gesis Leibniz Institute for the Social Sciences

Coverage Bias in Election Research using Data from Social Media

Authors:
Hannah Bucher (hannah.bucher@gesis.org)
Marie Kühn(marie.kuehn@gesis.org)
Joss Roßmann (joss.rossmann@gesis.org)

Member of Leibniz Association

Social Media Data in Election Research

- **predicting election outcomes** by looking at the frequency of party mentions on **twitter** or using more sophisticated methods (e.g. Tumasjan, Sprenger et al. 2010 or Oliveira, Bermejo et al. 2017 for Spain and the USA, for an overview, see Gayo-Avello 2013 or Phillips, Dowling et al. 2017)
- **predicting public opinion or political orientation** analyzing **twitter** content (for an overview, see Phillips, Dowling et al. 2017)

Previous Research

- some mainly descriptive research investigating the **composition of demographics, opinions and personality traits** on **different social media platforms** (e.g. Greenwood 2016, Barberá and Rivero 2015, Ruths and Pfeffer 2014 or Ryan and Xenos 2011) with a focus on **twitter** (for an overview see Jungherr 2016) using social media data
- some **collections of possible data quality issues and concerns for social media data** (e.g. Ruths and Pfeffer 2014 or Gayo-Avello 2013)
- attempts to characterize users of different social media platforms using survey data:
  - Rainie, Smith et al. 2012 looking at political participation on social media in the USA amongst mostly social media users only
  - Blank and Lutz 2017 looking only at internet users and mostly at demographic variables in Great Britain
  - Blank 2017 looking at twitter users and nonusers

Research Questions

1. who is underrepresented in analyses based on social media data?
2. to what extent does this misrepresentation of certain parts of the population bias common estimates in election research?

Data

- data: GLES cross-sectional surveys before and after the German federal election 2017 (Roßteutscher et al. 2018, doi:10.4232/1.13139)

Year	mode	n	Sampling strategy	Social media
2017	CAPI	4294	stratified multistage sample (oversampling for East Germany)	Facebook, WhatsApp, YouTube, Twitter, Google+

gesis Leibniz Institute for the Social Sciences

Research Design

Model	Subset analyzed	Model type	Dependent Variable	Independent Variables
1	all	multinomial logistic	internet and social media usage	gender
				age
2	social media users	logistic	social media usage by platform: Facebook, WhatsApp, YouTube, Twitter, Google+	educational level
				marital status
				household size
				net income
3	social media users	linear OLS	days of internet usage per week	region (east/west)
				migration
				economic situation (prospective)
4	social media users	linear OLS	number of social media platforms used	justice (ego)
				recall (turnout 2013)
				party identification
5	social media users	logistic	form of social media usage: commenting and posting	left-right self-assessment
				ego positions
				populism (Akkerman et al.)

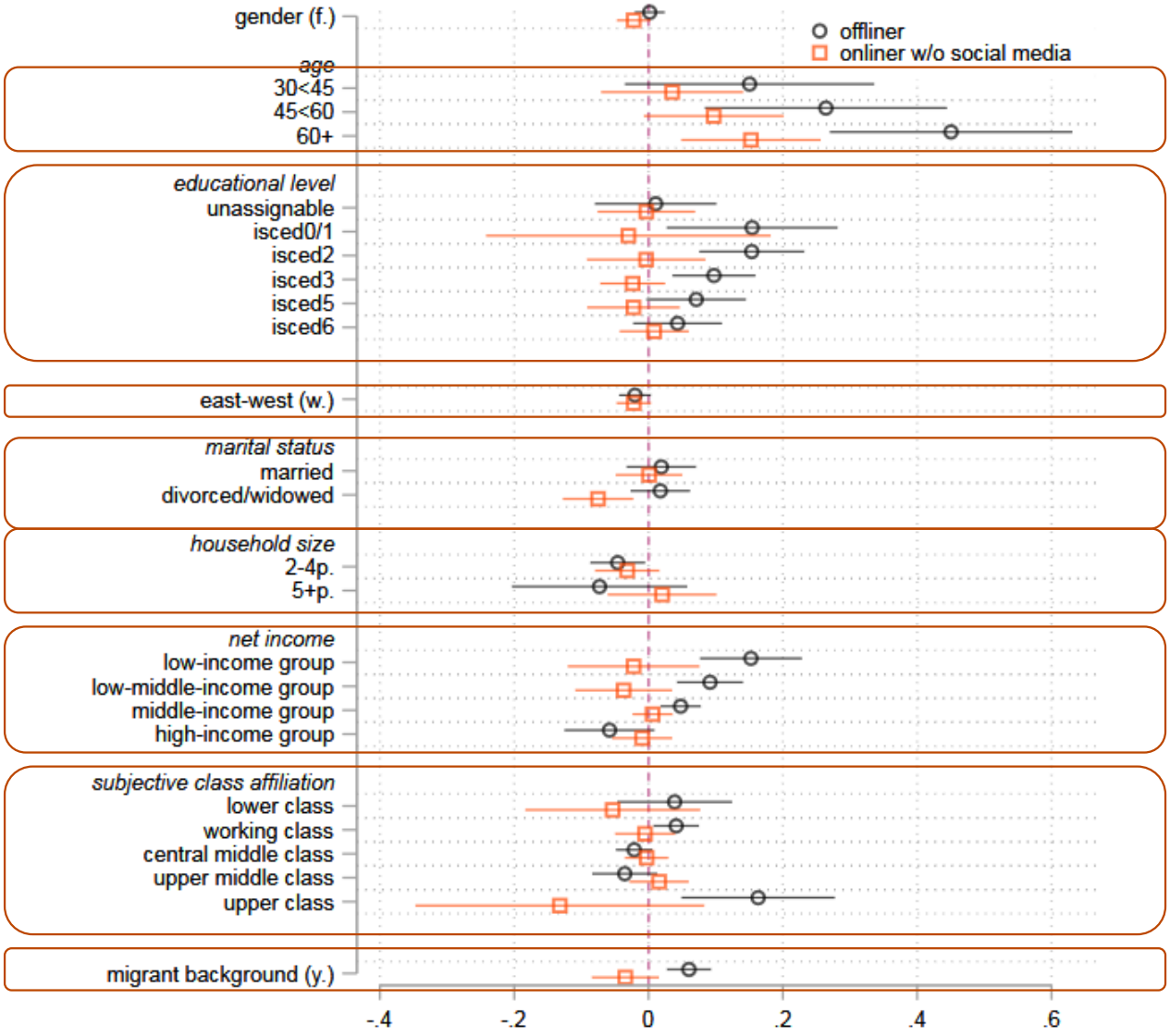
gesis Leibniz Institute for the Social Sciences

Research Design

Model	Subset analyzed	Model type	Dependent Variable	Independent Variables
1	all	multinomial logistic	internet and social media usage	gender
				age
2	social media users	logistic	social media usage by platform: Facebook, WhatsApp, YouTube, Twitter, Google+	educational level
				marital status
				household size
				net income
3	social media users	linear OLS	days of internet usage per week	region (east/west)
				migration
				economic situation (prospective)
4	social media users	linear OLS	number of social media platforms used	justice (ego)
				recall (turnout 2013)
				party identification
5	social media users	logistic	form of social media usage: commenting and posting	left-right self-assessment
				ego positions
				populism (Akkerman et al.)

gesis Leibniz Institute for the Social Sciences

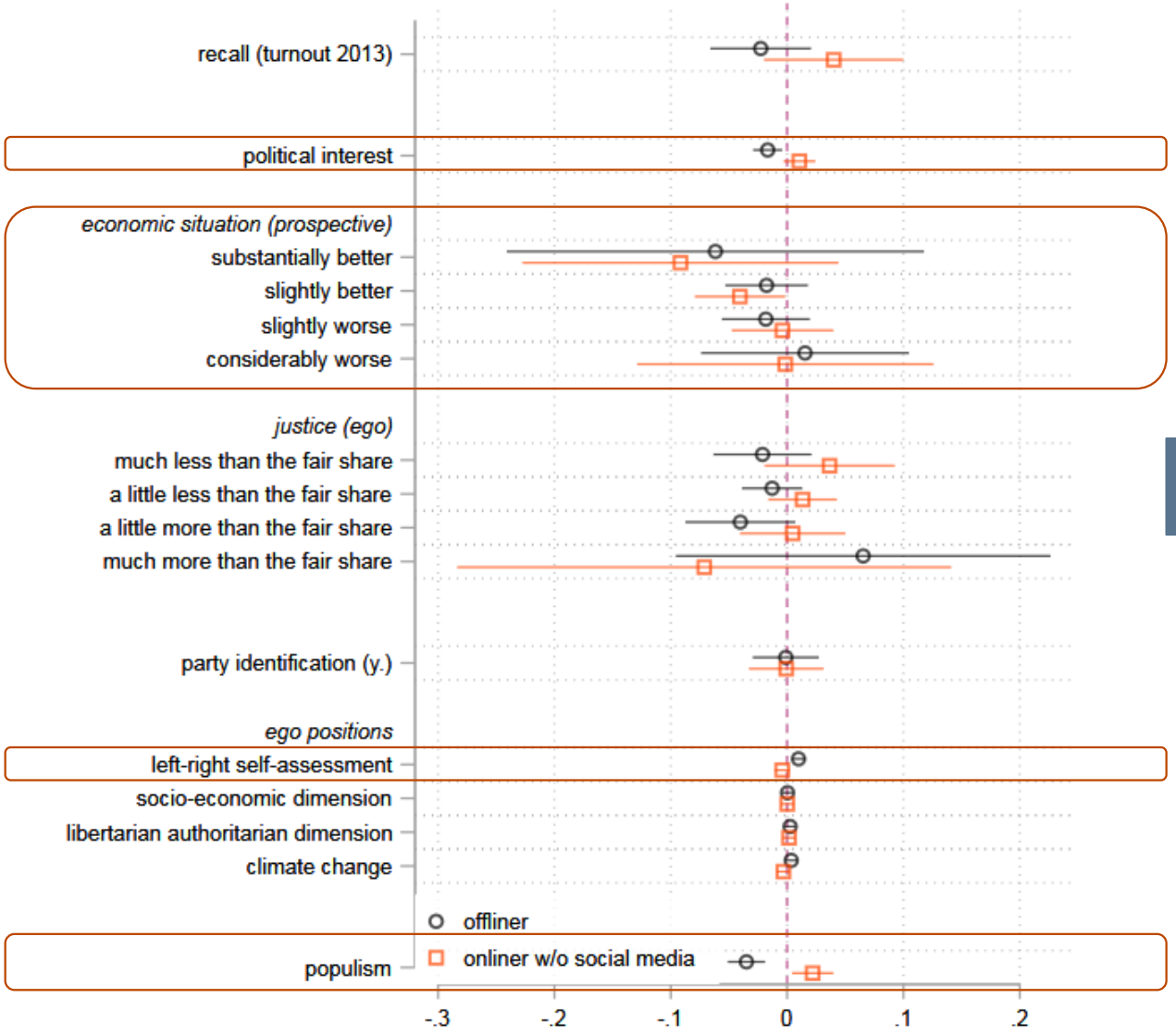
Differences in sociodemographics Model 1



Ref.: social media user
N= 3120
Pseudo R^2 = 0.29

gesis Leibniz Institute for the Social Sciences

Differences in attitudes Model 1



Ref.: social media user
N= 3120
Pseudo R^2 = 0.29

Summary Model 1

- the populations differ on key demographics
- different mechanisms driving internet usage and social media usage
- evidence for different attitudes between online and offline population and different assessment of own prospective economic situation between social media users and nonusers

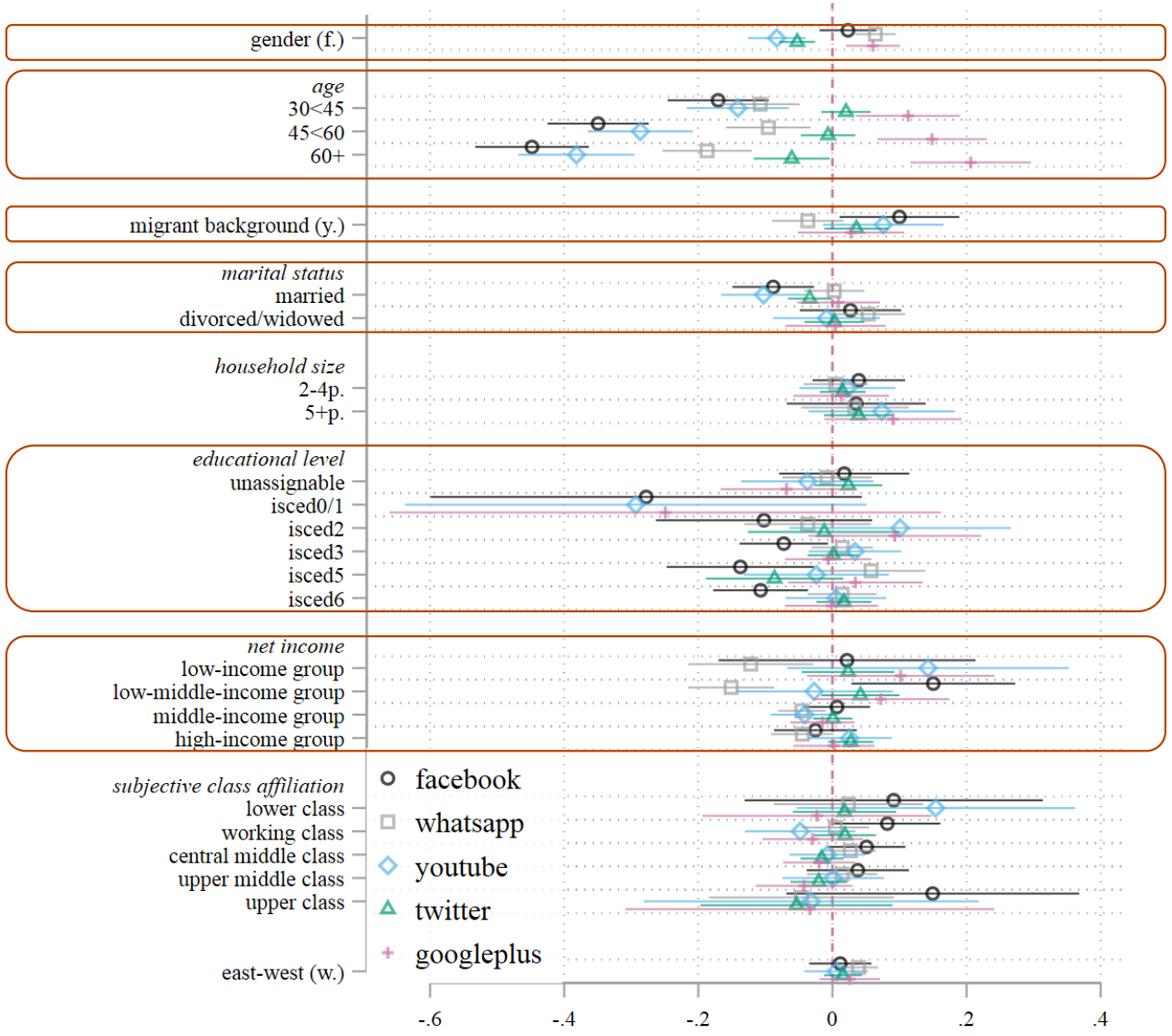
gesis Leibniz Institute for the Social Sciences

Research Design

Model	Subset analyzed	Model type	Dependent Variable	Independent Variables
1	all	multinomial logistic	internet and social media usage	gender
				age
2	social media users	logistic	social media usage by platform: Facebook, WhatsApp, YouTube, Twitter, Google+	educational level
				marital status
				household size
				net income
3	social media users	linear OLS	days of internet usage per week	region (east/west)
				migration
				economic situation (prospective)
4	social media users	linear OLS	number of social media platforms used	justice (ego)
				recall (turnout 2013)
				party identification
5	social media users	logistic	form of social media usage: commenting and posting	left-right self-assessment
				ego positions
				populism (Akkerman et al.)

gesis Leibniz Institute for the Social Sciences

Differences in sociodemographics Model 2



Subpopulation Social Media Users

DV: facebook (y/n)
N= 2145
Pseudo R^2 = 0.15

DV: whatsapp (y/n)
N= 2136
Pseudo R^2 = 0.11

DV: youtube (y/n)
N= 2145
Pseudo R^2 = 0.11

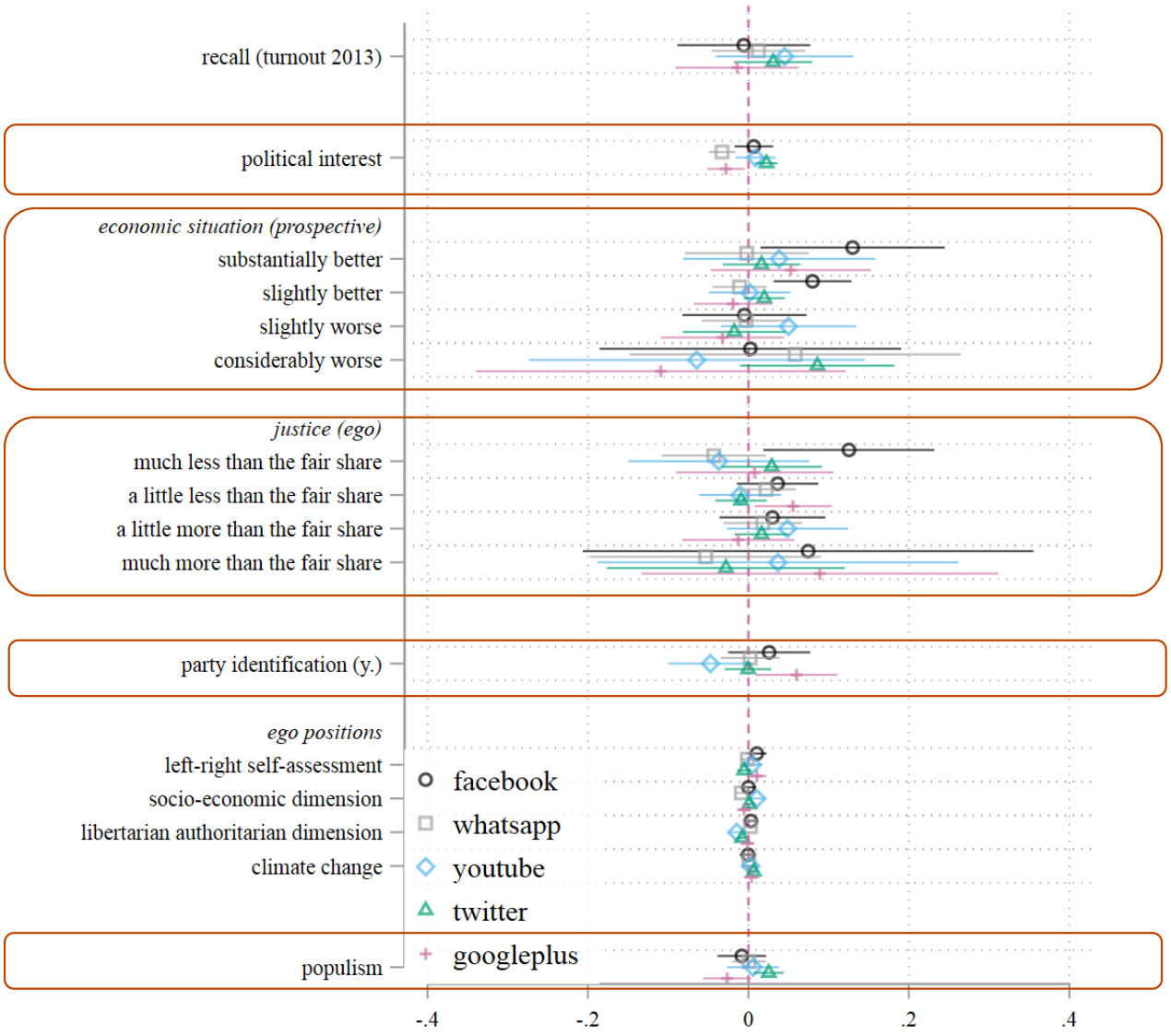
DV: twitter (y/n)
N= 2136
Pseudo R^2 = 0.13

DV: google+ (y/n)
N= 2145
Pseudo R^2 = 0.04

Member of Leibniz Association

gesis Leibniz Institute for the Social Sciences

Differences in attitudes Model 2



Subpopulation Social Media Users

DV: facebook (y/n)
N= 2145
Pseudo R^2 = 0.15

DV: whatsapp (y/n)
N= 2136
Pseudo R^2 = 0.11

DV: youtube (y/n)
N= 2145
Pseudo R^2 = 0.11

DV: twitter (y/n)
N= 2136
Pseudo R^2 = 0.13

DV: google+ (y/n)
N= 2145
Pseudo R^2 = 0.04

Summary Model 2

- different demographic selectivities for different social media platforms
- possible difference in attitude measures especially among facebook, whatsapp and google+ users

gesis Leibniz Institute for the Social Sciences

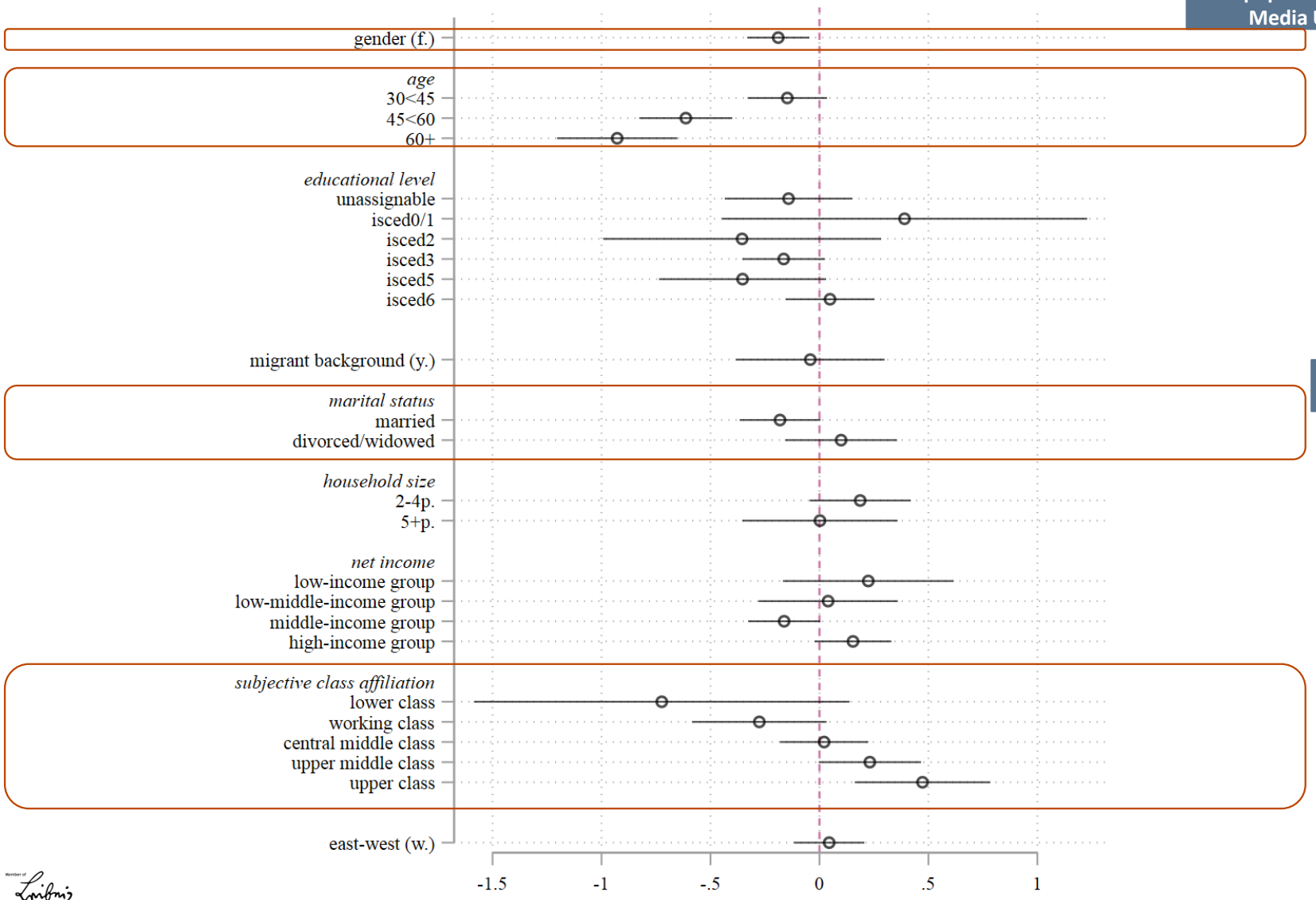
Research Design

Model	Subset analyzed	Model type	Dependent Variable	Independent Variables
1	all	multinomial logistic	internet and social media usage	gender
				age
2	social media users	logistic	social media usage by platform: Facebook, WhatsApp, YouTube, Twitter, Google+	educational level
				marital status
				household size
				net income
3	social media users	linear OLS	days of internet usage per week	region (east/west)
				migration
				economic situation (prospective)
4	social media users	linear OLS	number of social media platforms used	justice (ego)
				recall (turnout 2013)
				party identification
5	social media users	logistic	form of social media usage: commenting and posting	left-right self-assessment
				ego positions
				populism (Akkerman et al.)

gesis Leibniz Institute for the Social Sciences

Differences in sociodemographics Model 3

Subpopulation Social Media Users

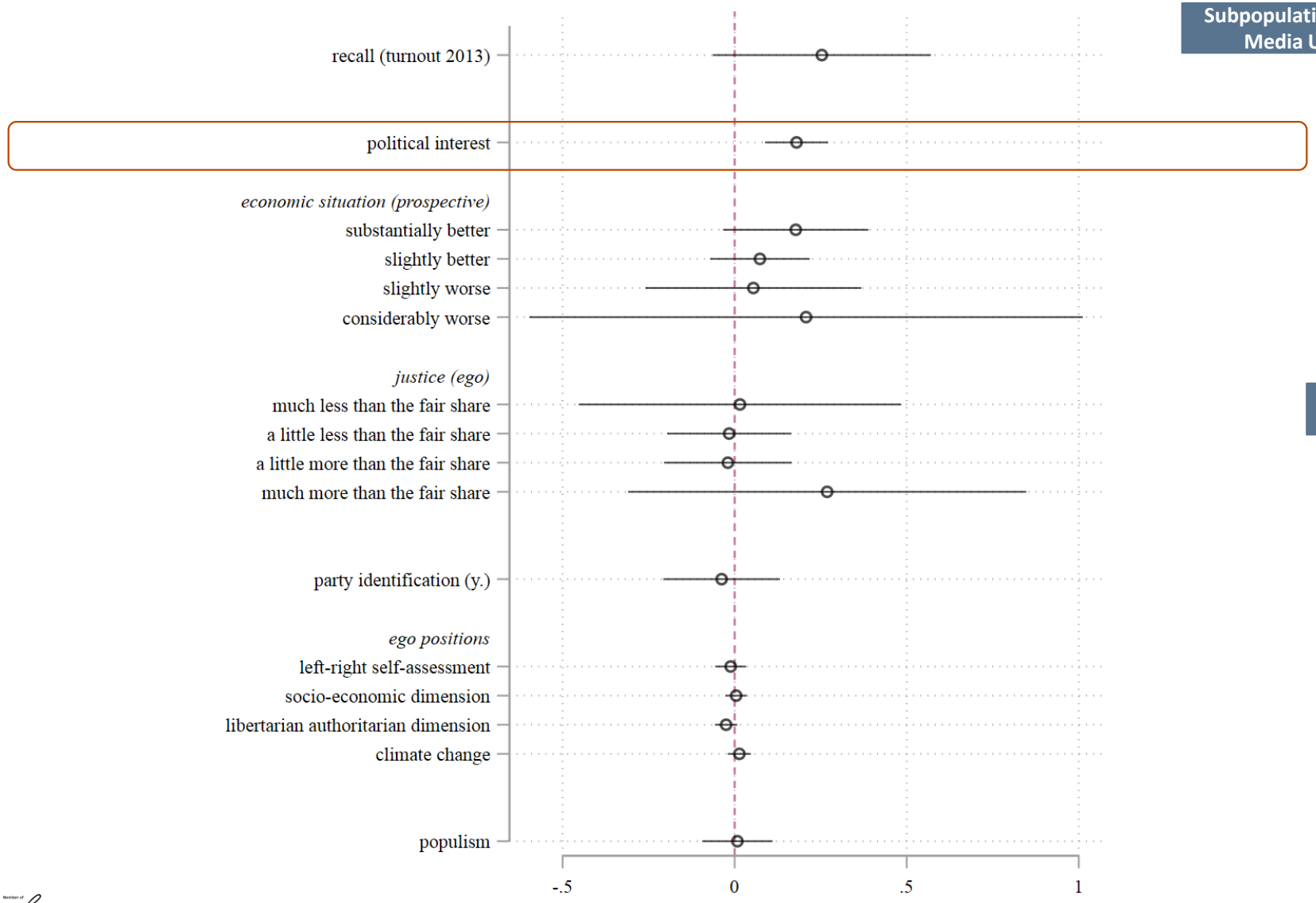


N= 2145
R^2 = 0.12

gesis Leibniz Institute for the Social Sciences

Differences in attitudes Model 3

Subpopulation Social Media Users



N= 2145
R^2 = 0.12

Summary Model 3

- internet usage subject to demographic selectivities
- risk of bias on political interest

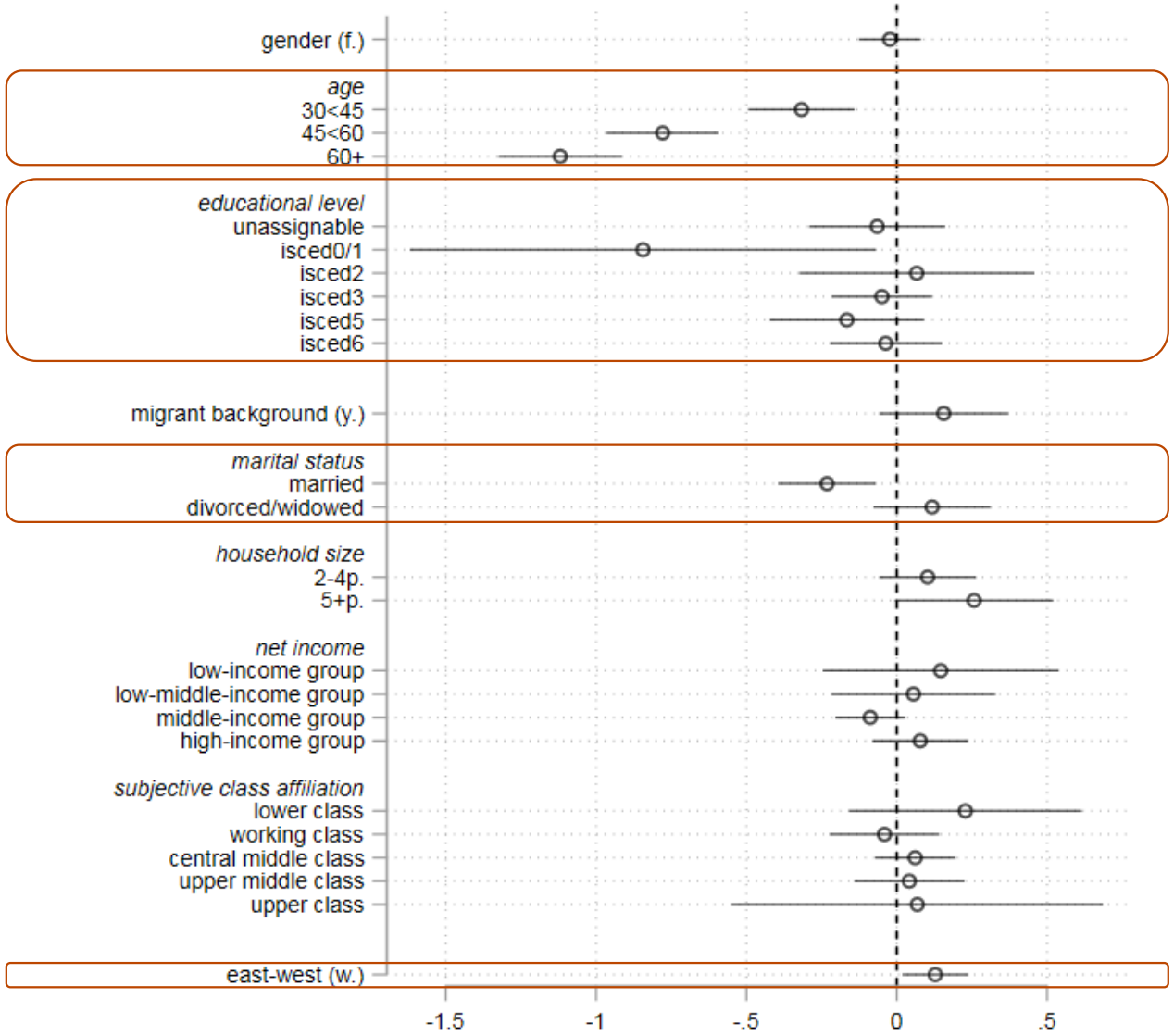
gesis Leibniz Institute for the Social Sciences

Research Design

Model	Subset analyzed	Model type	Dependent Variable	Independent Variables
1	all	multinomial logistic	internet and social media usage	gender
				age
2	social media users	logistic	social media usage by platform: Facebook, WhatsApp, YouTube, Twitter, Google+	educational level
				marital status
				household size
				net income
3	social media users	linear OLS	days of internet usage per week	region (east/west)
				migration
				economic situation (prospective)
4	social media users	linear OLS	number of social media platforms used	justice (ego)
				recall (turnout 2013)
				party identification
5	social media users	logistic	form of social media usage: commenting and posting	left-right self-assessment
				ego positions
				populism (Akkerman et al.)

gesis Leibniz Institute for the Social Sciences

Differences in sociodemographics Model 4



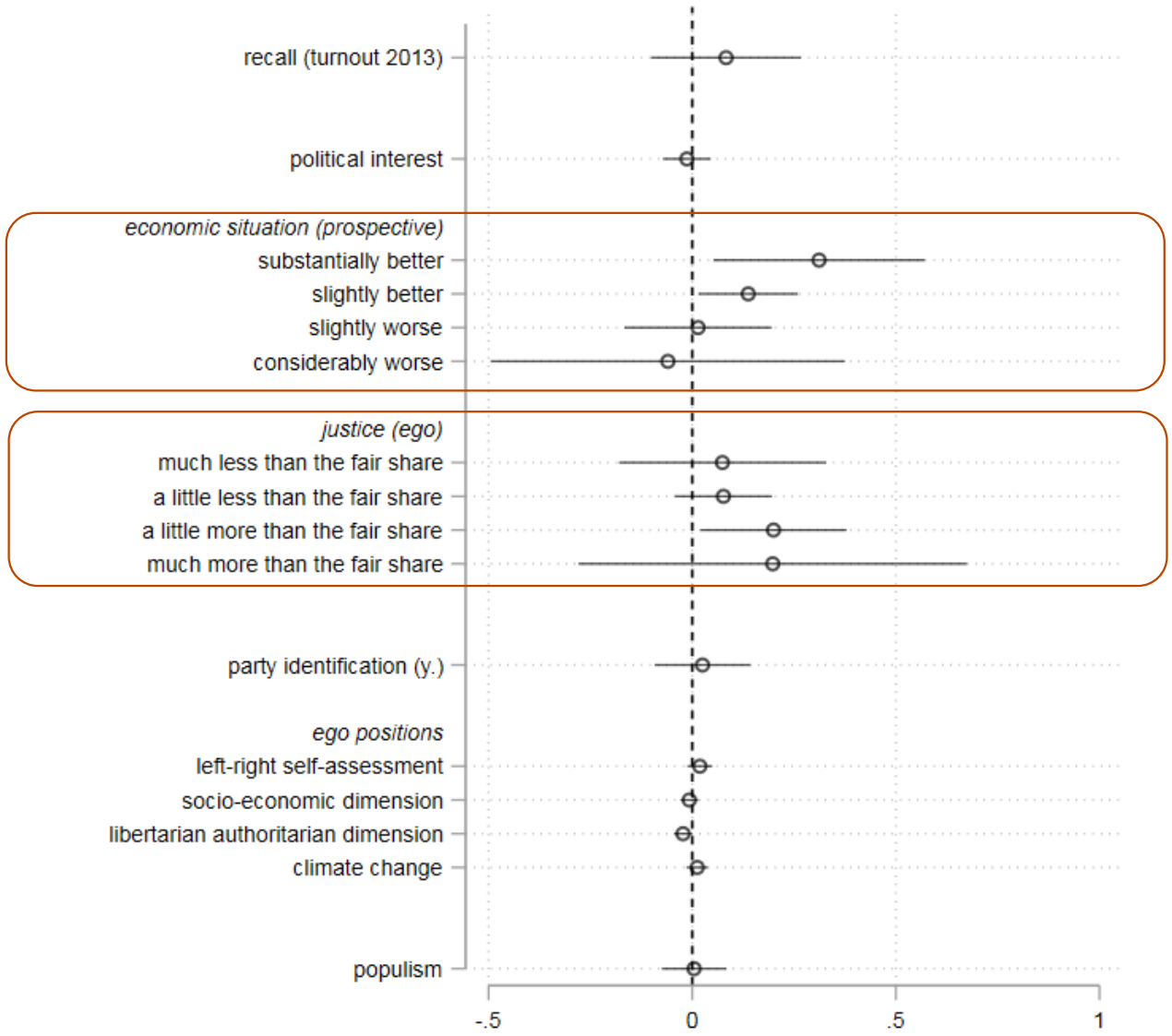
Subpopulation Social Media Users

N= 2145
R^2 = 0.18

gesis Leibniz Institute for the Social Sciences

Differences in attitudes Model 4

Subpopulation Social Media Users



N= 2145
R^2 = 0.18

Summary Model 4

- demographic selectivities
- possible bias on self perception items

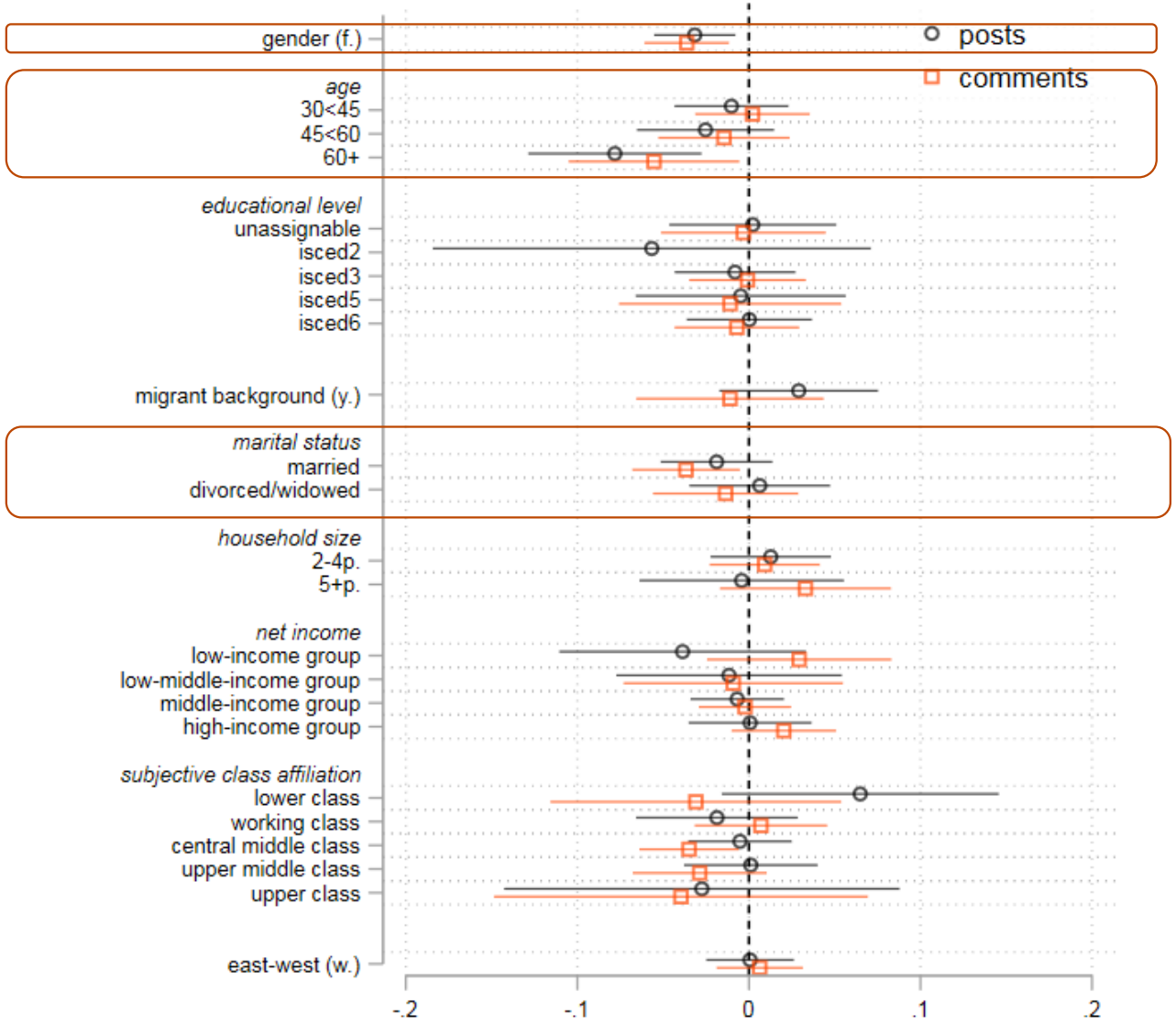
gesis Leibniz Institute for the Social Sciences

Research Design

Model	Subset analyzed	Model type	Dependent Variable	Independent Variables
1	all	multinomial logistic	internet and social media usage	gender
				age
2	social media users	logistic	social media usage by platform: Facebook, WhatsApp, YouTube, Twitter, Google+	educational level
				marital status
				household size
				net income
3	social media users	linear OLS	days of internet usage per week	region (east/west)
				migration
				economic situation (prospective)
4	social media users	linear OLS	number of social media platforms used	justice (ego)
				recall (turnout 2013)
				party identification
5	social media users	logistic	form of social media usage: commenting and posting	left-right self-assessment
				ego positions
				populism (Akkerman et al.)

gesis Leibniz Institute for the Social Sciences

Differences in sociodemographics Model 5



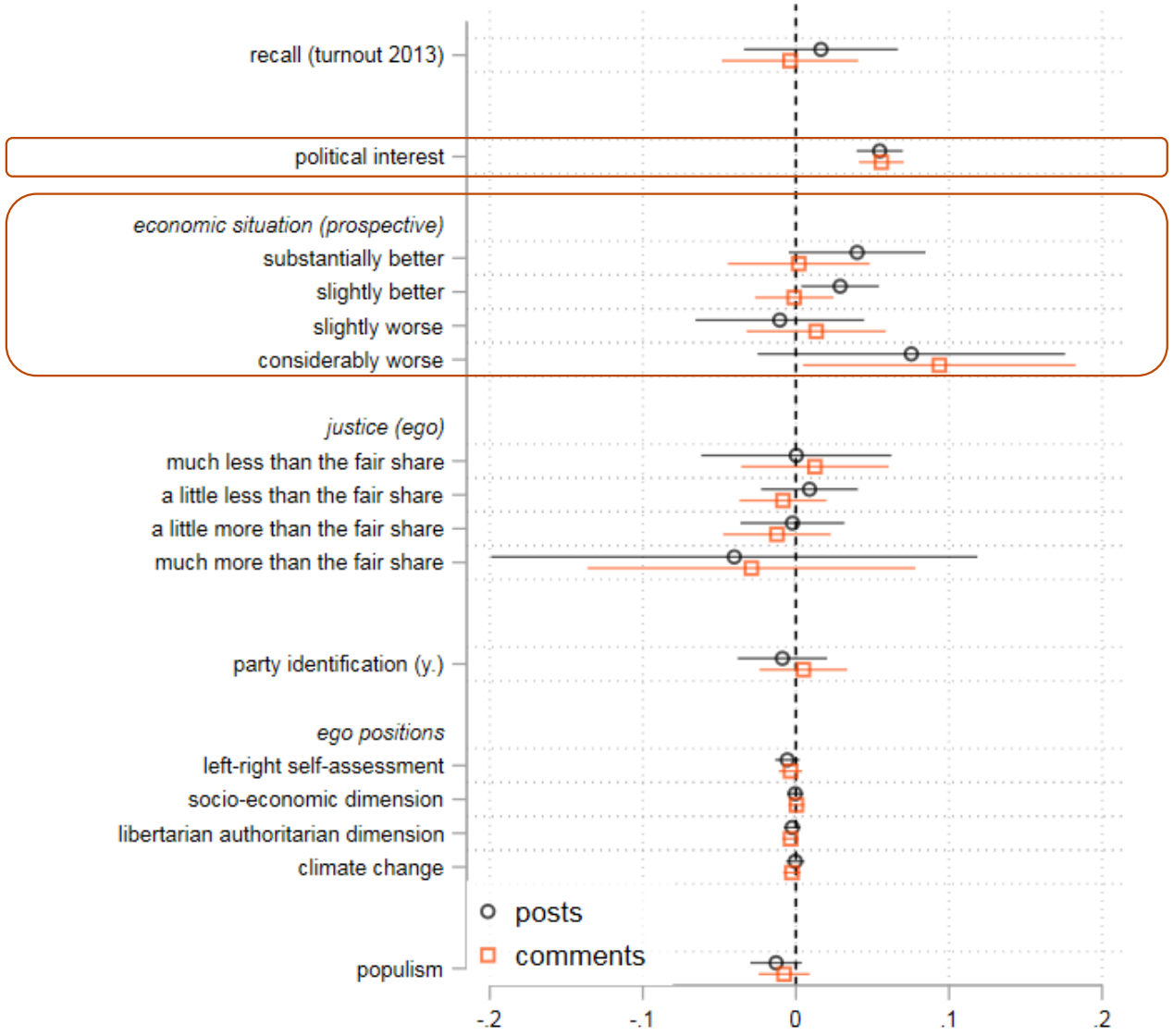
Subpopulation Social Media Users

DV: postings (y/n)
N= 2136
Pseudo R^2 = 0.14

DV: comments (y/n)
N=2088
Pseudo R^2=0.16

gesis Leibniz Institute for the Social Sciences

Differences in attitudes Model 5



Subpopulation Social Media Users

DV: postings (y/n)
N= 2136
Pseudo R^2 = 0.14

DV: comments (y/n)
N=2088
Pseudo R^2=0.16

Summary Model 5

- demographic selectivities
- possible bias in economic self-perception and political interest

Overall Results

- demographic selectivities in all considered selection steps
- this concerns especially age, gender, marital status and educational level
- possible bias in attitudinal variables mainly due to non-coverage of offliners, selection of social media platform and frequency and kind of internet usage
- possible bias in self perceived economic situation due to social media usage (general , specific and number of social media platforms)
- all in all, especially younger persons with a good actual and/or perceived economic situation and politically interested persons are likely to be overrepresented
- nature and direction of bias varies between different selection steps

Conclusions

- significant risk of bias due to non-coverage of offliners on demographic and attitudinal variables
- further demographic selectivities due to exclusion of onliners who do not use (certain) social media platforms
- additionally, possible bias due to different selectivities in certain attitudinal variables and measures of self-perception controlling for demographic differences

Limits and Discussion

- the data used only includes adults elgible to vote in Germany in 2017 -> inferences can only be made to that population
- limits inherent in available variables
- low incidences for some characteristics
- next steps
  - look at more variables
  - try to identify different groups of online and social media usage
  - fit common models in election research and identify differences in estimates due to online and social media usages

Literature

Barberá, P. and G. Rivero (2015). "Understanding the political representativeness of Twitter users." Social Science Computer Review 33(6): 712-729.

Blank, G. (2017). "The digital divide among Twitter users and its implications for social research." Social Science Computer Review 35(6): 679-697.

Blank, G. and C. Lutz (2017). "Representativeness of Social Media in Great Britain: Investigating Facebook, LinkedIn, Twitter, Pinterest, Google+, and Instagram." American Behavioral Scientist 61(7): 741-756.

Gayo-Avello, D. (2013). "A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data." Social Science Computer Review 31(6): 649-679.

Greenwood, S. P., Andrew; Duggan, Maeve (2016). Social Media Update 2016, Pew Research Center.

Jungherr, A. (2016). "Twitter use in election campaigns: A systematic literature review." Journal of Information Technology & Politics 13(1): 72-91.

Literature

Oliveira, D. J. S., et al. (2017). "Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls." Journal of Information Technology & Politics 14(1): 34-45.

Phillips, L., et al. (2017). "Using social media to predict the future: a systematic literature review." arXiv preprint arXiv:1706.06134.

Rainie, L., et al. (2012). "Social media and political engagement." Pew Internet & American Life Project 19: 2-13.

Ruths, D. and J. Pfeffer (2014). "Social media for large studies of behavior." Science 346(6213): 1063-1064.

Ryan, T. and S. Xenos (2011). "Who uses Facebook? An investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage." Computers in Human Behavior 27(5): 1658-1664.

Tumasjan, A., et al. (2010). "Predicting elections with twitter: What 140 characters reveal about political sentiment." ICWSM 10(1): 178-185.

Data

Roßteutscher, Sigrid; Schmitt-Beck, Rüdiger; Schoen, Harald; Weßels, Bernhard; Wolf, Christof; Bieber, Ina; Stövsand, Lars-Christopher; Dietz, Melanie; Scherer, Philipp; Wagner, Aiko; Melcher, Reinhold; Giebler, Heiko (2018): Pre- and Post-election Cross Section (Cumulation) (GLES 2017). GESIS Data Archive, Cologne. ZA6802 Data file Version 3.0.0, doi:10.4232/1.13139