// NO.24-037 | 06/2024

DISCUSSION PAPER

// ZAREH ASATRYAN, CARLO BIRKHOLZ,
AND FRIEDRICH HEINEMANN

Evidence-Based Policy or Beauty Contest? An LLM-Based Meta-Analysis of EU Cohesion Policy Evaluations

Leibniz Association

ZEW

Evidence-Based Policy or Beauty Contest?

An LLM-Based Meta-Analysis of EU Cohesion Policy Evaluations

Zareh Asatryan (ZEW Mannheim)[1]

Carlo Birkholz (ZEW Mannheim, University of Mannheim)

Friedrich Heinemann (ZEW Mannheim, University of Heidelberg)

April 2024

Abstract

Independent and high-quality evaluations of government policies are an important input for designing evidence-based policy. Lack of incentives and institutions to write such evaluations, on the other hand, carry the risk of turning the system into a costly beauty contest. We study one of the most advanced markets of policy evaluations in the world, the evaluations of EU Cohesion Policies by its Member States (MS). We use large language models quantify the findings of about 2,300 evaluations, and complement this data with our own survey of the authors. We show that the findings of evaluations are inconsistent with those of the academic literature on the output impacts of Cohesion Policy. Using further variation across MS, our analysis suggests that the market of evaluations is rather oligopolistic within MS, that it is very fragmented across the EU, and that there is often a strong involvement of managing authorities in the work of formally independent evaluators. These factors contribute to making the findings of the evaluations overly optimistic (beautiful) risking their overall usefulness (evidence-based policy). We conclude by discussing reform options to make the evaluations of EU Cohesion Policies more unbiased and effective.

JEL: A11, C45, D83, H43, H54.

Keywords: Policy Evaluation, EU Cohesion Policy, Large Language Model.

[1] Corresponding author: asatryan@zew.de

We thank the German Federal Ministry of Finance for sponsoring this project. We are grateful to Julia Bachtrögler-Unger, Maximilian von Ehrlich and Maxime Fajeau for comments, as well as to Yanxi Hou, Hana Jomni and Patrick Büscher for valuable research assistance.

1 Introduction

Cohesion Policy, which accounts for around a third of the EU's budget and funds over 10% of all public investments in the EU, is the most evaluated of all EU policies (Darvas et al. 2019, Heinemann etal. 2024). In fact, with the mandatory nature of these evaluations since the 2014-2020 programming period (Pellegrin et al. 2020), this evaluation system is advanced, with the EU scoring far ahead of any OECD country according to OECD's index of the strength of performance budgeting frameworks (Downes et al. 2017).

The aims of this evaluation system are clear, and they generally follow those of other systems of performance budgeting. High-quality evaluations can potentially improve policy design by basing them on evidence, and they may also induce learning externalities and increase the transparency of the budget.

These goals are important for any society, but there is a trade-off. Evaluations are not costless, they include direct monetary costs and, perhaps more importantly, they induce indirect costs by setting compliance rules and increasing bureaucracy. Thus, the question is whether the Cohesion Policy evaluation system provides the correct incentives to systematically produce high-quality evaluations, so as to provide a solid basis for better policy design.

Such incentives should promote the establishment of competitive markets of independent and capable evaluators who are able to write impartial and generally high-quality evaluations. Lack of such incentives, on the other hand, carries the risk of turning the system into a costly beauty contest, where the good performance of policies is simply stamped by the evaluations without any serious implications for improving future policy.

To answer this question, we, for the first time in the literature, quantitatively analyse the Cohesion Policy evaluations performed by MS in the last two programming periods. Apart from providing the first methodological basis for systematically analysing the evaluation system, our work is relevant for thinking about reform priorities that improve the evaluation system of EU Cohesion Policy. More generally, our work, which is based on the experience accumulated so far from the EU's well-developed evaluation system and which exploits the unique variation in evaluation markets across the EU MS and regions for its quantitative analysis, can inform the design of evidence-based policies elsewhere. Examples may include the impact of development aid, which is very often evaluated but where the so-called micro-macro paradox is pervasive (Mosley 1986, Doucouliagos and Paldam 2009), the national systems of evaluations in both developing and developed countries many of which are trying

to improve their frameworks of performance budgeting (Downes et al. 2017), or efforts to learn from and scale up successful policy experimentations, where having credible ex-ante evaluations of policy effectiveness are crucial but which are often shaped by political and institutional incentives (Hirsch 2016, Wang and Yang 2021).

The first step of our analysis is to measure the findings of evaluations. We quantify the findings of about 2,300 evaluations that have been written since 2007 by applying a Large Language Model (LLM) to run automated textual analysis of the evaluations' abstracts. This approach lets us estimate a sentiment score for each evaluation, which is a numerical index summarizing how positive or negative the finding of an evaluation about the performance of a specific Cohesion Policy intervention is presented. We validate these estimates by comparing them to findings independently assessed by humans, and work with the assumption that the measurement error in the AI-based estimates is not systematically correlated with our explanatory variables of interest. With this work we contribute to a fast-growing field in economics using LLMs to turn text into data in various application (for a review, see Korinek 2023). We complement this data with observational data on cohesion programmes and details about evaluations, and we also conduct our own survey on a sample of individual authors of evaluations. The survey collects further characteristics about the authors and the institutions they work at, and also asks questions about authors' views on the evaluation system and its bottlenecks.

Using these measurements, we show what the past evaluations have found about the performance of Cohesion Policy on aggregate. Overall, our results suggest that evaluations are, in general, very optimistic about the cohesion programmes they evaluate. We then decompose the variations in these findings and show the dimensions that contribute to the heterogeneity in the findings of evaluations. This decomposition suggests that the most important source of heterogeneity comes from cohesion programmes. However, after controlling for programme specific effects, there is still a substantial degree of heterogeneity across the MS as well as across the individual authors of evaluations.

Second, we compare these evaluation findings to those of the large and growing academic literature in economics on the growth and employment impacts of Cohesion Policy. We perform this exercise at both the MS and more disaggregated NUTS2 levels, as well as for a sub-set of evaluations that target growth and employment as their objective. This exercise suggests that the findings of policy evaluations do not square well with those of the academic literature.

Third, given the diverging results of evaluations and the economic literature, we study the incentives implicit in the evaluation markets and study if certain market-level frictions drive the findings of evaluations. Firstly, we study the competitiveness of markets for evaluations both across and within the MS. Secondly, we study the independence of evaluators from the managing authorities. Our data suggests that, overall, the evaluation markets are highly segmented across the EU, and are fairly oligopolistic within most of the MS, while the managing authorities often exert substantial control over the evaluators, thus, risking their independence. Our empirical analysis suggests that the larger these frictions the more skewed are the findings of evaluations towards showing more optimistic results.

Fourth, and finally, we present evidence from our own survey of evaluators on the more general bottlenecks of the evaluation system from the perspective of evaluators. The survey helps us rank the bottlenecks in terms of their relative importance and discuss some viable policy reform options that could potentially improve the functioning of the evaluation system. A fundamental challenge that stands out is the apparent disconnect between evaluations and decision-making. This, in the opinion of evaluators, may adversely affect the quality of evaluations by reducing the incentives to write high-quality evaluations since they do not matter for policy anyway. Our empirical analysis confirms the absence of policy impacts of evaluations by showing that cohesion funds are not less likely to flow to MS which have received the worst evaluations in the past programming period.

The rest of this paper is structured as follows. Section 2 describes the institutions governing the market of evaluations. Section 3 presents our observational and survey-based data, and describes the meta-analytical methods. Section 4 presents the main empirical results. Section 5 presents a descriptive analysis of the author survey regarding the main bottlenecks in the evaluation system with some ideas on possible reform options. Section 6 concludes with a summary of our main findings.

2 Institutions Governing the System of Evaluations

The Cohesion Policy evaluation framework is aimed at assessing the effectiveness, efficiency, and impact of Cohesion Policy interventions funded under the European Regional Development Fund (ERDF), the Cohesion Fund (CF), and the European Social Fund (ESF). The main legal basis defining the formal rules and procedures of the evaluation process is the Common Provision Regulation (CPR) (European Union 2006, 2013, 2021), which are further

accompanied by fund-specific regulations. For a detailed descriptions of the institutions governing the evaluation system, see Heinemann etal. (2024).

The focus of our analysis is the evaluations by the MS, and it does not include the ex-post evaluations performed by the European Commission. The national evaluations target individual investments and other projects that are part of operational programmes. These are commissioned by managing authorities which are typically the regional authorities, national ministries or local units of the central government (Pellegrin et al. 2020).

In the 2014-2020 programming period, all three main types of evaluations, that is ex-ante, ongoing and impact evaluations, have become mandatory for the MS. In the current programming period of 2021-2027, ex-ante evaluations ceased to be mandatory in an effort to simplify the system (European Commission 2021), while the Commission is now required to also carry out mid-term evaluations (European Union 2006, 2013, 2021).

As to the supply side, the important stakeholders that conduct the evaluations are research institutions, private consultancies, individual experts, but also internal evaluators such as civil servants. They must be functionally independent from the managing authorities which prepare and implement the cohesion programme (European Union 2013). The Commission provides guidance for the MS on how they should outsource evaluations, mentioning an assignment of the evaluation to external experts or a different organization than the one responsible for implementing the programme as best practices (European Commission 2013). To strengthen independence and impartiality, evaluators are also required to disclose potential conflicts of interest. The de-facto independence and impartiality of the evaluation system, however, faces significant challenges while the effectiveness of such ethical and best-practice-type measures arguably remains questionable given the potential high-stakes conflicts involved in the system (Naldini 2018).

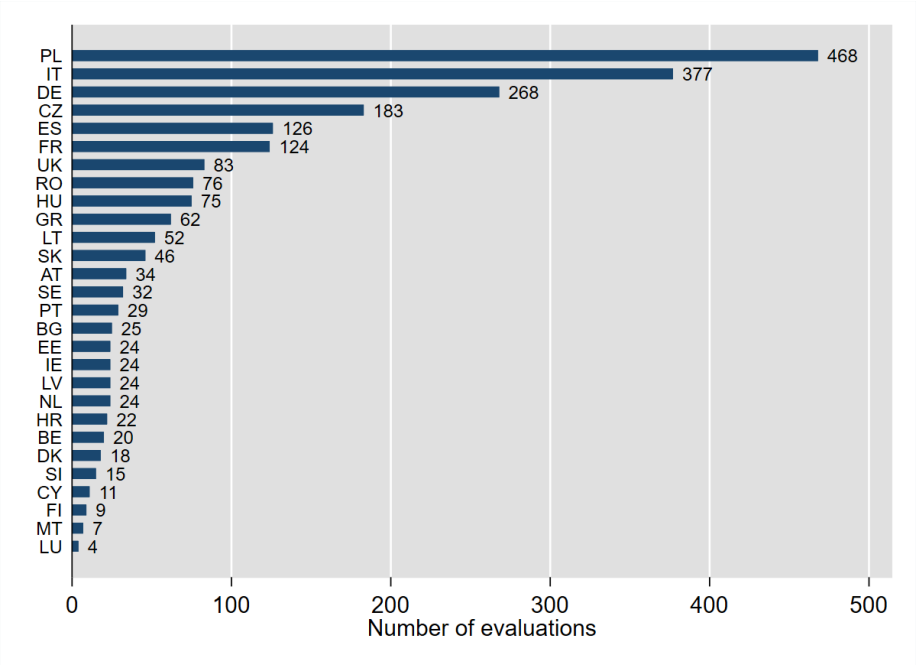
3 Data and Methodology

3.1 Data on Evaluations

Our main source of data is Cohesion Policy evaluations conducted by the 27 MS plus the UK as former MS. The data covers all evaluations conducted in the 2014-2020 programming period, the period when the three types of evaluations first became mandatory, and it extends to impact evaluations done in the 2007-2013 programming period. The data is available publicly at the Cohesion Open Data Platform. The data includes a total of 2538

evaluations, of which textual abstracts are available for 2259 evaluations. The abstracts are in English and they typically follow a standard structure. Other variables in this data include, the title of the evaluation, cohesion programme identifier (called CCI), country code, fund type, evaluation type, evaluation method, and thematic objective. The number of evaluations per MS is presented in Figure 1. Evaluations cover projects of different monetary size, which explains the differences in the number of evaluations even for MS receiving similar amounts of cohesion funds.

Figure 1: Number of evaluations by MS



3.2 Data on Cohesion Programmes and Authors of Evaluations

We merge this data on evaluations to two further datasets. First, we merge the main dataset to data on the budgets of cohesion programmes using the CCI identifiers and the fund type. This helps us capture the total cost of programmes and other details such as the national co-financing shares. The data on budgets is available for only 1765 evaluation abstracts.

Second, we manually collect data on the authors of evaluations. We use the full names of authors to identify authors who have written multiple evaluations.[2] We then use data on evaluations with multiple authors to create international and national co-authorship networks. This data helps us measure the degree of cross-border cooperation in the evaluation market, and also the concentration of the markets within MS.

The idea behind the concentration variable is to measure whether evaluations are written by few or many author clusters. We define author clusters to consist of all the authors that share at least one direct link to a joint co-author.[3] There are several reasons behind our choice to focus on individual authors rather than firms and institutions to construct concentration measures. First, we can precisely identify the individuals, whereas firms and institutions might consist of different branches and teams acting independently, forming different relationships with managing authorities and potentially changing over time too. Second, especially smaller firms might be run by the same ultimate owner, which we cannot systematically identify. One potential drawback of our choice is that authors, especially across institutions and firms, who collaborate on evaluations in some cases, might still compete for evaluation opportunities in the future. Given our choice, we then calculate the number of evaluations written by each cluster and construct the Herfindahl-Index (HHI) by MS, which is a measure of market concentration frequently used in the literature on industrial economics, and is constructed as follows:

$$HHI_{MS} = \sum_{i=1}^{N} \left(\frac{x_i}{\sum_{j=1}^{N} x_j} \right)^2,$$

where x_i is the number of evaluations written by co-author cluster i. Intuitively, the HHI is given by the sum of the squared market shares of each cluster of co-authors in the evaluation market of the respective MS.

3.3. Coverage of Data on Evaluations

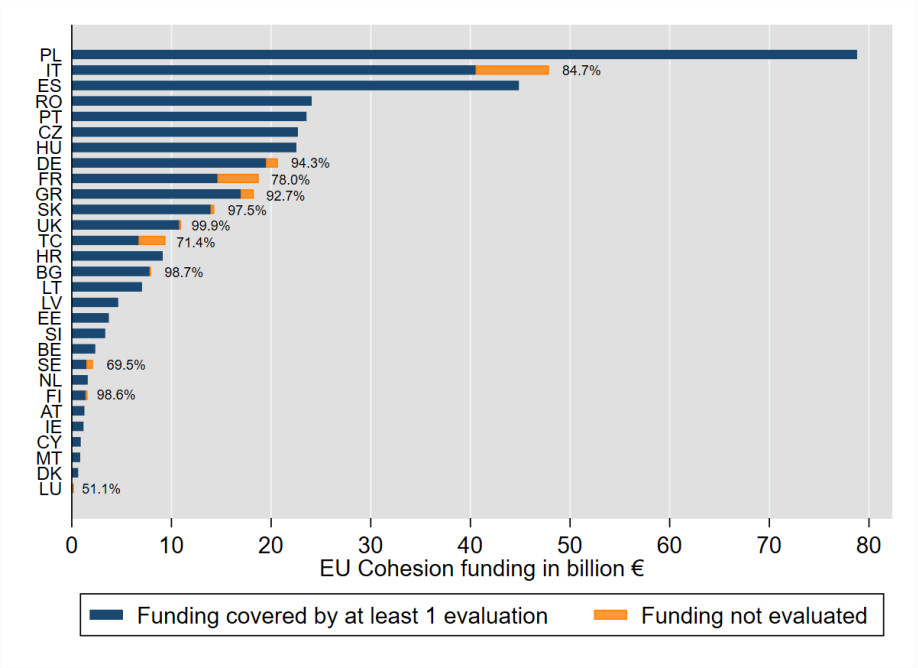
Given the mandatory nature of evaluations, the expectation is that all cohesion programmes are evaluated. We provide evidence in line with this expectation. Figure 2 presents data on

[2] In the unlikely case that two authors share the exact first and last name, we would mistakenly treat them as a single author.

[3] In this exercise, we drop cross-border programmes from this analysis to avoid constructing co-authorship networks across MS since the aim here is to construct measurements of concentration at the level of MS.

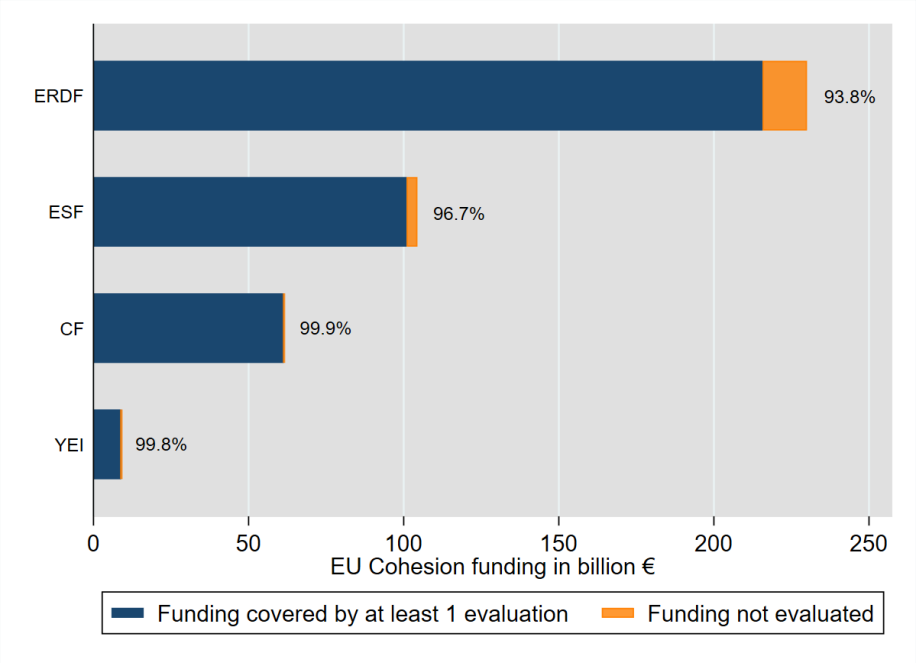
the volume of total and evaluated cohesion funds per country for the 2014-2020 programming period. This data suggests that with few exceptions nearly all of cohesion programmes have been evaluated. This helps reject the concern that there may be selection of the programmes that are being evaluated or not. In Figure 3, we then show the coverage of evaluations by fund. As above, we observe that evaluations nearly fully cover each main type of fund. The funds covered by the order of their total size are ERDF, ESF, CF, and YEI.[4]

Figure 2: Coverage of evaluations by MS



[4] In this classification we also list the Youth Employment Initiative (YEI) as a separate category, although we note that this is not a stand-alone fund and in 2021-2027 it has been fully integrated into the ESF.

Figure 3: Coverage of evaluations by fund



3.4 Meta-Analytical Methods: Quantifying the Findings of Evaluations

The key next step for our meta-analysis is to create a numerical variable that measures the findings of a given evaluation as described in the textual abstract of the evaluation.

While some evaluations present precise numbers on the evaluated performance of the programme, many of these evaluations are qualitative exercises that interpret the performance of programmes verbally. Thus, our approach is to create a score that is informative on whether a given evaluation finds a programme to be more or less successful. We call this score the "sentiment" as expressed in the abstract, and interpret it as capturing the direction and tonality of a given evaluation's finding for the performance of the evaluated cohesion programme.

Given that the definition of what makes a programme more or less successful is not well defined as well as heterogeneous in many directions, we suspect that our measurement of sentiment includes substantial noise. We first define transparently how we measure it using automated text-analysis techniques, then provide a validation exercise that compares the AI-coded sentiment to a manually assessed sentiment.

Our measurement utilizes the large language model, GPT 3.5, and conducts a sentiment analysis on the 2259 abstracts available in the evaluation database. The sentiment analysis is implemented in Python through OpenAI's Application Programming Interface (API) that allows us to interact with the GPT 3.5 model in a consistent and efficient way.

The core part of the code in Python is the prompt, i.e., the instructions provided to the model to obtain the desired response. The prompt should be precise and concise, because the results can be sensitive to how it is written. In our case, we asked the model to rate the sentiment of the abstracts from -1 to 1, with 1 being highly positive, 0 being neutral and -1 being highly negative. In Info Box 1 below we display the prompt used in our analysis.

Info Box 1: The prompt instructing GPT 3.5

{"role": "system", "content": "You are a helpful assistant that conducts a sentiment analysis on abstracts of Cohesion Policy evaluations. "},

{"role": "user", "content": "Rate the sentiment of the abstract from -1.000 to 1.000, -1.000 being highly negative, 0.000 being neutral and 1.000 being highly positive. Provide a three decimals rating and do not round up. Instead of replying with a text, please only state a number with no text. The abbreviations and the objective of the abstract will help you analyse the sentiment of the abstract better. Focus on the sentiment of the final result of the projects/support in your total rating, if available. Here is the abstract: '{}'".format(abstract)},

{"role": "assistant", "content": "Here are some abbreviations that can be found in the abstract:

`OP´ is `Operational Programme´,

`ERDF´ is `European Regional Development Fund´,

`ESF´ is `European Social Fund´,

`YEI´ is `Youth Employment Initiative´,

`CF´ is `Cohesion Fund´,

`TO´ is `Thematic Objective´,

`PaCE´ is `Parents Childcare and Employment´,

`PA´ is `Priority Axis´,

`IP´ is `Investment Priority´,

`SME´ is `Small and Medium Enterprises´.

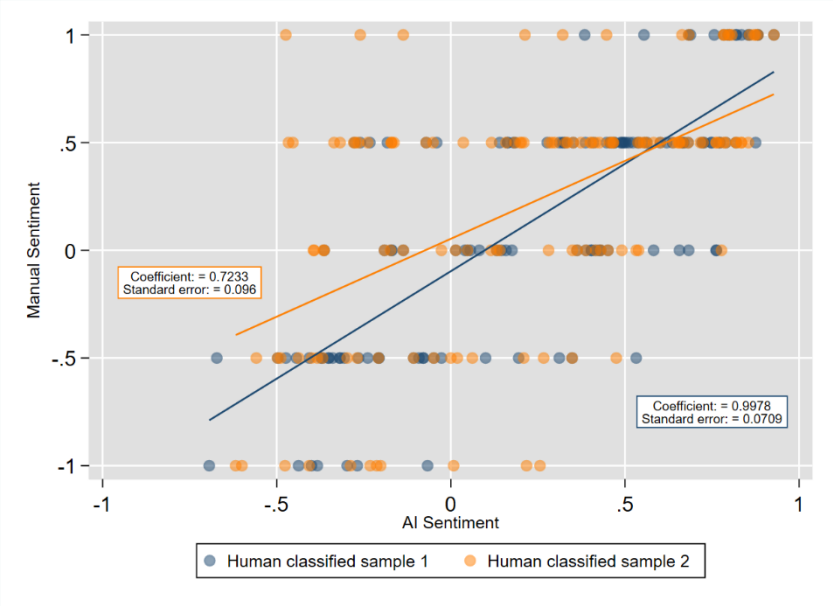
```
This is the objective of the abstract: '{}'".format(objective)},
```

One important choice parameter is the temperature of the model. The temperature parameter sets the volatility of the randomness of the text generated by the model. It ranges from 0 to 2, whereby a higher temperature value results in more diverse and creative output, while a lower temperature value makes the output more deterministic and focused (OpenAI 2023). We make use of the non-deterministic nature of the models output by implementing a bootstrap approach. That means we run the model 50 times for each evaluation with a temperature value of 1. This allows us to generate a measure of certainty about the model's prediction. The intuition is the following: More ambiguous evaluation abstracts will receive a wider range of sentiment scores over the 50 model runs, leading to a higher standard deviation of the predicted evaluation sentiment. For each evaluation we calculate the mean over the 50 runs which serves as our main variable of interest.

To test the accuracy of this method, we manually assess the sentiment of two samples of 132 abstracts. For the first sample we draw 132 abstracts at random. For the second sample we fix half of the initially drawn abstracts and independently reassess them, whereas the other half of the second sample is again randomly drawn. We code the sentiment in five categories: "highly negative", "negative", "neutral", "positive" or "highly positive". We convert this categorical sentiment to a numerical one (-1, -0.5, 0, 0.5 or 1, respectively) and test its correlation with the AI sentiment. For both samples we obtain strong correlation coefficients of 0.998 and 0.723, as depicted in Figure 4.

This exercise gives us confidence that the AI-based score delivers a reliable measure for the abstracts' sentiments, as it would be assessed by a human. Nevertheless, we do not claim that the sentiment is not a noisy measure. Instead, our assumption in the rest of the analysis is that this measurement error is not systematically correlated with the dimensions of our interest, such as across MS.

Figure 4: Manually coded sentiment versus AI-coded sentiment



Notes: The correlation is conducted for two samples of 132 observations. The AI sentiment variable is calculated as the average of 50 runs with temperature 1 and is plotted on x-axis. The manual sentiment is plotted on y-axis. It is a categorical variable, where highly positive is equal to 1, positive is 0.5, neutral is 0, negative is -0.5, and highly negative is -1.

Info Box 2: The methodology behind AI-coded sentiment scores using GPT 3.5

GPT (Generative Pre-trained Transformer) models are state-of-the-art Large Language Models (LLM) with promising applications in the field of meta-analysis. They typically acquire their ability to understand and generate general (as opposed to field-specific) language by training on very large quantities of textual data through machine learning algorithms. Being a recently emerging technology, there is only limited published literature on its role in advancing scientific research. Amin et al. (2023) compare the performance of ChatGPT, OpenAI's chatbot based on GPT, to three baseline methods: RoBERTa language model, Word2VEc word embedding and Bag-of-Words (BoW). The baseline models are specifically fine-tuned for the downstream classification tasks at hand, namely sentiment analysis, personality traits and suicide tendency assessment. The results show that the RoBERTa model is the best performer for the personality and suicide tendency tasks, while ChatGPT achieves the best performance for sentiment analysis. The worse performance of the baseline models is attributed to the noisy nature of twitter data. The authors infer that ChatGPT is a generalist model that can conduct different tasks without specific training, but training is necessary for achieving the best results on specific downstream tasks.

Bang et al. (2023) quantitatively evaluate ChatGPT using 23 publicly available datasets with 8 different Natural Language Processing (NLP) application tasks and find that ChatGPT outperforms other LLMs on several tasks and even achieves better results than fine-tuned models on some tasks. They also find that ChatGPT is better at deductive than inductive reasoning and that its interactive ability allows humans to improve its performance with prompt engineering. However, ChatGPT still produced failed results on each task, and like other LLMs, it suffers from hallucination problems.

Gilardi et al. (2023) use the same model as we do (the ChatGPT API with the gpt-3.5-turbo model) and compare the performance of Mturk crowd-workers to ChatGPT on several annotation tasks and use the human annotations of research assistants as their benchmark. The authors implement several text classification tasks of a large twitter dataset and find that ChatGPT outperforms Mturk crowd-workers on four out of five tasks while being twenty times cheaper than hiring Mturk workers.

Wang et al. (2023) examine whether ChatGPT can serve as a universal sentiment analyser by comparing its performance with the trained BERT and the state-of-the-art (SOTA) models. The authors find that ChatGPT has an impressive zero-shot sentiment analysis capabilities, even corresponding with the BERT and SOTA models that are specifically trained for the tasks at hand. They add that few shot prompting can significantly improve its performance on downstream tasks, datasets and domains, surpassing the fine-tuned BERT but it still performs below SOTA. Wang et al. (2023) deduce that that ChatGPT has powerful open domain sentiment analysis capabilities, yet its performance can be limited for certain specific domains. On the other hand, Kocoń et al. (2023) compare ChatGPT and GPT-4 to SOTA by analysing more than 49 thousand responses and find that ChatGPT exhibits a 25% quality loss on average compared to SOTA, but the loss is significantly lower for GPT-4. The authors also indicate that the ChatGPT quality loss increases the more difficult the task is.

Another study by Zhong et al. (2023) compares the understanding abilities of ChatGPT with four fine-tuned BERT models and show that ChatGPT exhibits comparable performance with BERT on sentiment analysis tasks, surpasses all BERT models on inference tasks, and that its understanding ability can be further improvement by adding advanced prompting strategies.

3.5 Survey of Authors

To enrich our results from the quantitative meta-analysis of Cohesion Policy evaluations, we conducted a survey of the authors of the evaluations. The general aim is to collect further relevant variables which we cannot collect using observational data, but also to measure the views of the authors, who are experts of the evaluation landscape, on various details of the evaluation system.

The two aims of the survey more specifically are as follows. First, we want to learn more about the people and institutions that conduct Cohesion Policy evaluations: What educational background do the evaluators typically have, what type of institutions are most commonly performing them and how reliant on these evaluations are they from a business perspective. Second, we are interested in understanding the experts' views on the EU and its policies in general, as well as on the Cohesion Policy and its evaluation landscape in particular. We asked up to a total of 16 questions. The invitation to participate in the survey and its introduction, as well as the exact questionnaire of the survey, can be found in Figures A.1, A.2 and A.3 of the Appendix.

The design of the survey is as follows. As a first step we manually collected publicly available email addresses of the authors through desk research. We managed to find a total of 1175 contacts, which is about half of the authors in our sample. The survey was sent out on September 27, 2023, and was in the field for four weeks until October 25, 2023.[5] Out of the 1175 emails we sent out, around 230 did not reach their intended recipient, either due to faulty email addresses or restrictive email filters of the recipients' email provider. Out of the 945 remaining potential participants, 213 completed the entire survey while 17 gave partial responses to the questionnaire. The fairly high response rate of almost 25% may be, for example, due to the close engagement of the participants with the topics of the survey.[6]

Table 1 below details the total number of unique authors, as well as the number of authors for whom we have successfully collected a contact email address and the number of respondents per MS. We received responses from almost all MS except those with very low evaluation activity due to the few unique authors these countries have.

5 The invitation email is displayed in Figure 13 in the Appendix.

6 To further increase the response rate a donation incentive was added whereby a donation of 5€ up to a cap of 1000€ was made for each full response towards disaster relief by the charity Aktionsbuendnis Katastrophenhilfe.

Table 1: Number of authors and survey participation by MS

Country Code	Evaluations	Unique authors	Invited to survey	Participated in survey	Response rate
AT	28	74	50	10	0.20
BE	9	25	13	3	0.23
BG	20	69	8	1	0.13
CZ	77	174	48	11	0.23
DE	249	326	180	50	0.28
DK	13	9	3	0	0.00
EE	16	99	35	3	0.09
ES	28	44	10	1	0.10
FI	10	29	11	2	0.18
FR	64	83	22	5	0.23
GR	15	23	4	2	0.50
HR	16	49	16	7	0.44
HU	34	97	22	4	0.18
IE	10	24	10	1	0.10
IT	205	280	120	29	0.24
LT	5	23	5	0	0.00
LU	4	8	6	2	0.33
LV	16	46	20	6	0.30
MT	1	1	0	0	-
NL	26	73	30	5	0.17
PL	288	611	172	39	0.23
PT	19	106	44	13	0.30
RO	61	179	52	13	0.25
SE	28	57	22	13	0.59
SI	11	35	14	5	0.36
SK	26	53	22	6	0.27
UK	44	78	31	6	0.19

Notes: The table depicts the number of evaluations and unique authors, as well as the response rate to the survey as the share of authors who participated in the survey out of the authors invited to the survey broken down by country.

Our design leads to two different types of selection issues. The first is stemming from the not full coverage of author contacts, and the second is coming from the below full response rates among the authors who have received the survey. To understand the representativeness of our sample of respondents we conduct two balance tests. First, we analyse balance across evaluation characteristics such as the type of fund and evaluation, the evaluation method, or the thematic objective. In Table A.1 of the appendix, we compare respondents to the underlying population of all evaluators, whereas in Table A.2 respondents are compared to all contacted authors. Systematic differences in the former would speak to authors' email addresses being differentially likely found, while differences in the latter would indicate differential response rates across observable characteristics. Importantly, we find no differences for the average sentiment or the programme size in either table. We find some minor differences, none of which suggest a systematic pattern which would bias our findings. Noteworthy are the differential contact finding and response

rates by budgeting period. This makes intuitively sense, as authors writing evaluations for earlier periods are more likely to have moved on to new institutions or jobs, or might have retired.

Second, we analyse author characteristics in Tables A.3 and A.4 of the appendix. We again compare survey respondents to all authors and to only those authors who were invited to participate in the survey. One clear difference in the respective samples is that for authors writing more evaluations contact email addresses were easier to find, and they were more likely to participate in the survey. The positive and significant difference in found email addresses by university affiliation is unsurprising, as universities commonly have public website profiles of their staff. However, the difference does not manifest in response rates.

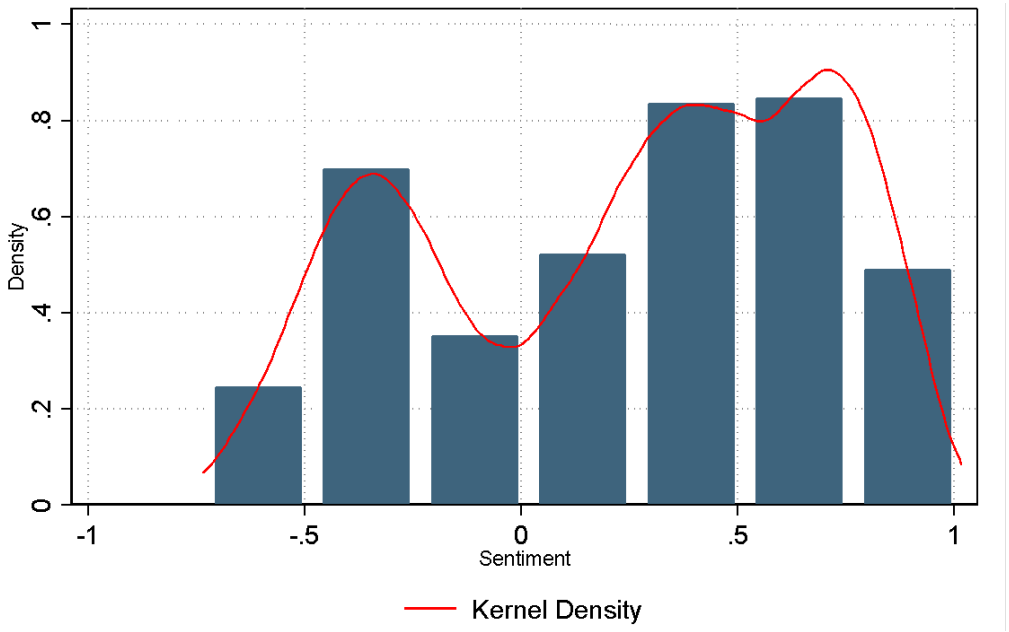
4 Results

4.1 Findings of Evaluations

In this section we present our measurements of the sentiment of the evaluations. We first present the evidence on aggregate and on the MS level, and then study the factors that explain the variation in these findings.

Figure 5 plots the distribution of sentiment for all evaluations. As discussed in Section 3.4, these scores are estimates using an AI analysis of abstracts of the evaluations, and they range from a very negative, -1, to a very positive, +1, score. Figure 5 documents three interesting facts. First, the sentiment is much more likely to be positive than negative, that is there are about twice as many evaluations with scores larger than 0, than evaluations with scores 0 or below. Second, within positive evaluations the scores are roughly normally distributed in their magnitude (i.e., there are many positive evaluations with an average magnitude and about equal number of very good and somewhat good evaluations), while within negative evaluations there are virtually no evaluations with very bad scores. Third, there is a relative lack of evaluations with sentiment close to 0, which are evaluations that either find null effects, or find both positive and negative effects which largely balance each other out.

Figure 5: Distribution of evaluation findings on aggregate

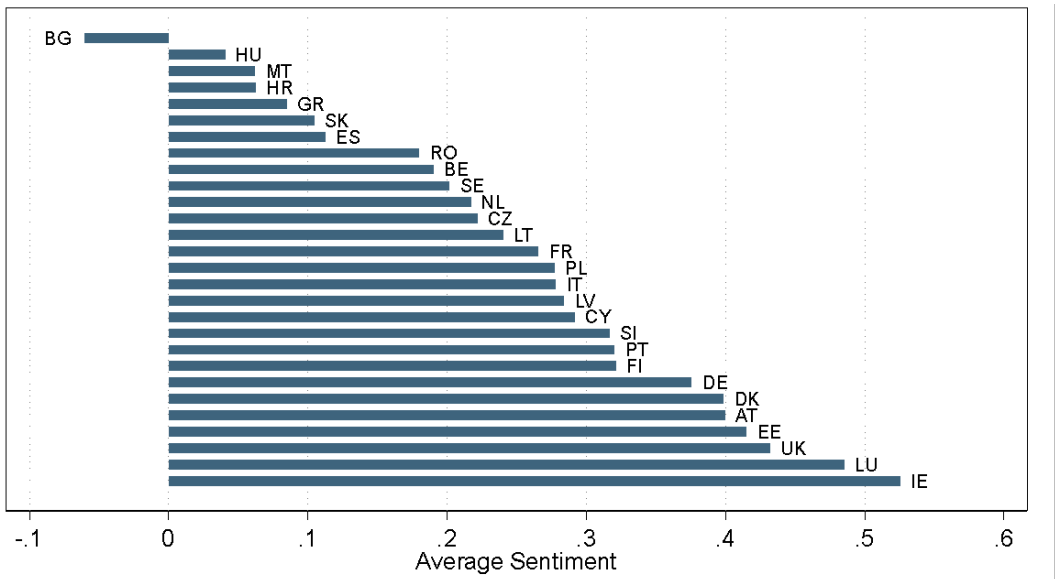


Notes: Number of underlying observations is 2,259. These are grouped into 8 equal bins. Densities are shown with the respective 8 bars. The sentiment variable is calculated as the average of 50 runs with temperature 1. The red line shows the estimates Kernel density using the underlying raw data.

Figure 6 presents the average sentiment score by MS. Overall, there are large differences in mean evaluation scores across MS. Bulgaria is a clear outlier with its negative mean score based on 25 programme evaluations, followed by Hungary, Malta, Croatia, Greece, Slovakia and Spain. In the upper tail, the leader is Ireland based on 24 evaluations, followed by Luxemburg, UK, Estonia, Austria and Germany. These differences may reflect real differences in the quality of projects across the MS, but they could also be driven by underlying differences in how strict or independent the evaluations are performed. In Figure B.1, we also show the distributions behind these average scores for every MS, in terms of the median value of the score within that MS, its minimum and maximum values, and the values at the bottom and top quartiles.[7]

[7] In Figure A. 5 of the appendix, we also check for heterogeneous evaluation scores by programme size. However, we do not detect statistically significant differences in this dimension.

Figure 6: Average unconditional evaluation result by MS



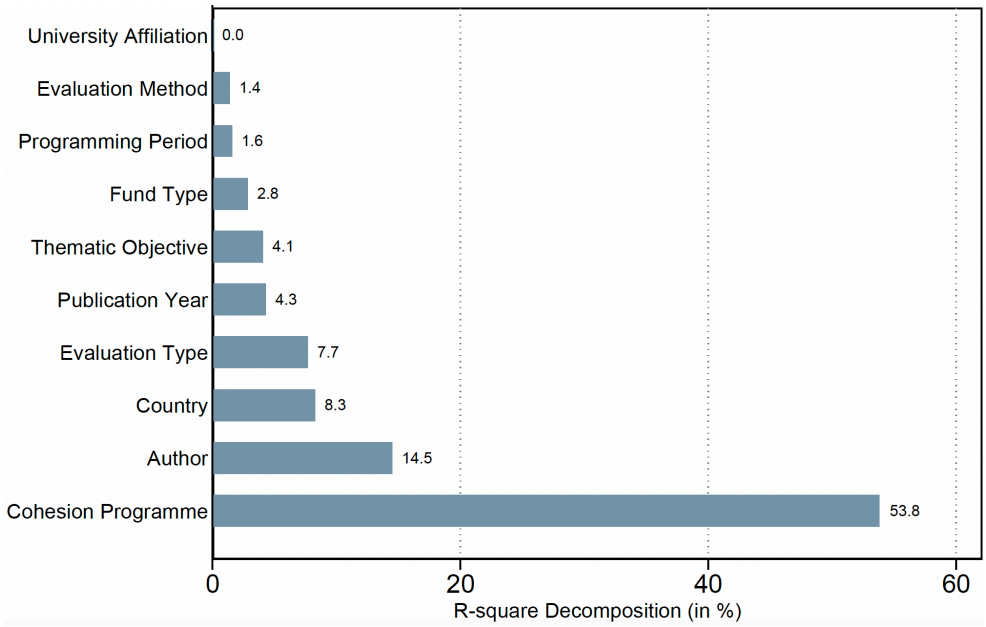
Notes: Number of observations (i.e. evaluations) per country is: Bulgaria (BG): 25, Hungary (HU): 75, Malta (MT): 7, Croatia (HR): 22, Greece (GR): 62, Slovakia (SK): 46, Spain (ES): 126, Romania (RO): 76, Belgium (BE): 20, Sweden (SE): 32, Netherlands (NL): 24, Czech Republic (CZ): 183, Lithuania (LT): 52, France (FR): 124, Poland (PL): 468, Italy (IT): 376, Latvia (LV): 24, Cyprus (CY): 11, Slovenia (SI): 15, Portugal (PT): 29, Finland (FI): 9, Germany (DE): 267, Denmark (DK): 17, Austria (AT): 34, Estonia (EE): 24, United Kingdom (UK): 83, Luxembourg (LU): 4, Ireland (IE): 24. Number of countries: 28. Total number of observations: 2,259.

Next, we investigate which factors predict the evaluation findings as captured by the sentiment score. To this end, we run a large linear regression of ten potential explanatory variables on the sentiment score. These variables are plotted on the y-axis of Figure 7. Overall, these variables jointly explain 41% of the variation in the sentient score (i.e., the R-squared of the regression), which is a fairly large number given our suspicion that the sentiment score includes substantial measurement error. We then perform a Shorrocks-Shapley decomposition to estimate the degree to which each of these ten variables contribute to explaining the variation in sentiment in relative terms.

From the ten initial regressors, cohesion programme fixed effects stand out as the most powerful predictor of the findings of evaluations. Dummies for the type of cohesion programme explain over half of the variation, which is more than all the other nine variables combined. In other words, evaluations performed on the same programmes are fairly

similar to each other in their findings. The next two variables ordered by their explanatory power are authors and countries. To understand the role of authors, we utilize the fact that single authors write many evaluations which allows us to estimate individual author fixed effects. In our data, from the 2,564 unique authors, 1,857 wrote two evaluations, with the average author writing 2.73 evaluations. The findings suggest that even after controlling for programme fixed effects and for the other explanatory variables, individual authors still have a considerable margin of impact on the findings of evaluations. Consistent with the evidence on the wide heterogeneity in the average unconditional sentiments across MS presented above, Figure 7 suggests that after controlling for the other explanatory variables there is still a substantial variation left across the MS. As an alternative specification, we include NUTS2 fixed effects instead of country fixed effects. This estimation presented in Figure B.3 of the appendix suggests that the role of programmes decreases, which is intuitive since programmes often coincide with NUTS2 regions, while the role of individual authors increases further explaining about 18% of the relative contribution of these variables. Several other of the remaining variables explain a non-negligible variation of the sentiment.

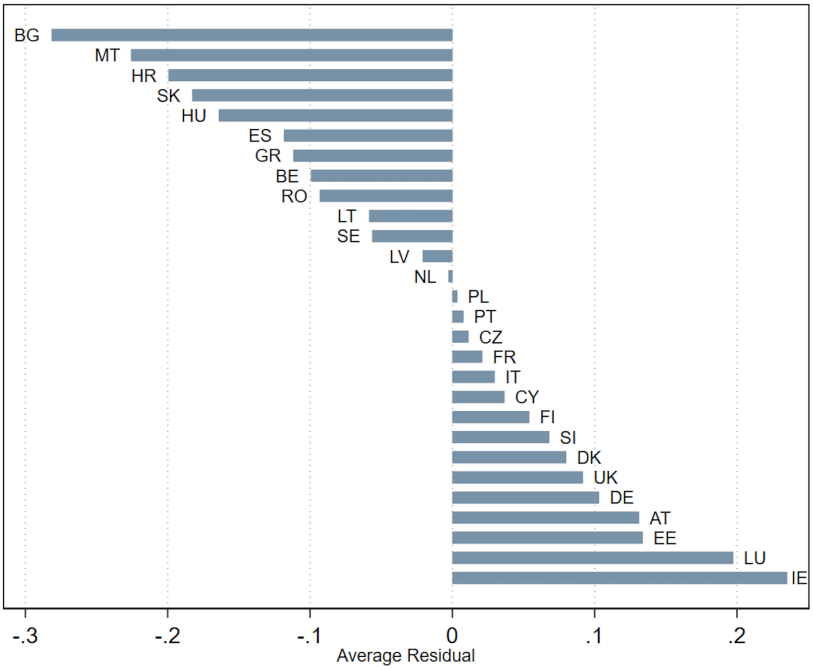
Figure 7: Explaining the variation in evaluation findings



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. n=1,363. R2=0.4114.

As a final exercise, in Figure 8 we show the MS level heterogeneity in sentiment but now taking the sentiment conditional on a number of evaluation level characteristics, rather than just averaging the raw data on sentiment as in Figure 6. This is an important exercise since the composition of evaluation characteristics will be different across the MS, and we want to make sure that the differences of MS level averages do not just reflect these compositional differences. Overall, the relative ranking of MS according to their average sentiment in Figure 8 is similar to the one in Figure 6, suggesting that composition differences in evaluations do not explain the substantial heterogeneities across the MS that we observe.

Figure 8: Average conditional evaluation result by MS



Notes: Figure presents the MS level sentiment similar to Figure 6, but now the sentiment is conditional on a number of evaluation characteristics: Fund, evaluation type, thematic objective, evaluation method, programming period, publication year. Thereby, we run a regression of sentiment on these control variables, and calculate the average of residuals at the MS level. As a result, the plotted sentiment score is in relative rather than absolute terms.

4.2 Comparison of the Findings of Evaluations to those from the Literature

Next, we investigate whether the evaluation findings square well with the findings from the economic literature on the impact of cohesion policies. We are aware of four different estimates of the impact of Cohesion Policy on either growth or employment that present its impact differentiated by the MS. Three of the papers are empirical and all of them use fairly sophisticated techniques of causal identification, and one estimate comes from the DSGE model of the European Commission used for simulating the impact of Cohesion Policy on growth and employment called the RHOMOLO model. We discuss these four estimates in some detail.

First, Di Caro und Fratesi (2022) use regional data from 1989 to 2015 covering four programming periods and apply a panel fixed effects model to examine the impact of ERDF funds on GDP growth. The authors use a heterogeneous coefficient approach and provide estimates of average impact at the level of MS as well as NUTS2 regions (see Figure 3 of the paper[8]). Second, Fidrmuc et al. (2019) employ regional data from 1997 to 2014 and apply a 2SLS strategy. Their spatial models lead to country-specific multipliers (as reported in Table 8 of the paper[9]). Third, Canova und Pappa (2021) construct regional data running over 30 years and implement an instrumental variable Bayesian approach. They estimate regional level dynamic multipliers separately for ERDF and ESF (see, respectively, Figures 4 and 5 of the paper). Fourth, and finally, we take the estimates of the RHOMOLO model from Crucitti et al. (2022) on the impact of the 2014-2020 Cohesion Policy on GDP per capita in 2021 by MS (see Table 4 of the paper). [10]

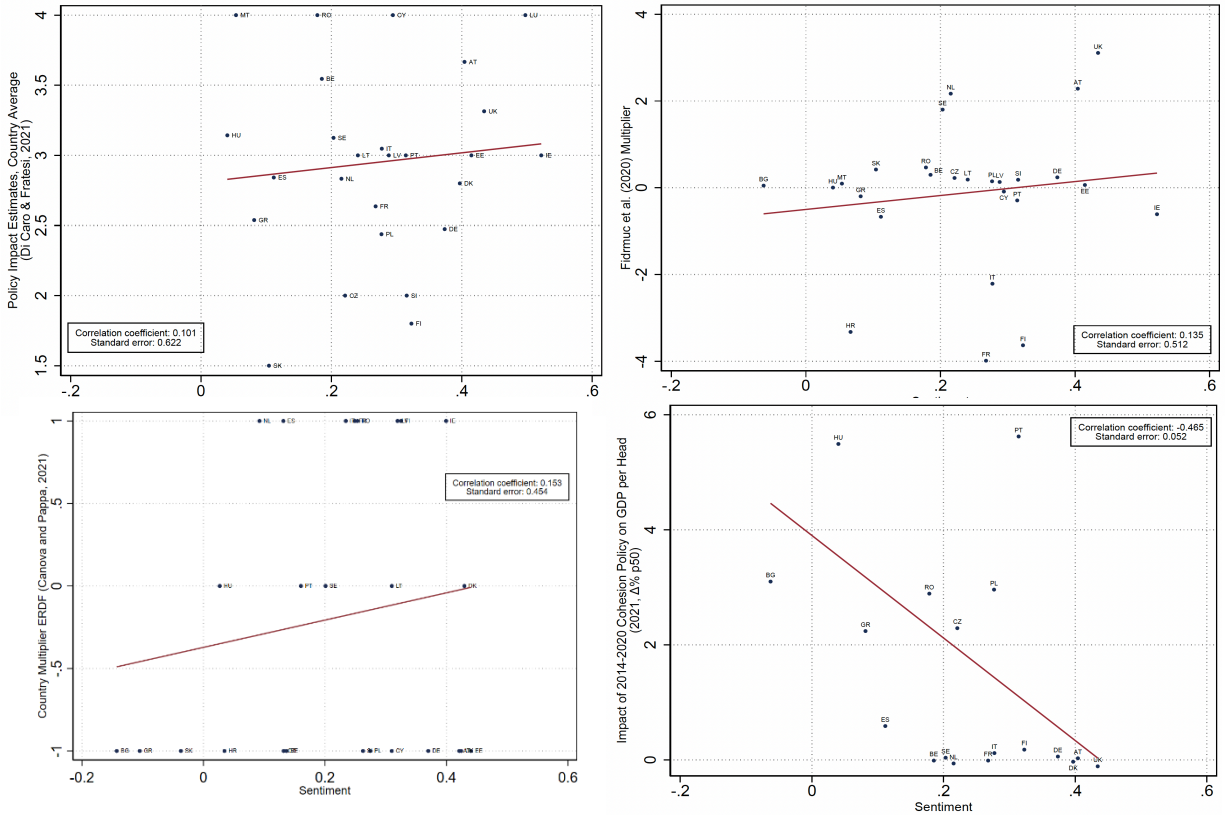
Figure 9 presents correlations of the findings from these four estimations with our sentiment score at the level of MS. Surprisingly, the results do not suggest any correlation with the three empirical papers, while the correlation with the findings of the RHOMOLO model is even negative. Assuming that the outcomes measured in the evaluations and in the literature are related, the absence of correlations suggests that the measurements of either the evaluations or the literature or both must be wrong.

[8] We are grateful to the authors for sharing with us the underlying data on NUTS2 level impact estimates.

[9] Note that these estimates are only present in the working paper version, but not in the published version of the paper.

[10] For each Member State the paper presents the distribution of regional estimates in terms of the median, bottom and top deciles of the magnitude of the impact. For our baseline analysis we take the median estimate per Member State, and in the appendix present robustness tests for the bottom and top deciles.

Figure 9: Comparison of MS specific evaluation sentiment with the output-impacts of Cohesion Policy as estimated by the economic literature



Notes: Sources for the Cohesion Policy impact measures are: Top-left: Di Caro und Fratesi (2022); top-right: Fidrmuc et al. (2019); bottom-left: Canova und Pappa (2021); bottom-right: Crucitti et al. (2022).

We perform three robustness tests to confirm this finding. First, at the MS level we have few observations, and thus the absence of observable correlation may potentially be driven by the lack of statistical power due to a limited sample size rather than a true absence of correlations. We reject this hypothesis by performing the analysis at the regional level using NUTS2-specific estimates of the effect of Cohesion Policy. Such estimates are available only in Di Caro und Fratesi (2022), and Figure B.4 of the appendix shows no correlation between their estimates and our sentiment data on the NUTS2 level. Second, it could be that the outcomes studied by the evaluations and the literature are very different. To reject this hypothesis, we look at a sub-sample of evaluations whose thematic objectives have economic growth or employment increases as the primary target, and repeat the analysis

for this sub-sample.[11] However, Figure C.1 of the appendix suggests that also in this sample the findings of the evaluations do not correlate with those in the literature, neither at the MS and nor at the NUTS2 levels. A third possibility is that the economic literature has done a poor job in estimating the impacts of Cohesion Policy at the MS level. In this case findings of the different papers in the literature should also be inconsistent with each other. We reject this third hypothesis by showing in Figure C.2 that the findings of the literature indeed correlate positively with each other.

Thus, we conclude that the national and regional variance of evaluation sentiments is unrelated to corresponding findings in the academic literature on the differentiated growth and employment impacts of Cohesion Policy. Of course, the evaluation sentiment – even for evaluations that focus on growth and employment effects – is an indirect measure of how an evaluation assesses a programme's growth effect. Nevertheless, this complete lack of correlation, and the even negative correlation in case of the estimates of the RHOMOLO model, shows that evaluations paint a rather different picture of Cohesion Policy performance compared to the academic papers.

4.3 Market Structure of Evaluations

In this section, we study the market structure of evaluations across and within MS. The aim is to understand how the markets for evaluations function and what their implications for the findings of evaluations are. Given the divergence between the findings of evaluations and the academic literature, it is helpful to study possible market imperfections (such as the potential oligopolistic power of evaluators or frictions arising from segmentation of markets across the MS) and ask whether these can help explain this divergence.

First, we ask if there is a single cross-border market for evaluations. Do authors frequently work on evaluating cohesion programmes in different MS, or are the markets segmented along national borders? Second, we ask if the markets in individual MS are competitive, that is whether there are many institutions and author clusters competing with each other to

[11] We select those thematic objective that target important input factors directly such as production technology with TO1 (research, technological development and innovation), infrastructure with TO2 (ICT access, use and quality) and TO7 (sustainable transport and network infrastructure improvement), human capital with TO8 (employment and labour mobility) and TO10 (human capital investments), firm subsidies with TO3 (SME competitiveness), or regulation with TO11 (efficient public administration).

win contracts and write evaluations or whether few firms and author clusters dominate the markets.

To get at these questions we make use of the evaluation author data in our database. As a first step, we identify how many authors have been involved in writing evaluation reports for multiple programmes implemented in different countries. As a second step, we measure the concentration of evaluation markets within MS. These measurements are discussed in detail in Section 3.2.

Table 2: The EU's "single market" for evaluations

Country	Authors	% Two or More MS	Country	Authors	% Two or More MS
AT	73	2.74%	IT	266	3.38%
BE	24	8.33%	LT	23	4.35%
BG	69	0.00%	LU	7	0.00%
CY	-	-	LV	47	4.26%
CZ	171	4.09%	MT	1	0.00%
DE	316	4.11%	NL	73	0.00%
DK	5	0.00%	PL	590	1.86%
EE	95	1.05%	PT	106	0.94%
ES	40	0.00%	RO	180	6.11%
FI	27	3.70%	SE	60	8.33%
FR	72	2.78%	SI	37	10.81%
GR	18	5.56%	SK	52	1.92%
HR	49	6.12%	UK	80	5.00%
HU	95	0.00%	CB	197	20.30%
IE	23	4.35%	Total	2517	3.26%

Notes: The table breaks down by country the share of authors who contributed to at least one evaluation from that country as well as at least one other country. When authors have worked on multiple countries, they are counted in all of these countries. The last row "CB" refers to cross-border and Interreg Europe programmes.

Table 2 presents the share of authors per MS that have contributed as a (co-)author to at least one evaluation report in at least one other MS. It shows that such authors are virtually absent. On aggregate, from 2,517 authors in our sample only 82 or 3.26% have contributed to evaluations in two or more MS. This low number suggests the absence of a single market in the EU for evaluations.

Of course, some programmes require knowledge of local context and language for proper evaluations, but on the other hand, most of the programmes should serve common

European goals and there must be a large element of learning externalities from programme to programme. The almost complete absence of cooperation across MS in writing evaluations is suggestive of the fact that the market of evaluations is very fragmented, and that probably substantial gains in terms of the quality of evaluations can be made in overcoming these barriers across country borders.

Table 3: Concentration of evaluation markets in MS

Country	HHI	CR3	Country	HHI	CR3
FI	1.000	1.000	BE	0.222	0.667
MT	1.000	1.000	GR	0.200	0.600
PT	0.773	1.000	RO	0.194	0.667
SK	0.633	0.929	LV	0.173	0.600
SI	0.630	1.000	NL	0.167	0.556
LU	0.625	1.000	FR	0.163	0.640
PL	0.595	0.879	HU	0.162	0.60
LT	0.556	1.000	BG	0.156	0.600
IT	0.524	0.888	UK	0.136	0.561
EE	0.438	0.875	ES	0.133	0.500
IE	0.388	0.857	AT	0.124	0.500
SE	0.361	0.813	DE	0.114	0.472
DK	0.333	1.000	CZ	0.056	0.300
HR	0.313	0.750			

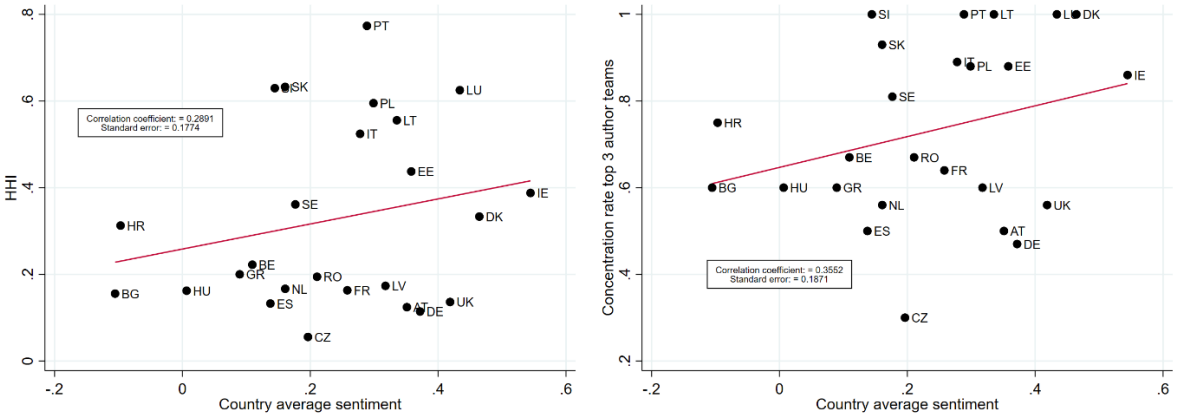
Notes: The table shows the Herfindahl-Hirschmann-index (HHI) and the concentration ratio of the largest 3 evaluation team clusters by country. HHI is calculated according to the formula in Section 3.2.

As to the competitiveness of markets within MS, Table 3 suggests an overall large degree of oligopolistic market structures. On average, the market share of top three author clusters across MS is at a stunning 75%. The leading countries with the least competitive markets are Finland and Malta, which is driven by the very small evaluation markets in these countries restricting participation by a wide group of potential evaluators. However, even looking at the most competitive markets at the bottom of Table 3, we see that the market share of top three clusters is still very large with 30% for the most competitive case, Czechia, and otherwise at about 50% and higher.

In Figure 10 we correlate the average sentiment of evaluations in countries with our measures of market competitiveness. In the left panel for the HHI, and in the right panel for the concentration rate of the top three author clusters (CR3) we find positive coefficients

for these correlations. In the case of CR3 this relationship is statistically significant. This suggests that, on average, evaluations in more oligopolistic markets tend to find more optimistic findings.

Figure 10: Correlation between market concentration and findings of evaluations



Notes: The left panel plots the correlation between Herfindahl-Hirschman-index of market concentration and average sentiment on the country level. In the right panel the correlation between the aggregate market share of the top 3 author teams and average sentiment on country level is depicted. Both concentration measures consider all evaluations for which we identify authors as individuals. Malta, Cyprus and Finland are excluded as we identify only a single author cluster for these countries.

A plausible interpretation of this result is that the few dominant evaluators of oligopolistic markets have strong ties with the managing authorities, which leads the evaluations to follow the interests of managing authorities more closely, and showing a more positive performance of cohesion programmes. This result is also consistent with the argument that lack of competition generally leads to lower quality evaluations (as well as higher prices, as predicted by economic theory) which then affects the direction of the findings of evaluations. Although we do not have direct measurements of the quality of evaluations, it is plausible to assume that the findings of low-quality evaluations are more prone to influences than the ones of high-quality evaluations. Consistent with the interests of the managing authorities, such influences might then lead to the sentiment scores to be skewed towards showing more optimistic findings.

However, formulating a policy conclusion from this exercise is less straightforward. More competition will not necessarily make evaluators more independent from the managing

authorities, since severe competition might lead evaluators to compete for winning contracts by being even less impartial.

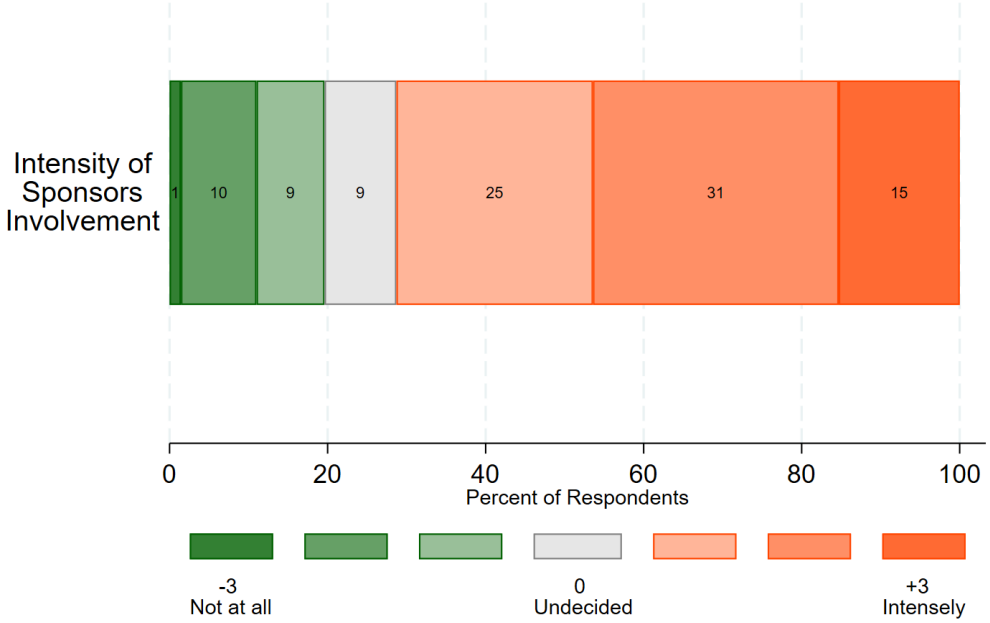
4.4 Impartiality

In this section, we study the question of how the involvement of the managing authority of a Cohesion Policy programme correlates with the sentiment of the resulting evaluations. A common feature is that the national or regional authorities that run cohesion programmes are also the ones that commission, monitor and approve the evaluations (Heinemann et al. 2024). Such an intense involvement of managing authority with the work of (formally independent) evaluators may have both favourable and unfavourable consequences. On the upside, a strong involvement through an intense communication may support the flow of information and the evaluator's understanding for the programme design and impact. On the downside, the involvement may limit the material independence of evaluators and imply pressure on the evaluator to deliver a preferred positive assessment at the expense of a truly impartial evaluation.

To study which of the possible directions dominates, we employ the data we collected from our survey of authors. In the survey we ask the following question: "How intensely are the sponsors of your EU programme evaluations typically involved in discussing your evaluation methods, results and policy conclusions". In their answers, the respondents had the option to choose the degree of involvement according to a seven-point Likert-scale or refrain from answering. We plot the responses to this question in Figure 11. Around 70% of responses indicate at least some involvement by the sponsors of the evaluation, which confirms that the managing authorities are heavily involved in discussing the methods, results and policy conclusions of evaluations.

In Table 4, we test whether the involvement of authorities in the evaluations process correlates with the evaluation sentiment on the performance of programmes. If a strong involvement of the managing authority serves as a positive input for the evaluation process such as by improving the information flow between the authority and the evaluator, we should not observe a systematic correlation of involvement and sentiment. On the other hand, a positive correlation would point in the direction of a bias-promoting effect where the managing authority uses its bargaining powers to steer the evaluation towards a more positive assessment.

Figure 11: Intensity of the involvement of authorities in the process of evaluations



Notes: The question asked in the survey is as follows: "How intensely are the sponsors of your EU programme evaluations typically involved in discussing your evaluation methods, results and policy conclusions?".

Our empirical exercise runs a regression of the sentiment found in the evaluation by the responding authors (or the average of the sentiments across evaluations, if there were more than one) and their response to the question on the degree of involvement of the managing authority in their work. We start with a simple correlation in column 1 and consequently add a number of control variables at the level of authors as well as fixed effects for MS in the consequent columns. The results suggest a robust positive correlation between authorities' involvement in the process and the findings of evaluations. That is, this evidence suggests that more involvement leads to evaluations finding more positive impacts, which is in line with the incentives of the managing authority and consistent with the hypothesis that their involvement leads to biased evaluations. The magnitudes are sizable. On average about 70% of evaluations find positive sentiment, while the cases where the authority is involved are 12-13% more likely to find a positive sentiment compared to cases where the authority is not involved. In Appendix Table C.1 instead of using the average sentiment score across evaluations of the author, we run this regression at the level of evaluations. The

results remain the same, with a noticeable improvement in statistical significance of the estimates likely due to the larger sample size.

Table 4: Involvement by managing authorities in the evaluation process and the findings of evaluations

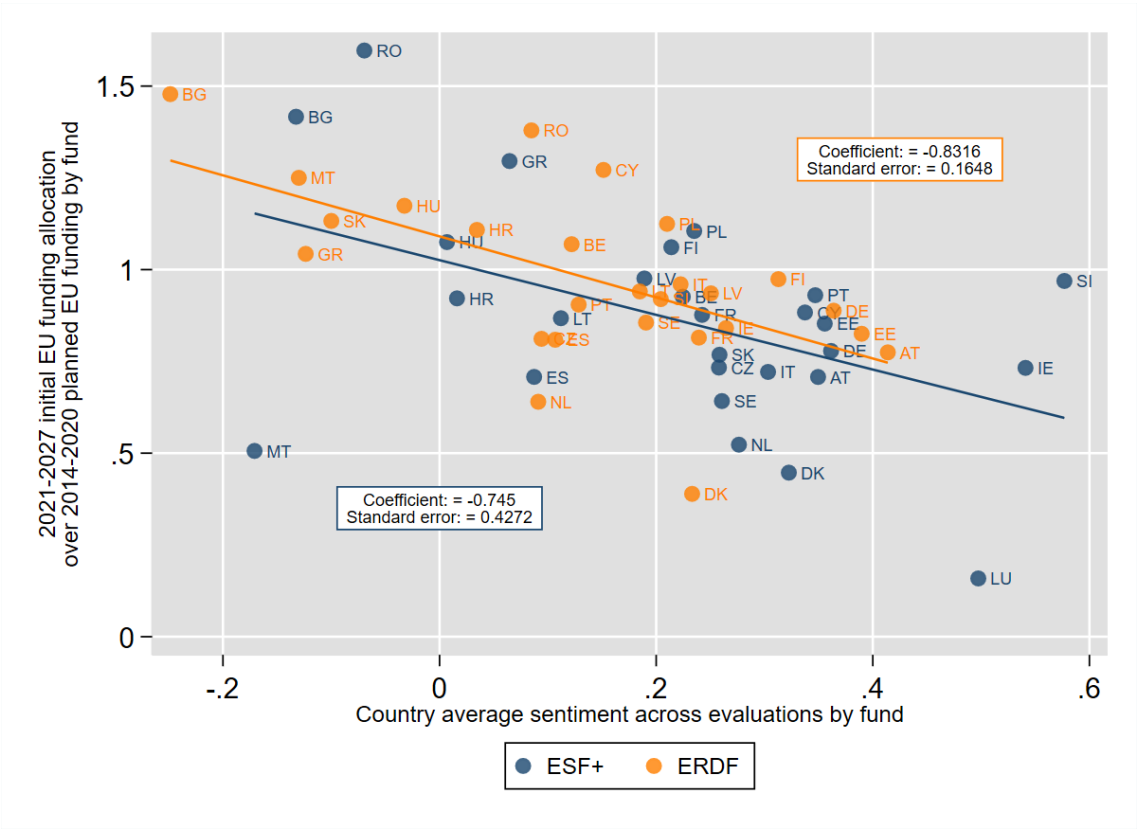
VARIABLES	(1) Positive avg. sentiment	(2) Positive avg. sentiment	(3) Positive avg. sentiment	(4) Positive avg. sentiment	(5) Positive avg. sentiment
At least somewhat intense involvement of sponsor	0.1398**	0.1348**	0.1329*	0.1386*	0.1290*
	(0.0665)	(0.0669)	(0.0677)	(0.0720)	(0.0740)
Evaluations are employers main activity		0.0472	0.0434	0.0062	0.0036
		(0.0608)	(0.0635)	(0.0704)	(0.0713)
University / public institute			-0.0218	-0.0854	-0.0890
			(0.0745)	(0.0850)	(0.0863)
Public sector			-0.0456	-0.0693	-0.0715
			(0.0932)	(0.1047)	(0.1065)
Impartiality is perceived at least somewhat of a bottleneck					-0.0338
					(0.0702)
Woman					0.0342
					(0.0678)
EU sceptic					0.0055
					(0.0991)
Constant	0.6863***	0.6715***	0.6850***	0.7131***	0.7168***
	(0.0568)	(0.0600)	(0.0664)	(0.0733)	(0.0800)
Country FE	No	No	No	Yes	Yes
Observations	189	189	189	189	189
R^2	0.0231	0.0262	0.0277	0.1737	0.1760
F	4.419	2.506	1.311	1.346	0.819

Notes: The table regresses author-level characteristics using data from the survey on the average sentiment score of the evaluations written by the respective author. The sentiment variable is transformed into a dummy variable for positive and non-positive sentiment scores. The main variable of interest, plotted in the first row, is the degree of involvement of managing authorities as measured in the survey and as described in the text in detail. This variable too is transformed into a dummy. Columns 1 to 5 consequently add more control variables. Columns 4 and 5 include fixed effects for the MS.

4.5 Evaluations and Decision Making

We study the question of whether evaluation findings matter for policy making. To do so we correlate the evaluation findings aggregated at the level of MS with the growth of cohesion funding planned to flow to MS in the 2021-27 programming period compared to the 2014-20 period. If evaluations matter for policy making, we would expect to see some relocation in funding across the MS by cutting and expanding the funds in MS with respectively bad and good evaluation results.

Figure 12: Evaluation findings of the past and planned funding amounts in the current budgetary period



Notes: The figure displays correlations between the country level average sentiment across ESF+ and ERDF fund evaluations and the ratio of the amount of funding in the ESF+ and ERDF initially allocated to countries in the 2021-2027 MFF to the amount of funding in these funds disbursed in the 2014-2020 MFF. The sample includes all evaluations that pertain to the ERDF or ESF/ESF+ funds, regardless of whether they also evaluate other funds.

Figure 12 implements the test separately for the ERDF and ESF+[12] funds. It does not find evidence for this hypothesis, if anything it suggests the opposite that is MS with worse average sentiment scores are planned to receive even more money in the future. Figure C.3 of the Appendix replicates the exercise by limiting the sample of evaluations to those that can be precisely mapped to evaluate either of the two funds only, as some evaluations pertain to more than one fund. A similar test would be to look at the regional level within MS, however this is left for future work as data on regional level cohesion funds for the 2021-27 programming period is not yet available.

5. Main Bottlenecks and Reform Options from the Perspective of Evaluators

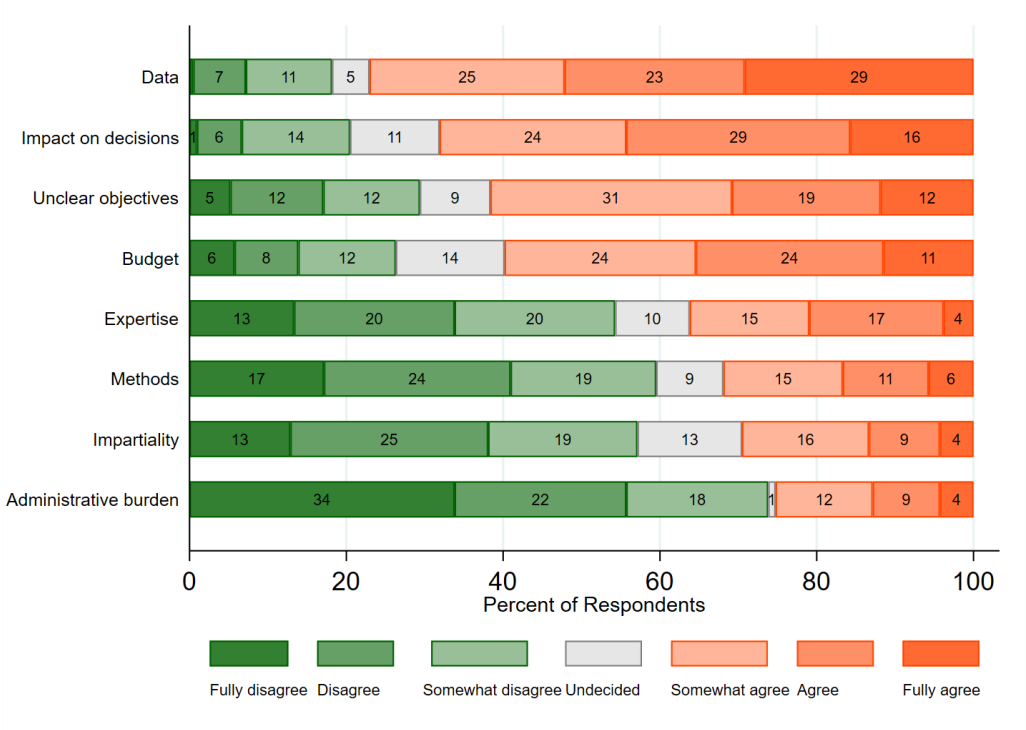
As a final exercise, we describe the responses of authors to a question in our survey on the importance of various bottlenecks implicit in the evaluation system. We show the views of the authors on bottlenecks ranked in their relative importance, along with describing some potential reform options on overcoming these bottlenecks. A much more detailed analysis of reform options is presented by Heinemann et. al. (2024), a paper that also abstracts from the perspective of evaluators which, as we argued in this paper, cannot be considered as fully impartial in the first place. Heinemann et. al. (2024) generally agrees with the main arguments of this paper that the vested interests of managing authorities and the uncompetitive markets for evaluations are significant barriers for high-quality evaluations, but it also makes the more general case for an unfavorable equilibrium characterized by limited evaluation capacities, poor methods, and a formalistic approach to evaluations.

Nevertheless, turning to the bottlenecks as expressed by the evaluators, Figure 12 highlights several important issues. There is a clear consensus among authors that access to data is a very large bottleneck. One policy response to this, a process that is ongoing from the side of EU authorities, is to provide data at high spatial granularity centrally. On methods, although modern and sophisticated methods, such as the use of randomized trials or counterfactual approaches, are important for credible evaluations, there are tradeoffs in imposing a tight methodological corset on all evaluations. Many evaluations cannot be purely quantitative exercises, and even for quantitative exercise a rigid European approach may fail to work, because one-size-fits-all type policies generally do not work well given the heterogeneous circumstances. A related question pertains to the transmission of knowledge generated by evaluations, since even well-measured impacts of a certain program on a specific outcome

[12] With the 2021-2027 programming period, the ESF has been renamed “ESF+”.

cannot always be easily be transferred to other settings. Authors also stress issues related to their capacities for high quality evaluations, as well as often ask for bigger budgets to be made available for their work on evaluations. This latter view is somewhat inconsistent with the view that the evaluation system does not impose significant administrative burden on Cohesion Policy. However, this is hardly surprising, given the respondents' vested interests in keeping the status-quo also related to the fact that for many of them writing the evaluations constitute their core source of revenue. Unfortunately, we neither have data on the direct costs of evaluations, nor on their indirect compliance costs. We suspect, however, that these costs are non-negligible and it would be a task for future research to collect such data, perhaps by starting from the measurement of direct monetary costs based on the procurements of evaluation requests.

Figure 12: Main bottlenecks according to authors of evaluations



Notes: The question asked in the survey is as follows: "Finally, we are interested in potential bottlenecks of the Cohesion Policy evaluation system. Please select for each of the following items whether you agree or disagree that they are a major obstacle to the success of the Cohesion Policy evaluation system."

Turning to the issue of impartiality, the dimension we have analyzed in Section 4.4, we see that authors still rank it an important bottleneck despite authors' potential interest to present themselves as being independent from the authorities. One policy response could be to create an independent body, perhaps a branch of the national auditing authority, which would commission the evaluations instead of the authorities that run the cohesion programmes. Importantly, there is a consensus that cohesion programmes have too many and ever-increasing number of objectives, making the job of evaluation difficult. A reform that simplifies the cohesion objectives, clearly assigns their goals and defines the intermediate indicators that measure the progress on the way to reach them would help the evaluation system become more effective. Heinemann et. al. (2024) views such a broad and imprecise definition of objectives of the policy as a key challenge and develops proposals to overcome it.

Finally, an important aspect is the question of the impact of evaluations on policy. Authors feel that there is a huge disconnect between evaluations and decision-making, a result that is consistent with our empirical evidence linking evaluation results and funding amounts across programming periods. Policy options at the one extreme are to make cohesion policies ex-ante conditional on the results of the evaluations. This is perhaps a too far-reaching reform, given the many bottlenecks in the ability to perform high quality evaluations with very certain outcomes, however the status quo is a policy at the other extreme: Evaluations have nearly zero impact on policy decisions. One plausible policy option is to force authorities to be more accountable by imposing a "comply-or-explain" principle. That is, if authorities do not follow the suggestions of evaluations, they have to explain their decisions publicly. Another even softer approach that implies less of a bureaucratic burden than the latter proposal, is to have better communication between evaluators and policy makers. This last reform option clearly comes with its own set of problems around monitoring and enforcement. These reforms are not only important because they can improve the quality of evaluations, but they have the potential to make Cohesion Policy as a whole and in each MS a better policy. This is because the practical absence of any possibility to impact policy turns the evaluation system into a beauty contest, thus weakening the incentives of putting effort into writing truly independent and high-quality evaluations.

6 Conclusion

In this chapter we use meta-analytical tools to quantitatively analyse about 2,300 evaluations written on Cohesion Policy starting in 2007. We apply an AI-based methodology to quantify the sentiment of Cohesion Policy evaluations with respect to the performance of programmes and show that this new measure ranks results consistently compared to human assessment. Merging the data on evaluations to data on cohesion programmes and their budgets reveals that the evaluations formally cover the cohesion programmes as they are supposed to. This methodological work provides the basis for our work on analysing the evaluation system of Cohesion Policy.

In terms of the results of evaluations, on the aggregate we show that the estimated sentiment scores are heavily skewed towards showing more positive impact of cohesion programmes, as well as towards showing either positive or negative effects rather than null or balanced effects. We uncover large variation in the performance of Cohesion Policy programmes as suggested by the evaluations, and by decomposing the drivers of these differences we find the individual MS but also the authors of evaluations to play a key role.

We compare the MS level scores of the evaluations to country-specific estimates of the growth and employment impacts of Cohesion Policy coming from the academic literature. This comparison shows that the two sources do not provide consistent pictures on the impact of Cohesion Policy. This conclusion is robust when we replicate the analysis on the level of regions as well as for a sub-sample of programmes which have growth and employment as their objective. These findings raise questions on the credibility of evaluations.

We then study several of the potential reasons that may explain the diverging results of the academic literature and the insights from the evaluations. In particular, our analysis suggests that the market of evaluations is rather oligopolistic, that it is very fragmented across the EU MS, and that there is often a strong involvement of managing authorities in the work of (formally independent) evaluators. We show that these strong interference as well as the uncompetitive nature of national evaluation markets correlate with, on average, more optimistic findings in the evaluations.

Finally, the author survey identifies some further key bottlenecks for high-quality and impartial evaluations from the perspective of the authors of the evaluations. These suggest the importance of more technical aspects of evaluations, such as the availability of data or the reliability of methods, which often do not have one-size-fits-all solutions and need more

detailed and context-dependent discussions. Responses of authors also highlight more fundamental challenges to the system, in particular related to the large disconnect between evaluations and decision-making. This disconnect is also consistent with our empirical evidence and it may adversely affect the quality of evaluations by further weakening the incentives to invest resources in writing good evaluations.

Overall, this work lays down the methodological groundwork for further formal analysis of cohesion evaluations, as well as for a more evidence-based understanding on the limits of evaluations and their reform priorities in the EU and other jurisdictions trying to establish systems of performance-based budgeting more generally.

References

Amin, Mostafa M.; Cambria, Erik; Schuller, Björn W. (2023): Will Affective Computing Emerge From Foundation Models and General Artificial Intelligence? A First Evaluation of ChatGPT. In: *IEEE Intell. Syst.* 38 (2), S. 15–23. DOI: 10.1109/MIS.2023.3254179.

Bang, Yejin; Cahyawijaya, Samuel; Lee, Nayeon; Dai, Wenliang; Su, Dan; Wilie, Bryan et al. (2023): A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. Online verfügbar unter https://arxiv.org/pdf/2302.04023.

Canova, Fabio; Pappa, Evi (2021): What are the likely macroeconomic effects of the EU Recovery plan? CEPR Discussion Paper No. DP16669.

Crucitti, Francesca; Lazarou, Nicholas-Joseph; Monfort, Philippe; Salotti, Simone (2022): The RHOMOLO impact assessment of the 2014-2020 cohesion policy in the EU regions. Seville: European Commission, Joint Research Centre (JRC) (JRC Working Papers on Territorial Modelling and Analysis, 01/2022). Available online at https://www.econstor.eu/handle/10419/265238.

Darvas, Zsolt; Mazza, Jan; Midões (2019): How to improve European Union cohesion policy for the next decade, Bruegel Policy Contribution, 8/May. Available online at https://www.bruegel.org/policy-brief/how-improve-european-union-cohesion-policy-next-decade, accessed on 19.06.2023.

Di Caro, Paolo; Fratesi, Ugo (2022): One policy, different effects: Estimating the region-specific impacts of EU cohesion policy. In: *Journal of Regional Science* 62 (1), S. 307–330. DOI: 10.1111/jors.12566.

Doucouliagos, H., & Paldam, M. (2009). The aid effectiveness literature: The sad results of 40 years of research. *Journal of economic surveys*, *23*(3), 433-461.

Downes, Ronnie; Moretti, Delphine; Nicol, Scherie (2017): Budgeting and performance in the European Union. In: *OECD Journal on Budgeting* 17 (1), S. 1–60. DOI: 10.1787/budget-17-5jfnx7fj38r2.

European Commission (2013): The Programming Period 2014-2020: Guidance document on monitoring and evaluation - European Regional Development Fund and Cohesion Fund. European Commission. Brussels. Available online at https://ec.europa.eu/regional_policy/sources/evaluation/2014/wd_2014_en.pdf.

European Commission (2021): Performance, monitoring and evaluation of the European Regional Development Fund, the Cohesion Fund and the Just Transition Fund in 2021-2027, Commission Staff Working Document SWD(2021) 198 final.

European Union (2006): Council Regulation (EC) No 1083/2006 of 11 July 2006 laying down general provisions on the European Regional Development Fund, the European Social Fund and the Cohesion Fund and repealing Regulation (EC) No 1260/1999. In: *Official Journal of the EU*. Available online at https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32006R1083.

European Union (2013): Regulation (EU) No 1303/2013 of the European Parliament and of the Council of 17 December 2013 laying down common provisions on the European Regional Development Fund, the European Social Fund, the Cohesion Fund, the European Agricultural Fund for Rural Development and the European Maritime and Fisheries Fund and laying down general provisions on the European Regional Development Fund, the European Social Fund, the Cohesion Fund and the European Maritime and Fisheries Fund and repealing Council Regulation (EC) No 1083/2006, OJ L. In: *Official Journal of the EU*. Available online at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32013R1303.

European Union (2021): Regulation (EU) 2021/1060 of the European Parliament and of the Council of 24 June 2021 laying down common provisions on the European Regional Development Fund, the European Social Fund Plus, the Cohesion Fund, the Just Transition Fund and the European Maritime, Fisheries and Aquaculture Fund and financial rules for those and for the Asylum, Migration and Integration Fund, the Internal Security Fund and the Instrument for Financial Support for Border Management and Visa Policy. In: *Official Journal of the EU*. Available online at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32021R1060.

Fidrmuc, Jan; Hulényi, Martin; Zajkowska, Olga (2019): The Elusive Quest for the Holy Grail of an Impact of EU Funds on Regional Growth. CESifo Working Paper, No. 7989. Ifo Institute – Leibniz Institute for Economic Research at the University of Munich. Available online at http://hdl.handle.net/10419/214991.

Gilardi, Fabrizio; Alizadeh, Meysam; Kubli, Maël (2023): ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. Available online at https://arxiv.org/pdf/2303.15056.

Heinemann, Friedrich, Zareh Asatryan, Julia Bachtrögler-Unger, Carlo Birkholz, Franceso Corti, Maximilian von Ehrlich, Ugo Fratesi, Clemens Fuest, Valentin Lang and Martin Weber (2024): Enhancing Objectivity and Decision Relevance: A Better Framework for Evaluating Cohesion Policies.

Hirsch, A. V. (2016). Experimentation and persuasion in political organizations. *American Political Science Review*, *110*(1), 68-84.

Kocoń, Jan; Cichecki, Igor; Kaszyca, Oliwier; Kochanek, Mateusz; Szydło, Dominika; Baran, Joanna et al. (2023): ChatGPT: Jack of all trades, master of none. In: *Information Fusion* 99, S. 101861. DOI: 10.1016/j.inffus.2023.101861.

Korinek, A. (2023). Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature*, *61*(4), 1281-1317.

Mosley, P. (1986). Aid-effectiveness: The micro-macro paradox. *Ids Bulletin*, *17*(2), 22-27.

Naldini, Andrea (2018): Improvements and risks of the proposed evaluation of Cohesion Policy in the 2021–27 period: A personal reflection to open a debate. In: *Evaluation* 24 (4), S. 496–504. DOI: 10.1177/1356389018804261.

OpenAI (2023): Create completion. Available online at https://platform.openai.com/docs/api-reference/completions.

Pellegrin, Julie; Colnot, Louis; Pedralli, Matteo (2020): The Role of Evaluation in Cohesion Policy, Study Requested by the REGI Committee.

Wang, Zengzhi; Xie, Qiming; Ding, Zixiang; Feng, Yi; Xia, Rui (2023): Is ChatGPT a Good Sentiment Analyzer? A Preliminary Study. Available online at https://arxiv.org/pdf/2304.04339.

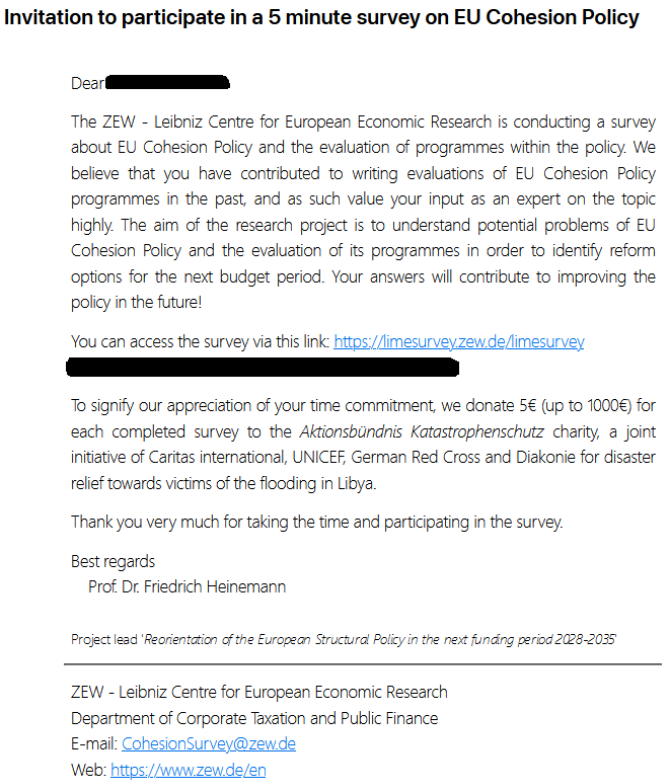
Wang, S., & Yang, D. Y. (2021). Policy experimentation in China: The political economy of policy learning. *National Bureau of Economic Research No. w29402*.

Zhong, Qihuang; Ding, Liang; Liu, Juhua; Du Bo; Tao, Dacheng (2023): Can ChatGPT Understand Too? A Comparative Study on ChatGPT and Fine-tuned BERT. Available online at https://arxiv.org/pdf/2302.10198.

APPENDIX

A Survey design

Figure A.1: Survey invitation email



Invitation to participate in a 5 minute survey on EU Cohesion Policy

Dear [redacted]

The ZEW - Leibniz Centre for European Economic Research is conducting a survey about EU Cohesion Policy and the evaluation of programmes within the policy. We believe that you have contributed to writing evaluations of EU Cohesion Policy programmes in the past, and as such value your input as an expert on the topic highly. The aim of the research project is to understand potential problems of EU Cohesion Policy and the evaluation of its programmes in order to identify reform options for the next budget period. Your answers will contribute to improving the policy in the future!

You can access the survey via this link: https://limesurvey.zew.de/limesurvey [redacted]

To signify our appreciation of your time commitment, we donate 5€ (up to 1000€) for each completed survey to the *Aktionsbündnis Katastrophenschutz* charity, a joint initiative of Caritas international, UNICEF, German Red Cross and Diakonie for disaster relief towards victims of the flooding in Libya.

Thank you very much for taking the time and participating in the survey.

Best regards
Prof. Dr. Friedrich Heinemann

Project lead '*Reorientation of the European Structural Policy in the next funding period 2028-2035*'

ZEW - Leibniz Centre for European Economic Research
Department of Corporate Taxation and Public Finance
E-mail: CohesionSurvey@zew.de
Web: https://www.zew.de/en

Figure A.2: Survey introduction

EU Cohesion Policy Survey

Dear [redacted]

Thank you very much for agreeing to participate in this survey, it should take about **5 minutes** of your time. Your answers serve as an important input for a scientific research project that analyses how programmes of the EU Cohesion Policy are evaluated, in order to improve the evaluation system in future funding periods. To signify our appreciation of your time commitment, we donate 5€ (up to 1000€) for each completed survey to the *Aktionsbündnis Katastrophenschutz* charity, a joint initiative of Caritas international, UNICEF, German Red Cross and Diakonie for disaster relief towards victims of the flooding in Libya.

ZEW is committed to comply with the EU General Data Protection Regulations, and as such you have the right to access (article 15 GDPR), rectification (article 16 GDPR) and erasure (article 17 GDPR) of your data. We process your data for scientific purposes only, in accordance with article 6 section 1 (f) GDPR. The survey is conducted as part of the project '*Reorientation of the European Structural Policy in the next funding period 2028-2035*'.

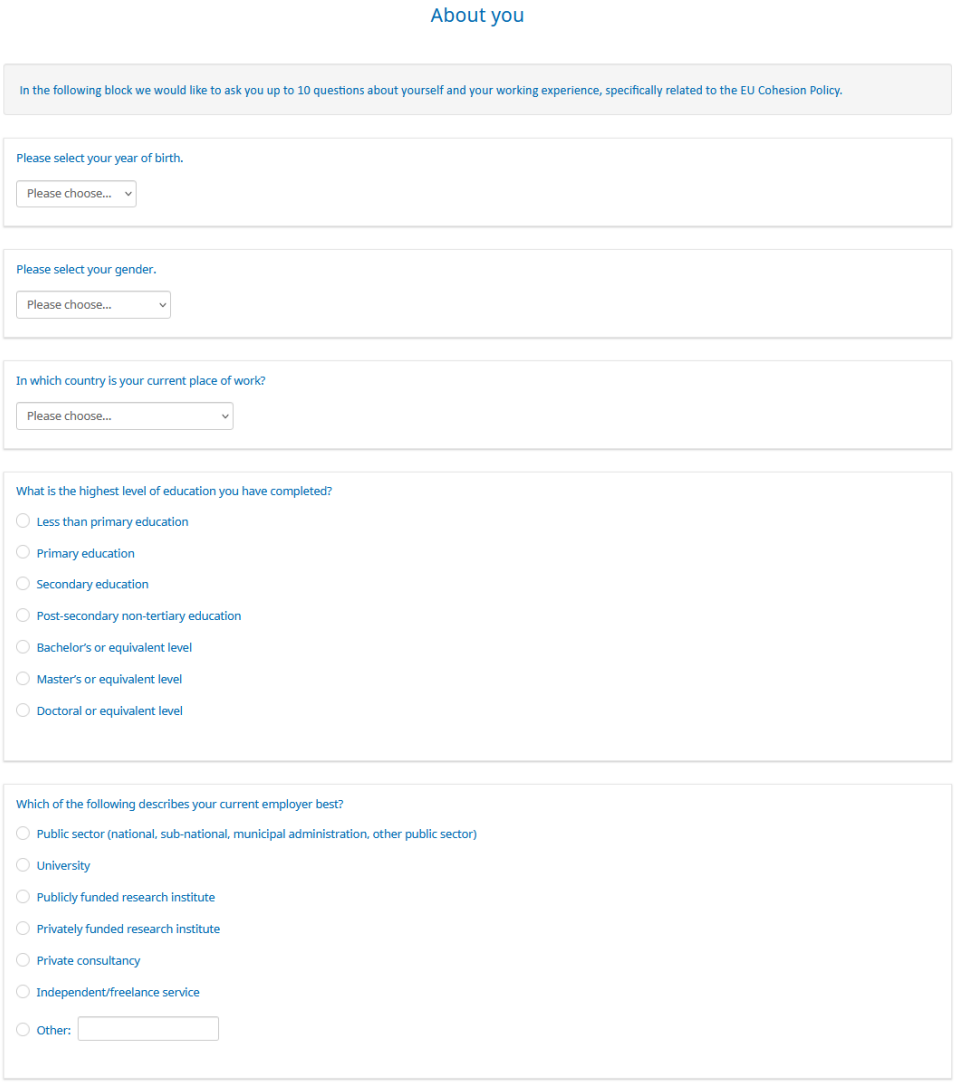
Contact person for project-specific enquiries is Prof. Friedrich Heinemann who can be reached via email at CohesionSurvey@zew.de.

The data protection officer Dr. Thomas Wirth can be reached via email at datenschutzbeauftragter@zew.de. Your data will be archived for 10 years to comply with ZEWs guidelines for good academic practice. For more information about ZEWs data protection commitment visit: https://www.zew.de/en/commitment-to-data-protection.

There are 16 questions in this survey.

Next

Figure A.3: Survey questionnaire



In what function are you working at your current employer?

- Researcher
- Project lead
- Management level
- Other:

Did you work for a different employer than your current when you contributed to Cohesion Policy evaluation reports?

If you wrote evaluation reports under your current and a former employer, please select "yes" only if the majority of evaluation reports you wrote were under a former employer.

Yes | No

Which of the following describes your former employer best?

- Public sector (national, sub-national, municipal administration, other public sector)
- University
- Publicly funded research institute
- Privately funded research institute
- Private consultancy
- Independent/freelance service
- Other:

In which country was your place of work under your former employer?

Please choose...

From your perception, were Cohesion Policy programme evaluations the main field of activity for your former employer?

Yes | No

Next

Views on the EU and economic policy

In this second and final block we would like to ask you 4 questions about your general views on the EU and your views on the Cohesion Policy evaluation system.

GENERAL VIEWS ON THE EU AND ECONOMIC POLICY

Please select the option that reflects your view the best:

	Fully disagree -3	-2	-1	Undecided 0	+1	+2	Fully agree +3	No answer
The EU budget should grow substantially in order to cope with Europe's challenges.	○	○	○	○	○	○	○	●
There should be more redistribution from richer to poorer EU Member States.	○	○	○	○	○	○	○	●
EU competencies on fields such as taxation, social standards, and labor market regulation should be extended.	○	○	○	○	○	○	○	●
The government (EU and Member States) should take a more active role in European industrial policies.	○	○	○	○	○	○	○	●

VIEWS ON COHESION POLICY EVALUATION SYSTEM

Please select the option that reflects your view the best:

	Not at all -3	-2	-1	Undecided 0	+1	+2	Intensely +3	No answer
How intensely are the sponsors of your EU programme evaluations typically involved in discussing your evaluation methods, results and policy conclusions?	○	○	○	○	○	○	○	●

Do you have any suggestions how to improve the evaluation system of the EU Cohesion Policy programmes?

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

	Fully disagree -3	-2	-1	Undecided 0	+1	+2	Fully agree +3	No answer
Unclear objectives to be evaluated against	○	○	○	○	○	○	○	●
No impact of evaluations on decision making	○	○	○	○	○	○	○	●
Lack of a budget for evaluations	○	○	○	○	○	○	○	●
Lack of appropriate methods	○	○	○	○	○	○	○	●
Evaluations create unnecessary administrative burden	○	○	○	○	○	○	○	●
Lack of impartial evaluations	○	○	○	○	○	○	○	●
Lack of expertise and capacities	○	○	○	○	○	○	○	●
Lack of good data	○	○	○	○	○	○	○	●

Would you like to receive a digital copy of the final report once it is published?

✔ Yes ⊘ No

Table A.1: Balance test - survey respondents versus all authors

	Control			Treatment			
	N	mean	sd	N	mean	sd	Diff
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

Notes: Observations are at the author-evaluation level. The Diff column is the coefficient of a simple regression of surveyed status on the variable, with clustered standard errors at the author level. Stars indicate whether this difference is significant. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.2: Balance test - survey respondents versus all contacted authors

	Control			Treatment			
	N	mean	sd	N	mean	sd	Diff
Average Sentiment	1429	0.30	0.45	626	0.29	0.46	-0.008
Evaluation has abstract	1532	0.93	0.25	677	0.93	0.26	-0.008
Fund: ERDF	1532	0.58	0.49	677	0.58	0.49	-0.004
Fund: CF	1532	0.11	0.31	677	0.08	0.27	-0.031
Fund: ESF	1532	0.60	0.49	677	0.56	0.50	-0.035
Fund: YEI	1532	0.11	0.31	677	0.07	0.26	-0.033*
Type: Impact	1532	0.49	0.50	677	0.43	0.49	-0.063**
Type: Process	1532	0.55	0.50	677	0.60	0.49	0.056**
Type: Monitoring	1532	0.58	0.49	677	0.62	0.49	0.038
Type: Summary	1532	0.04	0.19	677	0.02	0.15	-0.015**
Type: Report	1532	0.04	0.18	677	0.05	0.22	0.016
MFF 2007-2013	1532	0.20	0.40	677	0.15	0.36	-0.047**
MFF 2014-2020	1532	0.84	0.37	677	0.87	0.34	0.030
Total Programme Budget (in billion €)	1531	2.30	4.49	676	1.93	4.12	-0.372
Estimated Co-financing Rate	1107	0.29	0.16	514	0.32	0.16	0.027
Thematic Objective: 1	1532	0.34	0.47	677	0.42	0.49	0.077**
Thematic Objective: 2	1532	0.22	0.42	677	0.22	0.41	-0.004
Thematic Objective: 3	1532	0.30	0.46	677	0.36	0.48	0.056*
Thematic Objective: 4	1532	0.29	0.45	677	0.27	0.45	-0.016
Thematic Objective: 5	1532	0.19	0.39	677	0.20	0.40	0.006
Thematic Objective: 6	1532	0.25	0.43	677	0.27	0.45	0.022
Thematic Objective: 7	1532	0.24	0.43	677	0.22	0.41	-0.022
Thematic Objective: 8	1532	0.47	0.50	677	0.49	0.50	0.024
Thematic Objective: 9	1532	0.47	0.50	677	0.44	0.50	-0.031
Thematic Objective: 10	1532	0.43	0.50	677	0.42	0.49	-0.015
Thematic Objective: 11	1532	0.25	0.43	677	0.26	0.44	0.015
Thematic Objective: mutliple	1532	0.36	0.48	677	0.39	0.49	0.035
Thematic Objective: all	1532	0.17	0.37	677	0.17	0.38	0.006
Method: Theory-based Impact Evaluation	1532	0.18	0.39	677	0.20	0.40	0.014
Method: Qualitative Analysis	1532	0.91	0.28	677	0.90	0.30	-0.013
Method: Quantitative Analysis	1532	0.88	0.32	677	0.84	0.36	-0.038*
Method: Cost-benefit Analysis	1532	0.04	0.19	677	0.03	0.16	-0.013
Method: Counterfactual Impact Evaluation	1532	0.18	0.39	677	0.15	0.36	-0.031
Method: Mod?	1532	0.06	0.23	677	0.04	0.20	-0.016

Notes: Observations are at the author-evaluation level. The Diff column is the coefficient of a simple regression of surveyed status on the variable, with clustered standard errors at the author level. Stars indicate whether this difference is significant. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.3: Balance test - survey respondents versus 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**

Notes: Observations are at the author level. The Diff column is the coefficient of a simple regression of surveyed status on the variable, with clustered standard errors at the author level. Stars indicate whether this difference is significant. * p < 0.10, ** p < 0.05, *** p < 0.01.

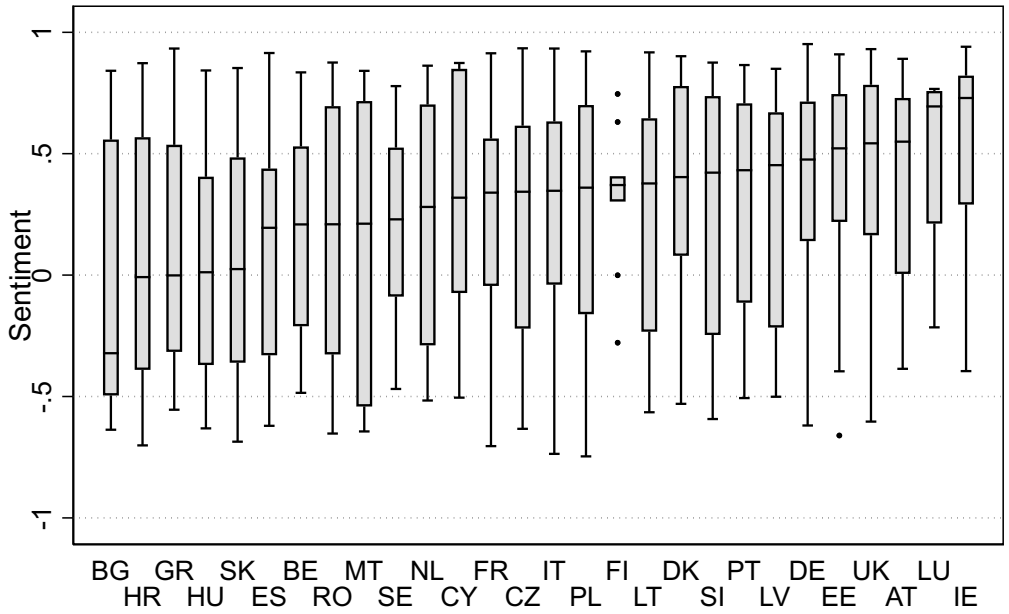
Table A.4: Balance test - survey respondents versus all contacted authors

	Control			Treatment			
	N	mean	sd	N	mean	sd	Diff
Average sentiment	682	0.31	0.40	219	0.29	0.37	-0.018
Number of evaluations	717	2.14	2.87	227	2.98	3.53	0.846***
University affiliated?	717	0.12	0.33	227	0.13	0.33	0.006

Notes: Observations are at the author level. The Diff column is the coefficient of a simple regression of surveyed status on the variable, with clustered standard errors at the author level. Stars indicate whether this difference is significant. * p < 0.10, ** p < 0.05, *** p < 0.01.

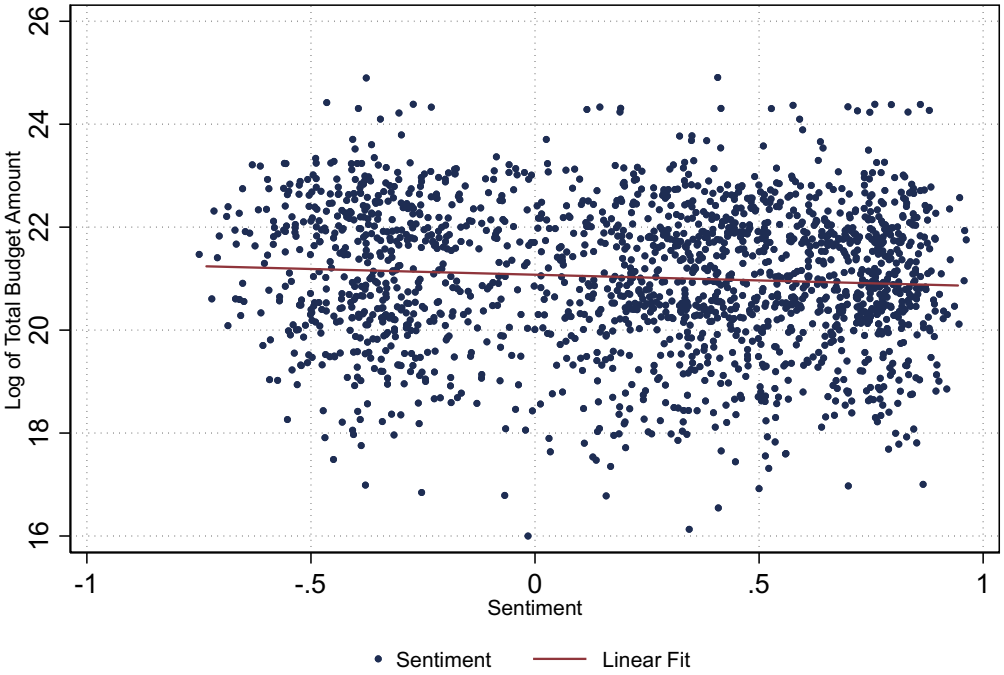
B Additional Results on Evaluation Sentiment

Figure B.1: Distribution of evaluation result by MS



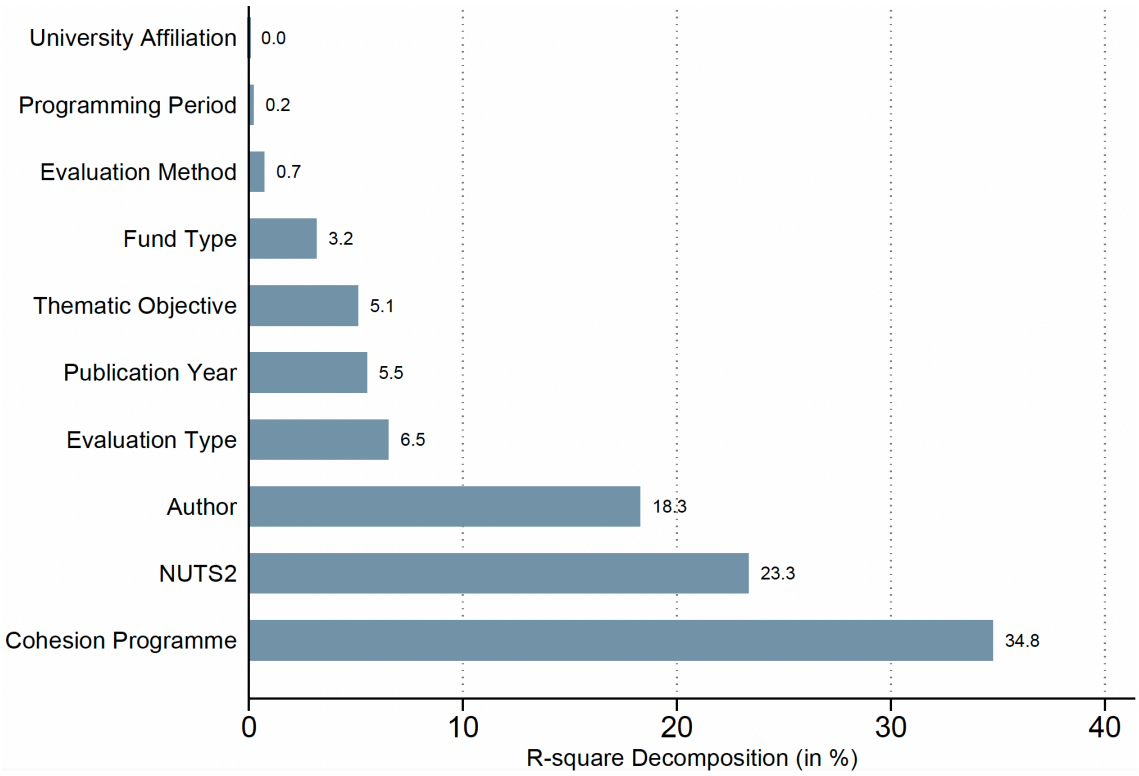
Notes: Number of observations per country is: Bulgaria (BG): 25, Hungary (HU): 75, Malta (MT): 7, Croatia (HR): 22, Greece (GR): 62, Slovakia (SK): 46, Spain (ES): 126, Romania (RO): 76, Belgium (BE): 20, Sweden (SE): 32, Netherlands (NL): 24, Czech Republic (CZ): 183, Lithuania (LT): 52, France (FR): 124, Poland (PL): 468, Italy (IT): 376, Latvia (LV): 24, Cyprus (CY): 11, Slovenia (SI): 15, Portugal (PT): 29, Finland (FI): 9, Germany (DE): 267, Denmark (DK): 17, Austria (AT): 34, Estonia (EE): 24, United Kingdom (UK): 83, Luxembourg (LU): 4, Ireland (IE): 24. Number of countries: 28. Total number of observations: 2,259.

Figure B.2: Average evaluation result versus size of projects



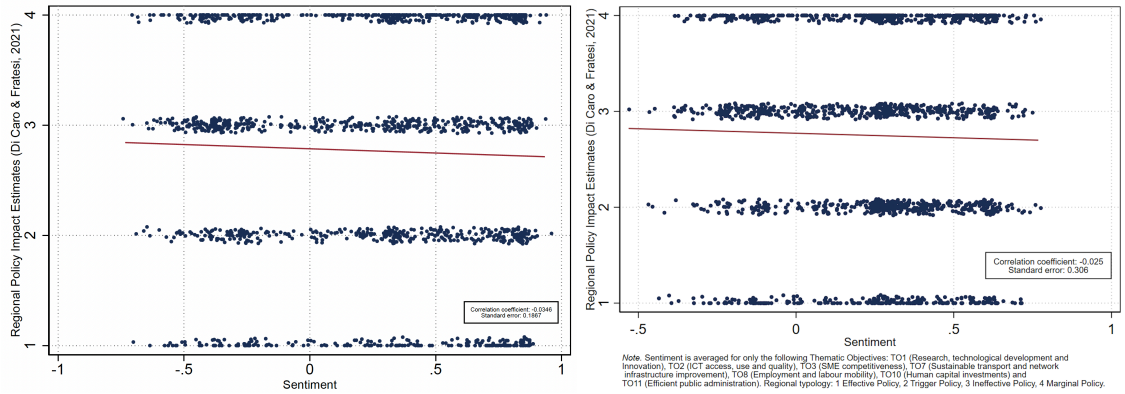
Notes: Figure correlates the monetary budget of programmes (EU funds and national co-financing) in logs with average sentiment for the programme. The number of observations is 1,881.

Figure B.3: Explaining the variation in evaluation findings



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. This figure is similar to Figure 7, with the exception that we plot NUTS2 fixed effects instead of MS fixed effects.

Figure B.4: Comparison of NUTS2-specific evaluation findings (evaluations with all TOs and only evaluations with growth-friendly TOs) with the effects of Cohesion Policy as estimated by the economic literature



Notes: This figure is similar to Figure 9 but performed at the NUTS2, rather than MS, level. Source of the NUTS2 level Cohesion Policy impact estimates is Di Caro und Fratesi (2022). The left sub-figure uses the whole sample of evaluations, while the sub-figure on the right restricts the sample of evaluations only to those which have growth friendly Thematic Objectives according to our classification.

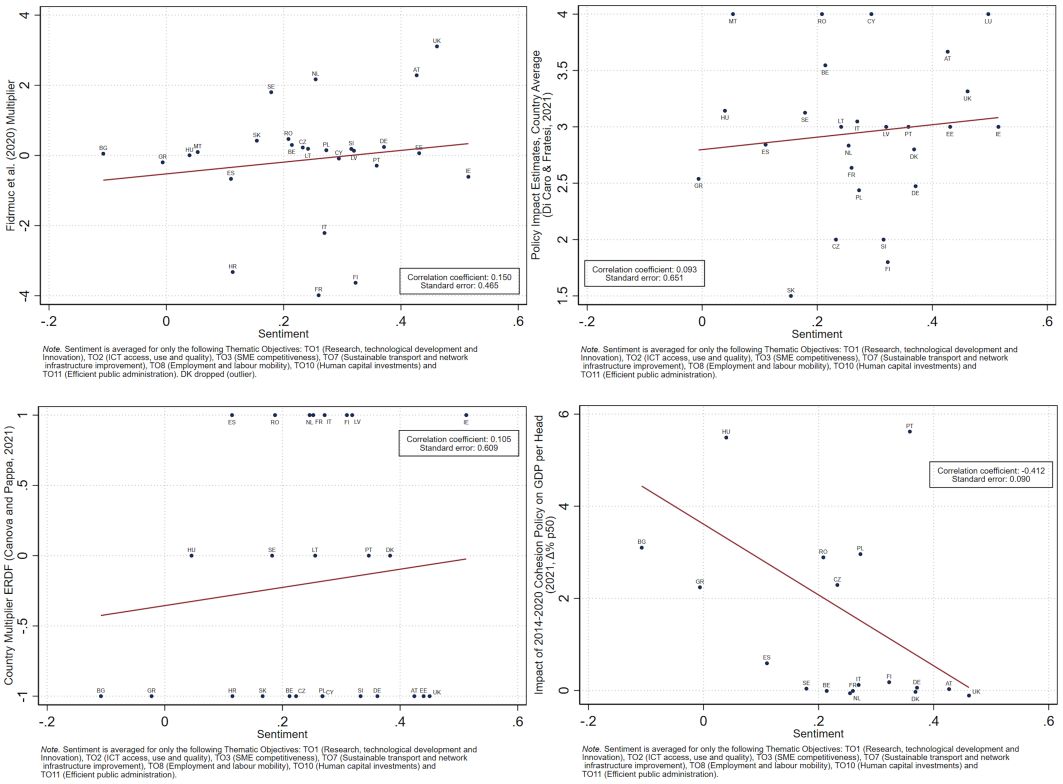
C Further robustness checks

Table C.1: Involvement by managing authorities in the evaluation process and the findings of evaluations (alternative specification)

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	0.1044**	0.1027**	0.1033**	0.1284***	0.1209**
	(0.0444)	(0.0443)	(0.0453)	(0.0490)	(0.0496)
Evaluations are employers main activity		-0.0683*	-0.0693*	-0.0305	-0.0297
		(0.0371)	(0.0390)	(0.0460)	(0.0474)
University / public institute			-0.0043	0.0246	0.0300
			(0.0502)	(0.0594)	(0.0601)
Public sector			-0.0004	0.0454	0.0428
			(0.0800)	(0.0894)	(0.0898)
Impartiality is perceived at least somewhat of a bottleneck					-0.0453
					(0.0465)
Woman					0.0555
					(0.0438)
EU sceptic					-0.0481
					(0.0621)
Constant	0.6159***	0.6521***	0.6530***	0.6055***	0.6100***
	(0.0390)	(0.0436)	(0.0460)	(0.0538)	(0.0586)
Country FE	No	No	No	Yes	Yes
Observations	610	610	610	610	610
R^2	0.0090	0.0145	0.0146	0.0518	0.0565
F	5.540	4.478	2.234	2.181	1.657

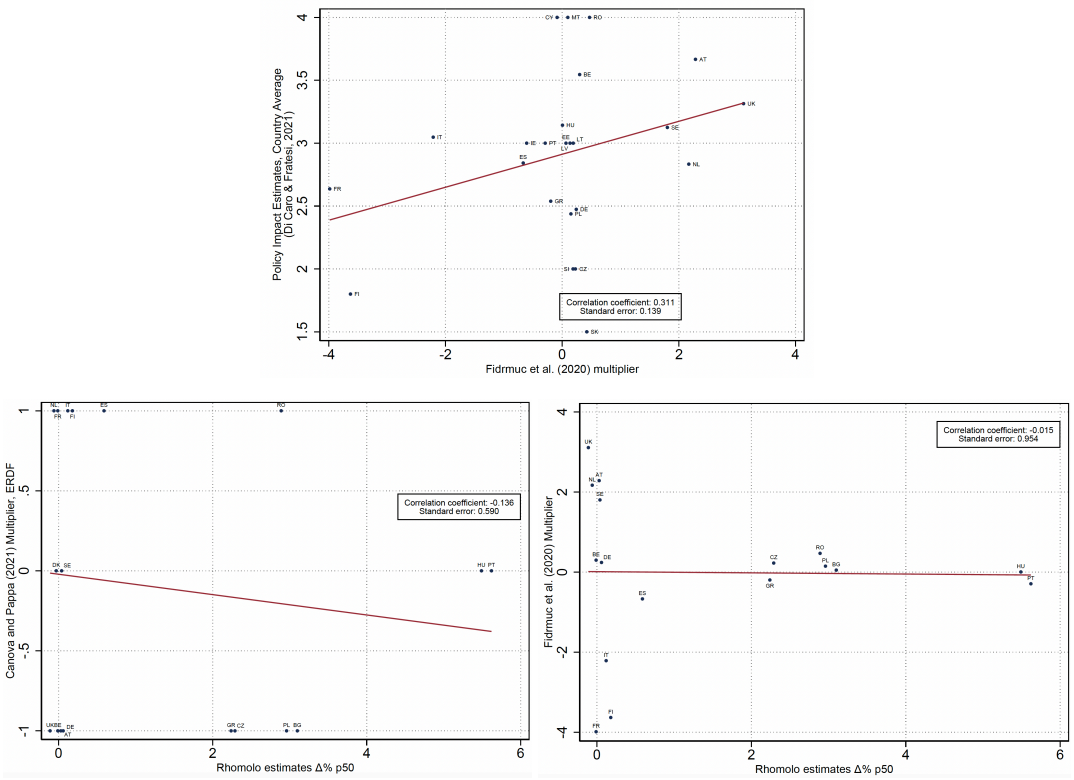
Notes: The table regresses author-level characteristics using data from the survey on the sentiment score of each evaluation written by the respective author. The sentiment variable is transformed into a dummy variable for positive and non-positive sentiment scores. The main variable of interest, plotted in the first row, is the degree of involvement of managing authorities as measured in the survey and as described in the text in detail. This variable too is transformed into a dummy. Columns 1 to 5 consequently add more control variables. Columns 4 and 5 include fixed effects for the MS.

Figure C.1: Comparison of MS specific sentiments from evaluations targeting growth friendly Thematic Objectives with the output-impacts of Cohesion Policy as estimated by the economic literature



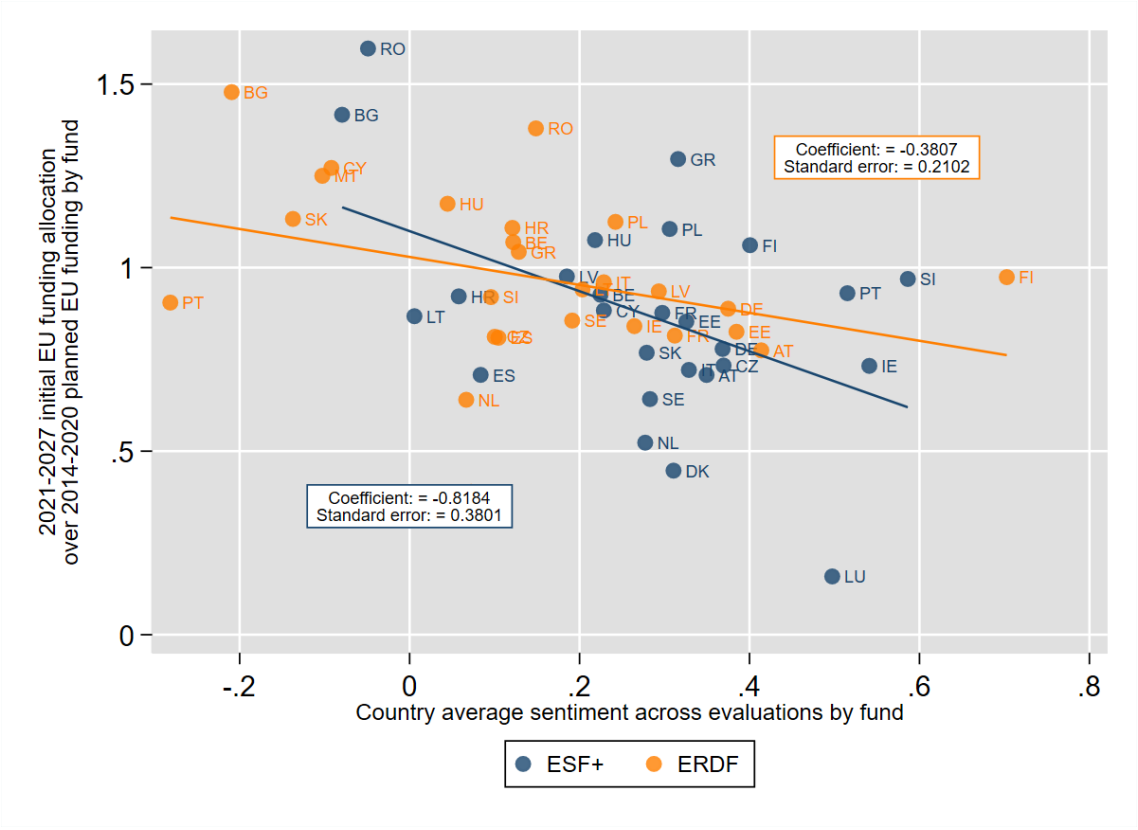
Notes: This figure is similar to Figure 9 but restricts the sample of evaluations only to those which have growth friendly Thematic Objectives according to our classification.

Figure C.2: Output-impacts of Cohesion Policy as estimated by several sources in the economic literature



Notes: This figure is similar to Figure 9 but correlates the findings of the economic literature with each other, rather than against the sentiment of evaluations.

Figure C.3: Output-impacts of Cohesion Policy as estimated by several sources in the economic literature



Notes: The figure displays correlations between the country level average sentiment across ESF+ and ERDF fund evaluations and the ratio of the amount of funding in the ESF+ and ERDF initially allocated to countries in the 2021-2027 MFF to the amount of funding in these funds disbursed in the 2014-2020 MFF. The sample includes evaluations that pertain to the ERDF or ESF+ funds only.

Download ZEW Discussion Papers:

https://www.zew.de/en/publications/zew-discussion-papers

or see:

https://www.ssrn.com/link/ZEW-Ctr-Euro-Econ-Research.html
https://ideas.repec.org/s/zbw/zewdip.html

IMPRINT

ZEW – Leibniz-Zentrum für Europäische Wirtschaftsforschung GmbH Mannheim
ZEW – Leibniz Centre for European Economic Research

L 7,1 · 68161 Mannheim · Germany
Phone +49 621 1235-01
info@zew.de · zew.de