EVIDENCE BASED POLICY OR BEAUTY CONTEST?

A LLM-BASED META-ANALYSIS OF EU COHESION POLICY EVALUATIONS

Zareh Asatryan   Carlo Birkholz   Friedrich Heinemann

(ZEW - Leibniz Centre for European Economic Research)

Project Sponsored by the BMF

2024-10-02
Fünfte Jahreskonferenz des
Netzwerks Bessere Rechtsetzung und Bürokratieabbau

RESEARCH QUESTIONS

- Cohesion evaluations as a prime example of performance budgeting:
  - Better policy: Evidence-based, learning externalities, transparency.
  - Tradeoff: More bureaucracy and (compliance) costs.
- Ambitious and noble goals, but no free-lunch.
- Our (meta-)analysis:
  1. What do the evaluations by MS find?
  2. How do these findings square (or not) with the existing evidence?
  3. Is the market competitive and impartial? Do these correlate with evaluations?
  4. Do evaluations impact decisions?
  5. What are the main bottlenecks of the evaluation system?

DATA

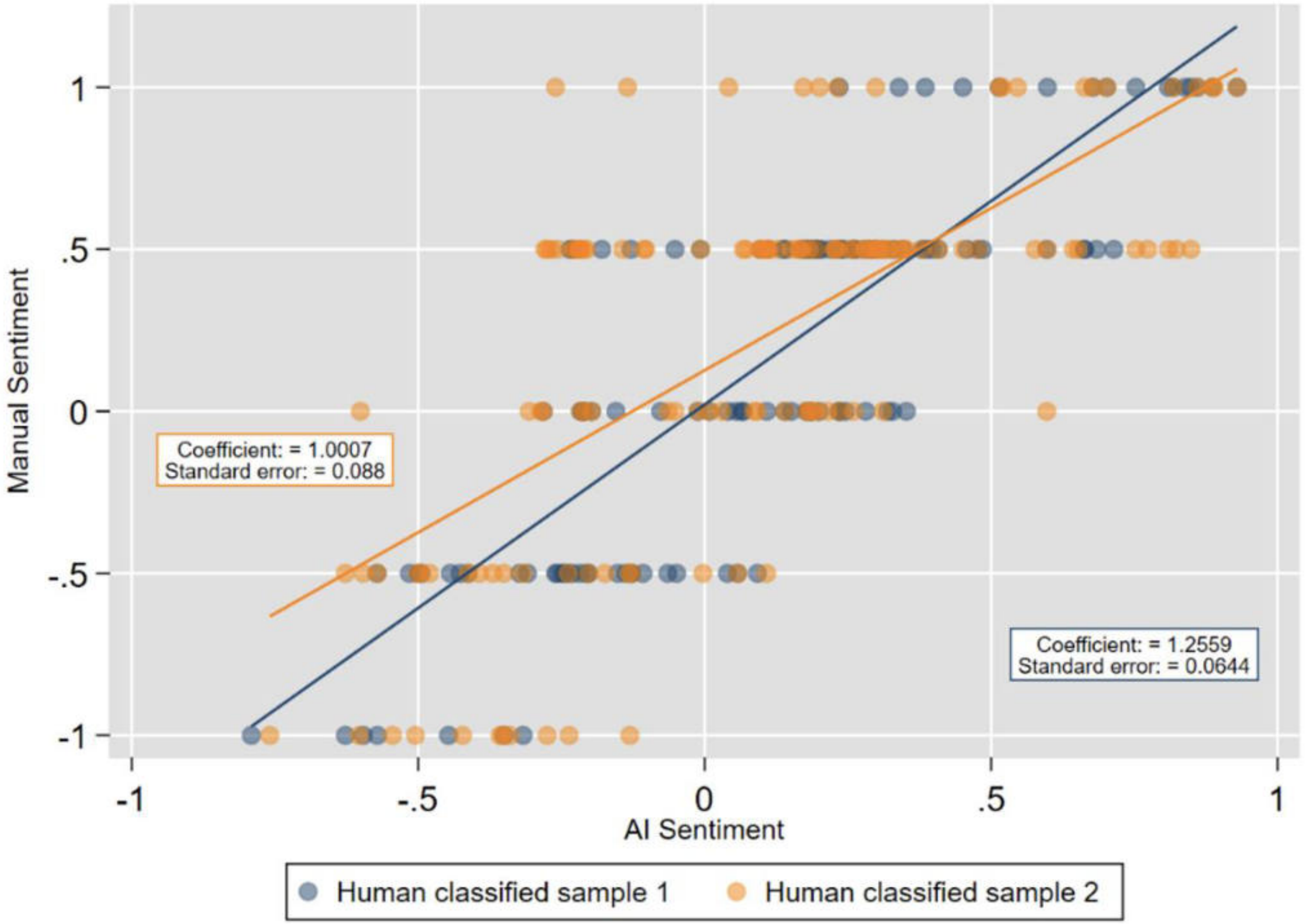
- Sample: Cohesion Policy programme evaluations by MS since 2007:
  - About 2,300 evaluations. Based on „library“ of evaluations provided by COM.
- Complemented with further data on:
  - Cohesion programmes, their budgets and other details.
- And, on the authors of the evaluations:
  - About 2,300 authors. On average 2.73 evaluation per author.
  - Co-authorship networks nationally and internationally.
  - Plus, own recent survey of about 200 individual authors.

METHODS

- Meta-methods:
  - Estimate the sentiment of each abstract using GPT’s large language model.
  - Run independently for each abstract. Bootstrap 50 x.
  - Test AI v.s. human assessments in two sub-samples (next slide).
  - Also, AI v.s. library based assessment on the whole sample (appendix).
  - Then also, abstract v.s. full report (appendix).
  - Main assumption: Measurement error not heterogenous.

ZEW

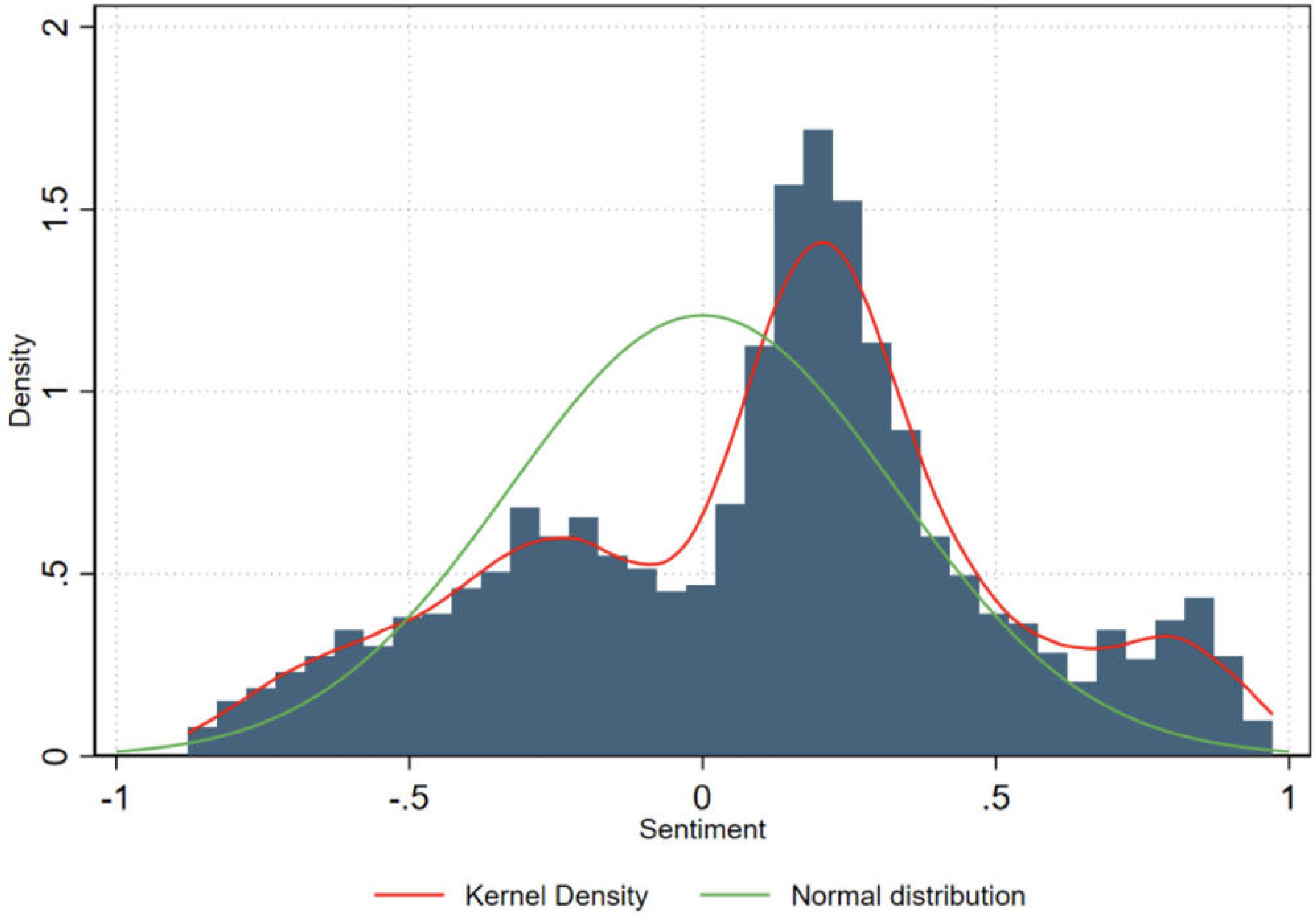
METHODS: AI V.S. HUMAN



WHAT DO THE EVALUATIONS FIND?

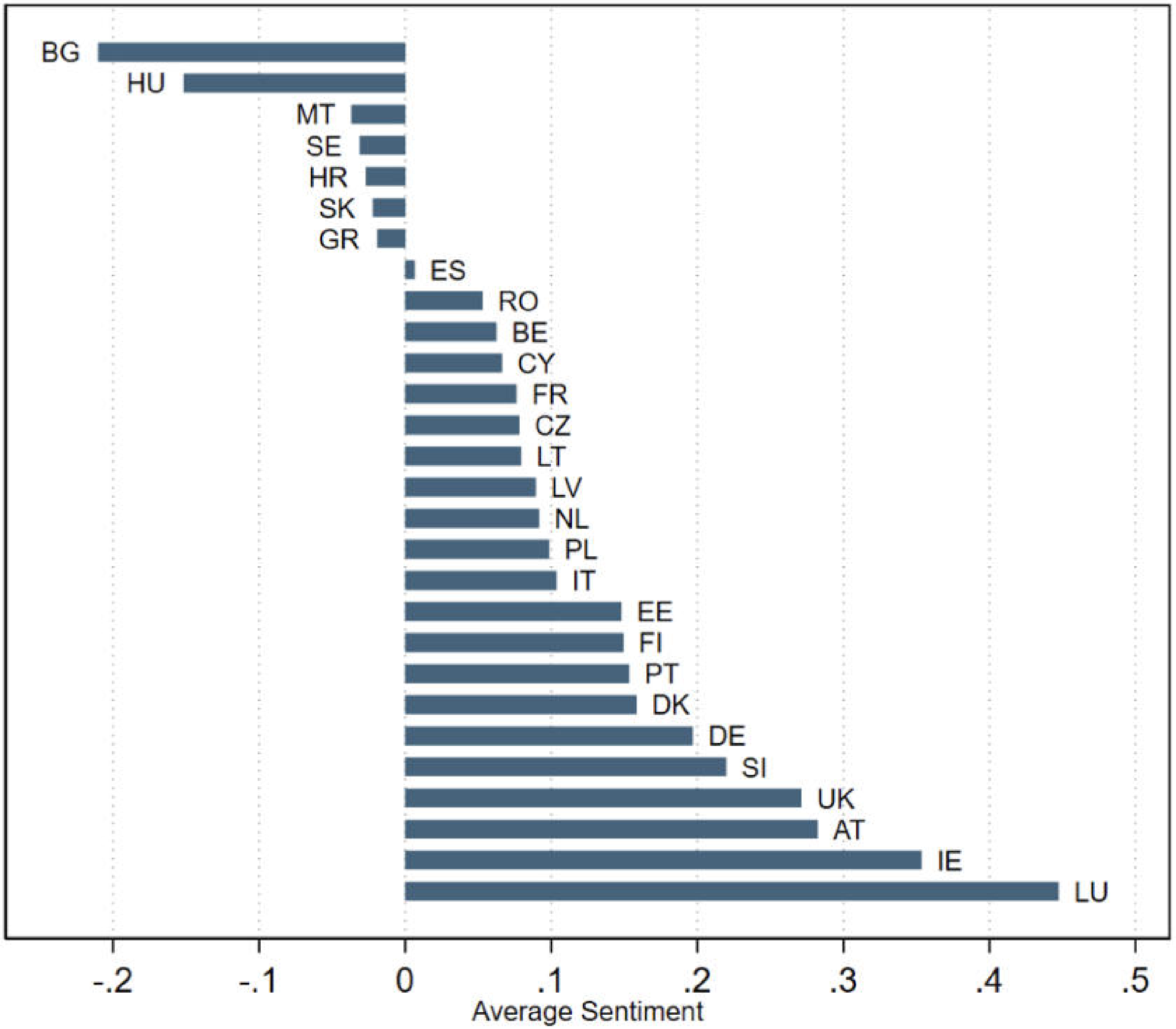
EVALUATIONS HAVE A POSITIVE TONE: P(70)>0

DISTRIBUTION OF SENTIMENT: INDEX -1 (VERY NEGATIVE) TO +1 (VERY POSITIVE)



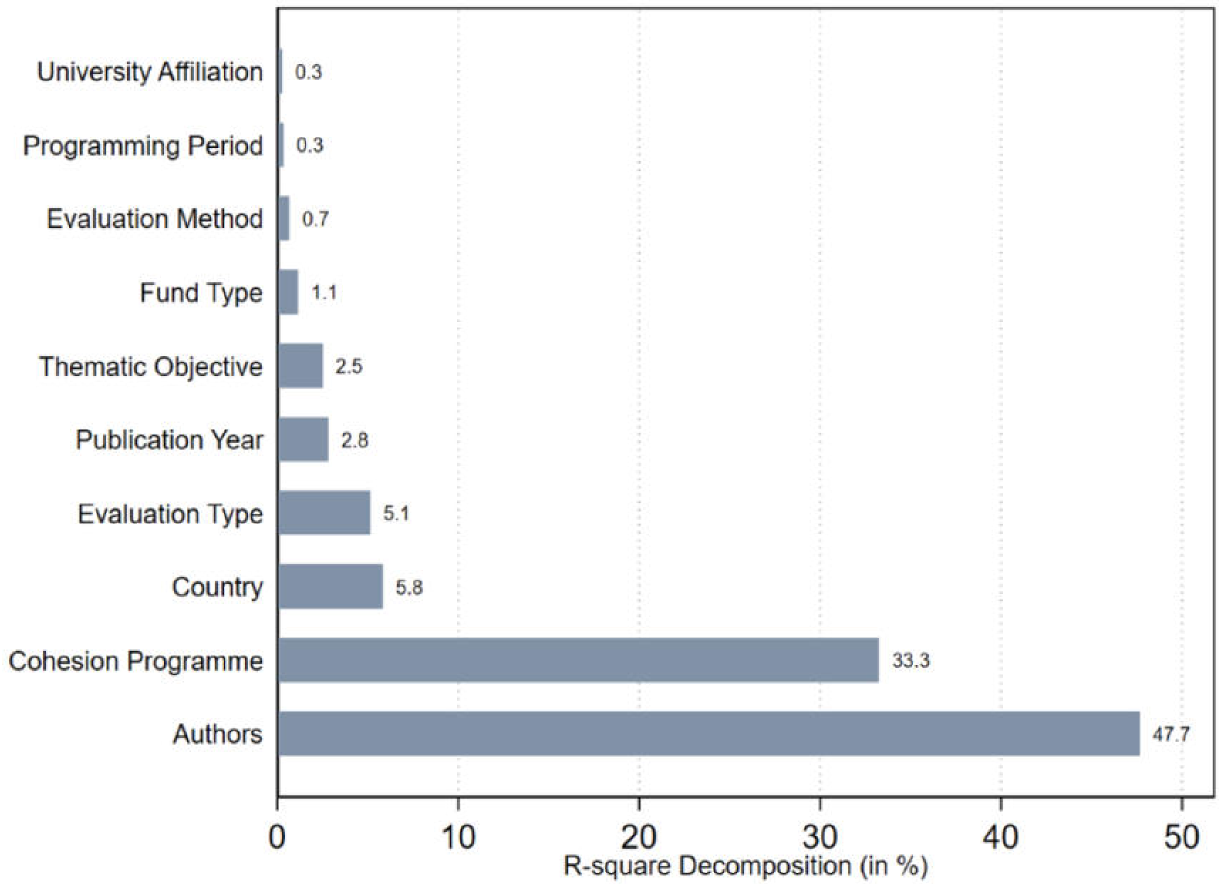
WHAT DO THE EVALUATIONS IN DIFFERENT MEMBER STATES FIND?

UNCONDITIONAL SENTIMENT BY MEMBER STATE



WHAT EXPLAINS THE VARIATION IN FINDINGS?

PROGRAMMES ARE VERY IMPORTANT… COUNTRIES AND AUTHORS STILL IMPORTANT…

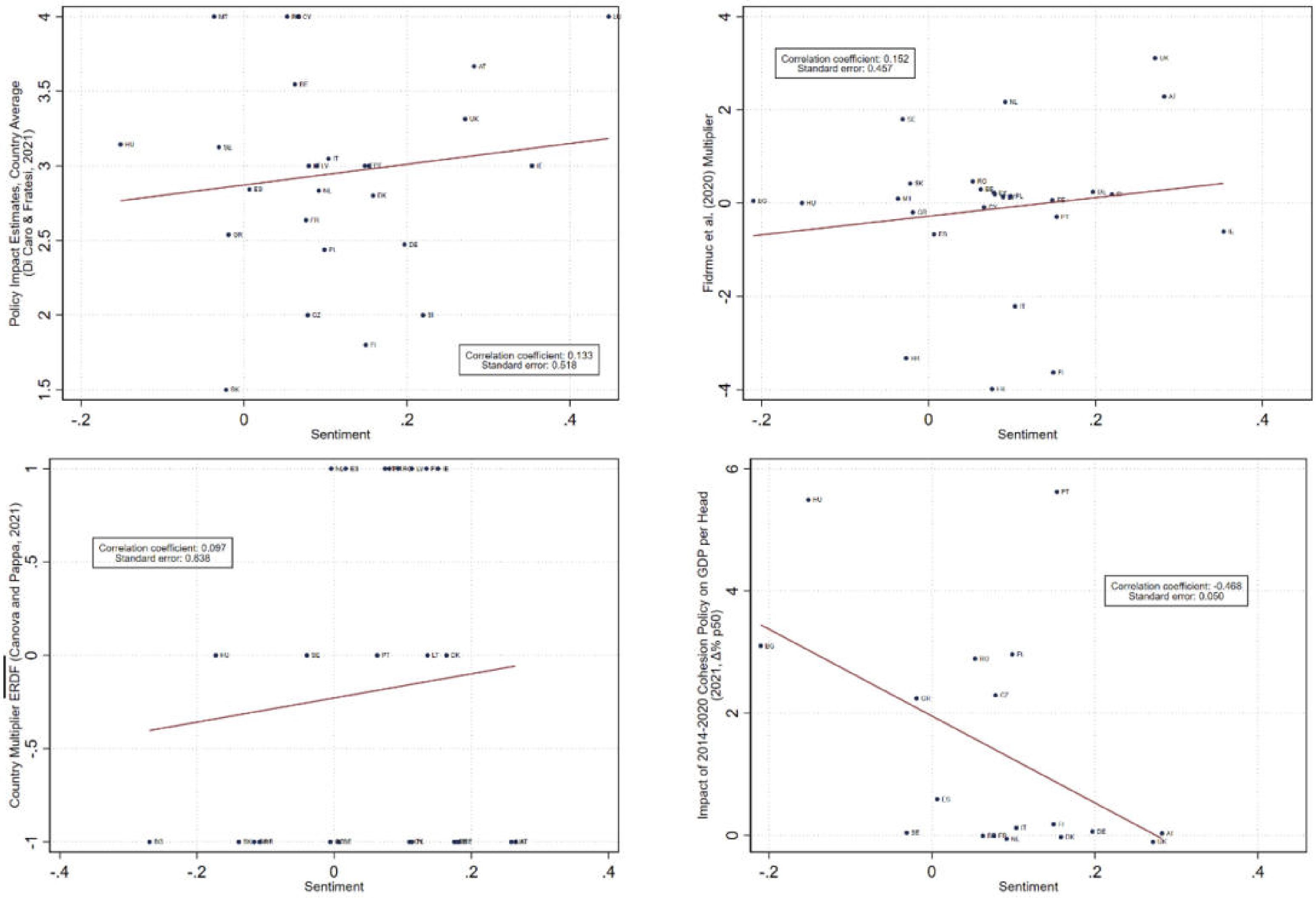


Notes: Bars present Shorrocks-Shapley decomposition of R-squared in a regression where the shown 10 variables (in their fixed effects specification) are jointly linearly regressed on the sentiment score.

HOW DO THESE FINDINGS SQUARE WITH THE EXISTING EVIDENCE?

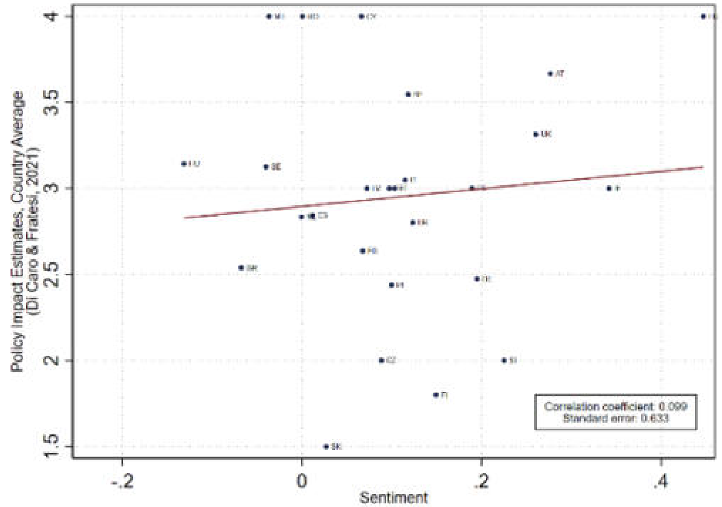
SENTIMENT V.S. MS-SPECIFIC GROWTH EFFECTS

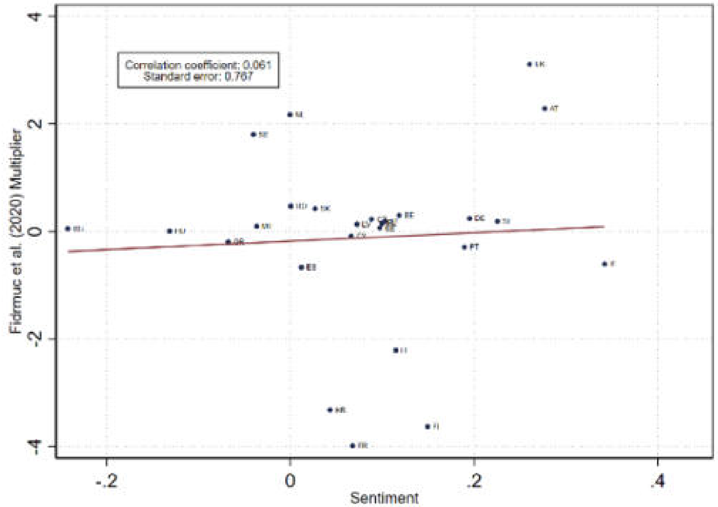
TOP: DI CARO & FRATESI (LEFT) & FIDRMUC ETAL (RIGHT)
BOTTOM: CANOVA AND PAPPA'S ERDF (LEFT) & COM' RHOMOLO-P50 (RIGHT)

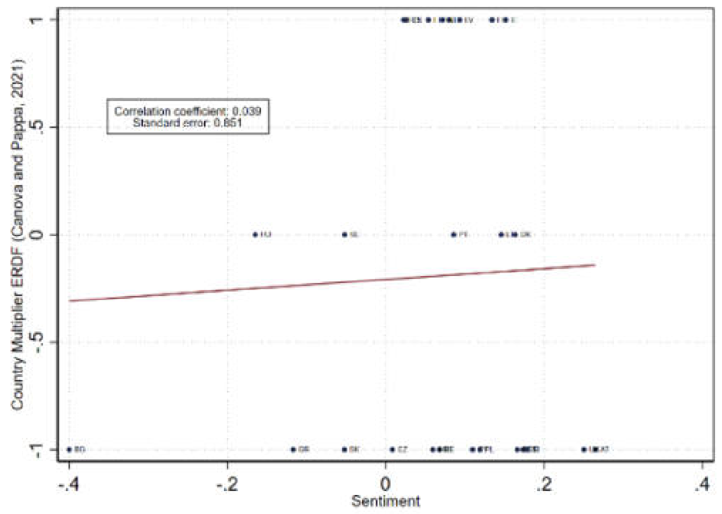


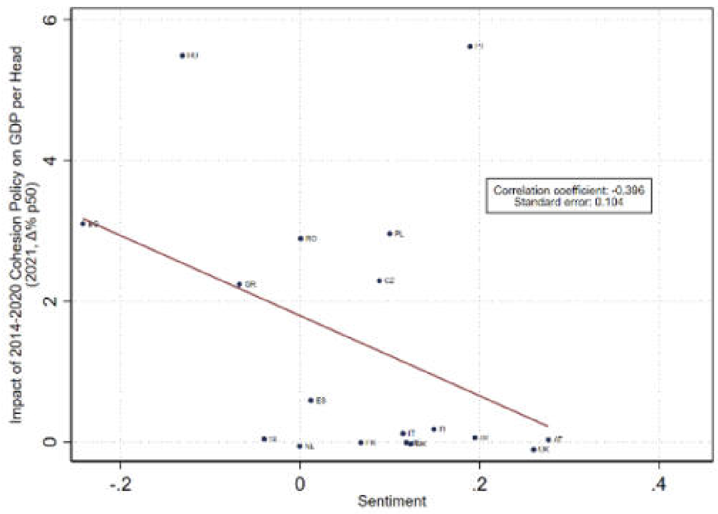
SENTIMENT V.S. MS-SPECIFIC GROWTH EFFECTS

ONLY GROWTH-FRIENDLY THEMATIC OBJECTIVES



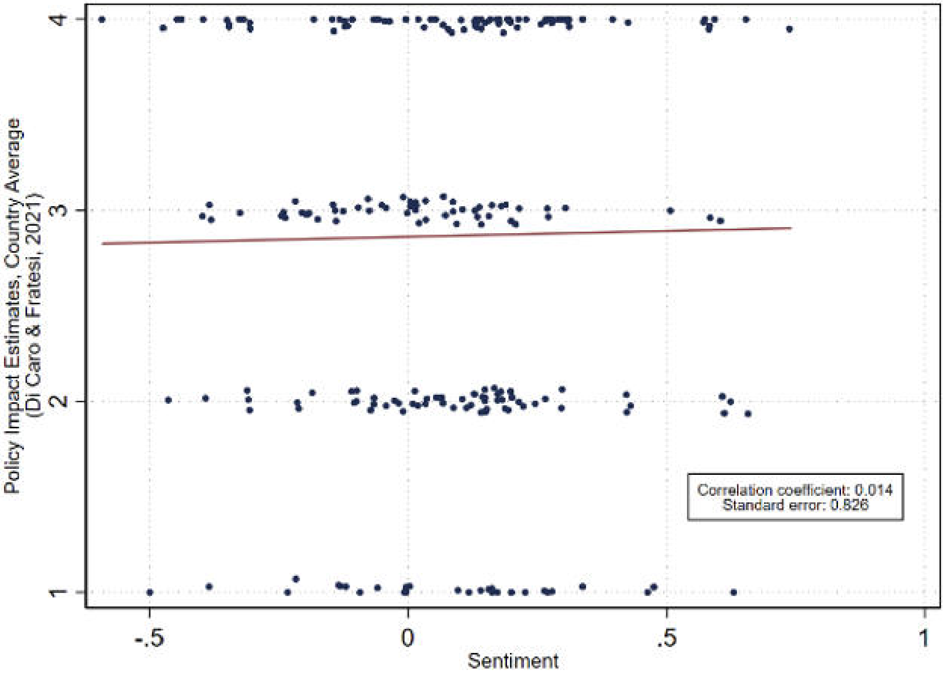


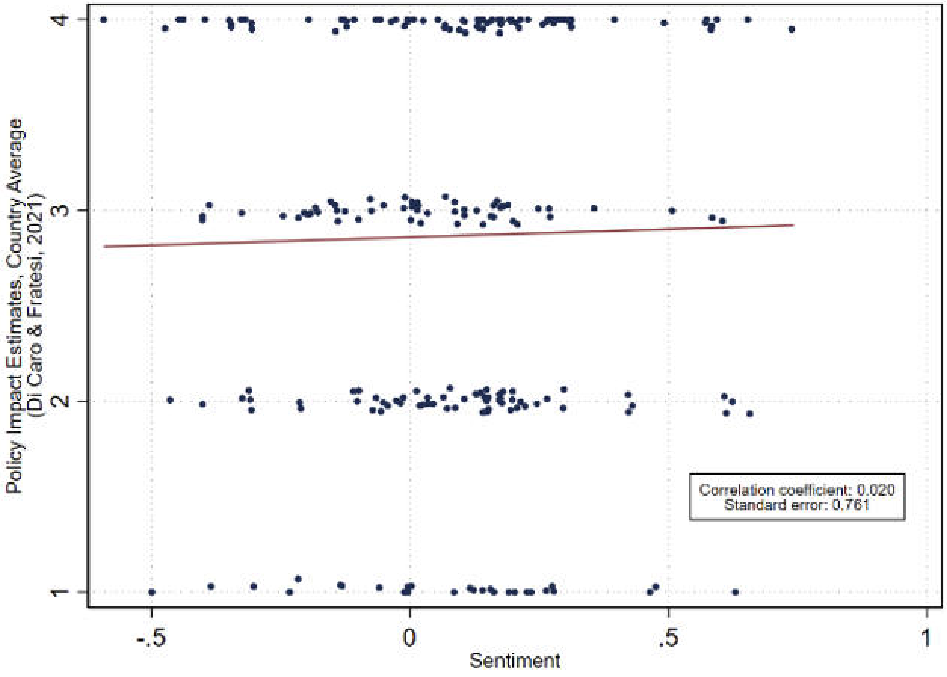




SENTIMENT V.S. NUTS2-SPECIFIC GROWTH EFFECTS

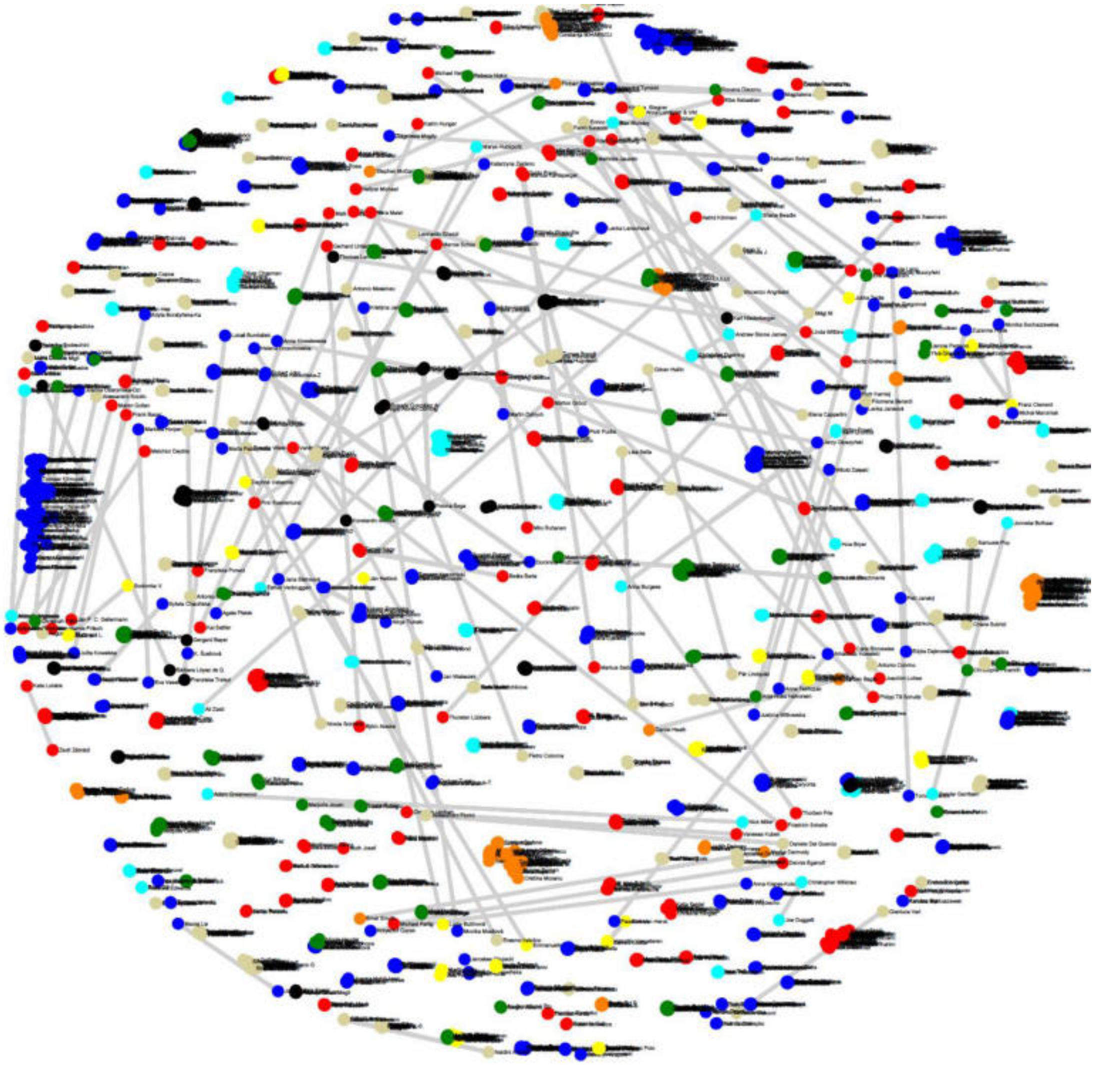
BY DI CARO & FRATESI: NUTS2 LEVEL ESTIMATES (*N*=260)

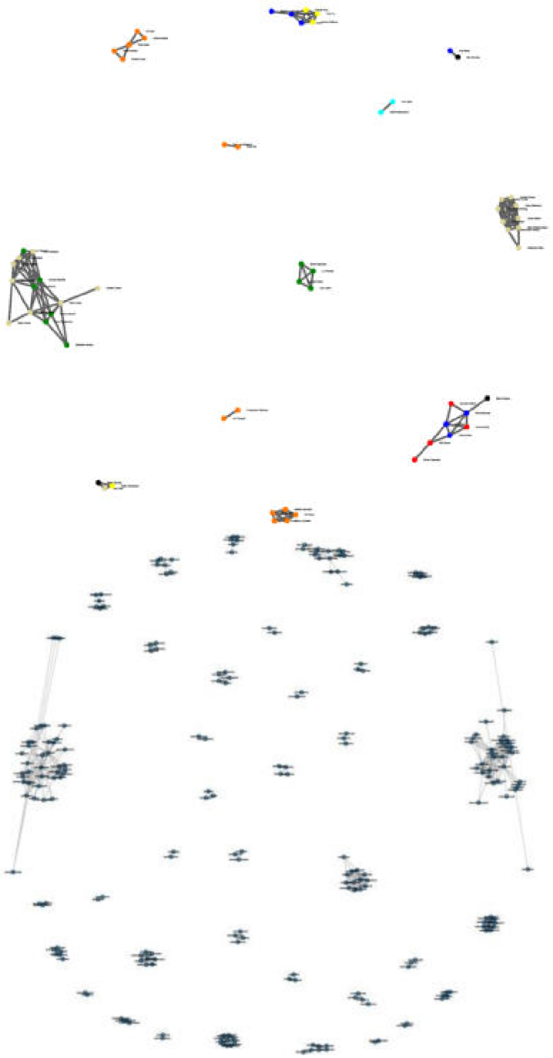




HOW COMPETITIVE IS THE EVALUATION MARKET?

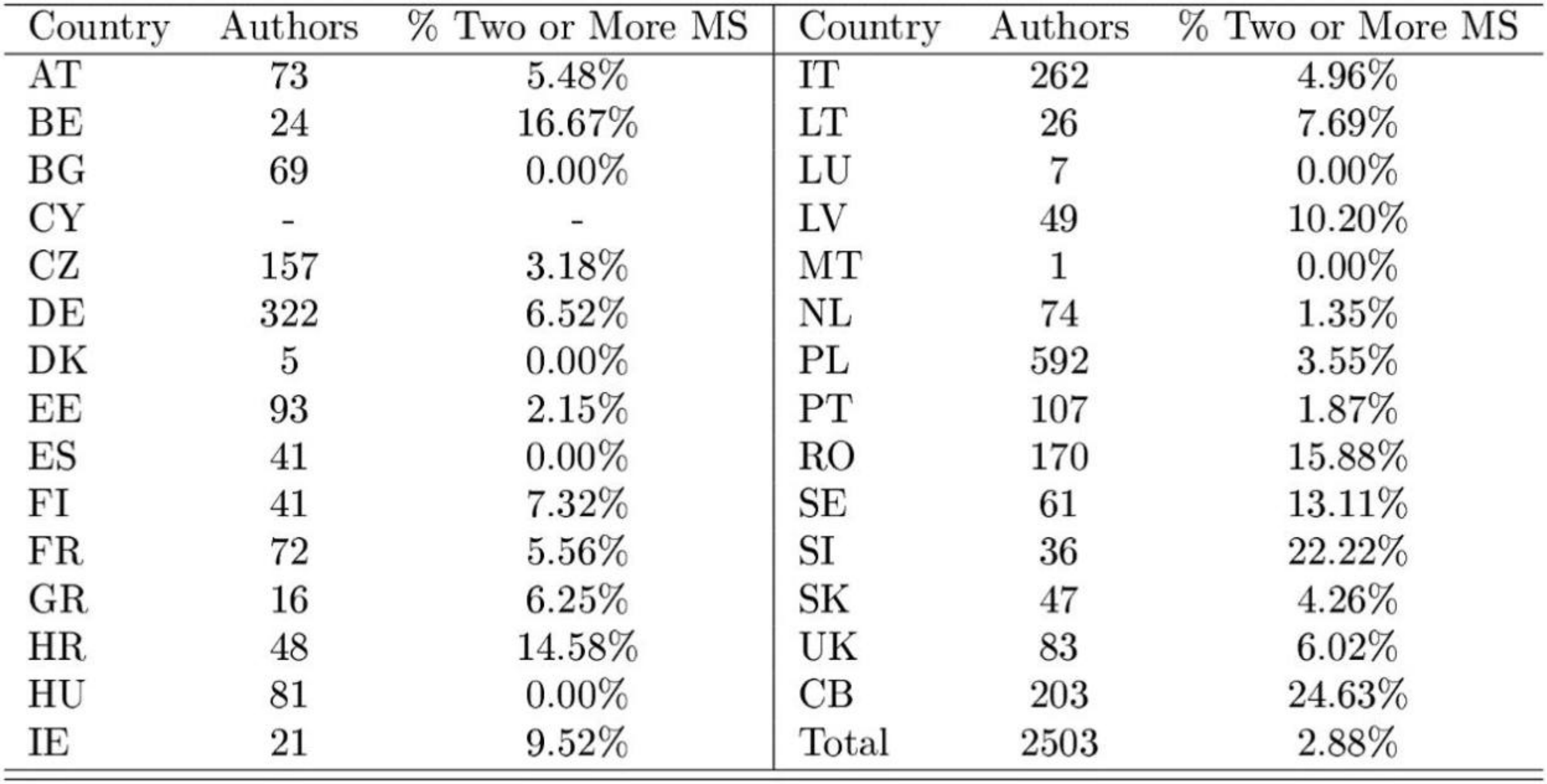
DATA: AUTHOR CLUSTERS IN EU AND UK / ITALY





EU’S “SINGLE MARKET” FOR EVALUATIONS…

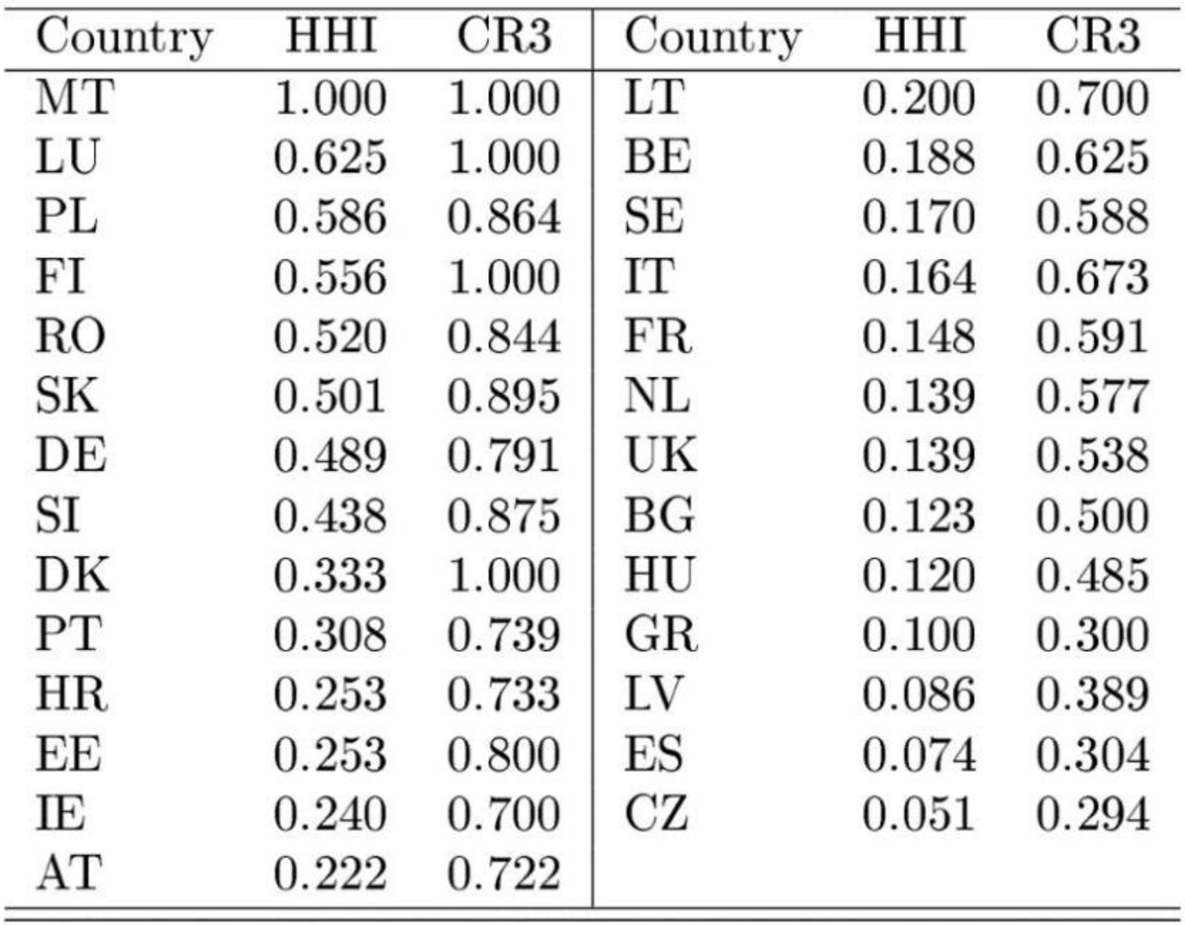
OUT OF 2,233 AUTHORS ONLY 2.88% HAVE WORKED IN TWO OR MORE MS!



Country	Authors	% Two or More MS
AT	73	5.48%
BE	24	16.67%
BG	69	0.00%
CY	-	-
CZ	157	3.18%
DE	322	6.52%
DK	5	0.00%
EE	93	2.15%
ES	41	0.00%
FI	41	7.32%
FR	72	5.56%
GR	16	6.25%
HR	48	14.58%
HU	81	0.00%
IE	21	9.52%
IT	262	4.96%
LT	26	7.69%
LU	7	0.00%
LV	49	10.20%
MT	1	0.00%
NL	74	1.35%
PL	592	3.55%
PT	107	1.87%
RO	170	15.88%
SE	61	13.11%
SI	36	22.22%
SK	47	4.26%
UK	83	6.02%
CB	203	24.63%
Total	2503	2.88%

MARKET CONCENTRATION IN MS

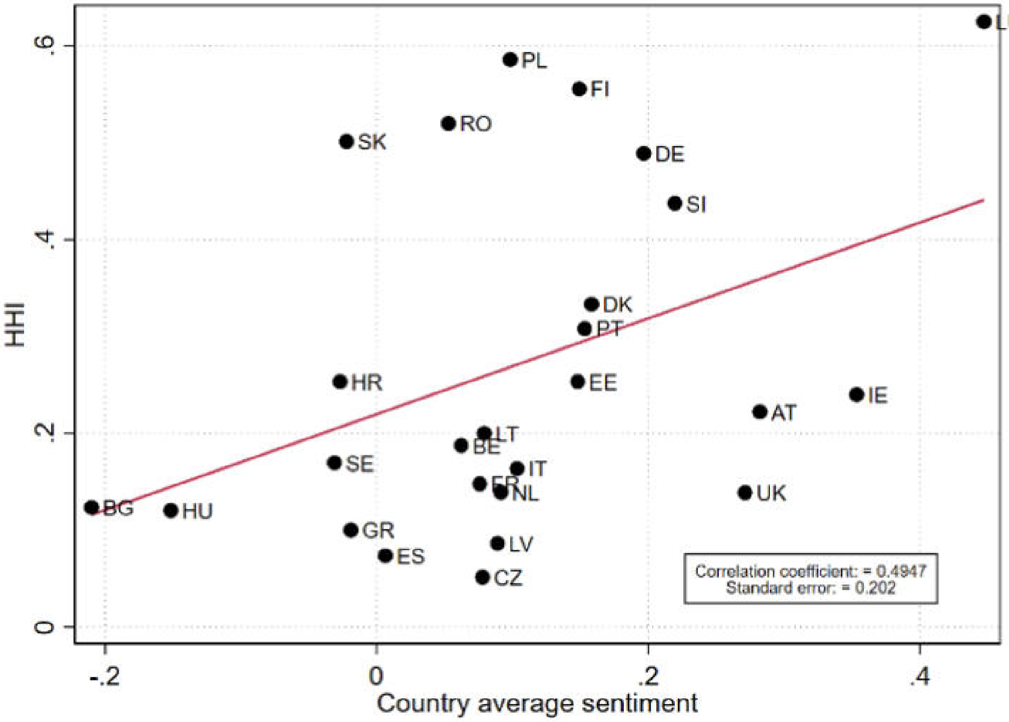
ACROSS MS, THE TOP-3 CLUSTERS WRITE >70% OF EVALUATIONS ON AVERAGE. THE TOP FIRM/CLUSTER IN GERMANY WRITES >60% OF EVALUATIONS.

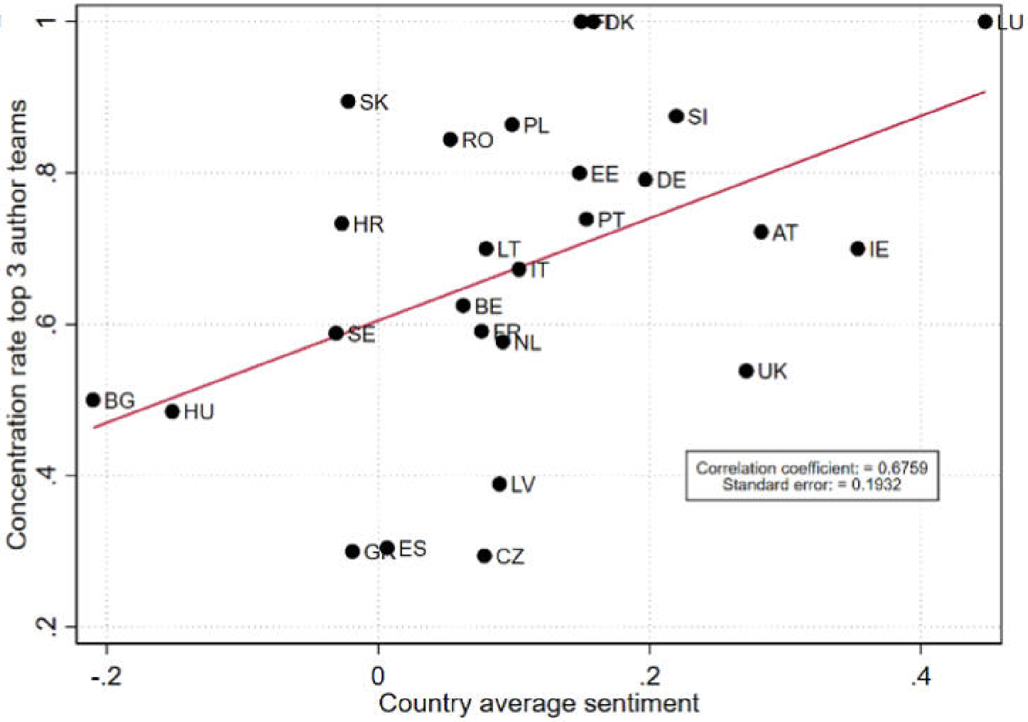


Country	HHI	CR3	Country	HHI	CR3
MT	1.000	1.000	LT	0.200	0.700
LU	0.625	1.000	BE	0.188	0.625
PL	0.586	0.864	SE	0.170	0.588
FI	0.556	1.000	IT	0.164	0.673
RO	0.520	0.844	FR	0.148	0.591
SK	0.501	0.895	NL	0.139	0.577
DE	0.489	0.791	UK	0.139	0.538
SI	0.438	0.875	BG	0.123	0.500
DK	0.333	1.000	HU	0.120	0.485
PT	0.308	0.739	GR	0.100	0.300
HR	0.253	0.733	LV	0.086	0.389
EE	0.253	0.800	ES	0.074	0.304
IE	0.240	0.700	CZ	0.051	0.294
AT	0.222	0.722			

Notes: HHI is the Herfindahl-Hirschman Index normalized by the number of firms; CR3 is the market share of top 3 clusters/firms.

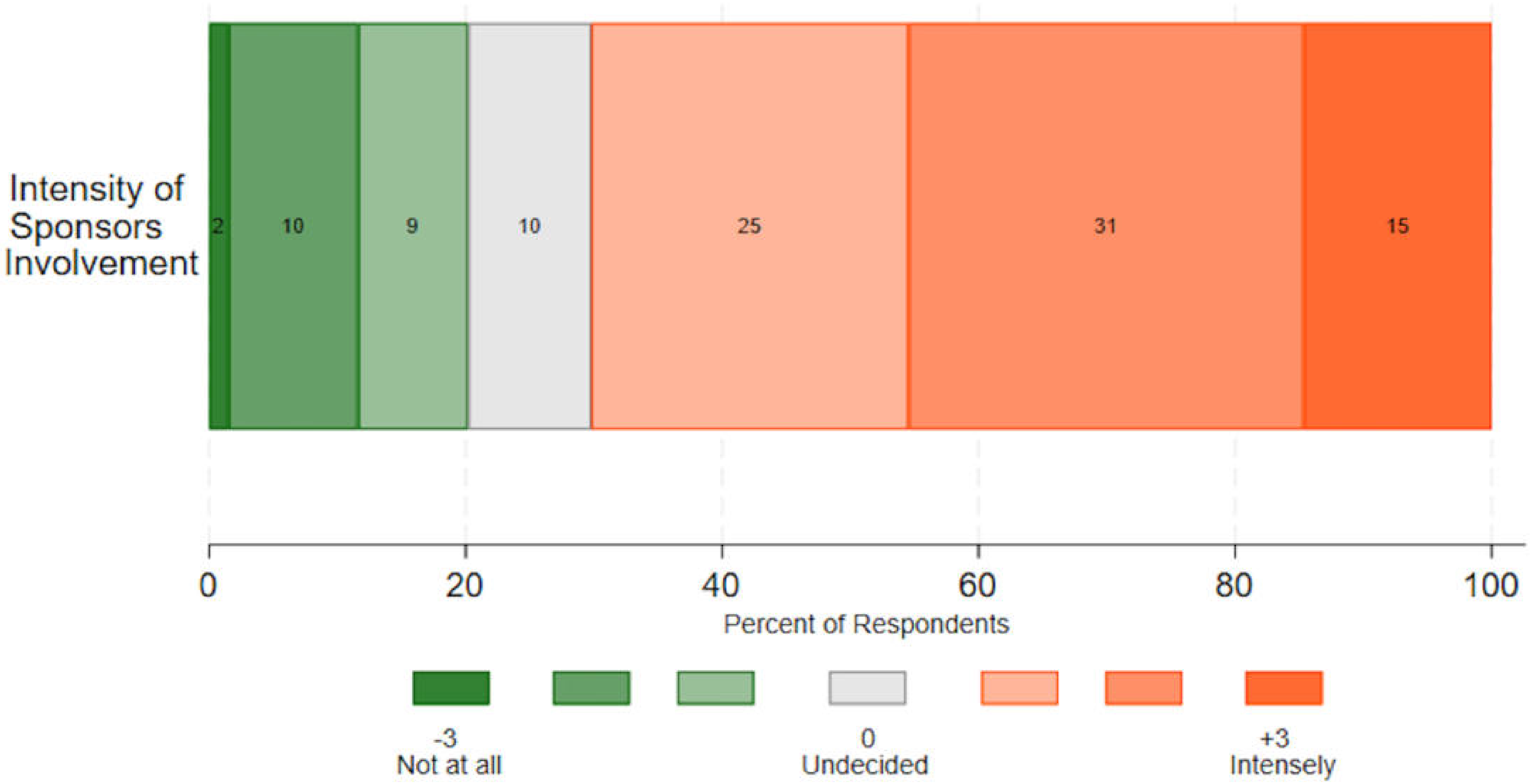
CONCENTRATION --> MORE OPTIMISTIC RESULTS





WHAT IS THE ROLE OF IMPARTIALITY?

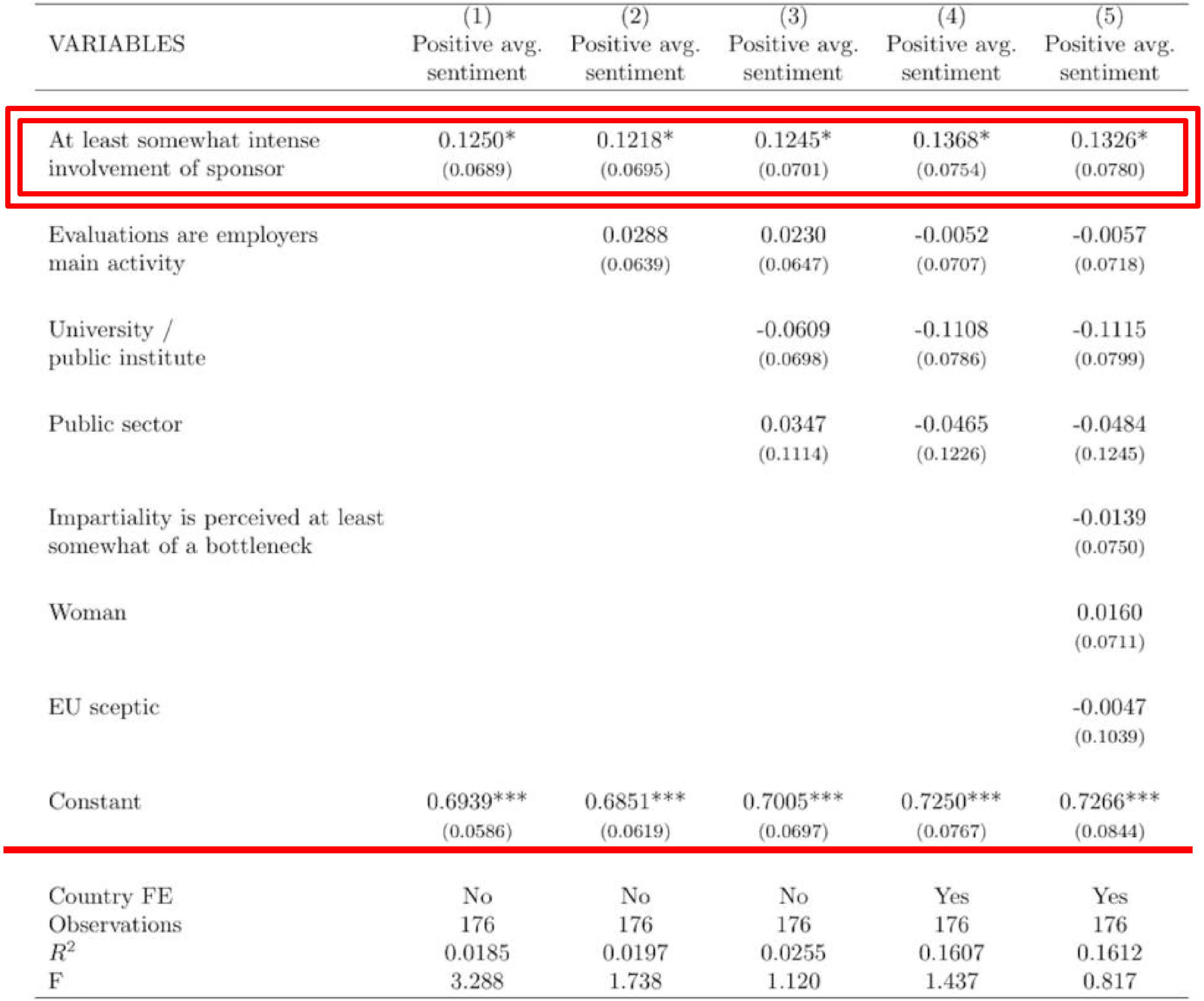
SURVEY OF AUTHORS: 80% OF AUTHORS CLAIM CLIENTS/AUTHORITIES INTERVENE IN THEIR WORK.



Survey: About 200 full responses (20% response rate). Authors: 43% female, 47% with PhD. Employer: 10% public sector, 27% universities and institutes, 63% private sector.

Question: How intensely are the sponsors of your EU programme evaluations typically involved in discussing your evaluation methods, results and policy conclusions?

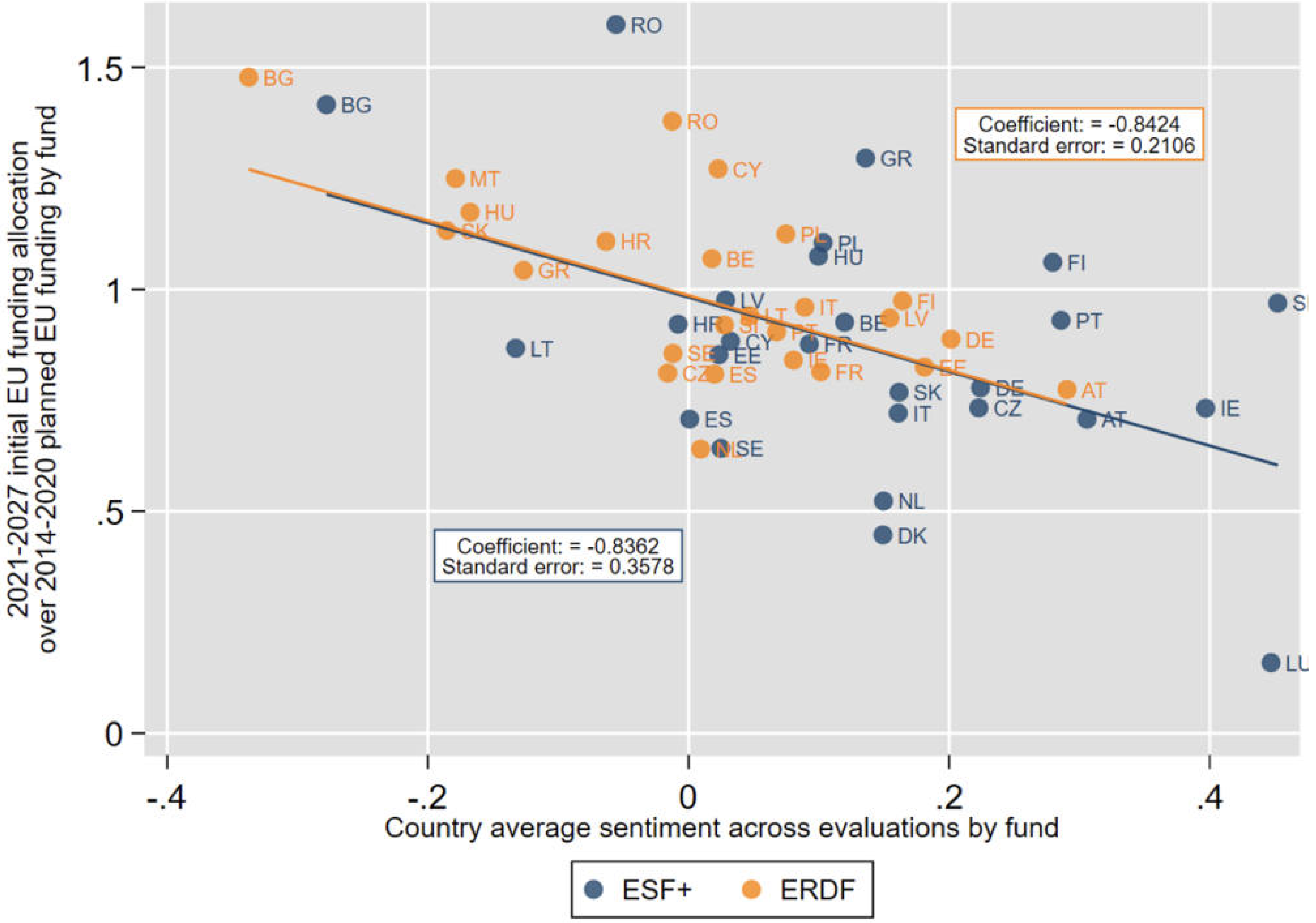
STRONGER INVOLVEMENT BY CLIENTS LEADS TO MORE OPTIMISTIC EVALUATION SENTIMENT: SIZEABLE AND ROBUST



VARIABLES	(1) Positive avg. sentiment	(2) Positive avg. sentiment	(3) Positive avg. sentiment	(4) Positive avg. sentiment	(5) Positive avg. sentiment
At least somewhat intense involvement of sponsor	0.1250* (0.0689)	0.1218* (0.0695)	0.1245* (0.0701)	0.1368* (0.0754)	0.1326* (0.0780)
Evaluations are employers main activity		0.0288 (0.0639)	0.0230 (0.0647)	-0.0052 (0.0707)	-0.0057 (0.0718)
University / public institute			-0.0609 (0.0698)	-0.1108 (0.0786)	-0.1115 (0.0799)
Public sector			0.0347 (0.1114)	-0.0465 (0.1226)	-0.0484 (0.1245)
Impartiality is perceived at least somewhat of a bottleneck					-0.0139 (0.0750)
Woman					0.0160 (0.0711)
EU sceptic					-0.0047 (0.1039)
Constant	0.6939*** (0.0586)	0.6851*** (0.0619)	0.7005*** (0.0697)	0.7250*** (0.0767)	0.7266*** (0.0844)
Country FE	No	No	No	Yes	Yes
Observations	176	176	176	176	176
R^2	0.0185	0.0197	0.0255	0.1607	0.1612
F	3.288	1.738	1.120	1.437	0.817

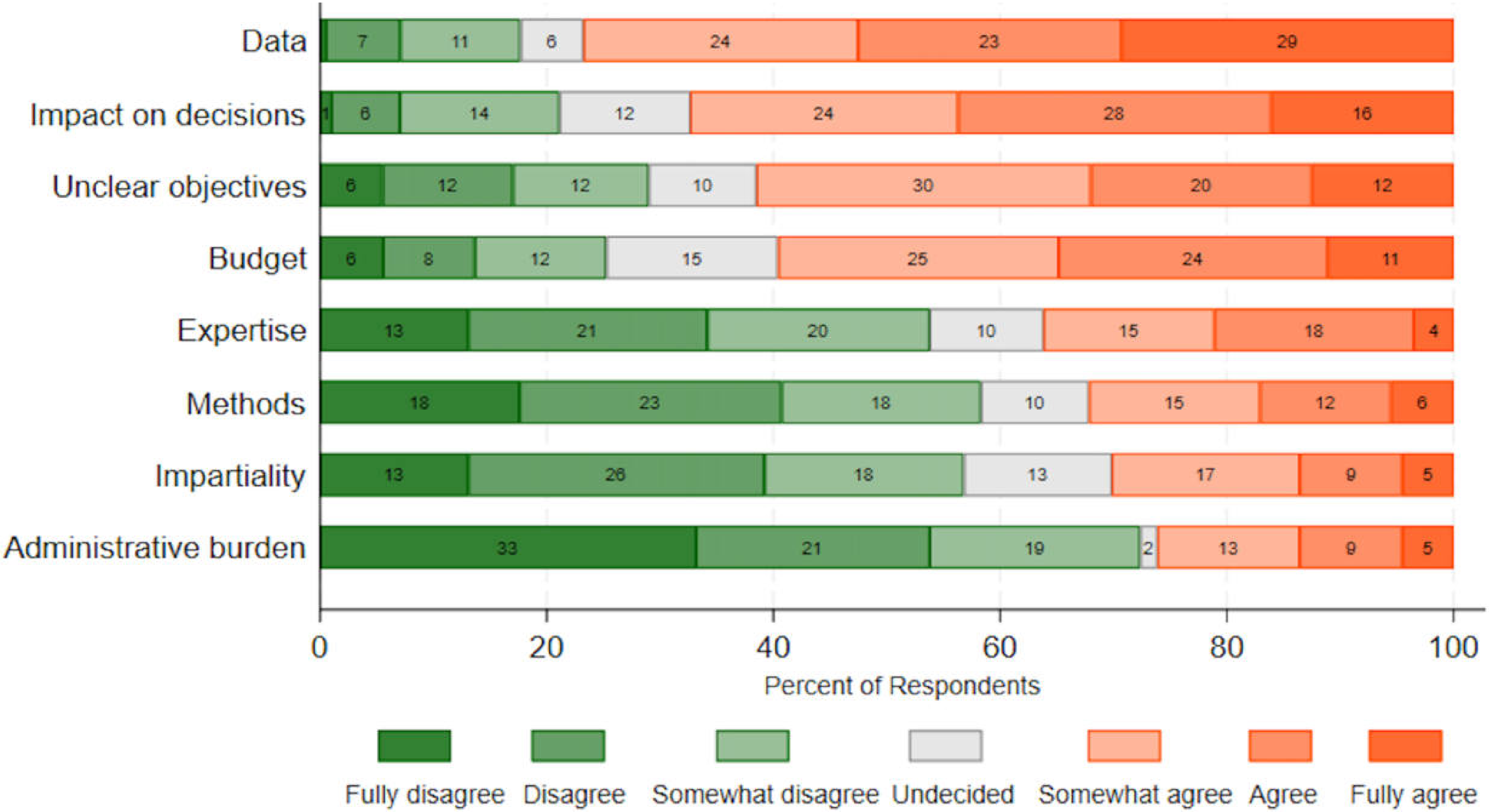
DO EVALUATIONS IMPACT DECISIONS?

AVERAGE SENTIMENT AND GROWTH OF FUNDING IN THE NEXT PERIOD



WHAT ARE THE OTHER MAIN BOTTLENECKS OF THE SYSTEM?

MAIN BOTTLENECKS ACCORDING TO AUTHORS



Question: Finally we are interested in potential bottlenecks of the Cohesion Policy evaluation system. Please select for each of the fdlowing items whether you agree or disagree that they are a major obstacle to the success of the Cohesion Policy evaluation system.

RECOMMENDATIONS

- Developed further in an accompanying paper

  Enhancing Objectivity and Decision Relevance: A Better Framework for Evaluating Cohesion Policies

  By: Heinemann, Friedrich, Zareh Asatryan, Julia Bachtrögler-Unger, Carlo Birkholz, Franceso Corti, Maximilian von Ehrlich, Ugo Fratesi, Clemens Fuest, Valentin Lang and Martin Weber.

CONCLUSIONS

CONCLUSIONS

- Cohesion programme evaluations find positive effects overall.
  - These depend on programmes, countries, but also individual authors.
- However, they do not square well with the existing evidence.
- Why? Can a re-design of evaluation markets fix evaluations?
  - Uncompetitive markets: Very local and, within MS, very concentrated.
  - Impartiality: Large involvement by managing authorities.
  - Both lead to substantially more optimistic findings.
- Technical constraints - data/methods/capacity - still important bottlenecks.
- But also, big disconnect from decision-making:
  - Just a beauty contest? May adversely affect quality of evaluations too.

THANK YOU!

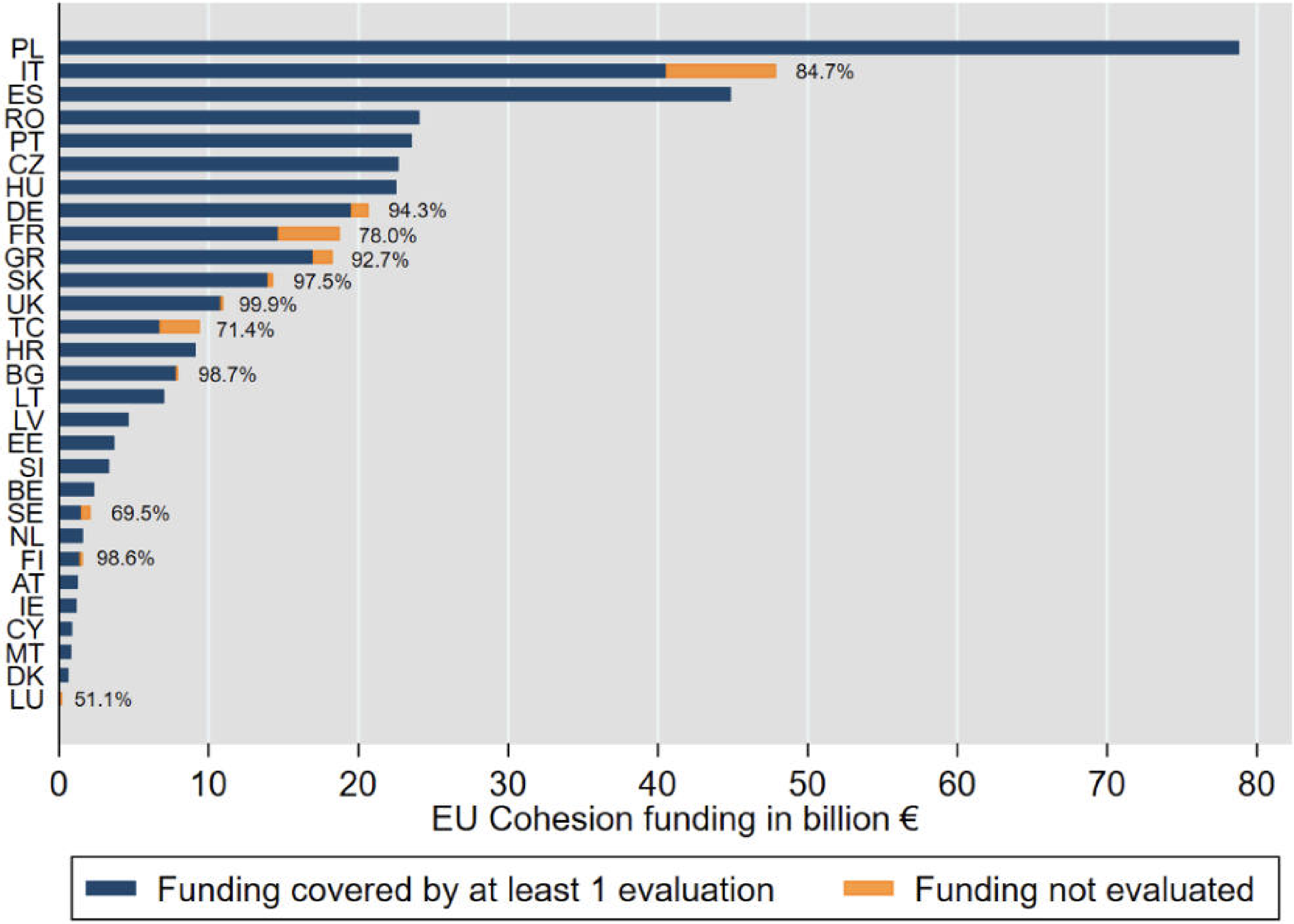
ASATRYAN@ZEW.DE

SITES.GOOGLE.COM/VIEW/ASATRYAN

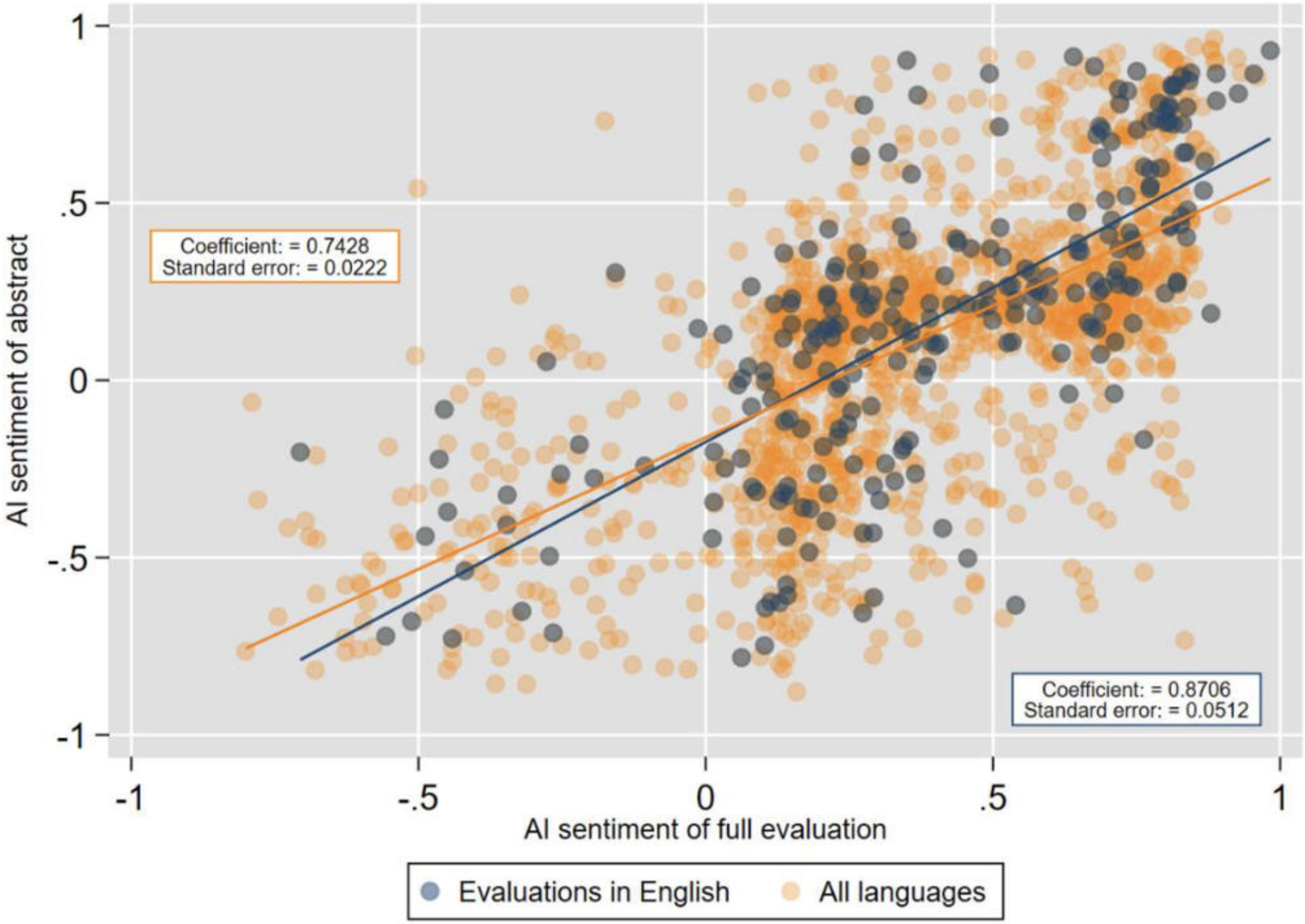
APPENDIX

APPENDIX A: DATA/METHODS

DATA: COVERAGE BY MS



METHODS: ABSTRACT V.S. FULL TEXT



TWO EXAMPLES: AI/HUMAN

0.88/+1

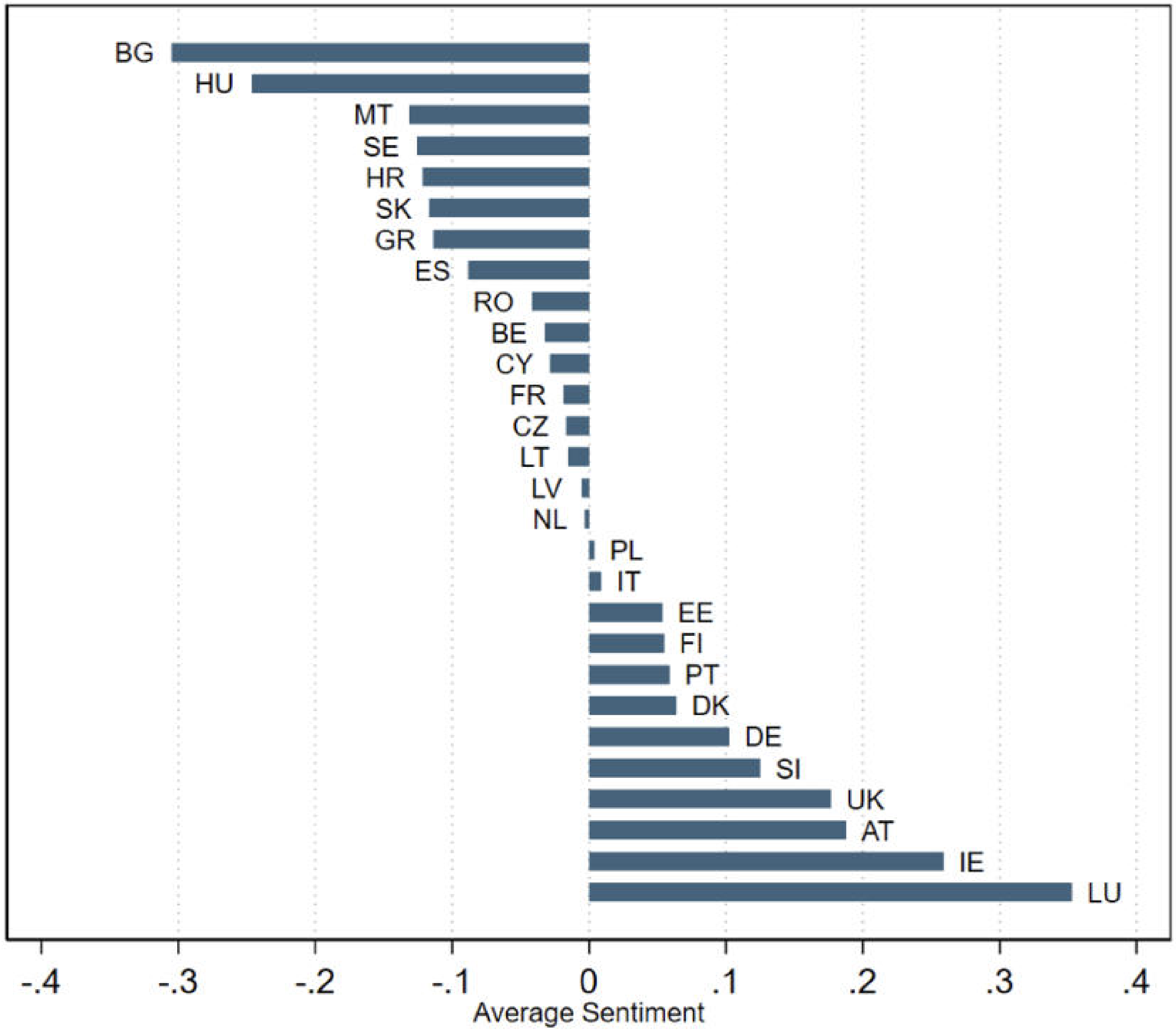
- Cooperation between project promoters and job centres was described as **result-oriented and satisfactory** by institutions and associations involved.
- In 2020 a total of 339 people took part in the employment activities of the 8 projects in line with the planned target.
- The participants were **largely satisfied** with support from project promoters and the work experience provided which had a large degree of flexibility both in terms of working time and content.
- The work experience is reported to have **helped improve social and professional skills** (75% of respondents) **motivation for work** (85%) and **chances to access the labour market** (50%).
- For over 70% of participants employment has also had a **positive impact on their living situation** and social participation.
- According to the survey projects have **benefited urban areas** by strengthening local cultural and support services improving their attractiveness and strengthening social participation and cohesion.

-0.49/-0.5

- The common output indicators are **in general relevant** in respect of the OP strategy but there are exceptions.
- Under IP 8vii "Modernisation of labour market institutions" most of the selected performance indicators have **a low level of consistency** with the logic of the measures implemented.
- In the case of IP 8.vii it will be **difficult to reach targets**.
- The target for IP 8.ii is **overestimated**.
- The result indicators CR05 and CR09 are **consistent with the objective** of IP 9v **but do not enable the results** of measures for strengthening entrepreneurship and the social economy **to be fully measured.**

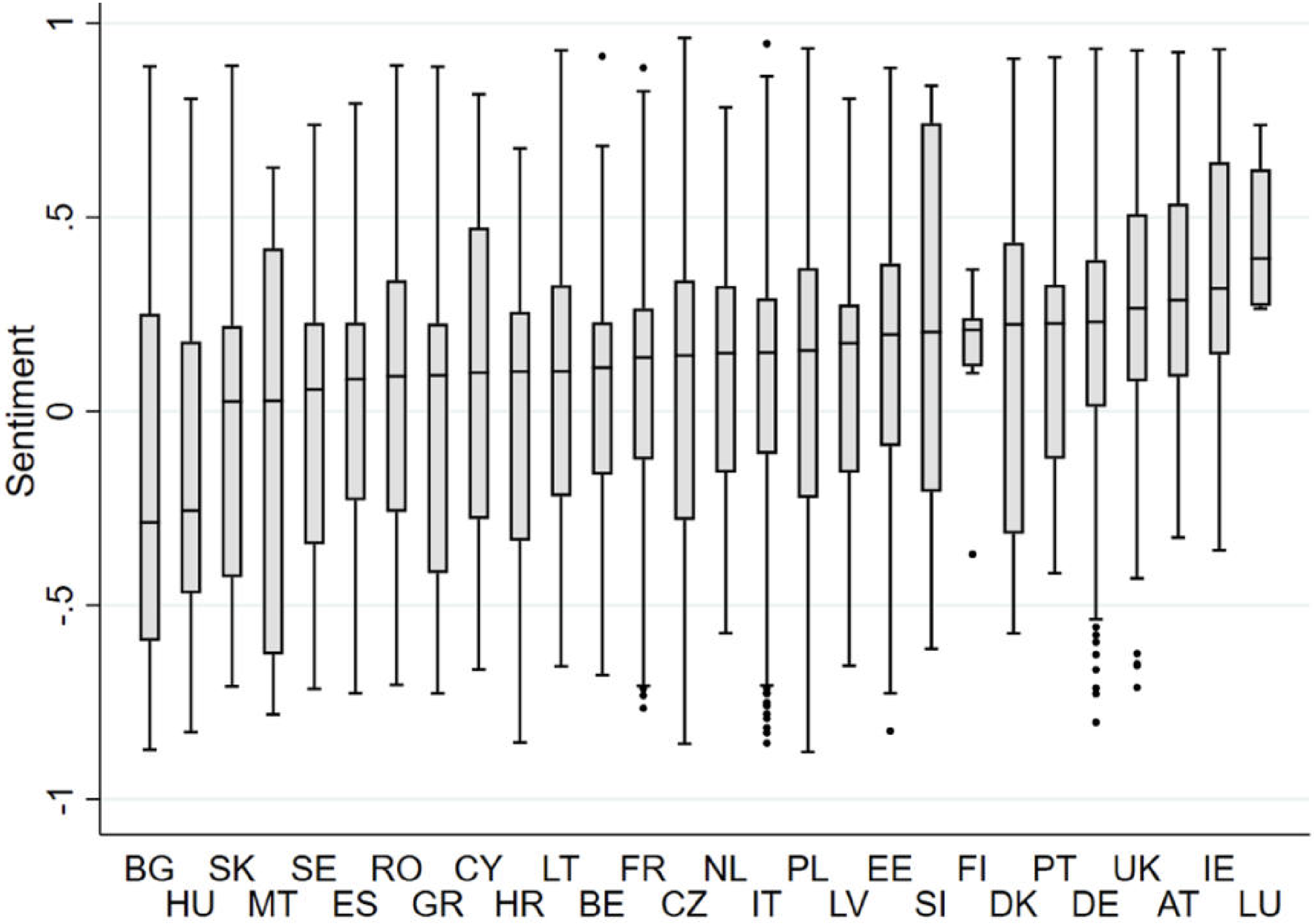
APPENDIX B: ROBUSTNESS

CONDITIONAL SENTIMENT BY MEMBER STATE



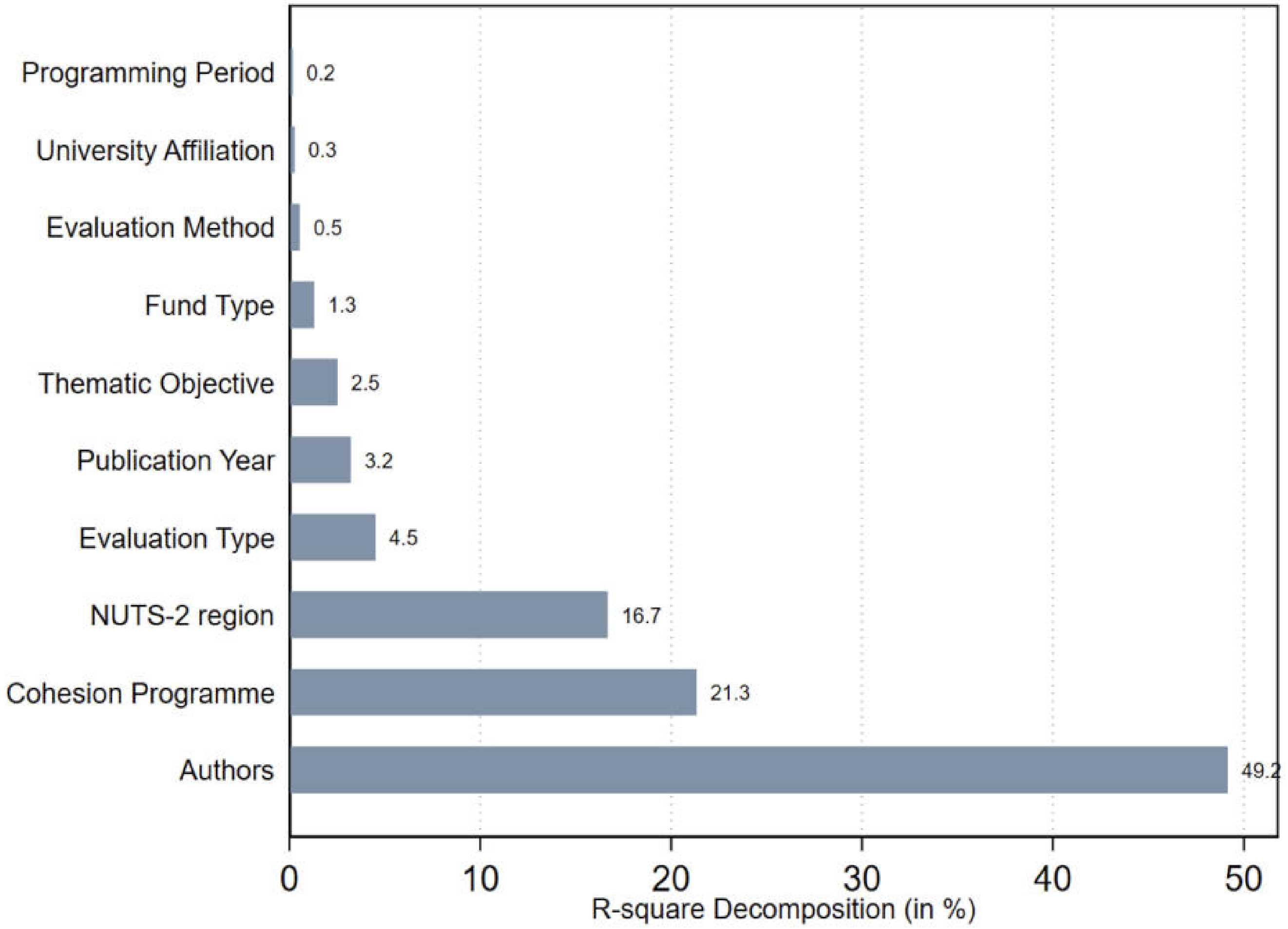
Quantitative Analysis of Cohesion Evaluations: Findings at the level of Member States.

DISTRIBUTION BY COUNTRY



Heterogeneities in evaluation sentiment

NUTS2 INSTEAD OF COUNTRY FIXED EFFECTS. ROLE OF AUTHORS EVEN MORE IMPORTANT…

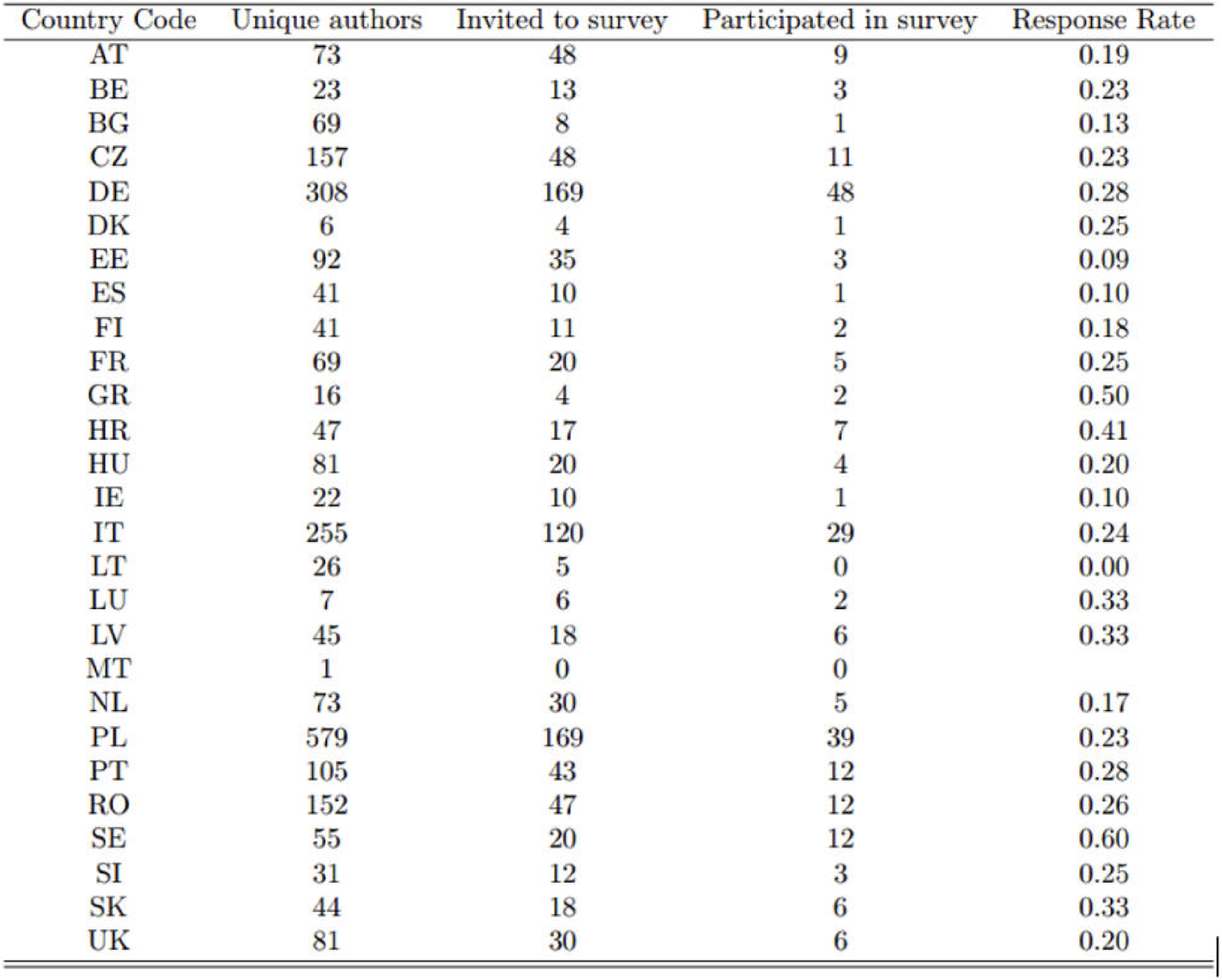


APPENDIX C: SURVEY

METHODS: SURVEY DESIGN

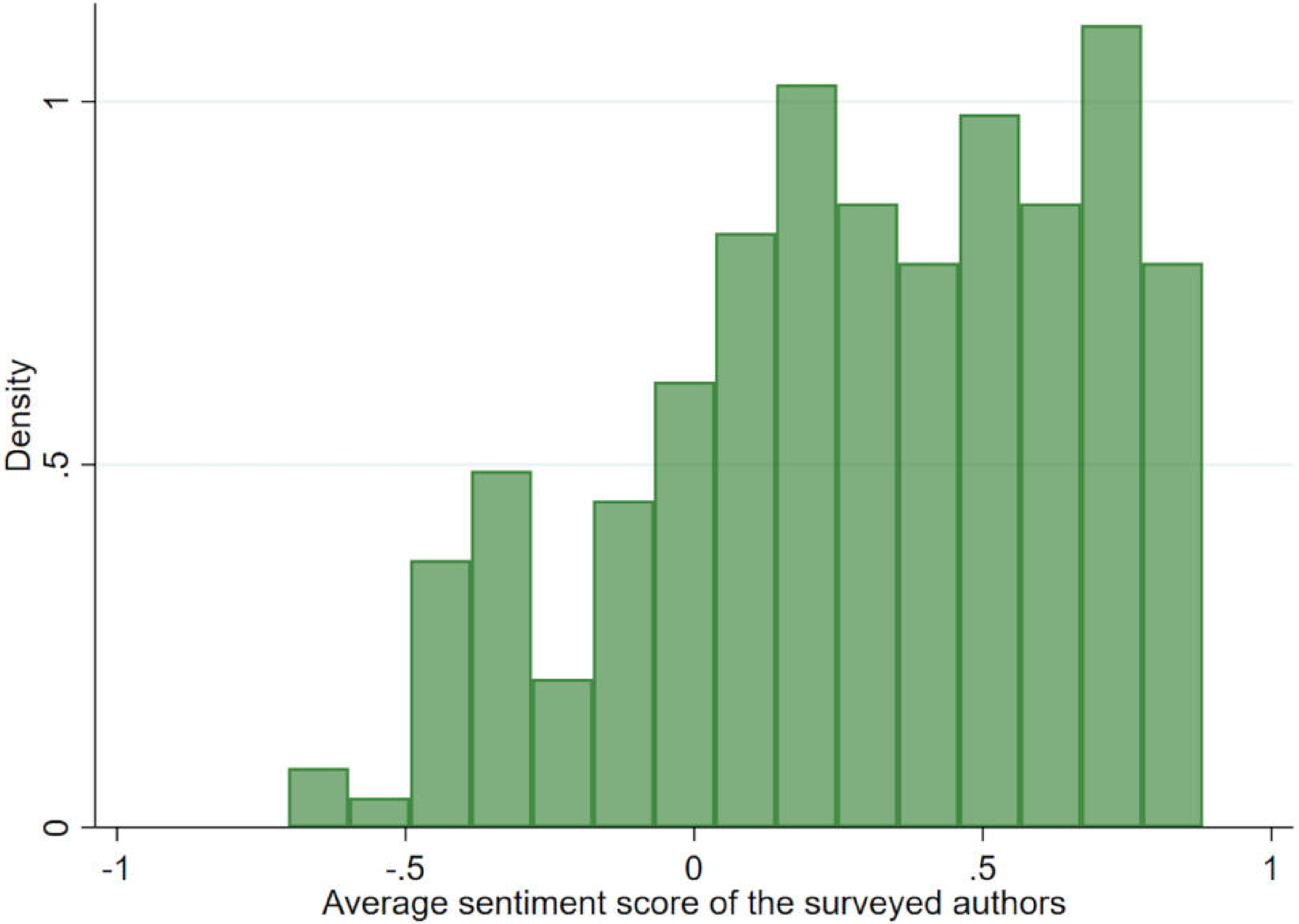
- Survey:
  - Short online survey of authors.
    - About 200 responses, 20% response rate.
  - Collect characteristics on authors and their institutions.
    - Authors: 43% female, 47% with PhD.
    - Employer: 10% public sector, 27% university/institute, 63% private sector.
  - Views on main bottlenecks of the evaluation system.
  - Alos, open-ended question on recommendations.

METHODS: SURVEY RESPONSE RATES

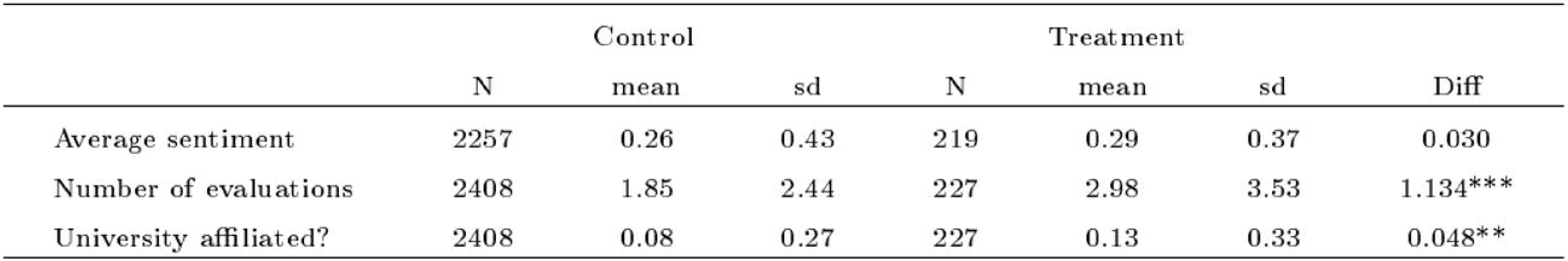


Country Code	Unique authors	Invited to survey	Participated in survey	Response Rate
AT	73	48	9	0.19
BE	23	13	3	0.23
BG	69	8	1	0.13
CZ	157	48	11	0.23
DE	308	169	48	0.28
DK	6	4	1	0.25
EE	92	35	3	0.09
ES	41	10	1	0.10
FI	41	11	2	0.18
FR	69	20	5	0.25
GR	16	4	2	0.50
HR	47	17	7	0.41
HU	81	20	4	0.20
IE	22	10	1	0.10
IT	255	120	29	0.24
LT	26	5	0	0.00
LU	7	6	2	0.33
LV	45	18	6	0.33
MT	1	0	0	
NL	73	30	5	0.17
PL	579	169	39	0.23
PT	105	43	12	0.28
RO	152	47	12	0.26
SE	55	20	12	0.60
SI	31	12	3	0.25
SK	44	18	6	0.33
UK	81	30	6	0.20

DISTRIBUTION OF SENTIMENT AMONG SURVEY RESPONDENTS: P(80)>0.

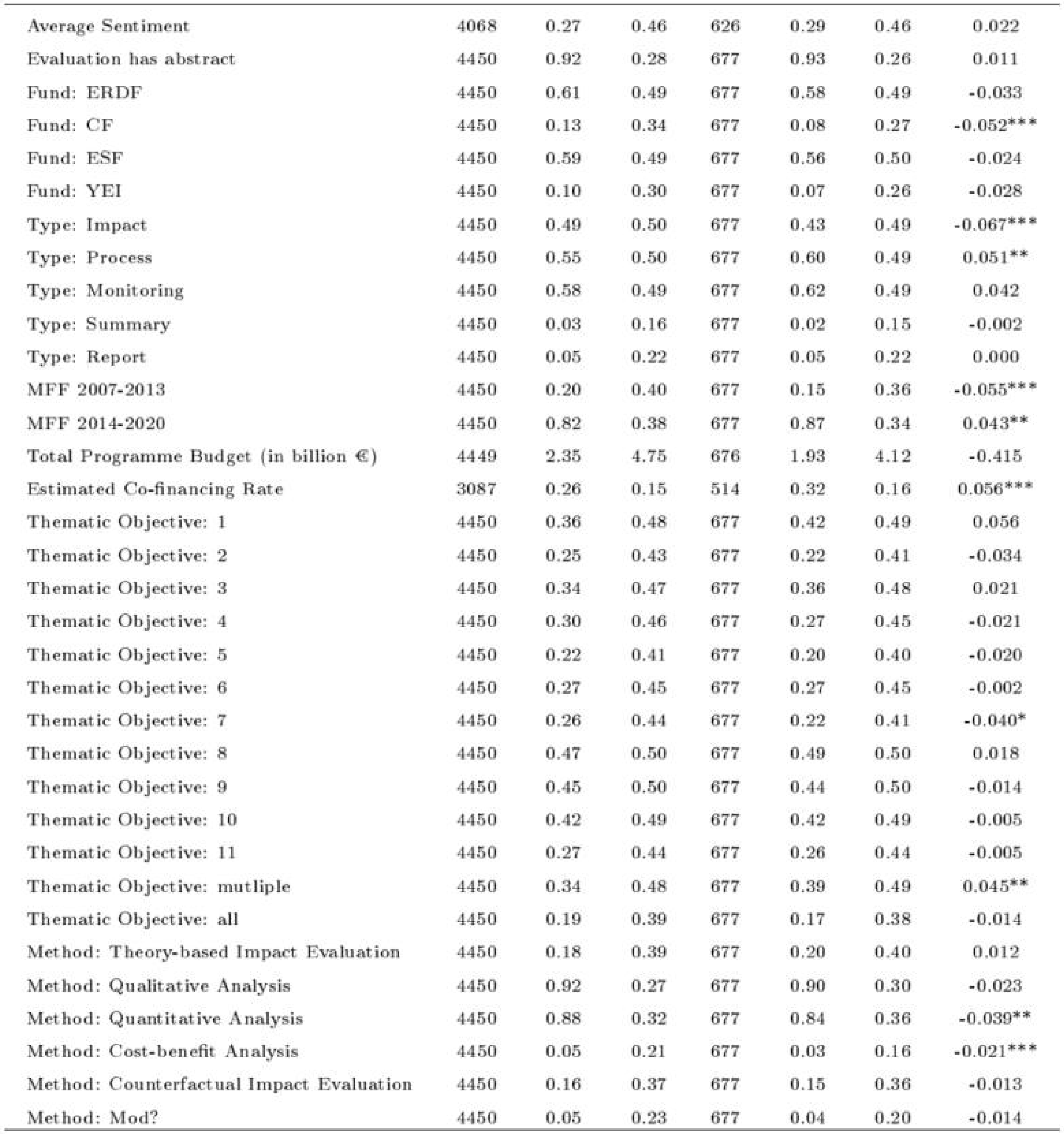


BALANCE TEST: SURVEY VS ALL AUTHORS



	Control			Treatment			
	N	mean	sd	N	mean	sd	Diff
Average sentiment	2257	0.26	0.43	219	0.29	0.37	0.030
Number of evaluations	2408	1.85	2.44	227	2.98	3.53	1.134***
University affiliated?	2408	0.08	0.27	227	0.13	0.33	0.048**

BALANCE TEST: SURVEY VS ALL AUTHORS, DETAILS



Average Sentiment	4068	0.27	0.46	626	0.29	0.46	0.022
Evaluation has abstract	4450	0.92	0.28	677	0.93	0.26	0.011
Fund: ERDF	4450	0.61	0.49	677	0.58	0.49	-0.033
Fund: CF	4450	0.13	0.34	677	0.08	0.27	-0.052***
Fund: ESF	4450	0.59	0.49	677	0.56	0.50	-0.024
Fund: YEI	4450	0.10	0.30	677	0.07	0.26	-0.028
Type: Impact	4450	0.49	0.50	677	0.43	0.49	-0.067***
Type: Process	4450	0.55	0.50	677	0.60	0.49	0.051**
Type: Monitoring	4450	0.58	0.49	677	0.62	0.49	0.042
Type: Summary	4450	0.03	0.16	677	0.02	0.15	-0.002
Type: Report	4450	0.05	0.22	677	0.05	0.22	0.000
MFF 2007-2013	4450	0.20	0.40	677	0.15	0.36	-0.055***
MFF 2014-2020	4450	0.82	0.38	677	0.87	0.34	0.043**
Total Programme Budget (in billion €)	4449	2.35	4.75	676	1.93	4.12	-0.415
Estimated Co-financing Rate	3087	0.26	0.15	514	0.32	0.16	0.056***
Thematic Objective: 1	4450	0.36	0.48	677	0.42	0.49	0.056
Thematic Objective: 2	4450	0.25	0.43	677	0.22	0.41	-0.034
Thematic Objective: 3	4450	0.34	0.47	677	0.36	0.48	0.021
Thematic Objective: 4	4450	0.30	0.46	677	0.27	0.45	-0.021
Thematic Objective: 5	4450	0.22	0.41	677	0.20	0.40	-0.020
Thematic Objective: 6	4450	0.27	0.45	677	0.27	0.45	-0.002
Thematic Objective: 7	4450	0.26	0.44	677	0.22	0.41	-0.040*
Thematic Objective: 8	4450	0.47	0.50	677	0.49	0.50	0.018
Thematic Objective: 9	4450	0.45	0.50	677	0.44	0.50	-0.014
Thematic Objective: 10	4450	0.42	0.49	677	0.42	0.49	-0.005
Thematic Objective: 11	4450	0.27	0.44	677	0.26	0.44	-0.005
Thematic Objective: mutliple	4450	0.34	0.48	677	0.39	0.49	0.045**
Thematic Objective: all	4450	0.19	0.39	677	0.17	0.38	-0.014
Method: Theory-based Impact Evaluation	4450	0.18	0.39	677	0.20	0.40	0.012
Method: Qualitative Analysis	4450	0.92	0.27	677	0.90	0.30	-0.023
Method: Quantitative Analysis	4450	0.88	0.32	677	0.84	0.36	-0.039**
Method: Cost-benefit Analysis	4450	0.05	0.21	677	0.03	0.16	-0.021***
Method: Counterfactual Impact Evaluation	4450	0.16	0.37	677	0.15	0.36	-0.013
Method: Mod?	4450	0.05	0.23	677	0.04	0.20	-0.014

MAIN BOTTLENECKS AND REFORM OPTIONS MENTIONED BY SURVEY RESPONDENTS

- Impartiality:
  - One survey respondent: "Wes' Brot ich ess', des' Lied ich sing".
  - Evaluations commissioned, monitored and approved by those who run Cohesion.
  - Should be an independent body, perhaps a branch of the national auditing authority.
- Impact on decisions:
  - Hugely disconnected from decision-making.
  - One extreme: Ex-ante conditionality.
  - At the least: Better communication between evaluators and policy makers.
- More technical aspects:
  - Data:
    - Consensus: More data, made available more easily.
    - E.g., centralize the burden of the data collection.
  - Methods:
    - Tradeoff: A more rigid European one size fits all approach v.s. comparability.
    - More precise objectives.
  - Capacity:
    - Technical capacity of evaluators but also of managing authorities.
    - Modest resources made available.