Can we detect bias in political fact-checking?

Evidence from a Spanish case study

Alejandro Fernández-Roldán, Carlos Elías,
Carlos Santiago-Caballero, David Teira

Forthcoming in *Journalism Practice*

https://www.tandfonline.com/doi/full/10.1080/17512786.2023.2262444

Abstract

Political fact-checkers evaluate the truthfulness of politicians' claims. This paper contributes to an emerging scholarly debate on whether fact-checkers treat political parties differently in a systematic manner depending on their ideology (bias). We first examine the available approaches to analyze bias and then present a new approach in two steps. First, we propose a logistic regression model to analyze the outcomes of fact-checks and calculate how likely each political party will obtain a truth score. We test our model with a sample of fact-checks from *Newtral,* a major Spanish fact-checker. Our model would signal bias under two assumptions: a) all political parties are on average equally accurate in their statements; b) the verification method gives precise instructions and is implemented systematically. We investigate this second assumption with a series of interviews with *Newtral* fact-checkers. We show that standard verification protocols are so loosely implemented that fact-checks reflect a set of journalistic decisions, rather than a bias in the statistical sense. We call for a more rigorous definition of verification methods as a pre-requisite for an unbiased assessment of politician's claims.

Keywords

Fact-checking, political parties, bias, noise, impartiality, public opinion.

Introduction

Political misinformation is pervasive and has been for decades, but its saliency in gaining widespread research interest is far more recent. To be sure, concerns have peaked in the last years as a result of how fast false information can spread online (Vosoughi et al., 2018), especially through social media (Shu et al., 2017). While much of the research on misinformation was initially aimed at conceptualizing *fake news* and explaining their dissemination (Weeks & Gil de Zuñiga, 2021), there is now a growing scholarship on the interventions that may help countering it, ranging from technological tweaks (e.g., flagging/removing content on digital platforms), to more 'traditional' approaches like *fact-checking* (Graves & Amazeen, 2019). Fact-checkers (FCs) analyze the factual accuracy of statements made by elites and institutions (Walter et al., 2020). Political fact-checking, our topic in this paper, has emerged as a standalone activity: fact-checkers single out claims made by politicians, gather evidence and they usually score the truthfulness and publish the verdict on websites and social media (Amazeen et al., 2018).

What is nowadays a global movement, formally started in 2003 with the launch of Factcheck.org, and gradually accelerated through the 2010s (Carr, 2012). In 2014 there were 44 active fact-checkers (Adair, 2014), while there were 378 active fact-checking outlets at the end of 2021 (Stencel, Ryan & Luther, 2022). Although the fact-checking movement is not monolithic (Graves, 2018), the roots of fact-checking are clearly grounded in journalistic practice (Graves et al., 2016), and influential outlets such as *The New York* Times, The *Washington Post*, and the *BBC* have launched their PFC arms to better inform their audiences. Moreover, fact-checking has been assumed by journalists as a signal of status (Graves & Cherubini, 2016), and what at the onset was a task confined to electoral periods has now become a full-time journalistic practice (Nieminen & Rapeli, 2019). Some fact-checking initiatives have likewise won prestigious awards, including the Pulitzer Prize[1].

While the fact-checking industry has experienced spectacular growth in recent years, evidence about the effects of verification[2] is still limited (Marietta et al., 2015). The main goal of fact-checking outlets is both holding politicians accountable (Amazeen, 2013) and promoting accuracy in public discourse (Humprecht, 2020). Yet, fact-checkers have frequently been criticized (Young et al., 2018) with accusations of political bias (better treatment of liberal/left-wing parties) (Gottfried et al., 2013). Fact-checkers have become themselves an object of academic investigation: the rationales for fact-checking (e.g., Graves & Cherubini, 2016), their organizational features (e.g., Graves, 2018), effectiveness (e.g., Walter, 2020), methodology (e.g., Uscinski and Butler, 2013), transparency (e.g., Humprecht, 2020), consistency (e.g., Lim, 2018) or differential treatment to political parties (e.g., Farnsworth & Lichter, 2019).

Our paper focuses on these last two points. Influential fact-checkers use protocols standardized according to the guidelines issued by an external audit institute like the International Fact-Checking Network (Mena, 2019). Fact-checks should be, to a certain extent, reproducible[3]: the same claim should receive a similar verdict when verified by different fact-checkers. And the score should not be different depending on the preferences (e.g., ideology) of the fact-checker. We present here a statistical model that could detect potential biases in the scores awarded by fact-checkers, and discuss under which conditions the model could be effective. We will first review the available literature on measuring consistency and bias in political fact-checking. We present then our own definition of bias and the statistical model to detect it. We test it with a sample of fact-checks from a Spanish fact-checker, *Newtral*. Our model would signal bias under two assumptions: a) all political parties are equally accurate; b) the fact-checking protocol is unambiguously defined and implemented systematically. We investigate this second assumption with a series of interviews with *Newtral* fact-checkers. We show that standard verification protocols are so loosely implemented that the verdicts of fact-checks reflect a random set of journalistic decisions, rather than a bias in the statistical sense. We call for a more rigorous definition of PFC methods as a pre-requisite for an unbiased assessment of politician's claims.

Literature review

Countering misinformation is a difficult task that remains only partly understood (Ecker et al., 2022), and research on debunking[4] strategies shows that efficacy is conditional on specific settings. Fact-checking seems to be most effective when people are provided with an alternative feasible explanation to an initial misperception (Chan et al., 2017), as simply encouraging people to reflect on a held belief can in fact reinforce the misperception (Lewandowsky et al., 2012). Moreover, timing also matters (Brashier et al., 2021) and so does format (Ecker et al., 2020; Amazeen et al., 2018; Young et al., 2018) and tone (Young et al., *ibid*). It is also important to consider the possible ephemeral effects of fact-checks (Carey et al., 2022) and the type of information addressed, i.e., whether it is politically laden or not. To be sure, evidence on journalistic-formatted fact-checks is inconclusive, as fact-checks on political issues could have from moderate to no effects (Amazeen et al., 2018, Thorson, 2016), or some effect on factual accuracy but not enough to change people's candidate evaluations or vote choices (Nyhan et al., 2020). Furthermore, results suggest that political fact-checking is often interpreted through *motivated reasoning* (Carnahan & Bergan, 2021), that is, that the acceptance of corrective information is highly moderated by political ideology. As stated, besides the study of its effectiveness, scholars have also shown interest in figuring out whether fact-checking organizations display any signs of political bias, which is the topic of this paper.

Newsroom fact-checking and media bias

Newsroom fact-checkers[5] are media companies staffed by journalists and funded by subscriptions, donations and/or advertising revenue (Graves, 2018). Many of these organizations have acquired public relevance for their fact-checks to political leaders, prompting social scientists to analyze their activity. The scrutiny of political bias in media outlets is far from recent (e.g., D'Alessio & Allen, 2000) and there have been various approaches that aimed to conceptualize and measure it. Bias is often defined as "any systematic slant favoring one candidate or ideology over another" (Waldman & Devitt, 1998: 302). As we will see below, it is difficult to measure this 'slant'. A popular approach is to estimate 'ideological scores' for different media outlets: Groseclose and Milyo (2005) point to an overall 'liberal bias' in US media outlets.

FCs usually follow verification protocols covering four main steps: claim selection, choice of the relevant source of evidence for fact-check, claim scoring and publication. We find a stylized illustration of this fact-checking process in the Code of Principles of the International Fact-Checking Network (IFCN), a supervising organization that evaluates whether its potential signatory members comply with their settled best practices. The code provides guidelines for articulating fact-checking guidelines in a transparent manner so that readers can "replicate" the work of FCs and, eventually, reach the same conclusions. The guidelines will foster consistency and impartiality in fact-checking, leaving no room for partisanship.

However, FCs in the US are frequently accused of displaying preferential treatment towards Democrats (Stencel, 2015). Few studies have addressed this accusation, probing whether there is any empirical evidence to support it. The main approach to bias detection have been qualitative comparisons between FCs, checking for consistency between their verification protocols. The studies focus on a particular step in the fact-checking process (mainly, claim selection or scoring) and compare the choices of different FCs. Bias is here understood as differential treatment: do agencies select a similar number of claims from the different parties under analysis? Do similar claims receive a similar score?

For instance, Marietta, Barker and Bowser (2015), using simple tabulations of claims on the same topics (broadly defined), assess the degree of agreement in claim selection between FCs. They conclude that FCs tend to select more often Democrat than Republican views –for similar approaches see Amazeen (2016), Farnsworth and Lichter (2016). Marietta, Barker and Bowser (2015) also assess the degree of agreement in the scoring processes, concluding it is, at best, moderate –see also Uscinski and Butler (2013), Farnsworth and Lichter (2016); also Amazeen, (2016), who finds higher agreement between FCs. Lim (2018) takes a different approach and finds low agreement.

However, the methodology for detecting bias through such comparison is far from consensual. In experimental fields with decades of experience in bias control (e.g., clinical trials in medicine), bias detection should be grounded in fair (*like with like*) comparisons: a difference between groups is only meaningful if the groups are exactly alike in every other respect. Although FCs follow similar methodologies, as we are going to see, they are incomparable in many respects. They do not use, for instance, the same rules for claim identification/selection and it has been observed that inconsistencies may simply arise when fact-checkers select statements containing with multiple claims but award just one truth score (Nieminen & Sankari, 2021; Walter & Salovich, 2021). They use different truth scales (see: table 0) for which there is no clear correspondence rule.

Score	Newtral	Politifact (US)	Pagella Politica
1	True	True	True
2	Half true	Almost true	Almost true
3	Misleading	Half true	Not clear
4	False	Mostly false	False
5	-	Fals	Crazy story
6	-	Pants on fire	-

Table 0

Moreover, the scores do not reflect directly the truth value of the claim, but rather the judgment of each fact-checker depending on the evidence she is using in the verification. Different scores for the same claims may not indicate differential treatment depending, e.g., on the ideology of the politician under analysis, but simply different verification methods. And, a priori, we do not know which of these methods is the unbiased one.

Comparing FCs for bias is less straightforward than we may think. We want to suggest an alternative approach that avoids the comparison problem, focusing on the verifications of a single fact-checking organization. Our concept of bias draws on an analogy with the literature on experimental design. Biases in statistically designed experiments are systematic errors in the measurement process: the outcome deviates from the true measurement value in a systematic manner, due to a flaw in the protocol. Whenever the protocol is implemented, there will always be a deviation in the same direction -not a random outcome distribution. We may say that a fact-checking protocol is *ideologically biased* if whenever the protocol is implemented the chances of obtaining a positive or negative score depend on the ideology of the politician under analysis and not of the truthfulness of her claim.

Assuming this definition of bias, our first research question would be as follows:

RQ1 Is it possible to identify ideological biases in the output of a single FC?

There are two potential obstacles for our concept of bias. On the one hand, we do not have a clear benchmark to test a fact-checking protocol: we do not know for certain whether politicians lie less or more depending on their ideology[6]. On the other hand, there are many potential sources for such ideological biases: they may occur at every stage in the verification process (claim selection, evidence selection, scoring, publication), but we are only observing published scores.

We are going to propose a statistical model for detecting potential political biases in FCs. Our method will work under two assumptions: a) politicians of mainstream parties (i.e., not populist parties) are, overall, equally accurate in their statements; b) FCs use their verification protocols (flawed or not) in a consistent manner. In this paper we will take assumption (a) for granted, examining in more detail assumption (b). Before presenting our method, let us introduce *Newtral*, on which verification output we will test our model and examine the validity of assumption (b).

Newtral, a Spanish fact-checking agency

Newtral is a Spanish media company established in 2018 by Ana Pastor, a prominent Spanish journalist and Newtral's single shareholder[7]. *Newtral* produces *El Objetivo*, a popular weekly show (conducted by Pastor) broadcasted on a national TV network, *La Sexta*. Focusing on political news and interviews, *El Objetivo* includes a regular fact-checking segment called 'Pruebas de Verificación'. *Newtral* has also launched its own website, where it offers additional fact-checks. *Newtral* relies on their social media to circulate their fact-checks and other journalistic pieces. Moreover, it is a signatory member of the IFCN Code of Principles. Thus, its activity follows a protocol that is foregrounded on the standards provided by the IFCN and undergoes an annual compliance audit[8]. Newtral's fact-checking protocol is explained as follows on its website: every day, members of their staff collect checkable claims from different sources. Then they select those they deem relevant using a "purely journalistic criterion": they assess the "significance" of the statement and the author, as well as whether the claim is "intentionally created to confuse" and whether it has "verifiable content with data"[9]. To check these claims, journalists at *Newtral* use official data sources and judgment from experts. If necessary, *Newtral* requests clarifications to the press offices of those politicians whose claims are under scrutiny [9(bis)]. Finally, they score the claim according to their own ordinal fact-checking scale: "Verdadero" (True), "Verdad a medias" (Half True), "Engañoso" (Misleading) and "Falso" (False).

Newtral's website does not provide further details about how this method is implemented (nor does the IFCN). As we already mentioned in the previous section, our bias detection method assumes that agencies implement their verification protocols in a consistent manner: if they do, our method will reliably detect a *prima facie* sign of political bias; if they do not, our method will only reveal an aggregation of random decisions.
Therefore, our second research question is:

> *RQ2.1.* What is *Newtral*'s fact-checking method?
>
> RQ2.2. Is Newtral's method applied in a consistent manner?

Let us now present the data that we will use to address our two research questions, before introducing our bias detection method.

Data

Data: Statistical analysis

To test our proposal we will proceed to the statistical analysis of 313[10] fact-checks coded from *Newtral*'s webpage. This database comprises all fact-checks of Spanish politicians published on *Newtral*'s website during its first year of fact-checking activity, that is, between October 3rd, 2018, and October 2nd, 2019. This lapse was chosen for various reasons. First, because in this period *Newtral* was already a signatory member of the IFCN[11], and hence fully compliant with its Code of Principles. Second, this period is interesting because it covers two electoral periods: the runup to the Spanish parliamentary elections held on April 28th, 2019, and the runup to the local and European elections held on May 26th, 2019. Ultimately, the number of observations is well above the bar literature has considered appropriate to run logistic regressions (Long, 1997: 54).

We created a database of claims, coding, for each claim, the party membership[12] of the speaker into a multinomial variable, and *Newtral*'s awarded truth score into an ordinal one[13]. These 313 fact-checks were developed by a team of 17 *Newtral* journalists[14] - who had signed at least one fact-check entry. In January 2021, we got in touch with *Newtral*'s higher education division to interview the authors of the fact-checks in our sample. Five of them participated. We also contacted some of the remaining journalists who had left *Newtral* since then. Two accepted, but others did not reply or were not available. The seven fact-checkers we interviewed were responsible for nearly 70% of the fact-checks in our sample[15].

Data: Interviews on the potential sources of noise

To check whether the assumptions supporting our model holds, we conducted a series of semi-structured interviews with the journalists who had contributed to generate our sample. Each interviewee received a questionnaire about a week in advance (see supplementary materials for the questionnaire). Each of the interviewees signed an informed consent form disclosing the goals of the study: to understand *Newtral*'s guidelines and how they implemented them[16]. Although we followed the order of the questionnaire, we deemed key not to interrupt and interviewees were free to jump from one topic to another. We sometimes brought follow-up questions. The questionnaire had nine items: the first three referred to the claim selection process: how the news sources for claims were chosen and monitored and how the "journalistic criteria" for selection were implemented. We also asked about how a claim was singled out for analysis (the wording and which parts were omitted). The following questions investigated the fact-checking process: how the source of evidence for a fact-check was chosen, how the score was decided and what were the criteria to publish the fact-checks. We also asked about their interpretation of the score and, finally, about the guidelines for dealing conflicts of interest. Two final questions were asked about their perception of the differences between fact-checking and traditional journalism.

We conducted the interviews on Zoom, in Spanish, between 9th February 2021 and 29th April 2021, recording audio and transcribing. Interviews took between half an hour and fifty minutes. Relevant excerpts were translated by the authors into English.

Method

Statistical analysis

We may say that a verification protocol is *ideologically biased* if whenever the protocol is implemented the chances of obtaining a positive or negative score depend on the ideology of the politician under analysis and not on the truthfulness of the claim. To implement this definition, we propose a model to calculate the probabilities for each political party to obtain a given truth score, drawing on a sample of verifications from a single fact-checker. Under the assumption that all mainstream parties are overall equally accurate (that we will take here for granted), a significant difference between these probabilities would be a *prima facie* sign of bias in the verification protocol.

We will illustrate our model using *Newtral*'s data. Below is an ordered logistic regression (*ologit*), where *Newtral*'s truth score is the dependent variable (Fullerton, 2009). This is an effective approach for comparing groups that is not affected by group differences and does not involve stringent assumptions (Long, 1997). The ordinal logit model is as follows:

$$(1)\quad log\left(\frac{Pr(y \leq m|X)}{Pr(y > m|X)}\right) = \tau_m - X\beta \quad (1 \leq m < M)$$

m is the truth score awarded at the end of each fact-check (dependent variable); X is the independent variable or predictor: our predictor is the party affiliation of the politician who stated each claim; τ represents the cutoff point and β is a vector displaying the relationship between m and X.

We take the Spanish socialist party (PSOE) as the model's arbitrary baseline[17] ($\beta 0$). It is important to note that PSOE was the party in government throughout the whole period we study. The other parties[18] present in the sample are: PP (conservative), Ciudadanos (liberal), Podemos (far left), Vox (alt-right) and Other, a variable that comprises regional parties with representation in the Spanish Parliament. Regarding party positions, PP, Ciudadanos, and Vox are considered right-wing, whereas PSOE and Podemos are considered left-wing (see: Simón, 2020).

The available literature suggests that political actors may have varying incentives for being more or less accurate (see: Davis & Ferrantino, 1996; Armstrong-Taylor, 2012) depending on various factors. In the US, for instance, former President Donald Trump was evaluated as making more false claims because he indeed made more false claims (Davis & Sinnreich, 2020). However, there aren't studies that address this issue for Spain. Therefore, we assume that all parties from the establishment, that is, excluding those that use populist rhetoric – Podemos and Vox (Vampa, 2020) – would have an almost equal propensity to being accurate in their statements. So in our model, had there been no differences between parties, we would find the *Newtral* scores of establishment parties around this baseline. Weighty deviations from this baseline signal, according to our interpretation, a potential differential treatment: either members of certain parties are lying significantly more or a glitch in the fact-checking process could be generating the deviations.

In our model, the β shows, for each party, the estimated change in the natural log odds of *Newtral*'s truth scores as a shift from the baseline (PSOE). However, as these β come in the form of log-odds, it is difficult to provide a straightforward interpretation of them (Ranganathan et al., 2017). Therefore, we calculate the predicted probabilities (Muller & MacLehose, 2014) of receiving each *Newtral* score by party, including the arbitrary baseline. These are computed by setting the independent variable to its mean values (Williams, 2012).

Results

Regression analysis

We firstly present a descriptive analysis of our sample. As Table 2 shows, for all the political parties under scrutiny, almost half of the fact-checked claims were considered “False” and barely one in five was considered “True”. Except for PSOE, each individual party receives significantly more “False” than “True” scores. Intermediate scores tend to be somewhat similar across parties and main differences seem to arise from the extremes.

Table 1: Totals by party and score

	TRUE	**HALF TRUE**	**MISLEADING**	**FALSE**	**TOTAL**
PSOE	27	8	23	29	87
PP	20	10	24	61	115
CS	4	5	10	22	41
PODEMOS	9	3	5	13	30
VOX	1	1	3	13	20
OTHER	2	2	7	11	22
TOTAL	63	29	72	149	313

Table 2: Average by party and score

	TRUE	**HALF TRUE**	**MISLEADING**	**FALSE**	**CUMULATIVE**
PSOE	31,03%	9,20%	26,44%	33,33%	100,00%
PP	17,39%	8,70%	20,87%	53,04%	100,00%
CS	9,76%	12,20%	24,39%	53,66%	100,00%

PODEMOS	30,00%	10,00%	16,67%	43,33%	100,00%
VOX	5,56%	5,56%	16,67%	72,22%	100,00%
OTHER	9,09%	9,09%	31,82%	50,00%	100,00%
W. AVG	20,13%	9,27%	23,00%	47,60%	100,00%

Table 3 shows the output of the *ologit* models. Model 1 is a ‘raw’ *ologit* with the outcome variable and the predictor. Model 2 includes two control variables: the dummy female politician (1 if female) to check if the evidence found in the literature concerning the existence of gender differences in the propensity to lie (Houser et al., 2012) is latent in this sample. Model 2 includes another dummy that controls for the use of statistics as evidence for the fact-check.

A positive β coefficient indicates that the party has more chances of receiving overall worse scores than PSOE. This is what happens for all the parties under *Newtral*’s scrutiny: Ciudadanos, PP, Podemos, Vox, plus a composite variable ‘Other’ capturing parties. Except for Podemos and Other (p > 0.10) these results are statistically significant for p < 0.05. Note that we focus on the interpretation of the interaction between the truth scores and political parties: following Keele et al. (2020), in Model 2 we consider control variables as mere moderators of the relationship between scores and political parties.

Table 3: OLM

y = scores (PSOE is baseline)	M1		M2	
Partido Popular	**0.77****	**(0.27)**	**0.83****	**(0.28)**
Ciudadanos	**0.90***	**(0.34)**	**0.96****	**(0.34)**
Podemos	0.27	(0.42)	0.26	(0.42)
Vox	**1.67****	**(0.54)**	**1.75****	**(0.61)**
Other	0.85	(0.40)	0.76	(0.38)

Female politician		-0.03 (0.26)
Fact-check with statistics		**-0.82*** (0.22)**
N	313	313

Note: ∗, ∗∗, and ∗∗∗ denote significance at 10, 5, and 1 percent levels, respectively. Robust standard errors between brackets.

Figures 1 and 2 show the predicted probabilities for each party of receiving each of the two truth scores on the extremes of Newtral's scale[19]: True (Figure 1) and False (Figure 2) (see intermediate scores in supplementary materials). These figures include the point estimate of the predicted probabilities with 90% CI- *L*: lower bound; *U*: upper bound. For instance, according to Figure 1, PSOE seems to have almost a 30% chance of receiving a True score. PP has 15,5% probability of receiving that verdict. The results for Ciudadanos are like those for PP. Values for Podemos are uncertain as the true value could lie above PSOE or below PP. The predicted probability for Vox is 6,8%, below PP and Ciudadanos. Figure 2 shows that PSOE is the least likely to receive a False score, and that PP, Ciudadanos and Vox significantly more of these.

Figure 1: Predicted probabilities of receiving 'True' by party

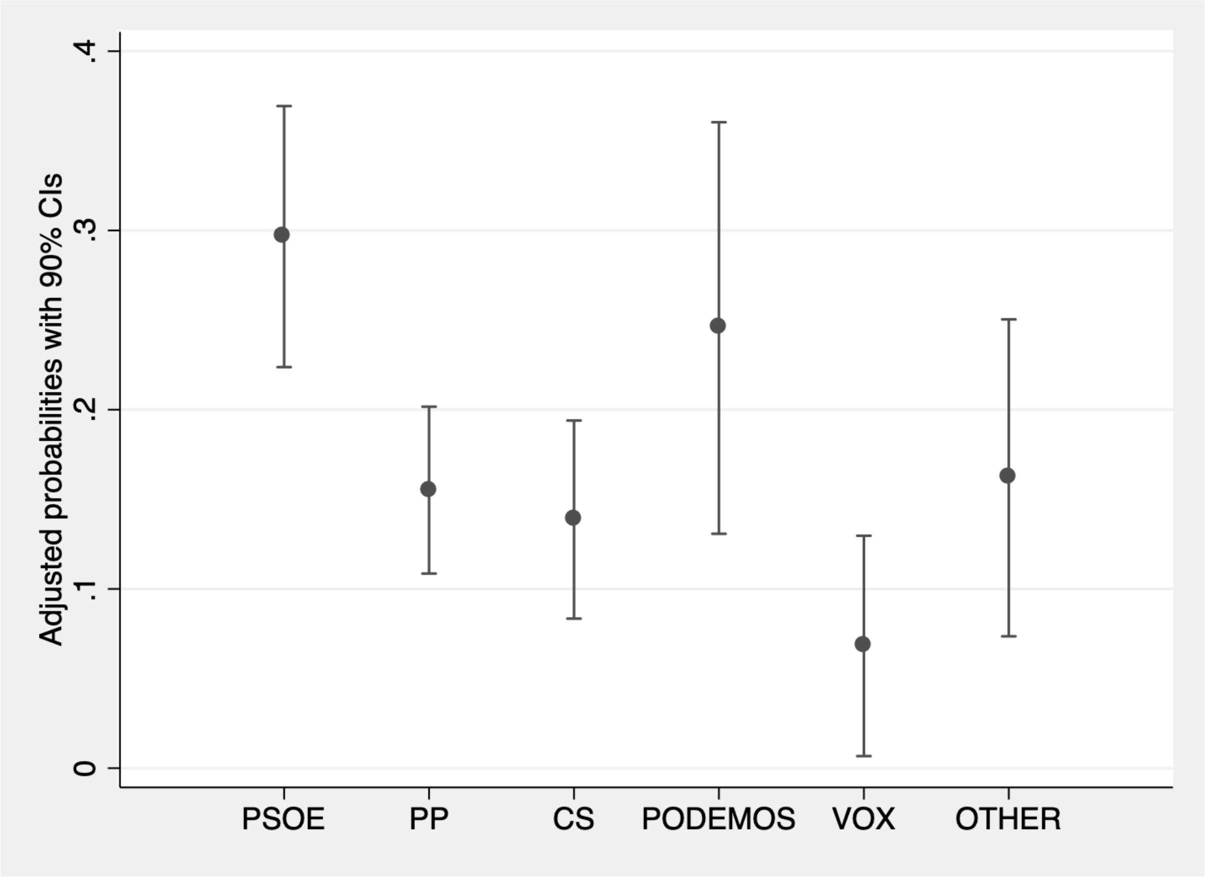
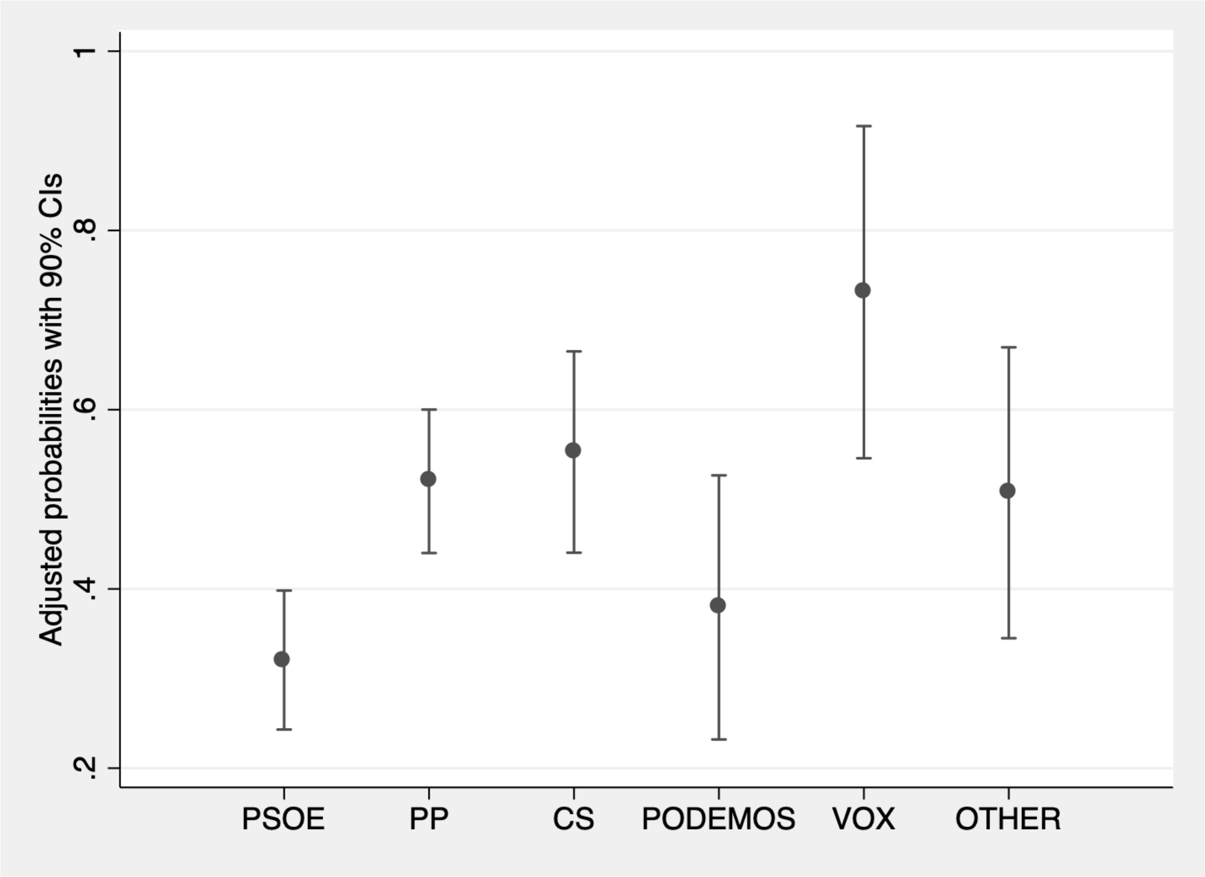
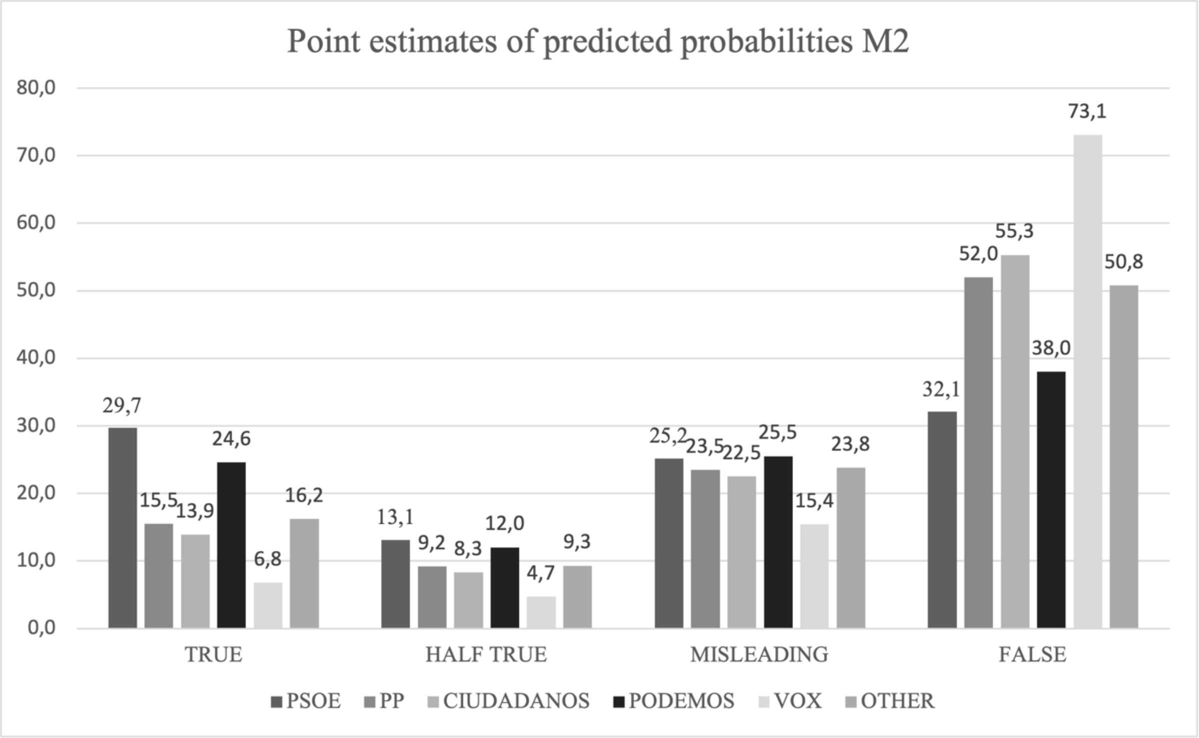


Figure 2: Predicted probabilities of receiving ‘False’ by party



By way of summary, figure 3 shows the point estimates of the predicted probabilities of receiving each truth-score for all political party under analysis. Our method detects a visible difference in Newtral's verifications: according to Newtral, left-wing politicians are more truthful than right-wing politicians. Of course, this conclusion only holds assuming (a) that politicians lie on average the same and (b) Newtral is implementing a well-defined verification method on a systematic manner. To find out whether this is the case, we conducted a series of interviews that we present below.

Figure 3: Point estimates of predicted probabilities of receiving each score by political party



Interviews

The goal of these interviews is to grasp Newtral's fact-checking method and the consistency it ensures. Had we identified a well-defined method, we would have organized an experimental analysis of its reliability (see: Crowder et al., 2017). As we are going to see next, our qualitative exploration revealed instead a lack of well-defined method with ample room for individual discretion at every stage.

Newtral's methodology indicates that the fact-checking team daily collects the statements of politicians from newspapers, radio and TV interviews, social media, and any other public platform[20]. Thus, we first asked how these claim sources were chosen and monitored. All our interviewees agreed on the answer: a senior team member prepared a daily agenda of events organized by the different political parties they were following: a spreadsheet in which each event is assigned to a different team member (FC3 - Personal interview, February 2021). This is a first selection point, as one of our interviewees explicitly acknowledged:

> *I was in charge of the agenda: I unified parties according to where they were going to be and what their spokespersons were going to do. As I was making that agenda, I was already selecting. At that point, you filter and go primarily to what's interesting on a national or regional level, purely for journalistic interest. I did it on an Excel in which*

> *I assigned a priority: I considered something as basic as how many journalists we had available at that moment.* (FC3 - Personal interview, February 2021)

The published selection criterion is newsworthiness and according to (FC3 - Personal interview, February 2021) this is implemented[21]. As to the allocation of events, each professional usually specialized in a political party –all the interviewees agreed on this point, but FC4 indicates that specialization was not limited to a single party. Depending on their availability, fact-checkers covered events.

> *There have been phases, especially during election campaigns, when we've divided ourselves by political parties. Not for any particular reason, but it's true that when you hear a political campaign many times, you are able to see their position; when they repeat a lie many times, you know where they're coming from and where they're going, so it's just a question of being agile in your work (...) When there's a change in the fact-checking team or something, we start listening to other parties again.* (FC1 – Personal interview, February 2021)

> *The team has had different phases, but during the electoral campaign, we started to realize that it was better to divide ourselves into parties; or rather, into [political] sides, because it wasn't just one party but rather right-wing or left-wing. (...) We've realized that when we specialize, everything is much clearer.* (FC4 – Personal interview, February 2021, [our emphasis])

Once the agenda is set, each fact-checker focuses on her assigned events, using a transcript of the politician's intervention (FC7 – Personal interview, April 2021). Again, all the interviewees agree on the process: they check the politician's intervention, identifying checkable statements. As tables 1 and 2 indicate, potential lies are more attractive than potential truths:

> *This is very personal, and the fact-checker's hearing can be trained. I must bear in mind that the statement can be verified, that it's not opinions. I write down everything that I think is susceptible to verification, because I myself have doubts… It's true that you're always looking for the lie, but there might be some truths that surprise you. And if it surprises you, it's going to be interesting.* (FC3 – Personal interview, February 2021)

However, there seem to be no clear guidelines for identifying a claim: how to single out the sentence to be checked, abstracting away the context and verbal nuances. For instance, rules to tell apart value and factual judgments:

> *We have something you could call principles on an internal level, that don't need to be written, because everybody knows them; we follow the principles set by the IFCN. (…) There are times that we disagree. Someone writes down a phrase and when we talk about it, we say, "the thing is that this is an opinion". There can't be a set of rules, so to say, because each situation varies*. (FC1 – Personal interview, February 2021)

Once each fact-checker has drafted a list of verifiable claims, they post them on a common platform (FC1, FC2, FC3, FC6, FC7). Here there seems to be ample room for personal selection:

> *This has changed over time, but now we upload all the statements that we have preselected onto [name of the platform] (…) Sometimes, even myself that I have more experience, I'll upload all those things that I think require work onto [name of the platform]. Basically, I do my own pre-selection (…) "The thing is that you voted against this-that-and-the-other", listen, I don't even waste my time, I go to the Congress website, and I look. And from there, we put it on [name of the platform]*. (FC3 – Personal interview, February 2021)

Once there is shared list of claims, a senior content editor leads the claim selection process (FC3, FC4, FC5, FC7). The editor filters the most interesting claims, from a journalistic perspective (FC4, FC7), and invites the fact-checkers in charge to contact the press office of the politician's party requesting a clarification, either of the context in which the claim was stated or the sources on which it was based. If there is no satisfactory response after 24-48 hours, then the fact-checking process starts. For this purpose, the fact-checker should find the relevant evidence for verification. We asked the fact-checkers whether *Newtral* had a written protocol to weigh what evidence sources should be prioritized. Our interviewees did not provide a straightforward answer but hinted instead to some informal guidelines.

> *The priority is that they are official sources, of course, as well as being public. That you can present them, and when it comes to checking them, you can also add it to the whole. (…) All the sources must be very serious and need to have experience over the*

> *data that they are providing (...) It can range from a Ministry report to university research. And then you can add other sources of data, like for example a report from a charity over a particular topic. Are they all valid? No. Valid are those that are backed by experience and trustworthiness.* (FC2– Personal interview, February 2021)

However, on a deeper layer, there wasn't consensus over whether an NGO report would qualify as robust evidence for a fact-check. Once selected the evidence source, each team member should score the truthfulness of the claim. Our questionnaire asked how they used their scale and, in particular, how could they tell apart the intermediate from the extreme values (e.g., true vs. half-true). Again, there were no formal guidelines for the scoring process. Some of the fact-checkers admitted that the scale was sometimes difficult to implement:

> *It works for us for now, because although it's true that the differences between 'Half-true' and 'Misleading' are that small, we've always thought that there should also be another category labelled 'Unchekable'"* (FC2– Personal interview, February 2021)

Another difficult point, for which there wasn't agreement, was how to deal with politician's intentions when stating the claim: was she aware or unaware of its falsehood? Depending on the interpretation fact-checkers may switch from "Misleading" to "False" (FC1, FC3, FC4, FC5, Personal interview, February 2021)

Once the claims are individually scored, they are shared with the team and there is debate around some scores (FC1, FC4 - Personal interview, February 2021). Ultimately, the senior team member leading the process has the last word over deciding the fact-checks to be published and the score.

> *Yes, we discuss it with whatever coordinator is there at the time. Precisely because even though a verification is not an opinion and is backed by facts, when it comes to providing the result, and precisely because there are those intermediate categories of "Half-truth" and "Misleading" before you reach the total "False", well, a debate takes place. It is often the case that when you speak to colleagues or to the person that's in charge, the scale tips in one direction or another.* (FC2– Personal interview, February 2021; confirmed also by FC5)

As Tables 1 and 2 suggest, the published fact-checks tend to have more negative scores. When asked about this imbalance, the answer is, again, that it is an editorial decision:

> *There's a selection process, since it's not just up to the fact-checkers but also the content supervisor (...) it's true that in the end, you keep the most 'interesting truths'– those that are editorially worthy publishing. (...) I think we find more 'false' [Newtral] because it has more interest and, of course, there's a criterion that is actually editorial (...) When you fact-check a whole event, you do try to think a bit more between 'Truths', 'Half-truths', 'Falses', so that there's a realer portrait of what has happened.* (FC4, Personal interview, February 2021, [our emphasis]; also FC5)

We also asked about how fact-checkers manage their own personal views or conflicts of interest. All the interviewees were confident that the fact-checking process and IFCN codes were rigorous enough to leave personal standpoints out of the equation.

> *There's no margin here [for personal opinions to have an effect], because even though the work is carried by people, we have to follow very clear guidelines: to look for sentences that are verifiable and verify them. There is no room for opinion.* (FC1, Personal interview, February 2021, [our emphasis]; also FC2)

At several points in the fact-checking process "journalistic criteria" play a crucial and explicit role, be it in the selection of events, claims, or publishable fact-checks. "Journalistic criteria" is mentioned, but not explained in Newtral's published guidelines. So, we asked the interviewees how they interpret and apply those criteria. The answers hint at an inconsistent interpretation. The most frequently mentioned feature is newsworthiness.

> *The basic journalistic criterion is that I prefer to listen to Pablo Casado [then PP's political leader] than a town councilor (...) But the journalistic criteria is very difficult to classify because it's journalistic: wherever the focus of attention is going to be, that's where I must be too.* (FC3 – Personal interview, February 2021, [our emphasis])

But then we find different operational criteria. E.g., how big a lie could be and how important is the politician or her political party:

> *The first criterion is relevance. It's not the same if it's the PM who lies or a Provincial Executive (...) There's also the topic, as well as the magnitude of the lie: It's different to say 'unemployment rates are at 4%' when the real number is 4.25%, than to say "Spain is leading the growth of the EU" when in reality our country is the 10th. It's a criterion that follows common sense, drawing from the premise that we fact-check everyone.* (FC1 – Personal interview, February 2021; also, FC2)

But this interpretation may be overrun depending on where the interest of the audience could be:

> *(...) If I'm still obsessed with the parliamentary representation, then maybe I'm moving away from where the attention is, which is where I think my work should be.* (FC3 – Personal interview, February 2021)

In summary, the fact-checking procedures at Newtral, as it emerges from our interviews, follow the published guidelines, but leave individuals ample room. The protocol is not specified in detail and there is little external supervision about how each step is carried out. It is far from obvious how could any independent third party could replicate[22] the process and reach the same conclusions. We consider the consequences in the following section.

Discussion

We have proposed a statistical model to detect political bias in the verification output of a FC. Our model suggests that there are *prima facie* signs of bias in Newtral's fact-checking protocol, since there is a noticeable difference between the probabilities of obtaining a positive/negative score depending on the ideology of the party under analysis. The answer to our first research question would be therefore positive: unlike other bias detection methods available in the literature, in which comparison between agencies are essential, our model allows us to detect bias in the output of a single fact-checking organization. However, our model depends on two key assumptions: (a) all politicians from mainstream parties are on average equally truthful - which is plausible in our case, although we have just taken it for granted; and (b) Newtral has a well-defined method that implements consistently. Testing this second assumption was our second research question and, as we have just seen, it is simply not granted. Newtral verification protocol seems ill-defined and leaves ample room for individual discretion.

Some of these verification decisions seem to have more weight on the final outcome (e.g., setting the agenda: choosing which claims should be scrutinized), but even if some of these decisions were somehow controlled for (e.g., introducing randomization in the selection of claims), there is no guarantee about the rest of the steps in the fact-checking process. A nuance in the claim wording, a different source of evidence for the check, a different understanding of the truth scale and the decision to publish or not: all these decisions may have easily generated different outputs. If the goal of an IFCN signatory is to deliver reproducible fact-checks, there is no guarantee about it in Newtral's *modus operandi*. A light external audit on a few verifications, like that the IFCN conducts, will not detect the magnitude of the problem.

We should also flag one final point about Newtral. It seems implausible that its verification method reliably tracks the number of inaccuracies Spanish politicians tell. The scores presented in Tables 1 and 2 are highly skewed towards falseness, suggesting that inaccuracies are pervasive in Spanish political discourse. Our statistical analysis probably suggests a combination of bias in Newtral (perhaps arising at the claim selection stage) and noisy individual decisions in the verification process.

However, our analysis presents various limitations. We have not used all *Newtral*'s fact-checks until today, but rather a sample of 313 fact-checks authored by a particular group of journalists. Widening the sample for the statistical analysis would have reduced the representativeness of our interviews. Our primary goal was to see how their fact-checking guidelines generated the data we used for the detection of potential differential treatment, not to assess *Newtral*'s fact-checking activity in general. Even within this sample, some of the truth scores were not used much and not all the parties were targeted with equal frequency: the reduced use of some scores (e.g., half true) implies limited variance for any meaningful interpretation of that particular verdict. Also, we were not able to interview all the fact-checkers that were responsible for our sample. Furthermore, our questionnaire was not completely closed and thus we faced evident tradeoffs. Moreover, the results from this case study don't allow for any generalization to other contexts.

Conclusion

The number of FCs keeps growing, and so does their influence in the public sphere. We need to better understand how their fact-checking methods work and whether they are more reliable than conventional forms of journalism. In this paper we have presented a concept of ideological bias and a method to detect it in the verdict output of political fact-checking agencies. However, our qualitative study shows that the verification methods of these agencies are probably too loose to admit a strict analysis, and this would affect not only our method, but any of the alternative approaches to consistency or bias detection examined in our literature review. If the goal is to attain reproducible verifications, instead of traditional journalistic fact-checks, audit bodies like the IFCN should strengthen their verification guidelines, reducing the amount of discretionary decisions they allow. Also, FCs should train their staff so that everyone use their methods in a systematic manner. Otherwise, political fact-checking would only contribute bias and noise to our public sphere.

References

Adair, B., 2014. *Duke Study Finds Fact-Checking Growing Around the World - Duke Reporters' Lab*. [online] Duke Reporters' Lab. Available at: <https://reporterslab.org/duke-study-finds-fact-checking-growing-around-the-world/> [Accessed 29 March 2022].

Amazeen, M. A. (2013). *A Critical Assessment of Fact-checking in 2012.* New America Foundation.

Amazeen, M. A. (2016). Checking the fact-checkers in 2008: Predicting political ad scrutiny and assessing consistency. *Journal of Political Marketing*, *15*(4), 433-464.

Amazeen, M. A., Thorson, E., Muddiman, A. & Graves, L. (2018). Correcting political and consumer misperceptions: The effectiveness and effects of rating scale versus contextual correction formats. *Journalism & Mass Communication Quarterly*, *95*(1), 28-48.

Armstrong-Taylor, P. (2012). When do politicians lie?. *The BE Journal of Economic Analysis & Policy*, *13*(3).

Brashier, N. M., Pennycook, G., Berinsky, A. J., & Rand, D. G. (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences*, *118*(5).

Carey, J. M., Guess, A. M., Loewen, P. J., Merkley, E., Nyhan, B., Phillips, J. B., & Reifler, J. (2022). The ephemeral effects of fact-checks on COVID-19 misperceptions in the United States, Great Britain and Canada. *Nature Human Behaviour, 6*(2), 236-243.

Carnahan, D., & Bergan, D. E. (2021). Correcting the Misinformed: The Effectiveness of Fact-checking Messages in Changing False Beliefs. *Political Communication,* 1-18.

Carr, D. (2012, November 6). A last fact check: It didn't work. *The New York Times*. Retrieved from http://mediadecoder.blogs.nytimes.com/2012/11/06/a-last-factcheck-it-didnt-work/

Chan, M. P. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological science*, *28*(11), 1531-1546.

Crowder, M. J., Kimber, A. C., Smith, R. L., & Sweeting, T. J. (2017). *Statistical analysis of reliability data*. Routledge.

D'Alessio, D., & Allen, M. (2000). Media bias in presidential elections: A meta-analysis. *Journal of Communication*, *50*(4), 133-156.

Davis, M. L., & Ferrantino, M. (1996). Towards a positive theory of political rhetoric: Why do politicians lie?. *Public Choice*, *88*(1), 1-13.

Davis, D. H., & Sinnreich, A. (2020). Beyond Fact-Checking: Lexical Patterns as Lie Detectors in Donald Trump's Tweets. *International Journal of Communication*, *14*, 24.

Ecker, U. K., O'Reilly, Z., Reid, J. S., & Chang, E. P. (2020). The effectiveness of short-format refutational fact-checks. *British Journal of Psychology, 111*(1), 36-54.

Ecker, U. K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou,P., Vraga, E., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13-29.

Farnsworth, S. J., & Lichter, S. R. (2019). Partisan targets of media fact-checking: examining President Obama and the 113th Congress. *Virginia Social Science Journal,* 53, 51-62.

Fullerton, A. S. (2009). A conceptual framework for ordered logistic regression models. *Sociological methods & research*, *38*(2), 306-347.

Gottfried, J. A., Hardy, B. W., Winneg, K. M., & Jamieson, K. H. (2013). Did fact checking matter in the 2012 presidential campaign?. *American Behavioral Scientist*, *57*(11), 1558-1567.

Graves, L., & Amazeen, M. (2019) Fact-Checking as Idea and Practice in Journalism. *Oxford Research Encyclopedia of Communication*.

Graves, L., & Cherubini, F. (2016). The rise of fact-checking sites in Europe. *Oxford Reuters Institute for the Study of Journalism.*

Graves, L., Nyhan, B., & Reifler, J. (2016). Understanding innovations in journalistic practice: A field experiment examining motivations for fact-checking. *Journal of Communication*, *66*(1), 102-138.

Graves, L. (2018). Boundaries not drawn: Mapping the institutional roots of the global fact-checking movement. *Journalism Studies*, *19*(5), 613-631.

Groseclose, T., & Milyo, J. (2005). A measure of media bias. *The Quarterly Journal of Economics*, *120*(4), 1191-1237.

Houser, D., Vetter, S., & Winter, J. (2012). Fairness and cheating. *European Economic Review*, 56(8), 1645-1655.

Humprecht, E. (2020). How do they debunk "fake news"? A cross-national comparison of transparency in fact checks. *Digital Journalism*, *8*(3), 310-327.

Lim, C. (2018). Checking how fact-checkers check. *Research & Politics*, *5*(3), 1-7.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, *13*(3), 106-131.

Long, S.J. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, California: Sage Publications.

Marietta, M., Barker, D. C., & Bowser, T. (2015). Fact-checking polarized politics: Does the fact-check industry provide consistent guidance on disputed realities? *The Forum*, 13(4), 577-596)

Mena, P. (2019). Principles and boundaries of fact-checking: Journalists' perceptions. *Journalism practice*, *13*(6), 657-672.

Muller, C. J., & MacLehose, R. F. (2014). Estimating predicted probabilities from logistic regression: different methods correspond to different target populations. *International journal of epidemiology*, *43*(3), 962-970.

Nieminen, S., & Rapeli, L. (2019). Fighting misperceptions and doubting journalists' objectivity: A review of fact-checking literature. *Political Studies Review*, *17*(3), 296-309.

Nieminen, S., & Sankari, V. (2021). Checking PolitiFact's Fact-Checks. *Journalism Studies*, *22*(3), 358-378.

Nyhan, B., Porter, E., Reifler, J., & Wood, T. J. (2020). Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior*, *42*(3), 939-960.

Ranganathan, P., Pramesh, C. S., & Aggarwal, R. (2017). Common pitfalls in statistical analysis: logistic regression. *Perspectives in clinical research*, *8*(3), 148.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, *19*(1), 22-36.

Simón, P. (2020). The multiple Spanish elections of April and May 2019: the impact of territorial and left-right polarisation. *South European Society and Politics*, *25*(3-4), 441-474.

Stencel, M., 2015. *The Weaponization of Fact-Checking*. [online] POLITICO Magazine. Available at: <https://www.politico.com/magazine/story/2015/05/fact-checking-weaponization-117915/> [Accessed 29 March 2022].

Stencel, M., Ryan, E., and Luther, J., 2022. *Fact-checkers extend their global reach with 391 outlets, but growth has slowed*. [online] Duke Reporters' Lab. Available at: <https://reporterslab.org/tag/fact-checking-census/> [Accessed 8 November 2022].

Thorson, E. (2016). Belief echoes: The persistent effects of corrected misinformation. *Political Communication*, *33*(3), 460-480.

Uscinski, J. E., & Butler, R. W. (2013). The epistemology of fact checking. *Critical Review*, *25*(2), 162-180.

Vampa, D. (2020). Competing forms of populism and territorial politics: the cases of Vox and Podemos in Spain. *Journal of Contemporary European Studies*, *28*(3), 304-321.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151.

Waldman, P., & Devitt, J. (1998). Newspaper photographs and the 1996 presidential election: The question of bias. *Journalism & Mass Communication Quarterly*, *75*(2), 302-311.

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, *37*(3), 350-375.

Walter, N., & Salovich, N. A. (2021). Unchecked vs. uncheckable: How opinion-based claims can impede corrections of misinformation. *Mass Communication and Society*, *24*(4), 500-526.

Weeks, B. E., & Gil de Zúñiga, H. (2021). What’s next? Six observations for the future of political misinformation research. *American Behavioral Scientist*, *65*(2), 277-289.

Williams, R. (2012). Using the margins command to estimate and interpret adjusted predictions and marginal effects. *The Stata Journal*, *12*(2), 308-331.

Young, D. G., Jamieson, K. H., Poulsen, S., & Goldring, A. (2018). Fact-checking effectiveness as a function of format and tone: Evaluating FactCheck. org and FlackCheck. org. *Journalism & Mass Communication Quarterly*, *95*(1), 49-75.

Endnotes

[1] Politifact received a Pullitzer prize in 2009 - https://www.pulitzer.org/winners/staff-69 (accessed: 21st March 2022)

[2] We use verification as a synonym to fact-checking.

[3] The IFCN mentions replicability: "Signatories want their readers to be able to verify findings themselves. Signatories provide all sources in enough detail that readers can replicate their work, except in cases where a source's personal security could be compromised." - https://ifcncodeofprinciples.poynter.org/know-more/the-commitments-of-the-code-of-principles (accessed: 21st March 2022)

[4] We use this term loosely, mostly as a synonym to fact-checking, as in Brashier et al. (2021)

[5] As in: Graves, 2018.

[6] There is no evidence that politicians in mainstream parties that represent mainstream ideologies are consistently more or less accurate in their statements. Therefore, we will asume that they have equal probability of being more or less accurate.

[7] Description gathered from Newtral: https://www.Newtral.es/quienes-somos/ (accessed: 21st March 2022)

[8] https://ifcncodeofprinciples.poynter.org (accessed: 21st March 2022)

[9] https://www.Newtral.es/metodologia-transparencia/ (accessed: 21st March 2022)

[10] Not all the fact-checks in the sample are independently published nor have a separate web link: Newtral also publishes compilations of fact-checks in a single piece. As a result, we counted as one fact-check each time Newtral issued a verdict according to their scale, regardless the fact-check was published isolated or in a compilation. However, we ruled out those fact-checks that were duplicate and those that targeted foreign politicians.

[11] https://ifcncodeofprinciples.poynter.org/application/public/newtral/08442D12-61C6-3050-33F7-C0F003C26588 (accessed: 21st March 2022)

[12] The party membership of the claim's author is adjudicated according to his/her then political appointment at the time the claim was made (e.g., member of parliament). Short of any appointment, the minimal requirement for affiliation was to be publicly known as member of the party. The affiliation was coded into a multinomial variable with unordered levels. The numeric code for each party is: 1- PSOE (social democratic), 2- PP (conservative), 3- Ciudadanos (liberal), 4- Podemos (far-left), 5- Vox (alt right), 6 – Other (an array of mainly regional parties with diverse ideologies)

[13] 1- "True", 2- "Half true", 3 - "Misleading" and 4 - "False".

[14] There are relevant differences in authorship: Some of these fact-checkers signed just one piece in the year.

[15] There were 24 fact-checks signed by "Newtral" without any reference to a particular author. If we assume that the interviewed fact-checkers were responsible for the same proportion (69'8%) of the 24 unsigned fact-checks, then we would see that the fact-checkers we interviewed are responsible for 75% of the verifications in the sample.

[16] This study obtained approval by the Ethics Committee at Universidad Nacional de Educación a Distancia (UNED) – Reference code: 1-2021 FSOF

[17] The arbitrary selection of an alternative baselines would not have changed any results.

[18] Ideologies are based on these parties' group affiliation in the European Parliament.

[19] Model 2

[20] https://www.newtral.es/metodologia-transparencia/ (accessed: 21st March 2022)

[21] "We choose all those statements that have interest or relevance from a purely journalistic criterion. We consider the relevance of the statement and the author, if it is repeated as an argument created intentionally to confuse and if it has verifiable content with data" (*ibid*.)

[22] One of the main goals stated in the IFCN Code of Principles: "Signatories want their readers to be able to verify findings themselves. Signatories provide all sources in enough detail that readers can replicate their work, except in cases where a source's personal security could be compromised." - https://ifcncodeofprinciples.poynter.org/know-more/the-commitments-of-the-code-of-principles (accessed: : 21st March 2022)