | Open Access

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

Oriol J. Bosch, Melanie Revilla

First published: 06 November 2022 | https://doi.org/10.1111/rssa.12956

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

SECTIONS PDF TOOLS SHARE

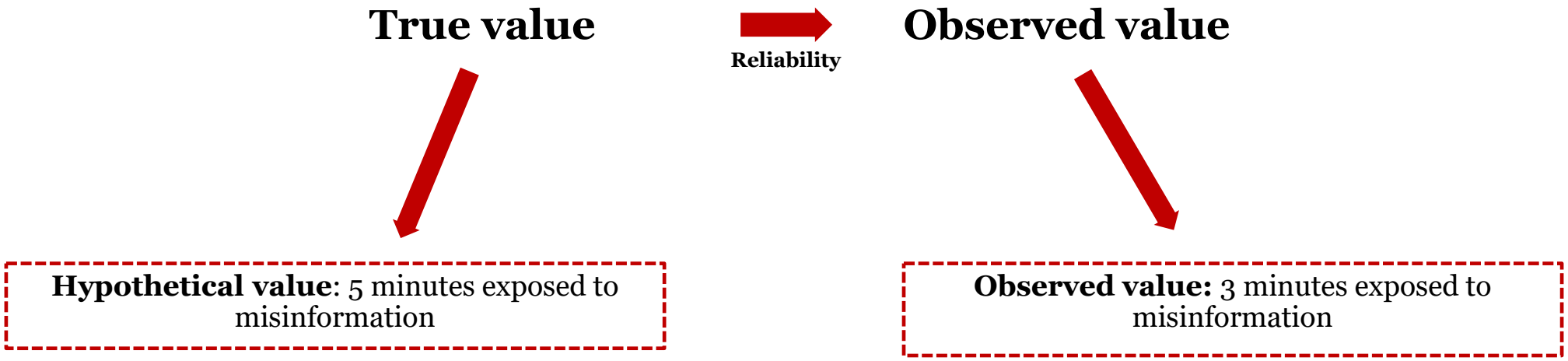
Abstract

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 foster the understanding of how and when survey and metered data can be combined.

Understanding the reliability of web tracking measures

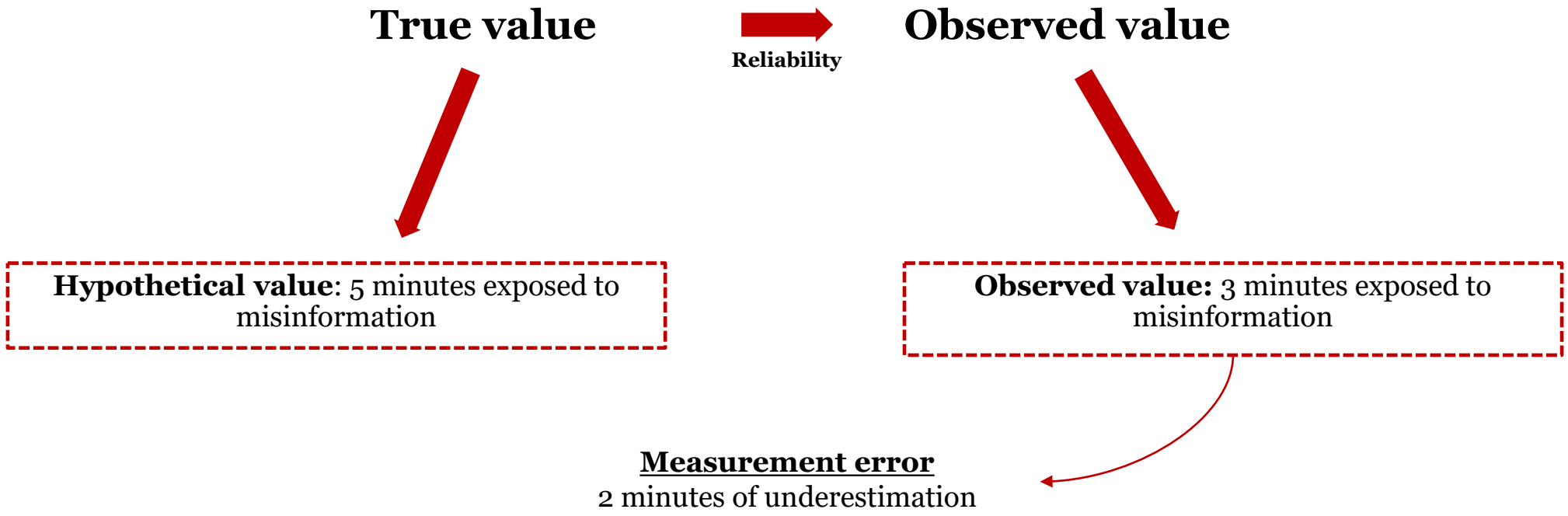
Reliability

- Regardless of how valid a measure is, does the observed values **reflects the hypothetical true value** of our measurement?



Reliability

- Regardless of how valid a measure is, does the observed values **reflects the hypothetical true value** of our measurement?



This study

Research questions

What is the overall reliability of news media exposure measures created using digital traces? (**RQ 1)**

Does the reliability of the news media exposure measures fluctuate across different measurements? (**RQ 2)**

What design choices increase the reliability of web tracking measures? (**RQ 3**)

THIS STUDY

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:** 1,289 (Spain)

- **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]

THIS STUDY

Creating the measurements

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

THIS STUDY

Creating the measurements

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

Task: operationalize this concept into a specific measurement

THIS STUDY

Creating the measurements

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

Task: operationalize this concept into a specific measurement

Objective: understand the reliability of the **whole universe of measurements** that could be used to measure this concept, **not one single** arbitrary measurement

THIS STUDY

Identify the available design choices

Characteristics	My choices

THIS STUDY

Identify the available design choices

1. **Metric:** what can best express variation in the “extent”?

Characteristics	My choices
Metric	Visits, Seconds, Days, Media

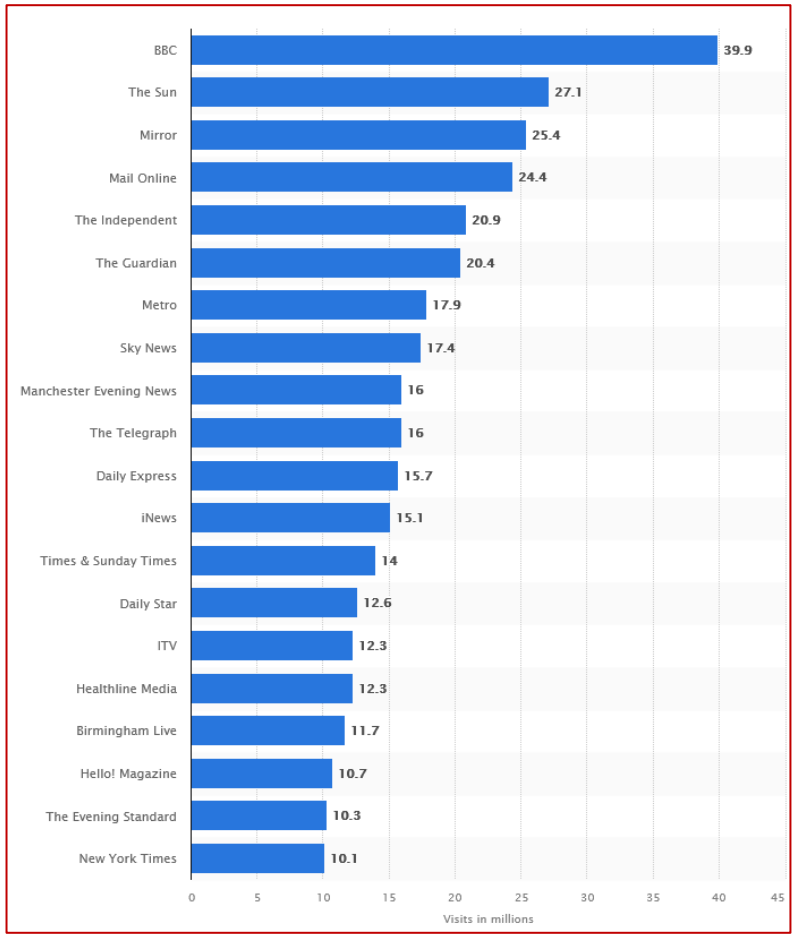
In surveys, the reliability is higher for days or media, is it the same for web tracking?

THIS STUDY

Identify the available design choices

2. List of traces: what is defined as “written news media”?

Characteristics	My choices
Metric	Visits, Seconds, Days, Media
List of traces	
List of media	Tranco, Alexa, Cisco, Majestic
Top media	10, 20, 50, 100, 200, All

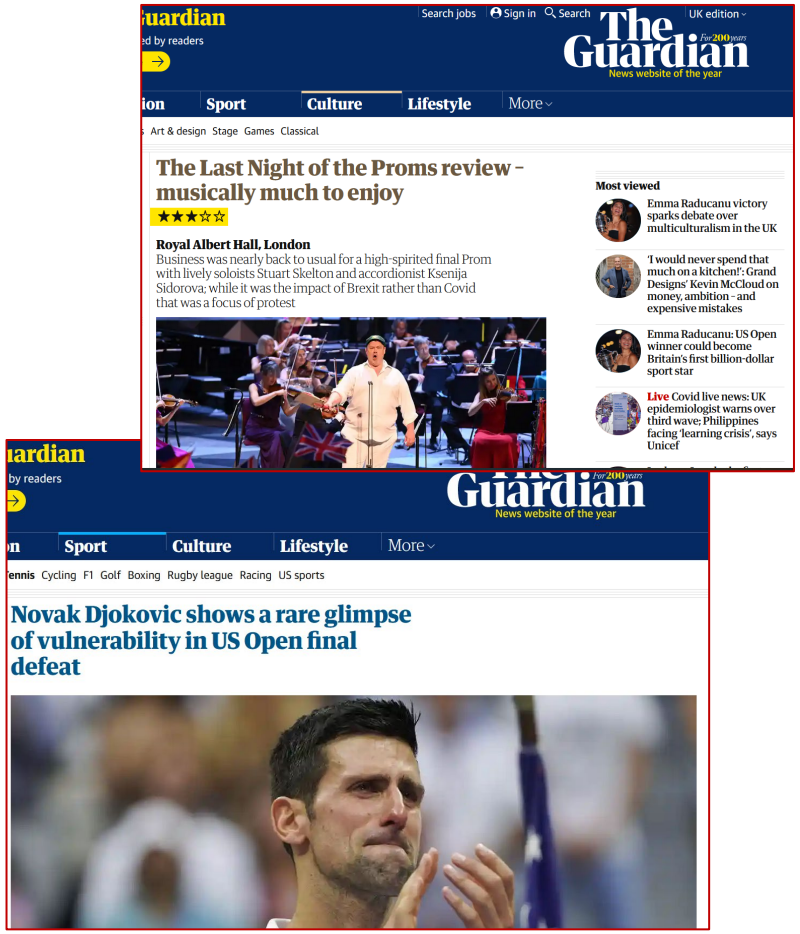


THIS STUDY

Identify the available design choices

2. **List of traces:** what is defined as “written news media”?

Characteristics	My 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	All URLs, only those identified as political



THIS STUDY

Identify the available design choices

3. Exposure: what events can be considered as "exposed"?

Characteristics	My 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	All URLs, only those identified as political
Exposure	
Time threshold	1 second, 30 seconds, 120 seconds

Exposure might mean just seeing something, or reading part / all of the article

THIS STUDY

Identify the available design choices

web data opp

3. Exposure: what events can be considered as "exposed"?

Characteristics	My 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	All URLs, only those identified as political
Exposure	
Time threshold	1 second, 30 seconds, 120 seconds
Devices	PC only, Mobile only, All, All without apps

Most research has focused only on behaviours through PCs, is this right?

Identify the available design choices

4. **Tracking period:** what time period allows to measure "normality"?

Characteristics	My 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	All URLs, only those identified 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

Longer tracking periods might be better, but also more expensive

THIS STUDY

Identify the available design choices

Characteristics	My 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	All URLs, only **those identified 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

Number of visits, lasting 1 second or more, to the political articles in the top 50 most popular news websites according to Alexa, through PCs, during the last 15 days

THIS STUDY

Identify the available design choices

Characteristics	My 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	All URLs, only those identified 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

8,070 potential variables*

* *Not 100% fully crossed. The time metric is not crossed with the 30 seconds and 120 seconds thresholds.*

THIS STUDY

Identify the available design choices

Characteristics	My 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	All URLs, only those identified 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

8,070 potential variables*

- I created **all** the potential variables
- Analyses are computed for each of the 8,070
- This would take **years and innumerable resources** to be replicated for surveys!

** Not 100% fully crossed. The time metric is not crossed with the 30 seconds and 120 seconds thresholds.*

Analytical approach

Assessing the **reliability** and its fluctuations (RQ 1 & 2)

web data opp

I focus on the true-score (TS) reliability

Assessing the **reliability** and its fluctuations (RQ 1 & 2)

web data opp

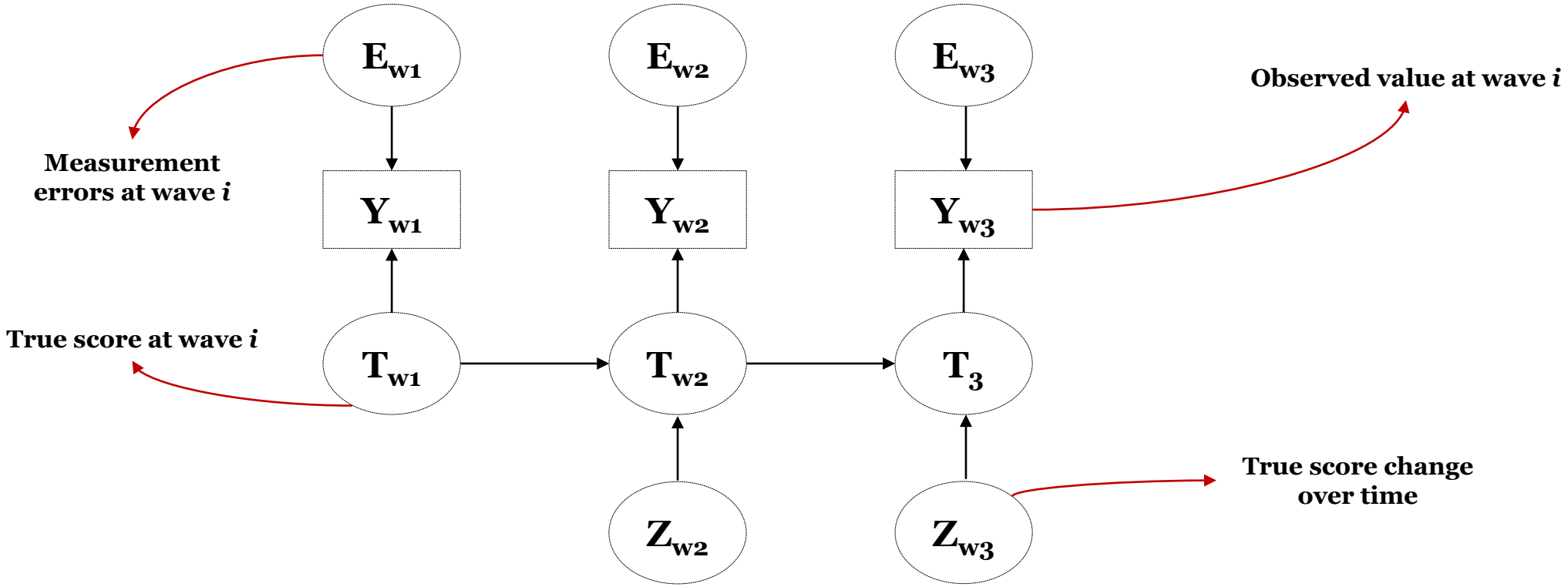
I focus on the true-score (TS) reliability

- Is the measure consistent across multiple observations?

If we account for the changes in the true value across waves, the true values should be stable independently of when we measure them

Assessing the **reliability** and its fluctuations (RQ 1 & 2)

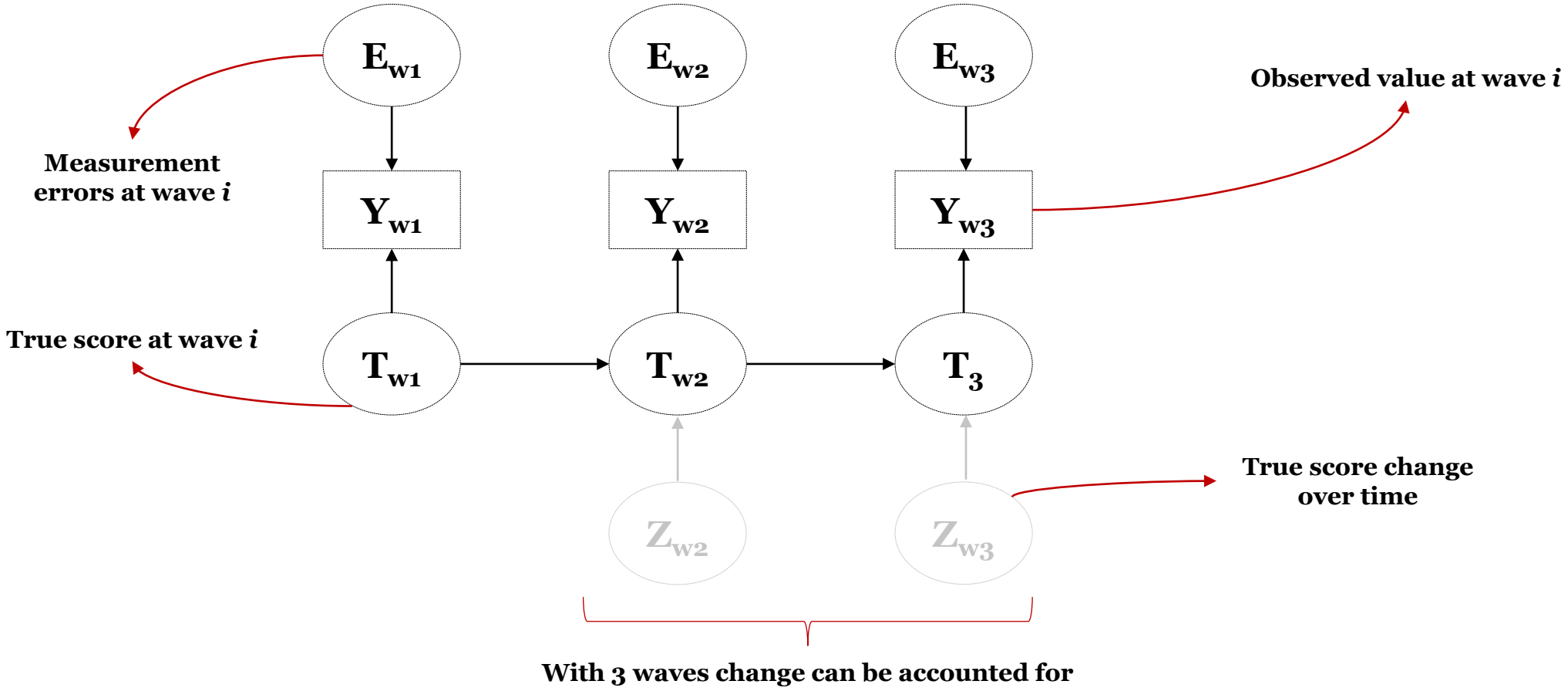
I use the **Quasi-Markov Simplex Model,** which allows separating reliability from stability



Heise, D. R. (1969). **Separating reliability and stability in test-retest correlation**. American sociological review, 93-101.
Alwin, D. F. (2007). **Margins of error: A study of reliability in survey measurement**. John Wiley & Sons.

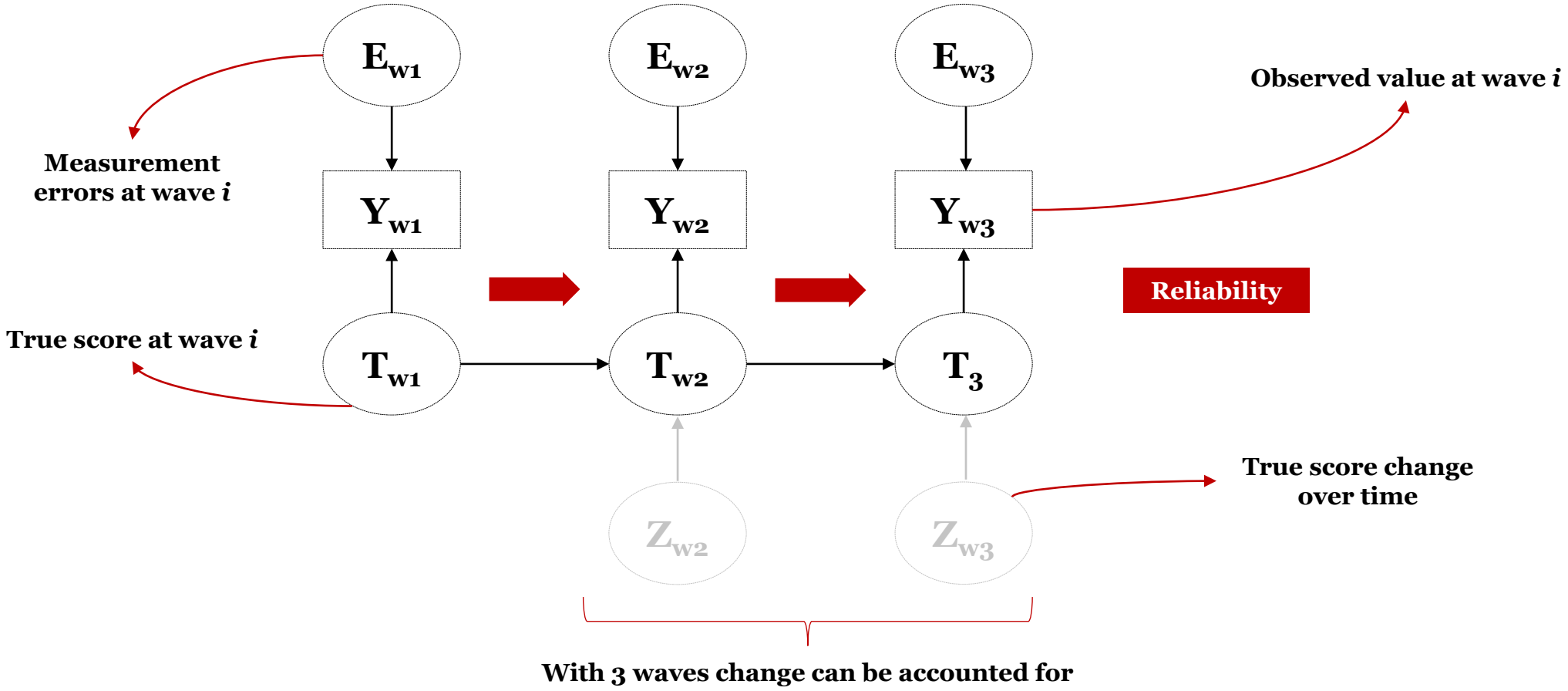
Assessing the **reliability** and its fluctuations (RQ 1 & 2)

I use the **Quasi-Markov Simplex Model,** which allows separating reliability from stability



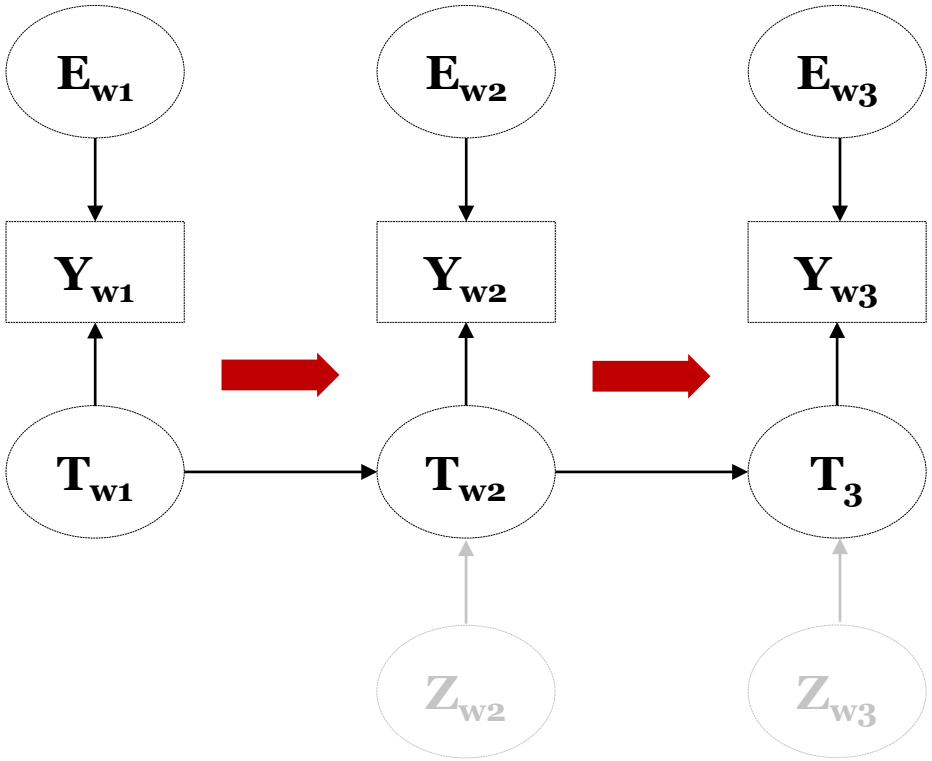
Assessing the **reliability** and its fluctuations (RQ 1 & 2)

I use the **Quasi-Markov Simplex Model,** which allows separating reliability from stability



Assessing the **reliability** and its fluctuations (RQ 1 & 2)

I use the **Quasi-Markov Simplex Model,** which allows separating reliability from stability



Reliability

8,070 unique coefficients*

RQ1.2: What is the average?

RQ2.2: Does it fluctuate?

*Computed using Heise's approach. Underlying correlation matrix created with using latent correlations for mixed data types, to account for the truncated and zero inflated nature of media exposure variables

ANALYTICAL APPROACH

The impact of design choices on **reliability** & **validity** (RQ 3)

web data opp

The impact of design choices on **reliability** & **validity** (RQ 3)

- After running the reliability analyses, I created a new dataset, with the following:
  - **Name** of the variables
  - Associated **reliability coefficient**
  - **Design choices** of the specific variable, for each **design characteristic**

corr_ALL

Filter

	variable	Coefficient	List	Top	Metric	Time_threshold	Tracking_period	Domain_Subdomain	Device
1163	PRE10_D_News_MobilePC_webapp_10A_120s	-0.1954861	Alexa	10	Day	120	PRE10	Domain	All devices
913	PRE10_D_News_MobilePC_webapp_ALL_1s	-0.1944916	ALL	222	Day	1	PRE10	Domain	All devices
828	PRE10_D_News_MobilePC_webapp_100T_1s	-0.1942604	Tranco	100	Day	1	PRE10	Domain	All devices
868	PRE10_D_News_MobilePC_webapp_200T_1s	-0.1942604	Tranco	200	Day	1	PRE10	Domain	All devices
908	PRE10_D_News_MobilePC_webapp_50T_1s	-0.1932236	Tranco	50	Day	1	PRE10	Domain	All devices
813	PRE10_D_News_MobilePC_webapp_100A_1s	-0.1911152	Alexa	100	Day	1	PRE10	Domain	All devices
853	PRE10_D_News_MobilePC_webapp_200A_1s	-0.1911152	Alexa	200	Day	1	PRE10	Domain	All devices
893	PRE10_D_News_MobilePC_webapp_50A_1s	-0.1911152	Alexa	50	Day	1	PRE10	Domain	All devices
832	PRE15_D_News_MobilePC_webapp_10A_1s	-0.1880830	Alexa	10	Day	1	PRE15	Domain	All devices
827	PRE15_D_News_MobilePC_webapp_100T_1s	-0.1856270	Tranco	100	Day	1	PRE15	Domain	All devices
867	PRE15_D_News_MobilePC_webapp_200T_1s	-0.1856270	Tranco	200	Day	1	PRE15	Domain	All devices
912	PRE15_D_News_MobilePC_webapp_ALL_1s	-0.1841421	ALL	222	Day	1	PRE15	Domain	All devices

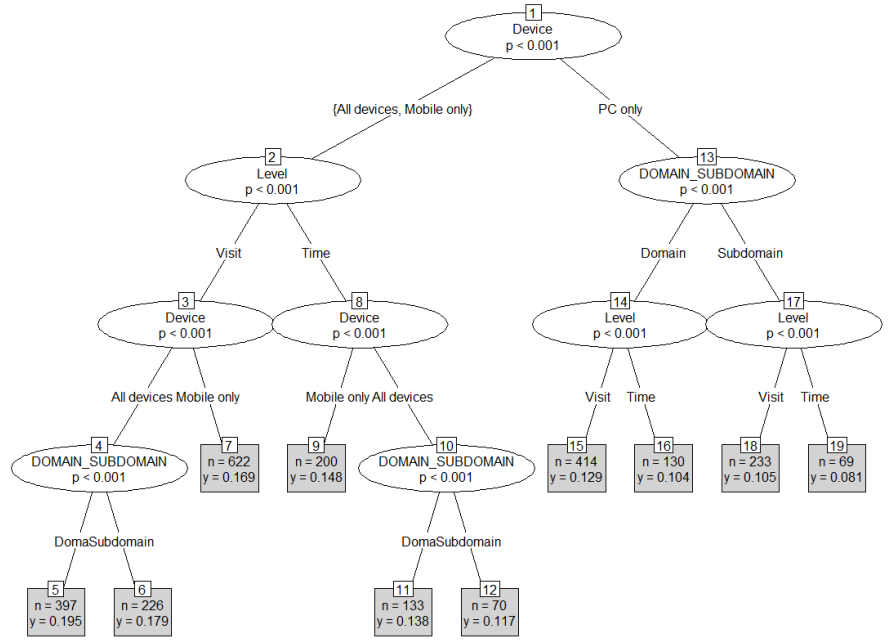
The impact of design choices on **reliability** & **validity** (RQ 3)

- After running the reliability analyses, I created a new dataset, with the following:
  - **Name** of the variables
  - Associated **reliability coefficient**
  - **Design choices** of the specific variable, for each **design characteristic**

With this dataset it is possible to **model the effect** of each **design choice** on the estimated **reliability,** using the **8,070 variables as observations**

The impact of design choices on **reliability** & **validity** (RQ 3)

- To predict the impact of each design choice, we used random forests of regression trees*

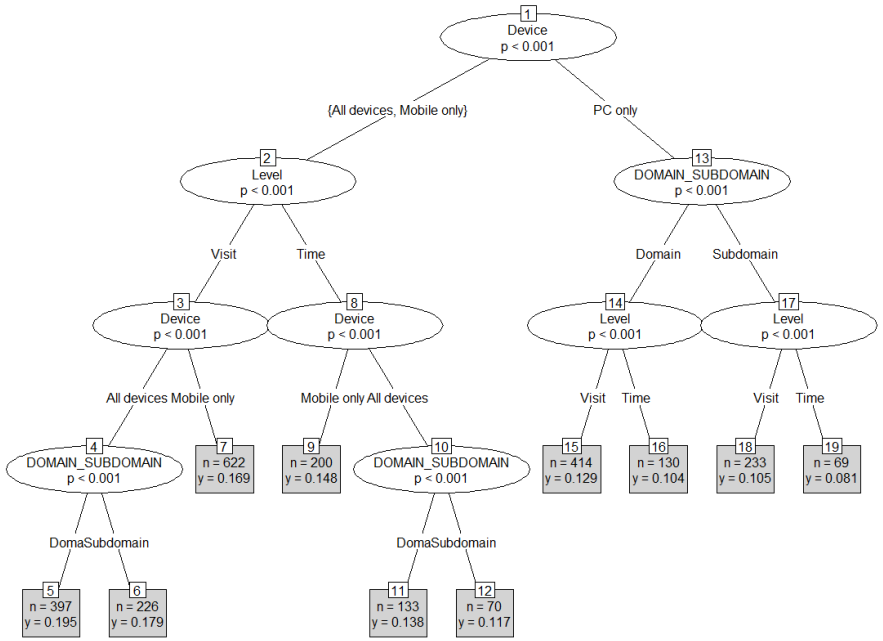


* R package *randomForest: Ntree*: 500 | *Mtry*: 4 | *Node size*: 3 | *Sample fraction*: 80%

The impact of design choices on **reliability** & **validity** (RQ 3)

- To predict the impact of each design choice, we used random forests of regression trees*

- I extract the following information:
  - The variable importance: % increase of MSE
  - And the marginal effect of each choice



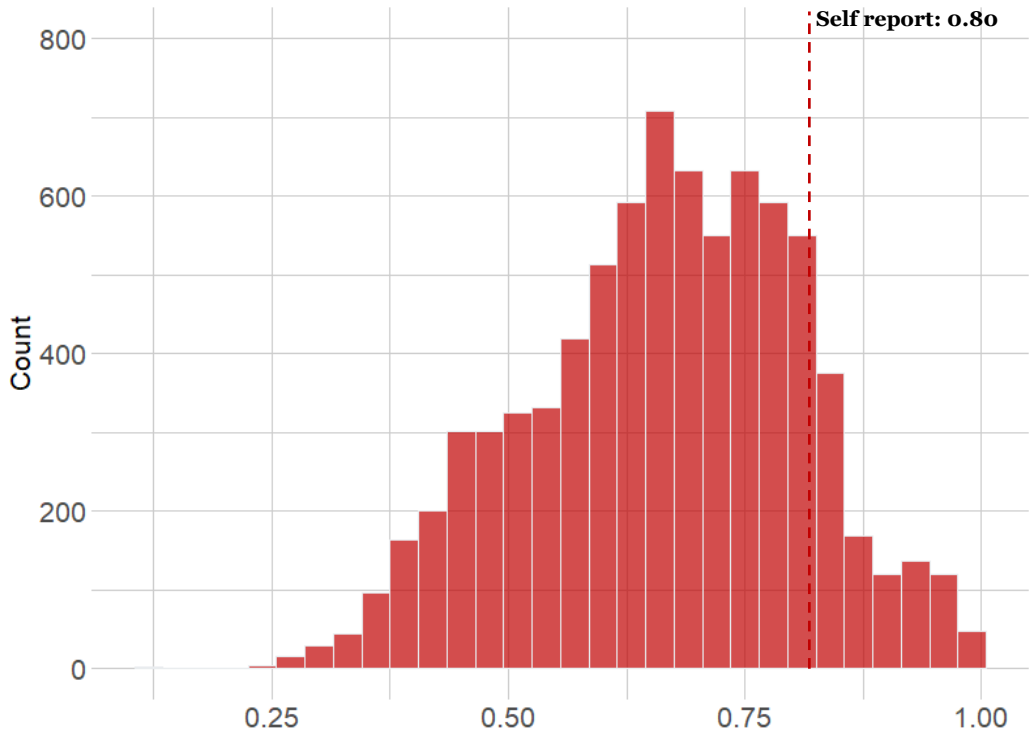
* R package *randomForest: Ntree*: 500 | *Mtry*: 4 | *Node size*: 3 | *Sample fraction*: 80%

What is the overall **reliability** of digital news media exposure created with digital traces? (**RQ1**)

And does it **fluctuate** across design choices? (**RQ2**)

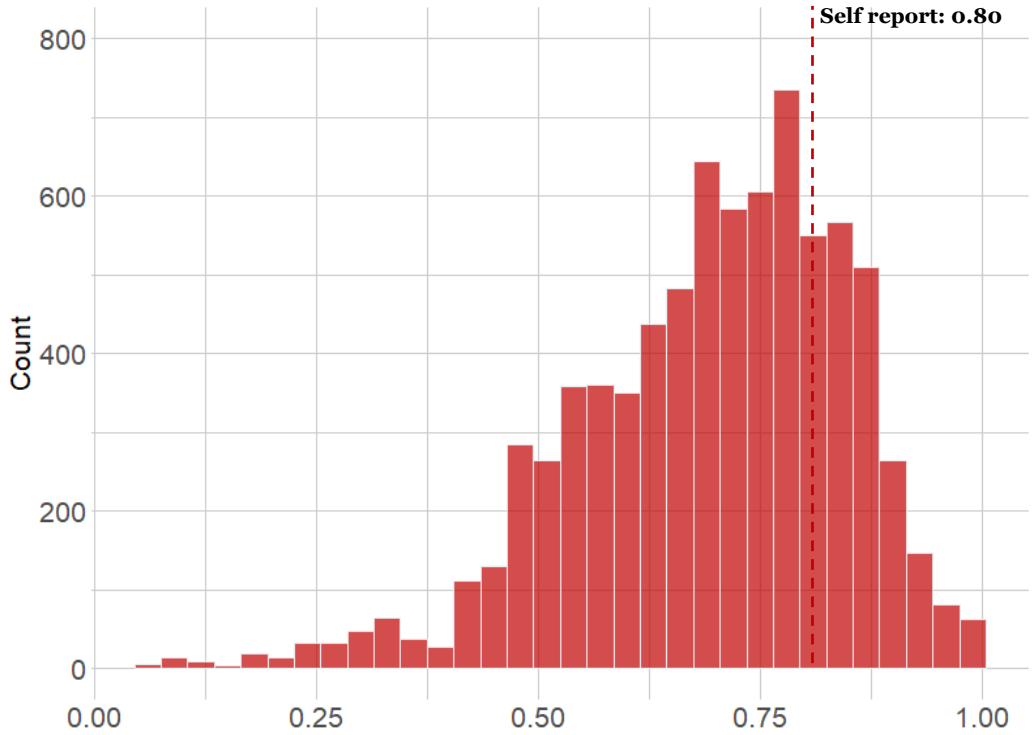
TS Reliability across different specifications

SPAIN



Mean: .66
Media: .67
1st Quart: .57
3rd Quart: .77

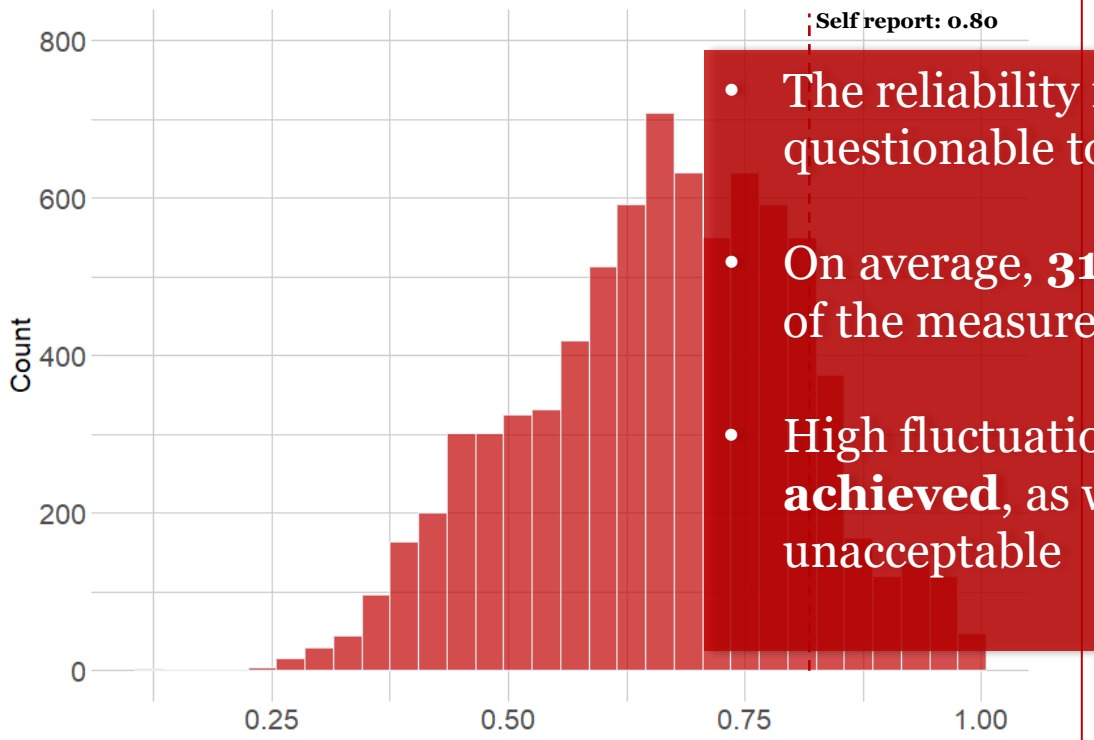
PORTUGAL



Mean: .69
Media: .71
1st Quart: .60
3rd Quart: .81

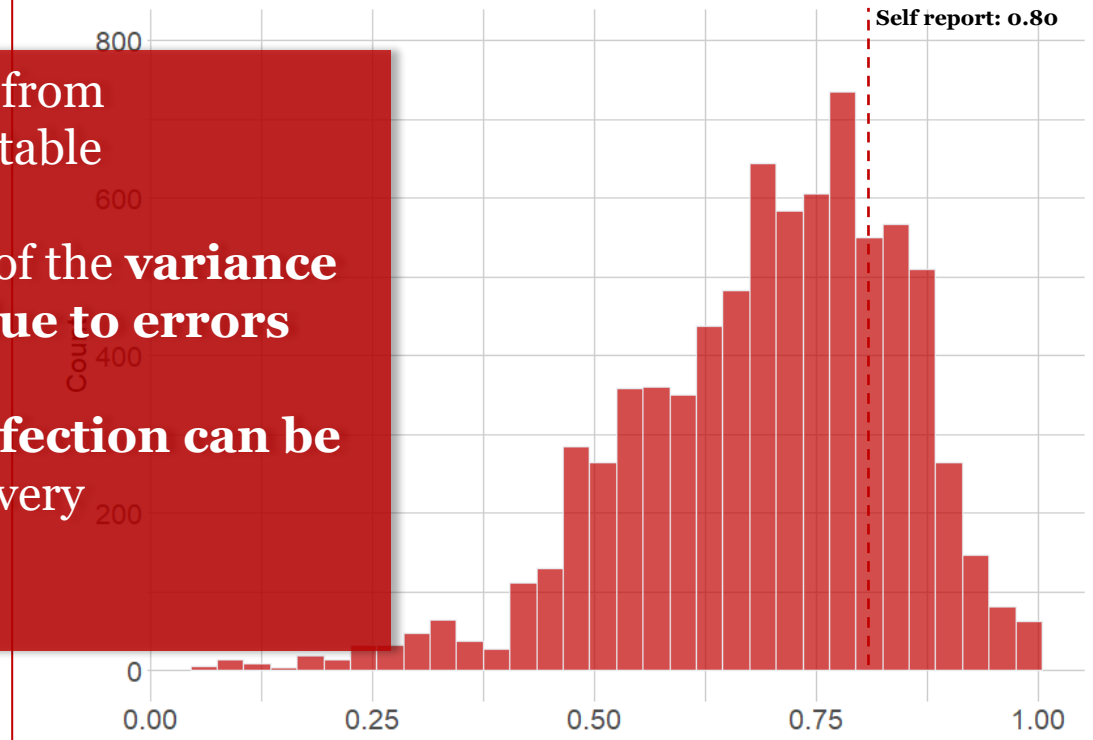
TS Reliability across different specifications

SPAIN



Mean: .66
Media: .67
1st Quart: .57
3rd Quart: .77

PORTUGAL

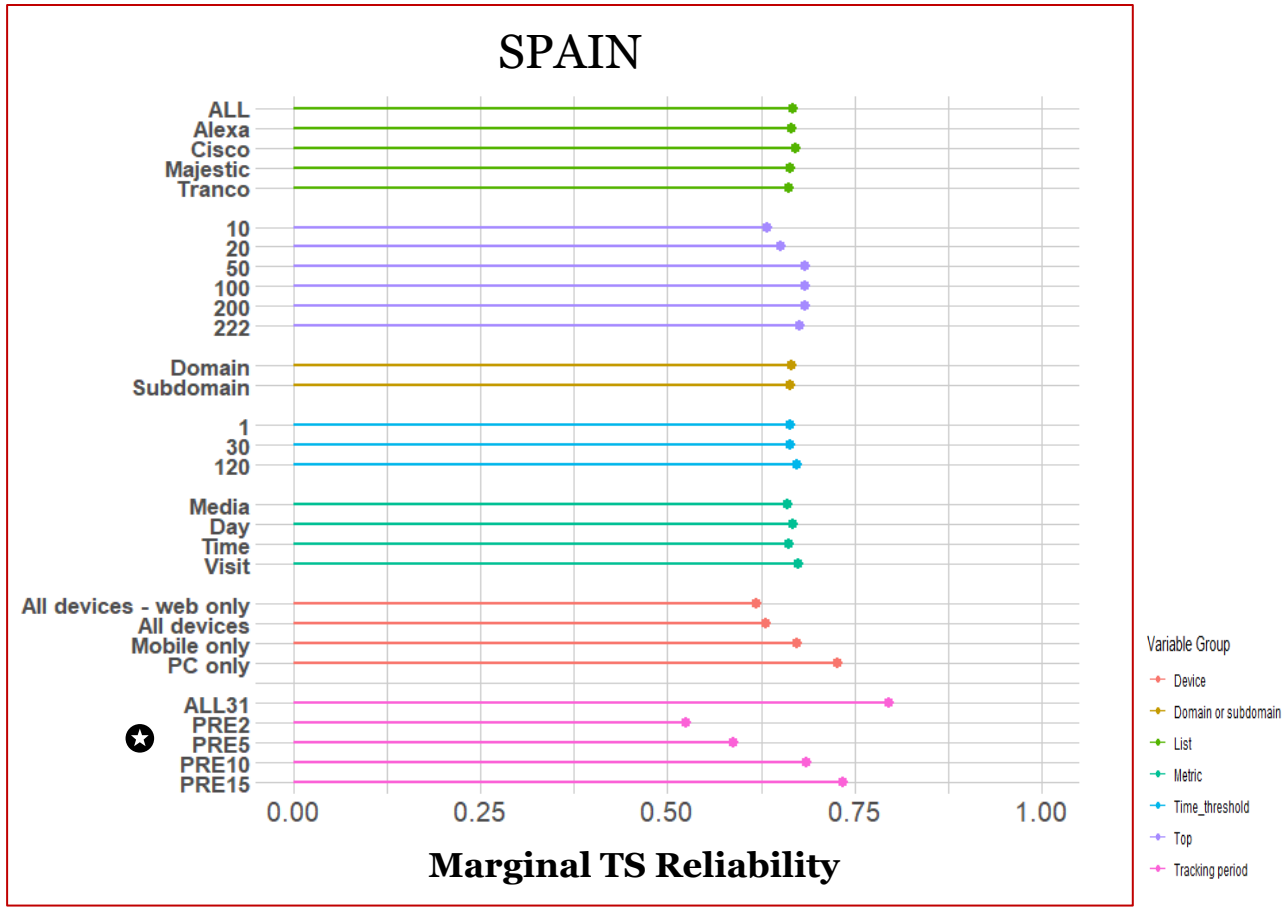


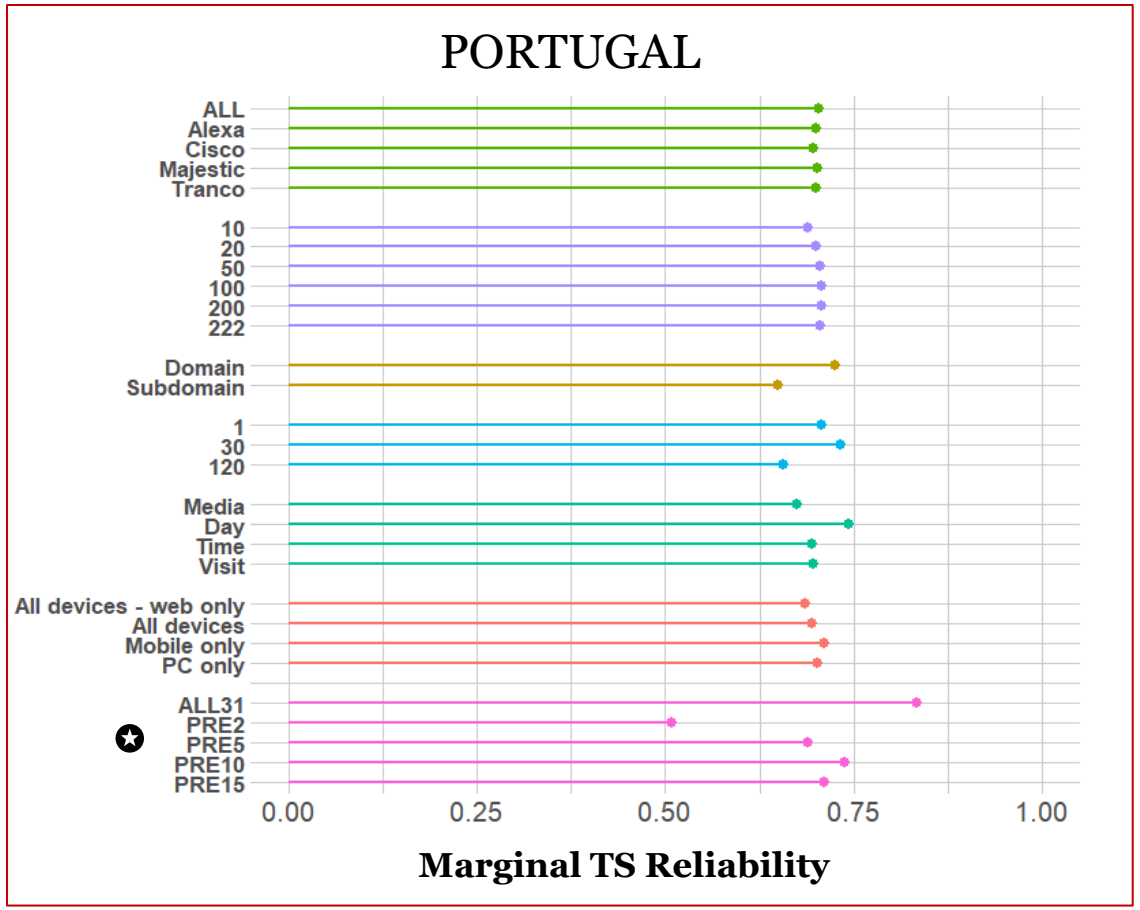
Mean: .69
Media: .71
1st Quart: .60
3rd Quart: .81

- The reliability ranges from questionable to acceptable
- On average, **31-34%** of the **variance** of the measures are **due to errors**
- High fluctuation, **perfection can be achieved**, as well as very unacceptable

What design choices increase the reliability of web tracking measures? (**RQ 3**)

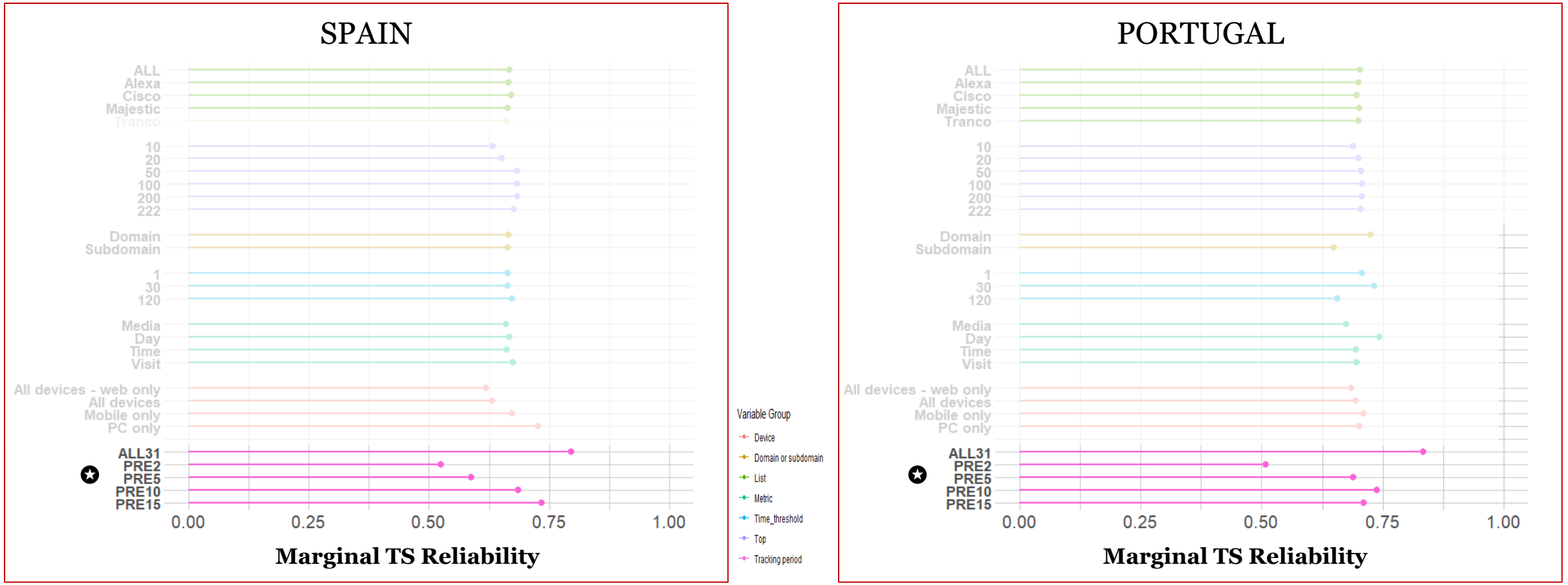
Marginal **reliability** for each design choice





✪ Indicates variable with the highest % increase of Mean Squared Error when excluded from the model

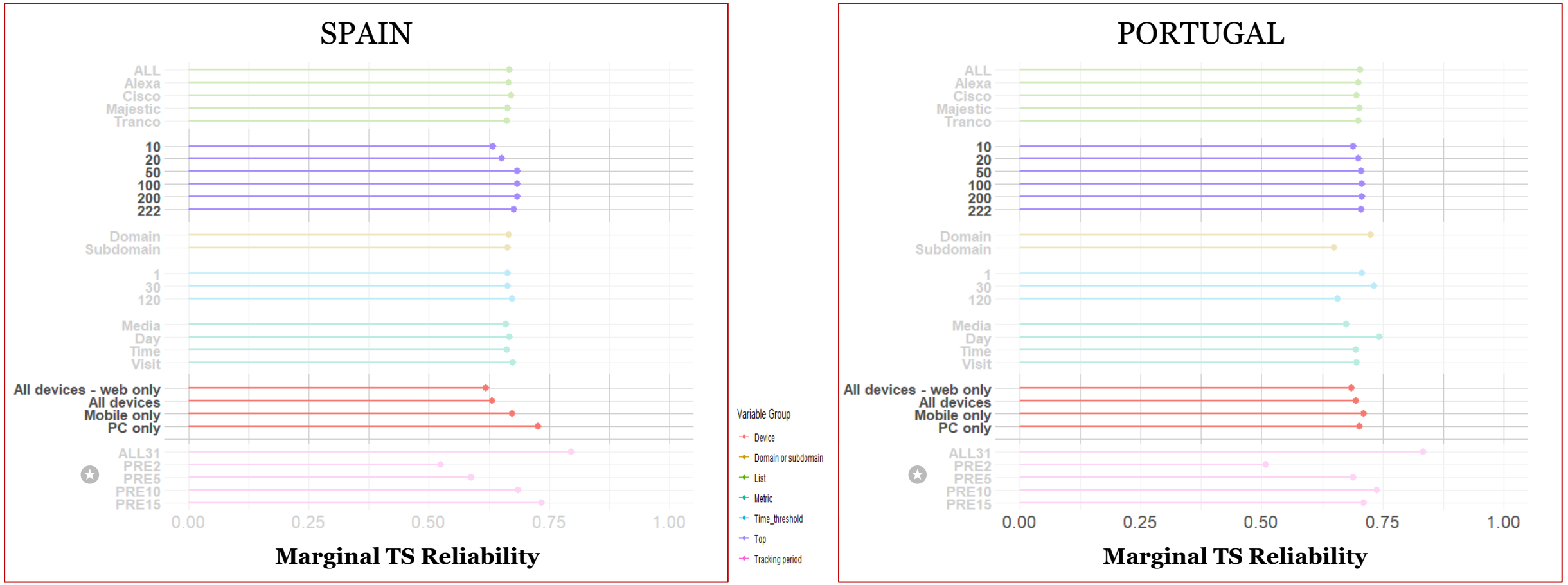
Marginal **reliability** for each design choice



The more days, the more stable the measures. Normality is key across countries

✪ Indicates variable with the highest % increase of Mean Squared Error when excluded from the model

Marginal **reliability** for each design choice



The other choices are less stable across countries. In Spain **only looking at PCs is best** but does not change much in Portugal. **Same with the number of media outlets.**

✪ Indicates variable with the highest % increase of Mean Squared Error when excluded from the model

CONCLUSIONS

Take-home messages

- Overall, the **reliability of the measures is average** (30% of the variance due to errors)

Take-home messages

- Overall, the **reliability of the measures is average** (30% of the variance due to errors)

All in all, **web tracking measures of media exposure are affected by errors**.

We should question its gold standard status.

The bigger picture

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

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

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

CONCLUSIONS

The bigger picture

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

- The paper shows that we can (1**) understand these errors**, (2) **quantify them**, and (3) **identify** which **design decision** might produce the **highest 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
  - Fertility
  - The relationship between misinformation and health outcomes

CONCLUSIONS

The bigger picture

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

- The paper shows that we can (1**) understand these errors**, (2) **quantify them**, and (3) **identify** which **design decision** might produce the **highest 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
  - Fertility
  - The relationship between misinformation and health outcomes

By helping researchers use digital trace data in the best possible way, we can foster our understanding of these pressing issues

Thanks!

Questions?

Oriol J. Bosch| PhD Candidate, The London School of Economics

o.bosch-jover@lse.ac.uk

orioljbosch

https://orioljbosch.com/

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

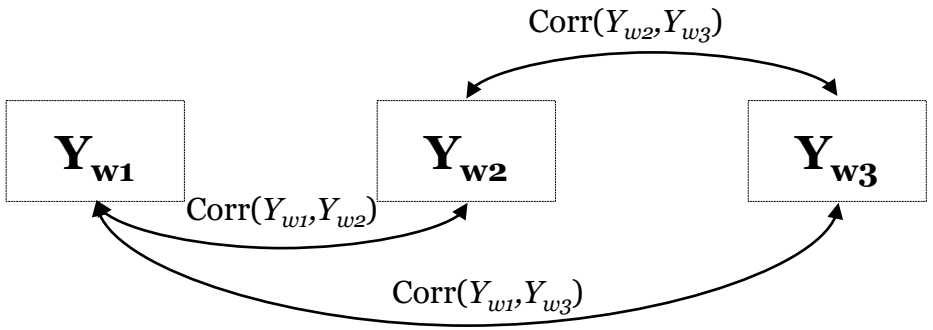
RECSM Research and Expertise Centre for Survey Methodology

web data opp

APPENDIX

Heise's approach

The (Quasi-Markov) Simplex Model



Heise's approach

$$\textbf{TS Reliability} = \frac{\text{Corr}(Y_{w1,}Y_{w2}) * \text{Corr}(Y_{w2,}Y_{w3})}{\text{Corr}(Y_{w1,}Y_{w3})}$$

APPENDIX

Main Heise's assumptions

- Measurement errors are not correlated across waves
- Reliability is constant across time periods
- True score change is not correlated in times 2 and 3, and not correlated with true score at time 1

APPENDIX

On the breach of assumptions

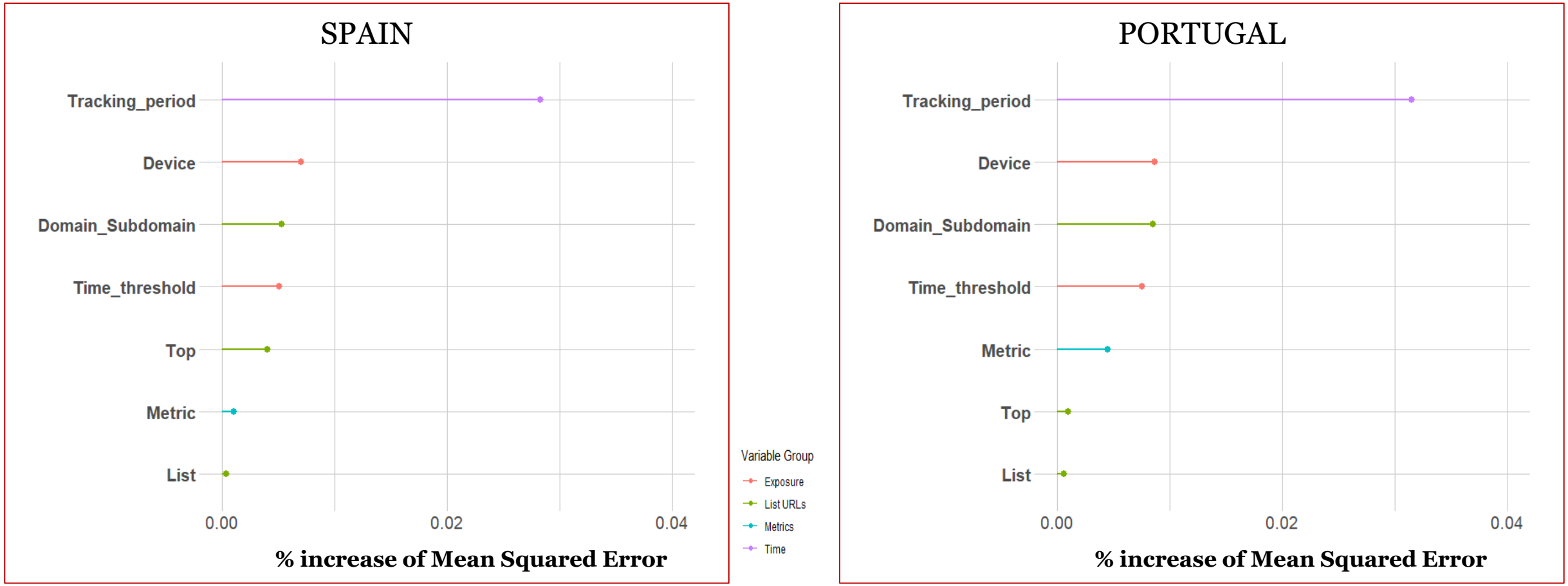
- As discussed by Prior (2013) and Torangeau, Sun and Yan (2021) breaching the assumptions when calculating TS Reliability, if anything, should inflate the reliability estimates, not deflate.

> Over-report bias, which no doubt varies across respondents and items, does not reduce calculated estimates of scale reliability. On the contrary, it tends to enhance apparent reliability. As long as survey respondents exaggerate with some degree of consistency from one exposure item to the next and one survey to the next, over-report bias cuts randomness and thereby enhances estimates of reliability.

Systematic & correlated measurement errors would inflate underlying correlation matrix

APPENDIX

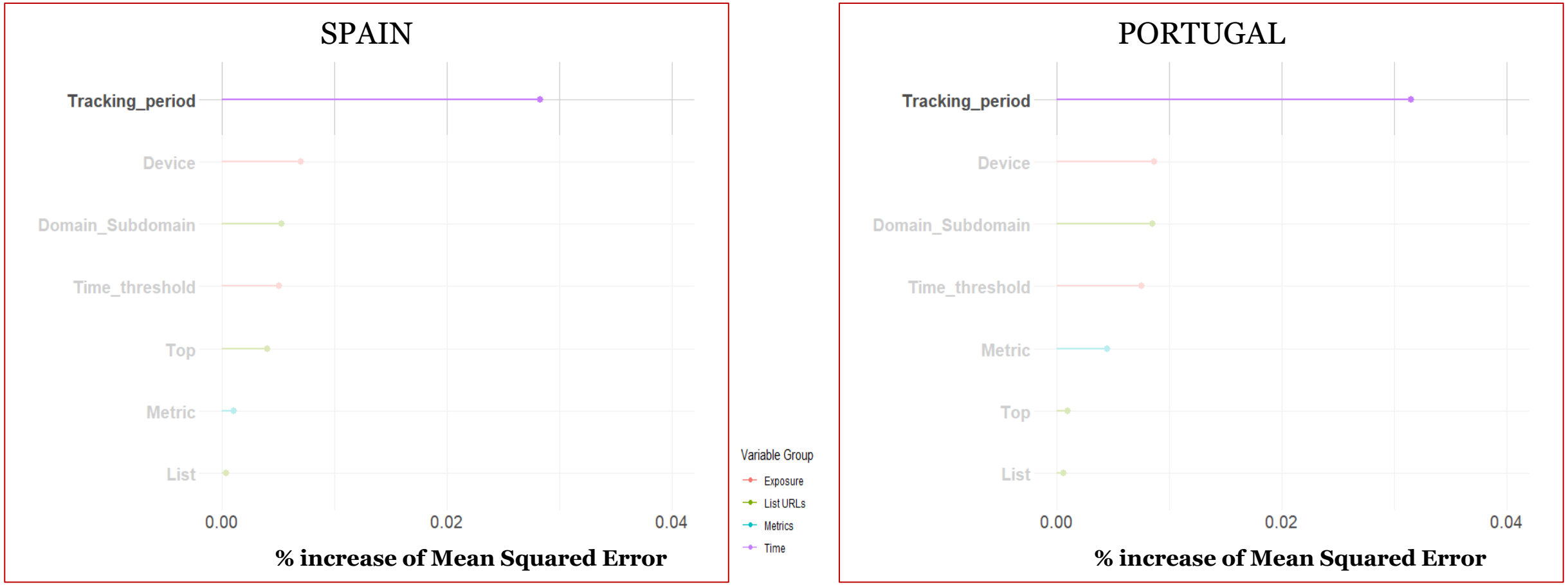
The importance of each design choice on **reliability**



* These results agree with the conditional (unbiased) important measures from cforest

APPENDIX

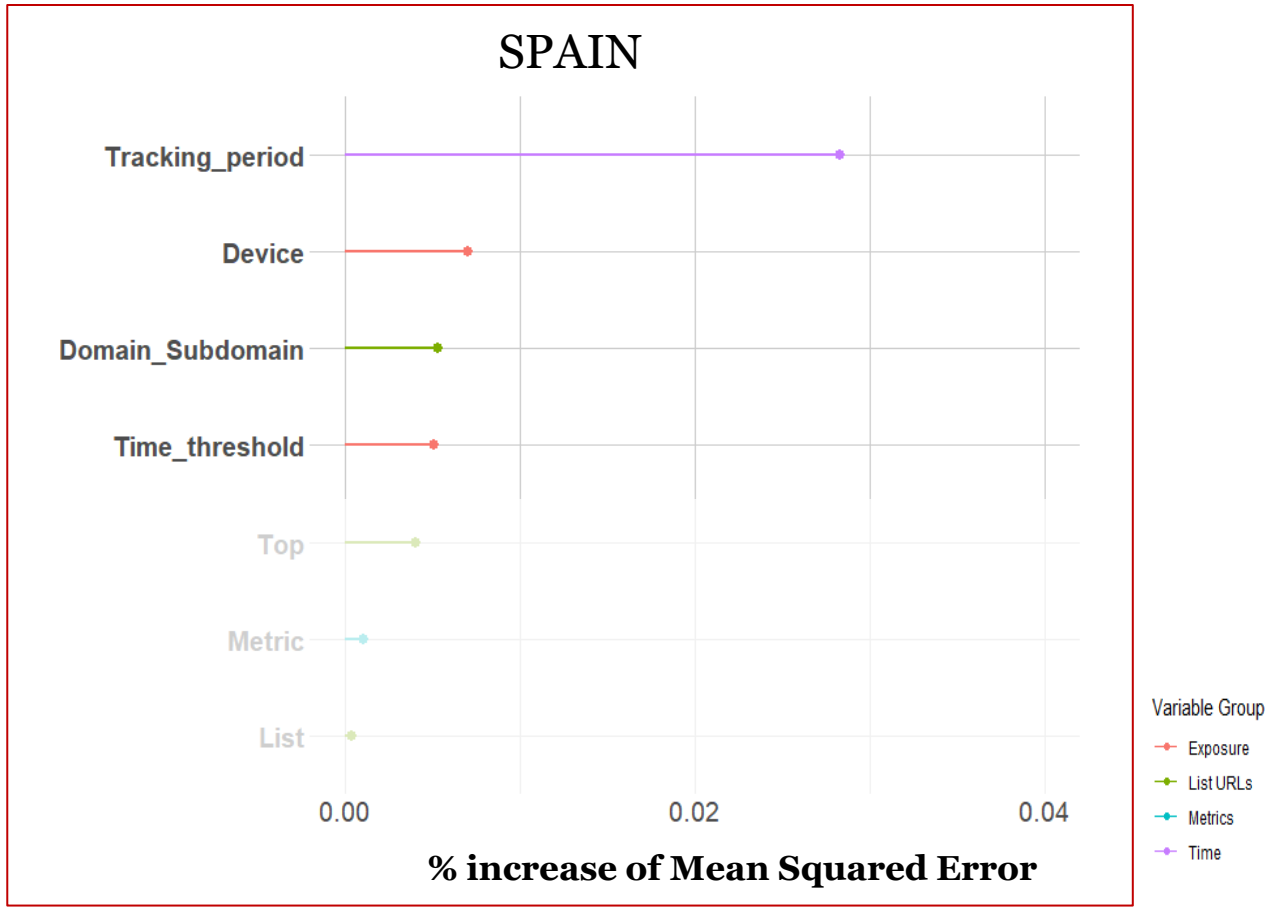
The importance of each design choice on **reliability**

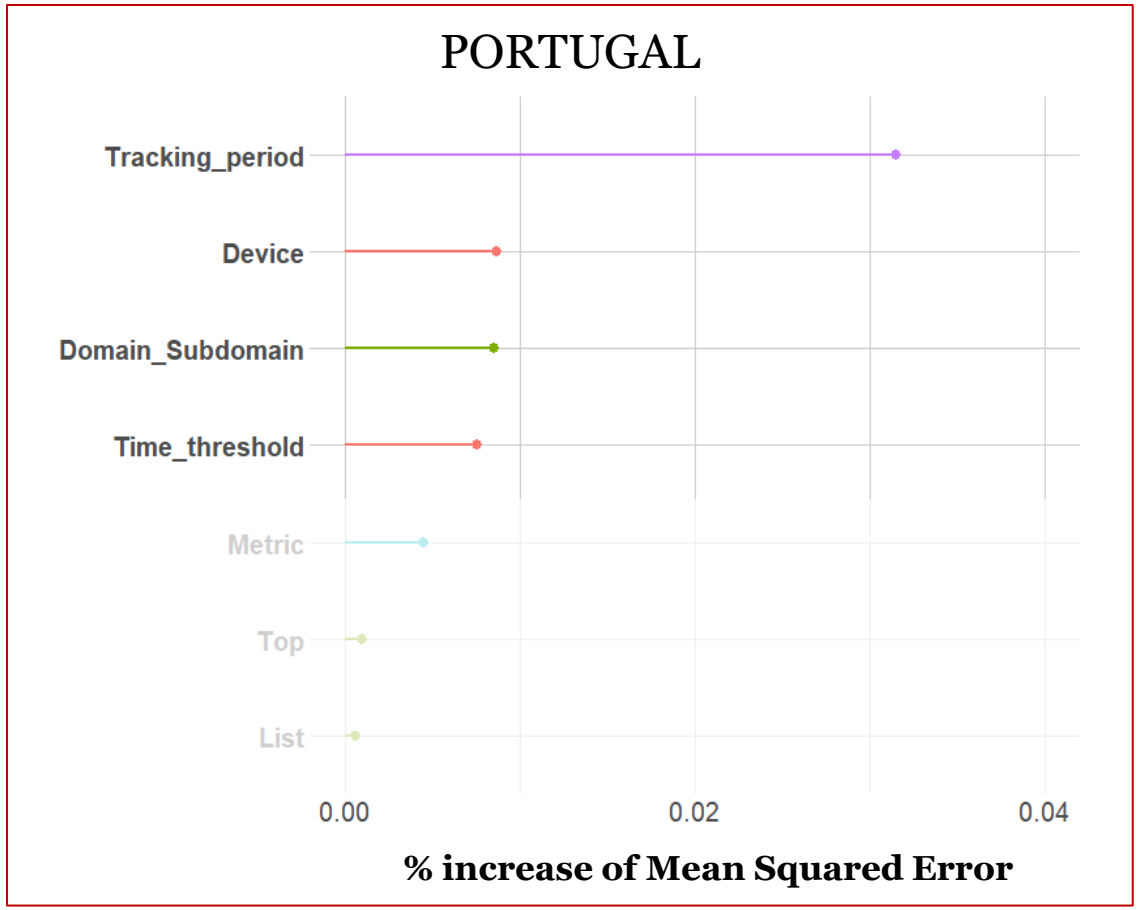


The number of days of information used is the best predictor, across countries

APPENDIX

The importance of each design choice on **reliability**





The device, type of information, and way of defining exposure also matter

APPENDIX

Model's performance

Predictive validity

Spain:

- Variance explained: 86%
- Mean Squared Residuals: .0002633571

Portugal:

- Variance explained: 93%
- Mean Squared Residuals: .0005002027

TS Reliability

Spain:

- Variance explained: 90%
- Mean Squared Residuals: .001975801

Portugal:

- Variance explained: 92%
- Mean Squared Residuals: .002011027