Tell me what you read, and I will tell you who you are: a novel method for measuring ideology using web browsing data

Oriol J. Bosch| University of Oxford

oriol.bosch-jover@demography.ox.ac.uk
orioljbosch
https://orioljbosch.com/

Acknowledgements: I would like to thank Melanie Revilla for her always insightful comments, and Yuanmo He for his brilliant computational expertise.

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). Oriol is supported by an ERC Advanced Grant (835079, PI M.C Mills), and Leverhulme Trust Large Centre Grant LCDS (RC-2018-003, PI M.C.Mills)

Can we measure ideology with web tracking data?

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, content that they saw…**

Web tracking data: a new source to measure ideology?

Web tracking data can be used to obtain **"objective" measures of participants' media diets**

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 over-reported 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

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.

Web tracking data: a new source to measure ideology?

web data opp

Web tracking data can be used to obtain **“objective” measures of participants’ media diets**

This might allow us to measure ideology

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 over-reported 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

AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

ARTICLE

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

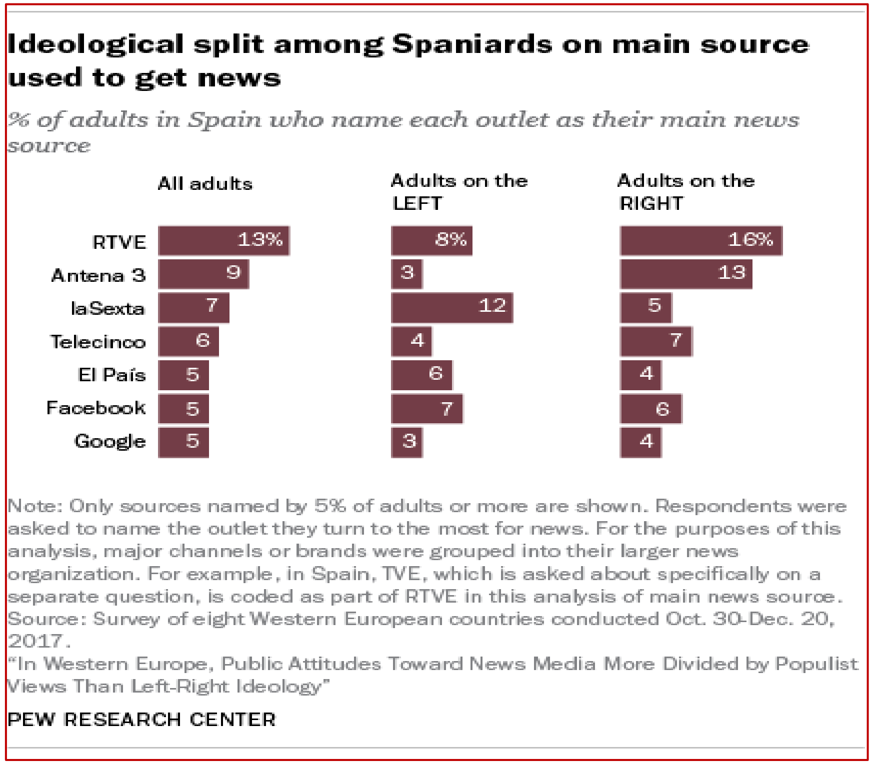
Andrew M. Guess

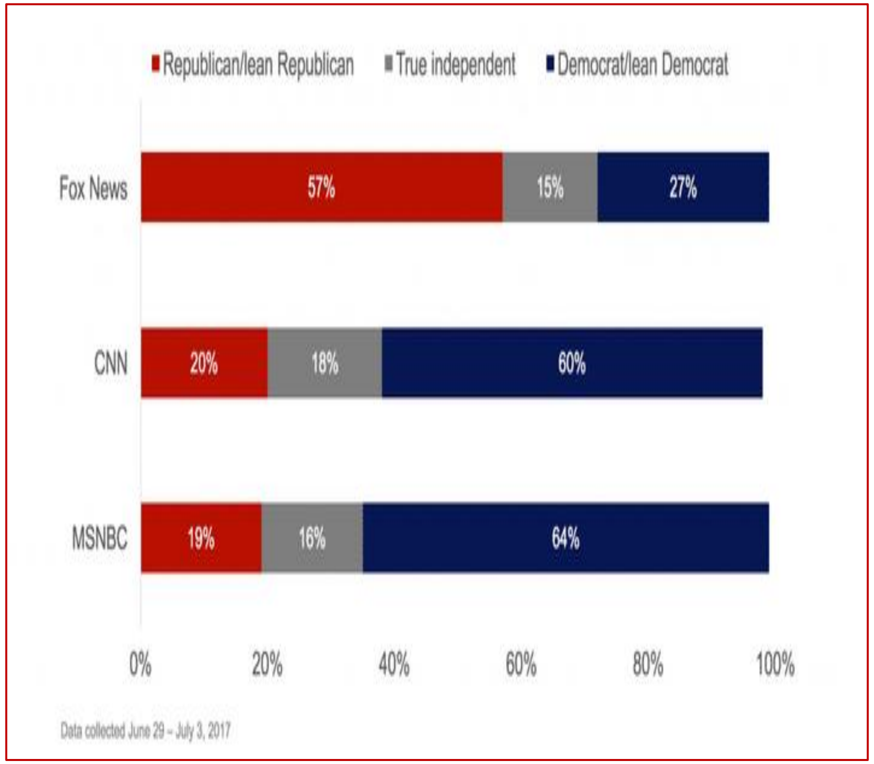
First published: 19 February 2021 | https://doi.org/10.1111/ajps.12589 | Citations: 13

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.

From observed media diets to ideology

We can assume that individuals prefer to read media outlets that they perceive to be **"close" to them in the (latent) left-right dimension**





https://www.surveymonkey.com/curiosity/cable-news-partisanship-drives-trust/
https://www.pewresearch.org/global/fact-sheet/news-media-and-political-attitudes-in-spain/

Why would we want to measure ideology with web tracking data?

Why would we want to measure ideology with web tracking data?

1. **Supplement (online) behavioural data** with attitudinal information without the need of self-reports (not always feasible)

Why would we want to measure ideology with web tracking data?

1. **Supplement (online) behavioural data** with attitudinal information without the need of self-reports (not always feasible)

2. **Measure media outlet's ideology** at a scale without relying on content analysis

Why would we want to measure ideology with web tracking data?

1. **Supplement (online) behavioural data** with attitudinal information without the need of self-reports (not always feasible)

2. **Measure media outlet's ideology** at a scale without relying on content analysis

Why would we want to measure ideology with web tracking data?

1. **Supplement (online) behavioural data** with attitudinal information without the need of self-reports (not always feasible)

2. **Measure media outlet's ideology** at a scale without relying on content analysis

3. Even if of lower quality than self-reports (my expectation), **combining self-reports and web-tracking** data could **improve** our **understanding of the errors of self-reports**, and the **overall quality** of the estimates we use

Why would we want to measure ideology with web tracking data?

1. **Supplement (online) behavioural data** with attitudinal information without the need of self-reports (not always feasible)

2. **Measure media outlet's ideology** at a scale without relying on content analysis

3. Even if of lower quality than self-reports (my expectation), **combining self-reports and web-tracking** data could **improve** our **understanding of the errors of self-reports**, and the **overall quality** of the estimates we use

   - Understand and quantify potential errors of self-reports: **problems in the centre and the extremes**

   - **Create a new, hopefully, better measure of ideology**

THIS STUDY

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]

Gobierno de España, Ministerio de Ciencia, Innovación y Universidades

Fundación BBVA

The University of Mississippi

upf. Universitat Pompeu Fabra Barcelona

RECSM Research and Expertise Centre for Survey Methodology

Georgia State University

Vanderbilt University

THIS STUDY

Case study for this presentation: Spain

1. The left-right dimension is very relevant in Spain

2. Spain has a highly partisan, pluralist media system

3. And a polarized multiparty system

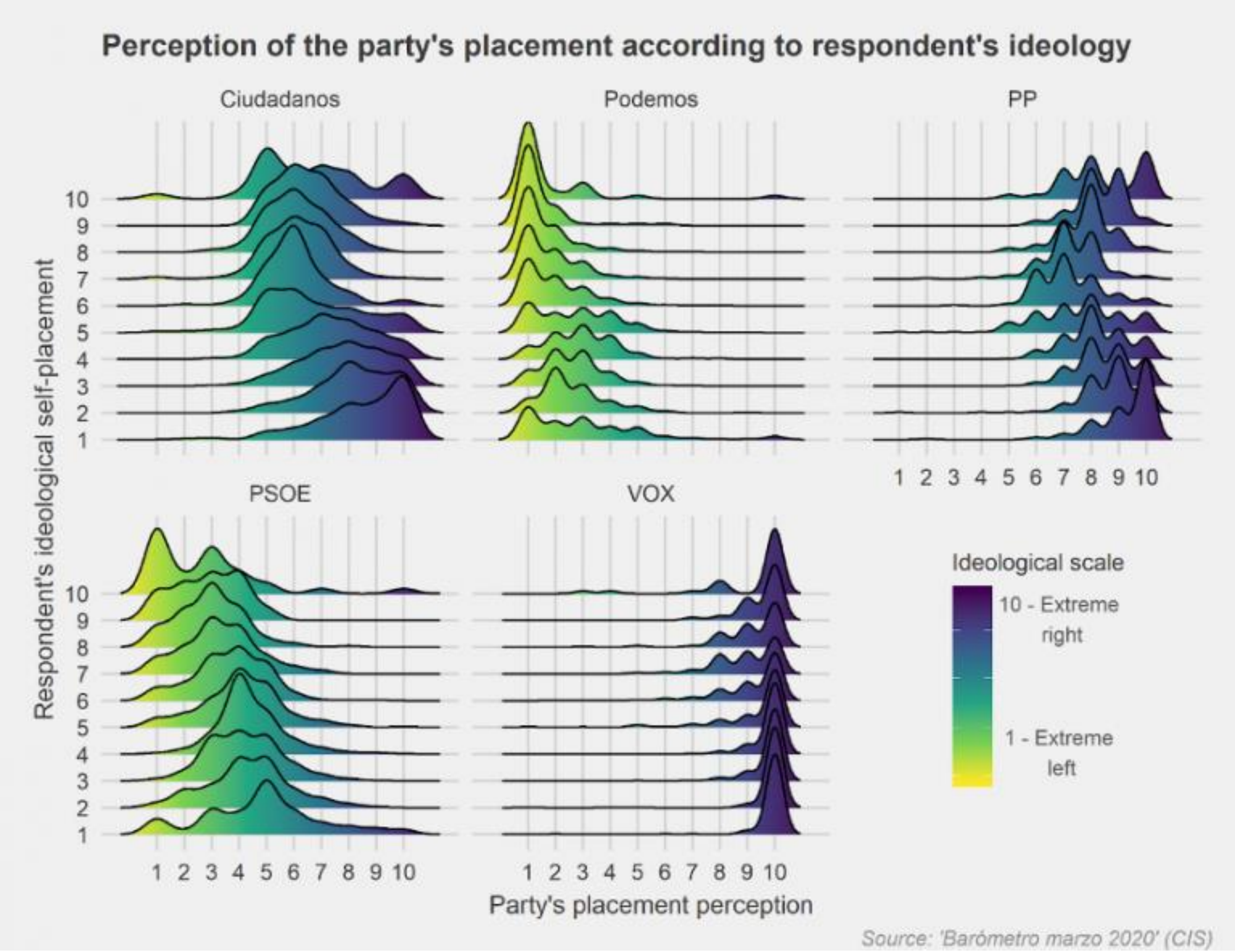
PSOE

PODEMOS

C's

PP

VOX



https://blogs.lse.ac.uk/eurocrisispress/2020/06/26/polarization-coronavirus/

ESTIMATING IDEOLOGY WITH WEB TRACKING DATA

The underlying model

An individual's (i) decision to read a specific media outlet (j) is a function of:

1. The ideological distance between them and the outlet (d_{ij}).
2. Plus some user- and media- random effects (α_i an β_j), to account for differences in political interest and popularity of media.

$$\boxed{\Pr(Y_{ij} = 1 | \alpha_i, \beta_j, d_{ij}) = \text{Logit}(\alpha_i + \beta_j - d_{ij})}$$

The underlying model

This approach has already been used to measure the ideology and socioeconomic status of individuals based on what accounts they follow on Twitter

General Article

Psychological Science
2015, Vol. 26(10) 1531–1542

DOI: 10.1177/0956797615594620
pss.sagepub.com
SAGE

Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?

Pablo Barberá[1], John T. Jost[1,2,3], Jonathan Nagler[3], Joshua A. Tucker[3], and Richard Bonneau[4]
[1]Center for Data Science, [2]Department of Psychology, [3]Department of Politics, and [4]Center for Genomics and Systems Biology, New York University

Abstract
We estimated ideological preferences of 3.8 million Twitter users and, using a data set of nearly 150 million tweets concerning 12 political and nonpolitical issues, explored whether online communication resembles an "echo chamber" (as a result of selective exposure and ideological segregation) or a "national conversation." We observed that information was exchanged primarily among individuals with similar ideological preferences in the case of political issues (e.g., 2012 presidential election, 2013 government shutdown) but not many other current events (e.g., 2013 Boston Marathon bombing, 2014 Super Bowl). Discussion of the Newtown shootings in 2012 reflected a dynamic process, beginning as a national conversation before transforming into a polarized exchange. With respect to both political and nonpolitical issues, liberals were more likely than conservatives to engage in cross-ideological dissemination; this is an important asymmetry with respect to the structure of communication that is consistent with psychological theory and research bearing on ideological differences in epistemic, existential, and relational motivation. Overall, we conclude that previous work may have overestimated the degree of ideological segregation in social-media usage.

Original Article

Sociological Methods & Research
1–36

DOI: 10.1177/00491241231168665
journals.sagepub.com/home/smr
SAGE

A Method for Estimating Individual Socioeconomic Status of Twitter Users

Yuanmo He
and Milena Tsvetkova

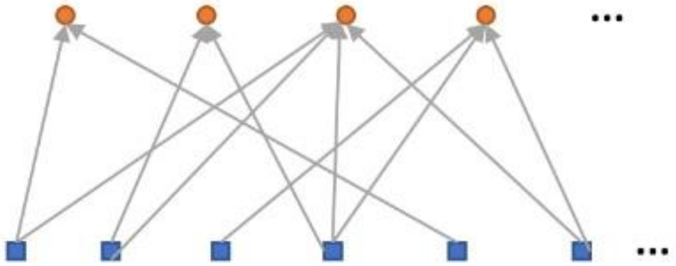
Abstract
The rise of social media has opened countless opportunities to explore social science questions with new data and methods. However, research on socioeconomic inequality remains constrained by limited individual-level socioeconomic status (SES) measures in digital trace data. Following Bourdieu, we argue that the commercial and entertainment accounts Twitter users follow reflect their economic and cultural capital. Adapting a political science method for inferring political ideology, we use correspondence analysis to estimate the SES of 3,482,652 Twitter users who follow the accounts of 339 brands in the United States. We validate our estimates with data from the Facebook Marketing application programming interface, self-reported job titles on users' Twitter profiles, and a small survey sample. The results show reasonable correlations with the standard proxies for SES, alongside much weaker or nonsignificant correlations with other demographic variables. The proposed method opens new opportunities for innovative social research on inequality on Twitter and similar online platforms.

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

From model to estimates: Correspondence Analysis

I adapt Pablo Barbera's approach to measure ideology based on who users follow on Twitter, using **Correspondence Analysis**

Media outlets

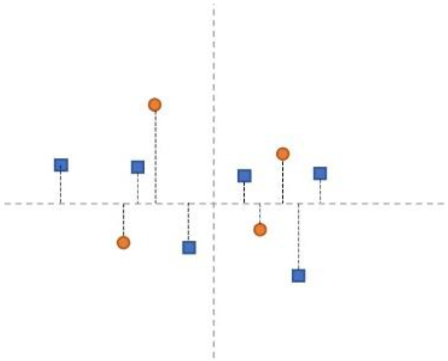


Participants

	Outlet$_1$	Outlet$_2$	Outlet$_3$	...
Participant$_1$	1	0	0	
Participant$_2$	1	1	0	
Participant$_3$	0	0	0	
Participant$_4$	1	1	1	
Participant$_5$	0	0	1	
...				

Correspondence Analysis
1. Compute matrix of standardized residuals
2. Use SVD to get orthogonal components
3. Project to get principal coordinates

Dimension 2

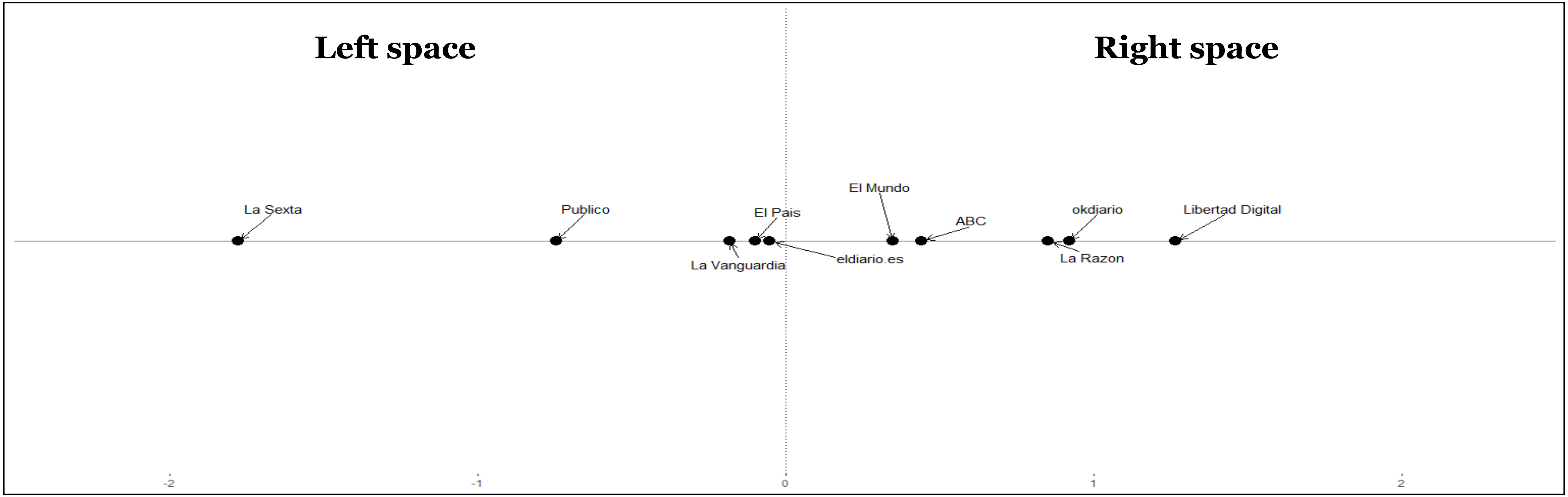


Dimension 1

The ideology of media outlets

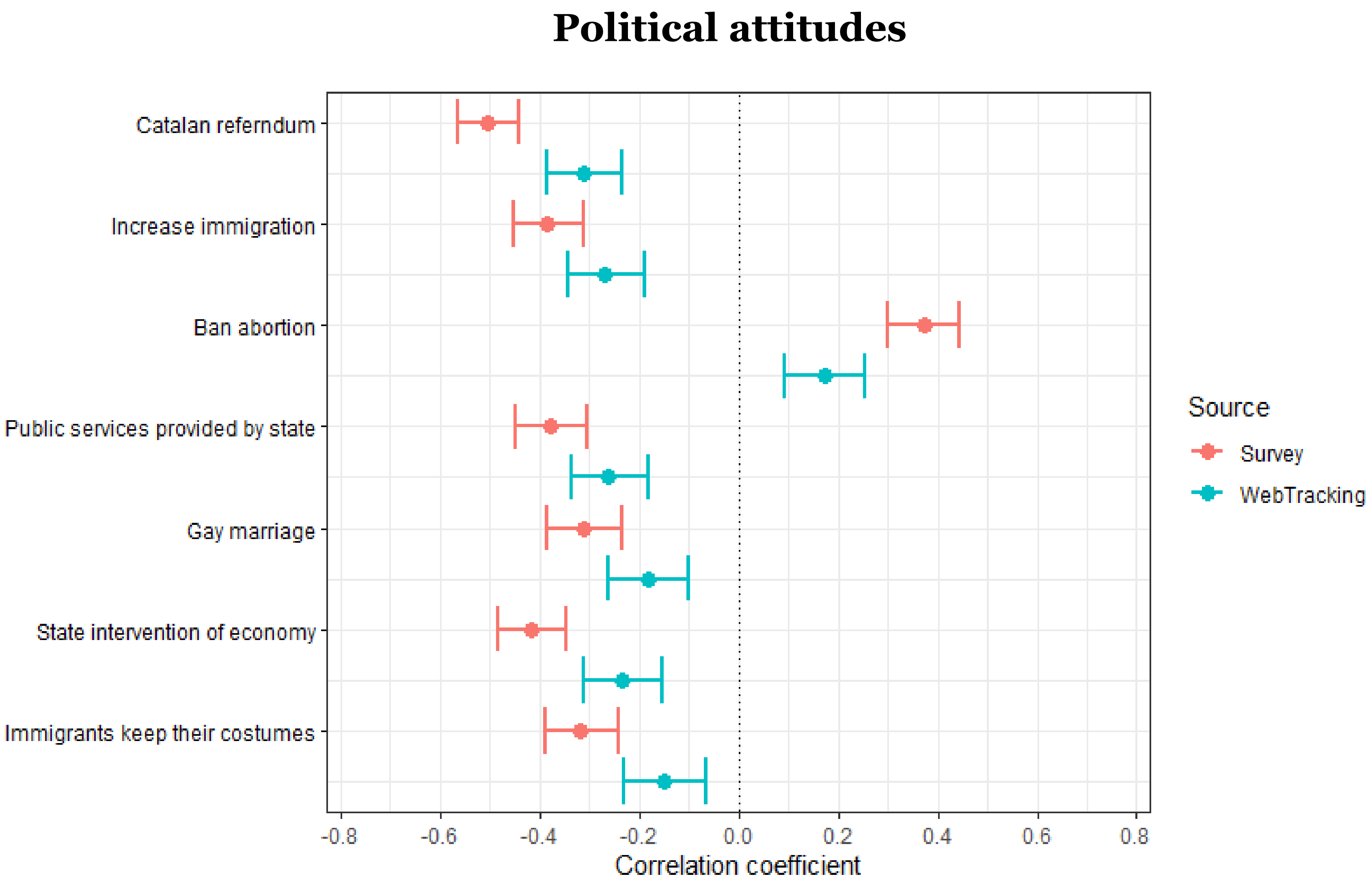
VALIDATING THE SCALE

The ideology of media outlets



Predictive validity

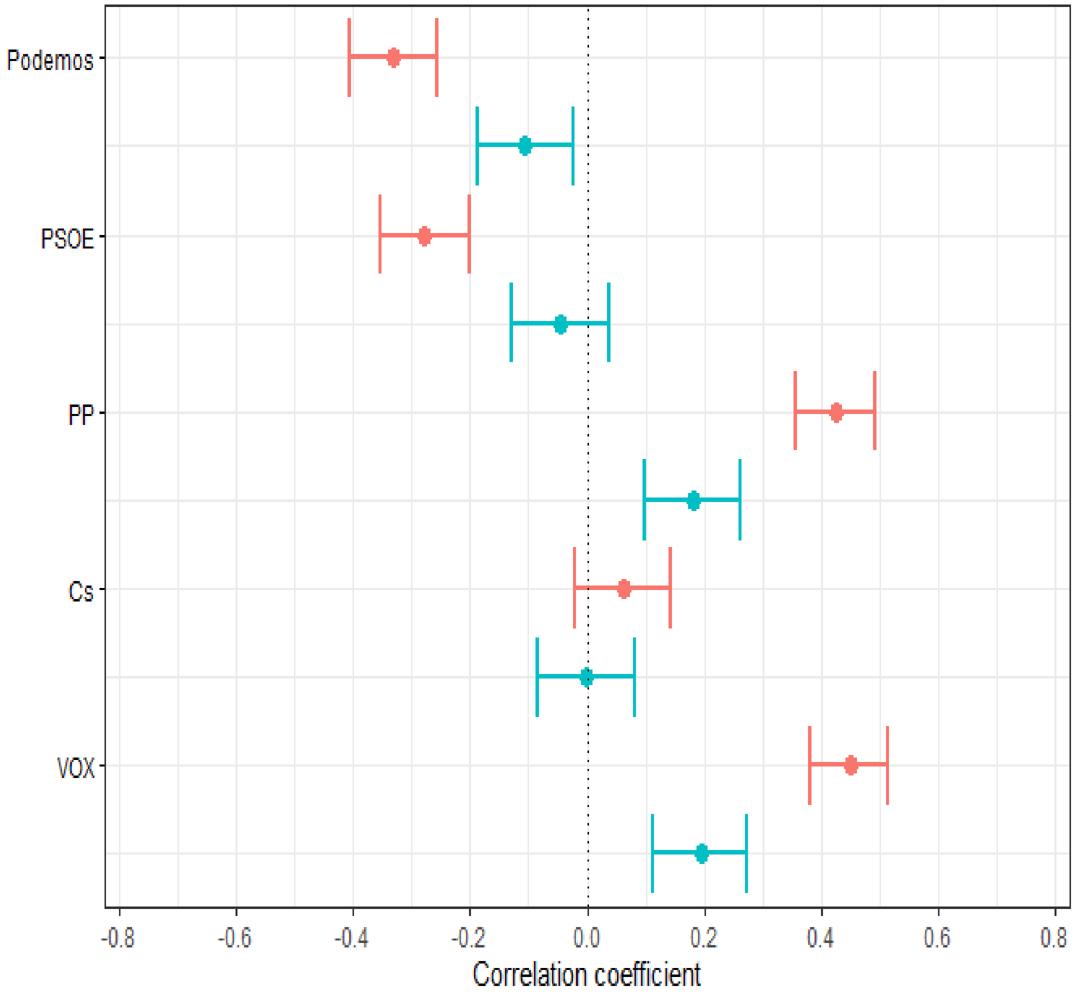
Predictive validity



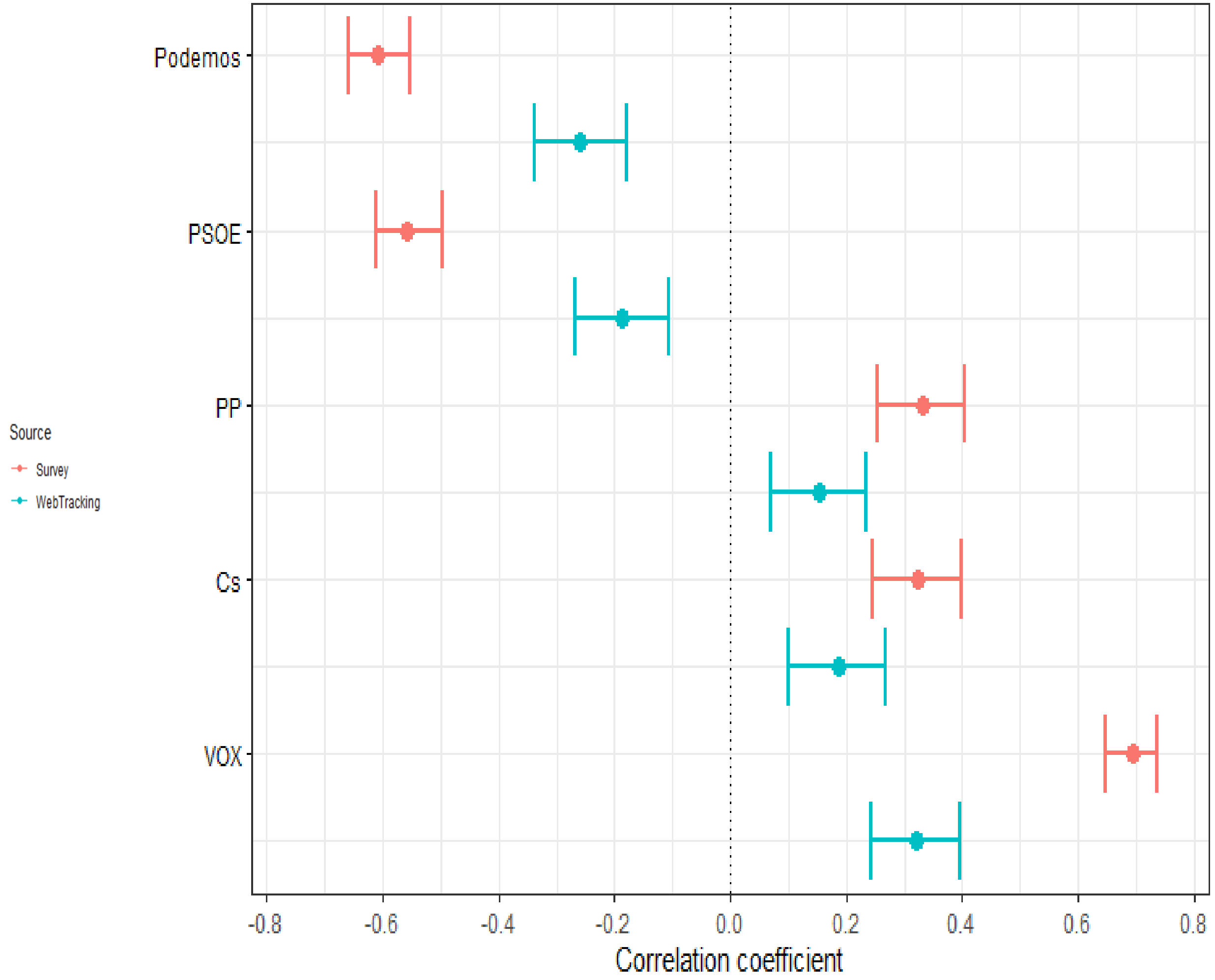
VALIDATING THE SCALE

Predictive validity

Voting intention



Attitudes towards candidates from…

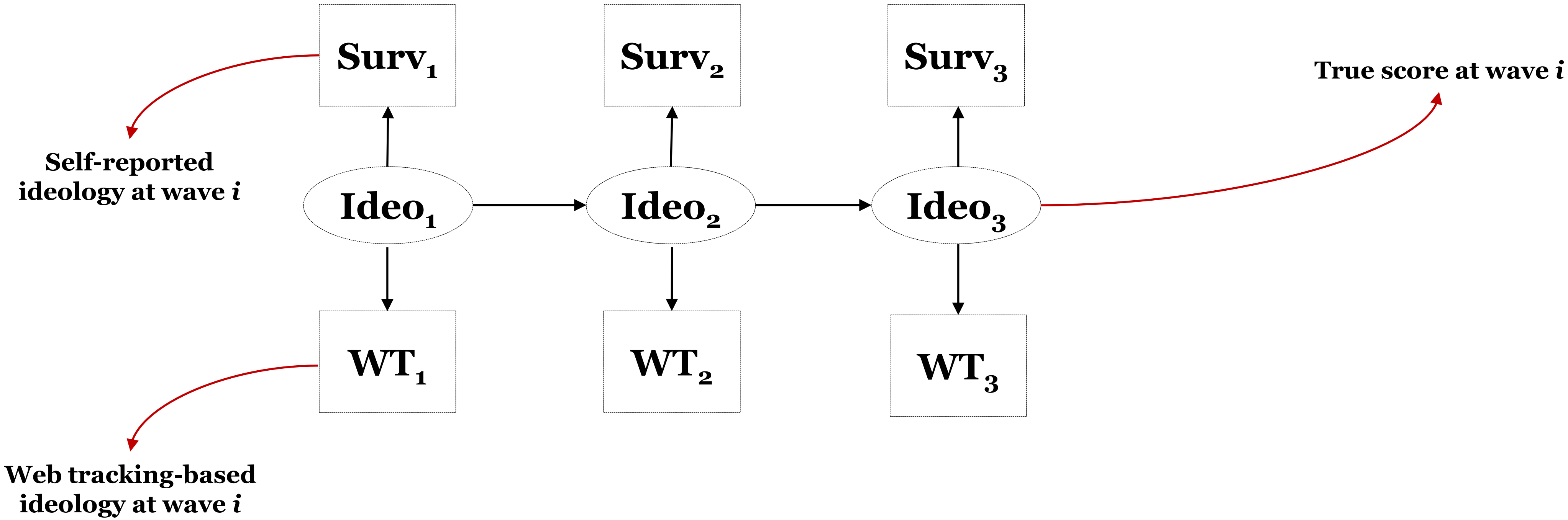


WHAT CAN WE LEARN BY COMBINING BOTH ESTIMATES?

Hidden Markov Models to estimate the quality of both sources

web data opp

- Group of latent class models used to **estimate and correct for measurement error** in categorical, longitudinal data
- Do **not require any of data sources to be error-free**



Misclassification error (5 categories)

web data opp

	Hidden classes				
	Class 1 (Far-left)	Class 2 (Left)	Class 3 (Centre)	Class 4 (Right)	Class 5 (Far-right)
Survey					
Far-left	**.82**	.03	.00	.00	.02
Left	.18	**.94**	.03	.02	.00
Centre	.00	.02	**.87**	.02	.00
Right	.00	.02	.09	**.94**	.09
Far-right	.00	.00	.01	.02	**.89**
Web tracking					
Far-left	.01	.01	.00	.00	.00
Left	**.55**	**.47**	.31	.23	.19
Centre	.14	.12	.16	.11	.15
Right	.30	.39	**.52**	**.64**	**.64**
Far-right	.00	.00	.01	.01	.02

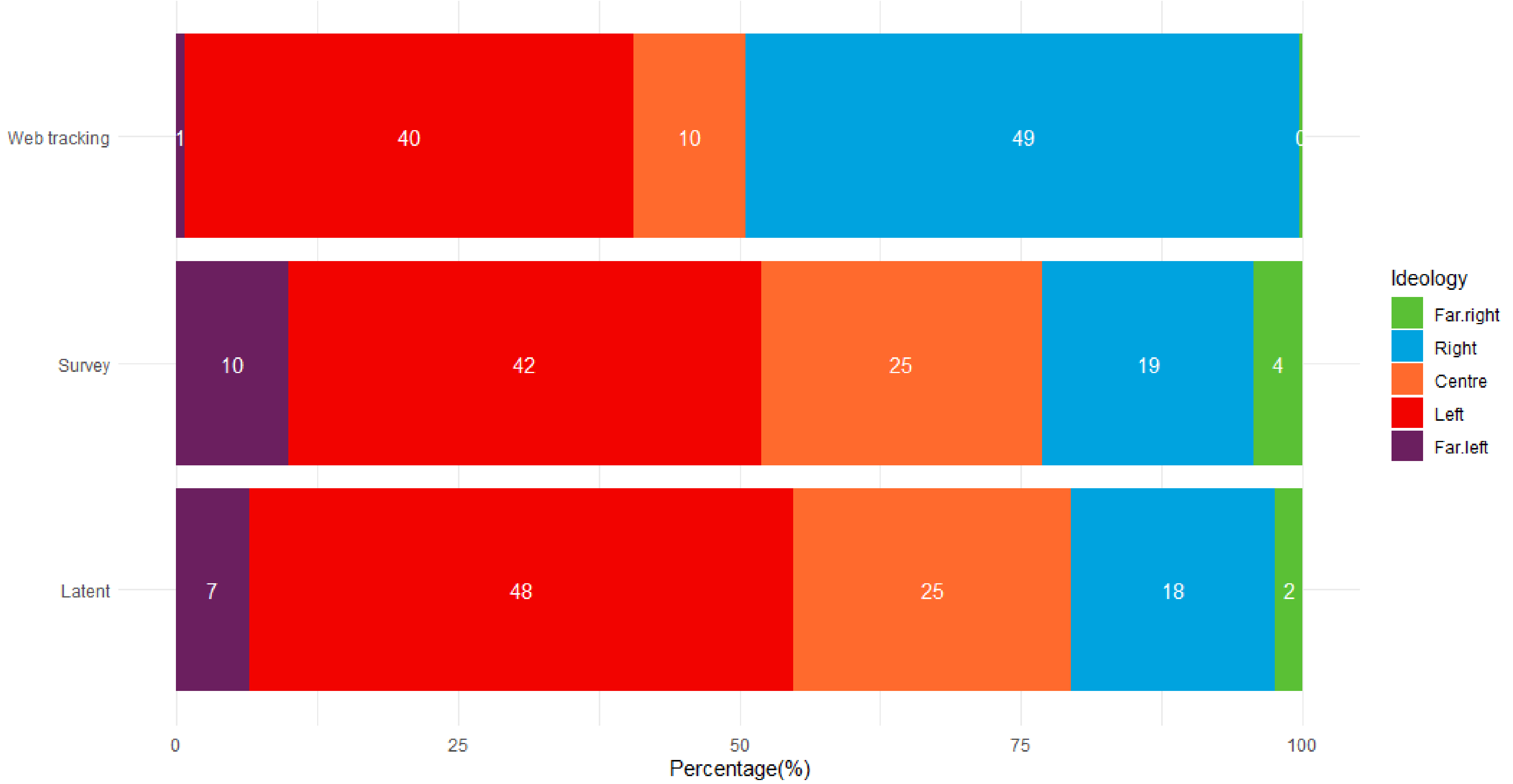
Misclassification error (5 categories)

	Hidden classes				
	Class 1 (Far-left)	Class 2 (Left)	Class 3 (Centre)	Class 4 (Right)	Class 5 (Far-right)
Survey					
Far-left	**.82**	.03	.00	.00	.02
Left	.18	**.94**	.03	.02	.00
Centre	.00	.02	**.87**	.02	.00
Right	.00	.02	.09	**.94**	.09
Far-right	.00	.00	.01	.02	**.89**
Web tracking					
Far-left	.01	.01	.00	.00	.00
Left	**.55**	**.47**	.31	.23	.19
Centre	.14	.12	.16	.11	.15
Right	.30	.39	**.52**	**.64**	**.64**
Far-right	.00	.00	.01	.01	.02

Misclassification error (5 categories)

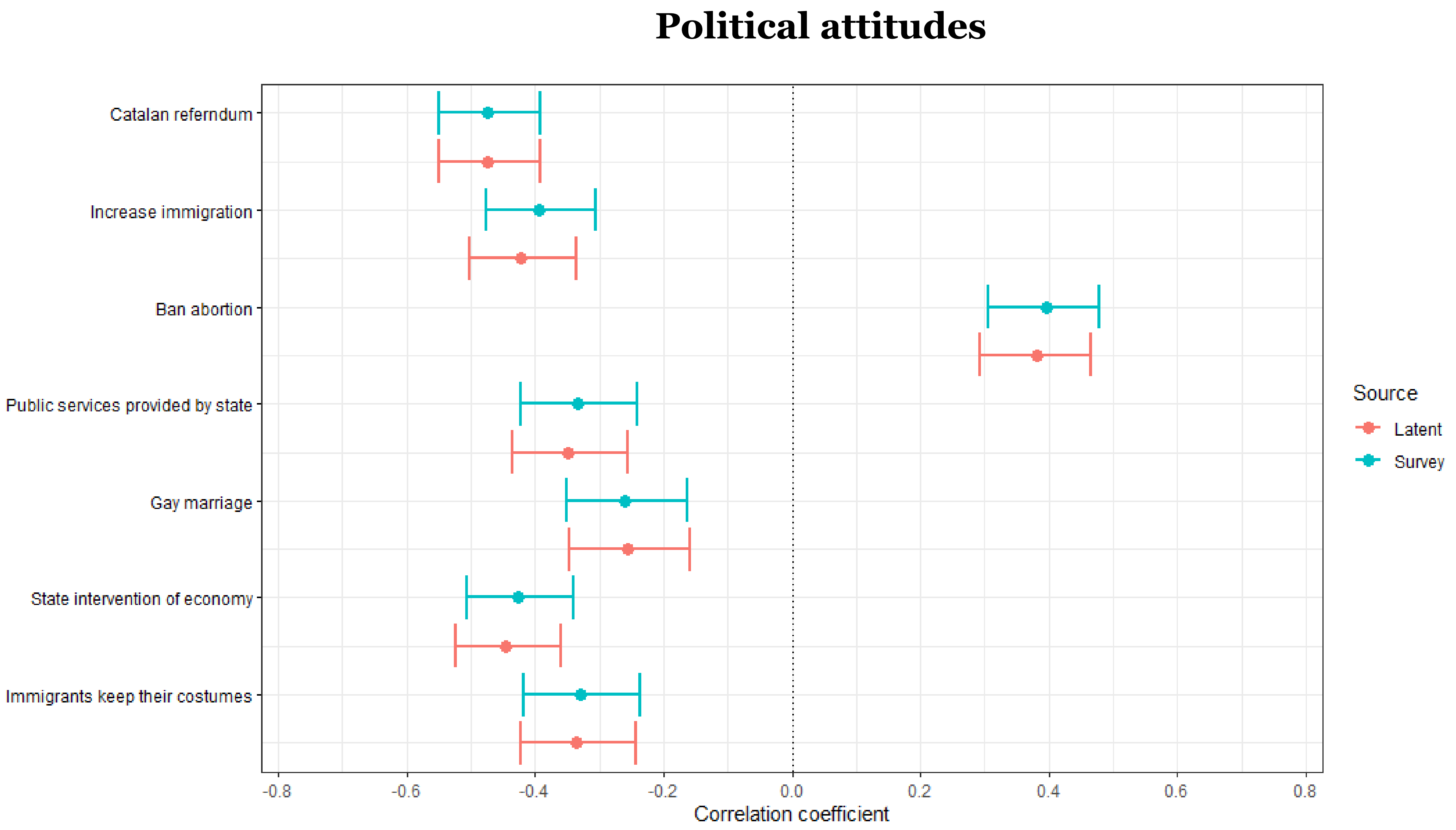
	Hidden classes				
	Class 1 (Far-left)	Class 2 (Left)	Class 3 (Centre)	Class 4 (Right)	Class 5 (Far-right)
Survey					
Far-left	**.82**	.03	.00	.00	.02
Left	.18	**.94**	.03	.02	.00
Centre	.00	.02	**.87**	.02	.00
Right	.00	.02	.09	**.94**	.09
Far-right	.00	.00	.01	.02	**.89**
Web tracking					
Far-left	.01	.01	.00	.00	.00
Left	**.55**	**.47**	.31	.23	.19
Centre	.14	.12	.16	.11	.15
Right	.30	.39	**.52**	**.64**	**.64**
Far-right	.00	.00	.01	.01	.02

How do they compare to the latent “true” ideology?



CAN WE IMPROVE THE SELF-REPORT?

Predictive validity



CONCLUSIONS

Take-home messages

- Promising approach to combine surveys and web tracking data
- It is possible to create a measure of ideology using web tracking data but far from perfect!
- Although survey self-reports do seem to have more problems identifying people on the extremes and the centre, the overall quality of the measure is very high
- There might be avenues for improvement, but the results suggest that surveys do a very good job

Thanks!

Questions?

Oriol J. Bosch | Postdoctoral Researcher, University of Oxford

oriol.bosch-jover@demography.ox.ac.uk

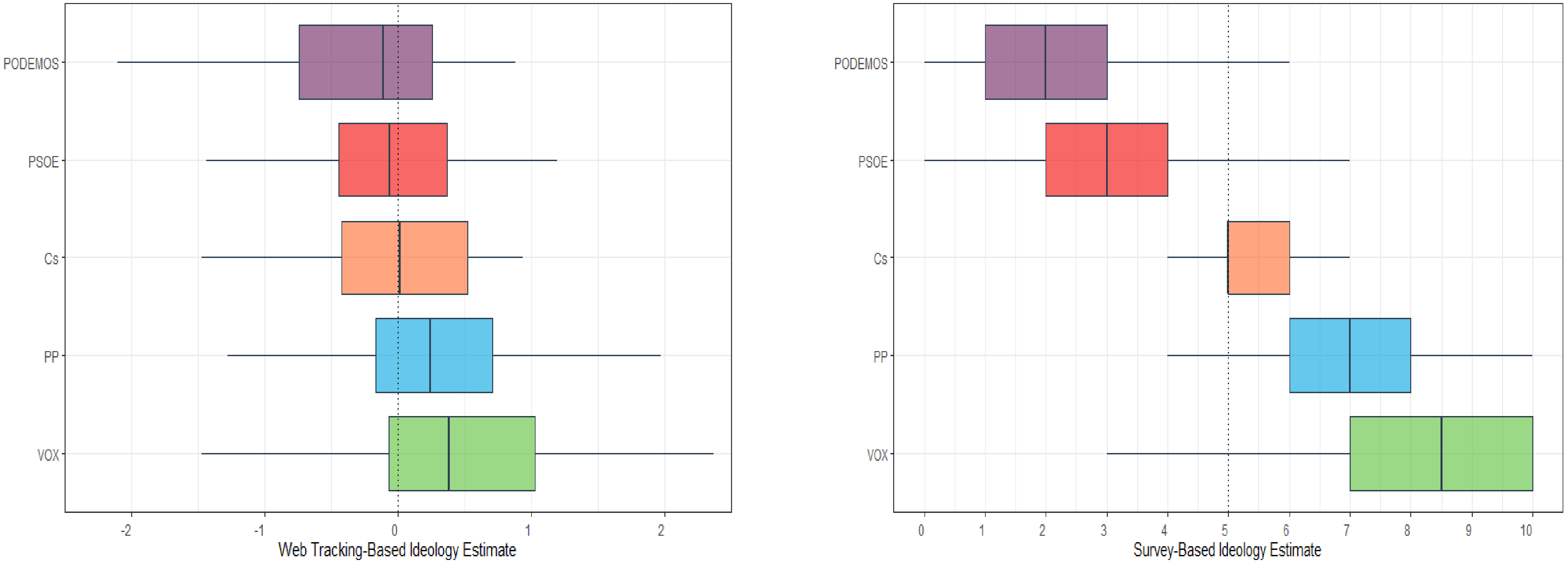
orioljbosch

https://orioljbosch.com/

Correspondence Analysis

Correspondence analysis considers $\mathbf{Y}$, the $n \times m$ adjacency matrix indicating whether user i (row) follows user j (column), as a representation of a set of points in a multidimensional space. This matrix is converted into the correspondence matrix $\mathbf{P}$ by dividing by its grand total, $\mathbf{P} = \mathbf{Y}/\sum_{ij} y_{ij}$, and used to compute the matrix of standardized residuals, $\mathbf{S}$, where $\mathbf{S} = \mathbf{D}_r^{1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^T)\mathbf{D}_c^{1/2}$, where $\mathbf{r}$ and $\mathbf{c}$ are the row and column masses, with $r_i = \sum_j p_{ij}$ and $c_j = \sum_j p_{ij}$, which are then used to construct the diagonal matrices $\mathbf{D}_r = \text{diag}(\mathbf{r})$ and $\mathbf{D}_c = \text{diag}(\mathbf{c})$. As described in Bonica (2013b), this step is equivalent to including the random effects α_i and β_j in the estimation. $\mathbf{S}$ is therefore a matrix of residuals between the observed and expected values based on the marginal distribution of the following matrix $\mathbf{Y}$; and correspondence analysis will scale the rows and columns under the assumption that these deviations respond to the distance between them on a latent multidimensional space.

Self-reported and predicted ideology, by party proximity



CAN WE IMPROVE THE SELF-REPORT?

Predictive validity

web data opp

Voting intention