Truth be told: Cognitive moderators of selective sharing of fact-checks on social media

Natalia Aruguete[a], Ingrid Bachmann[b], Ernesto Calvo[c], Sebastián Valenzuela[bd]*, and Tiago Ventura[c]

[a]Departamento de Ciencias Sociales, Universidad Nacional de Quilmes (UNQ), Bernal, Argentina; [b]School of Communications, Pontificia Universidad Católica de Chile, Santiago, Chile, [c] University of Maryland, College Park, USA, [d]Millennium Institute for Foundational Research on Data (IMFD), Santiago, Chile.

Word count: 7,962 words

*Corresponding author. Mail: Alameda 340, Santiago, Chile, CP 8115500. Email: savalenz@uc.cl

Natalia Aruguete (PhD, Universidad Nacional de Quilmes, 2009), is Professor of the Universidad Nacional de Quilmes and the Universidad Austral, and a member of CONICET (Argentina's Federal Agency for the development of Science and Technology). nataliaaruguete@gmail.com

Ingrid Bachmann (PhD, University of Texas at Austin, 2011) is Associate Professor in the School of Communications at Pontificia Universidad Católica de Chile. ibachmann@uc.cl

Ernesto Calvo (PhD, Northwestern University, 2001), is Professor of Government and Politics at the University of Maryland and Director of the interdisciplinary Lab for Computational Social Science (iLCSS). ecalvo@umd.edu

Sebastián Valenzuela (PhD, University of Texas at Austin, 2011) is Associate Professor and Research Chair in the School of Communications at Pontificia Universidad Católica de Chile, and Associate Researcher at the Millennium Institute for Foundational Research on Data (IMFD). savalenz@uc.cl

Tiago Ventura is a Ph.D. candidate at the University of Maryland and researcher at the interdisciplinary Lab for Computational Social Science (iLCSS). venturat@umd.edu

Truth be told: Cognitive moderators of selective sharing of fact-checks on social media

When do users share fact-checks with their peers? We describe a survey experiment (N = 2,041) conducted during the 2019 presidential election in Argentina measuring the propensity of voters to share corrections to political misinformation that randomly confirm or challenge their initial beliefs. In line with processes of motivated reasoning, we find evidence of selective sharing—the notion that individuals prefer to share pro-attitudinal rather than counter-attitudinal fact-checks. This directional effect, however, is regulated by the type of adjudication made by the fact-checking organization, such that sharing increases for attitude-consistent validations (i.e., 'true' ratings) but decreases for attitude-consistent refutations (i.e., 'false' ratings). Experimental results are partially confirmed with a regression discontinuity analysis of observational data of Twitter shares collected during a televised debate of the same election. Our findings suggest that fact-checking organizations could selectively increase exposure to their verifications on social media by validating correct information (e.g., 'It is true that vaccines prevent COVID-19') or reduce exposure to them by refuting incorrect claims (e.g., 'It is false that vaccines do not prevent COVID').

Keywords: misinformation; fact-checking; motivated reasoning; experiment

Introduction

Political fact-checking has become a central effort against the prevalence of misinformation. Defined as "the practice of systematically publishing assessments of the validity of claims made by public officials and institutions with an explicit attempt to identify whether a claim is factual" (Walter, Cohen, Holbert, & Morag, 2020, p. 350), fact-checking is a global phenomenon, with scores of initiatives spearheaded by news organizations, independent media, and NGOs (Graves, 2018). The popularity of fact-checking stems, in part, from its efficacy as a remedy against misinformation. A recent meta-analysis found that fact-checking messages are successful at reducing misperceptions, even after a single exposure (Walter et al., 2020). However, the same meta-analysis found that this positive effect is conditioned by

context, audience, and message characteristics.

In the current article, we expand on the literature of moderators of fact-checking effects by focusing on the practice of sharing fact-checks. This is important because the more people rely on social media for political information, the more likely it is that the reception of fact-checking messages depends on their social media visibility (Shin & Thorson, 2017). Likewise, message diffusion has become an important complement to studies on message exposure and selection (Amazeen, Vargo, & Hopp, 2019; Valenzuela, Piña, & Ramírez, 2017). Thus, the current study examines the type of fact-checks that are more likely to be shared, including the way in which they process corrections. To do that, we conducted two studies in Argentina during the 2019 elections: an online survey experiment (N = 2,041) and an observational study of Twitter data.

Our results confirm the existence of partisan selective sharing, that is, individuals prefer to spread pro-attitudinal fact-checks over counter-attitudinal fact-checks (Ekstrom & Lai, 2020; Shin & Thorson, 2017). However, we find an important boundary condition for this directional bias: pro-attitudinal confirmations (e.g., 'It is TRUE that vaccines PREVENT against COVID-19') are more likely to be shared than pro-attitudinal refutations (e.g., 'It is FALSE that vaccines DO NOT PREVENT against COVID-19'). Furthermore, we show that this happens because validations of prior congruent beliefs yield considerably more cognitive engagement than refutations, in line with processes of motivated reasoning. These results have an important practical implication: to make their fact-checks more visible, fact-checking organizations should evaluate the advantages of using 'true' ratings over 'false' ratings.

The organization of this article is as follows. We first review work on fact-checking effects on attitudes and behaviors, with a focus on motivated reasoning as a theoretical framework. We then describe the experimental design, which exposes respondents to both claims and different verifications of these claims, all of which are either cognitively congruent

or incongruent with one of the major political coalitions in Argentina. The next section presents the results of the experiment, followed by several robustness checks. Finally, we replicate part of the experimental findings with observational data from Twitter during a presidential debate using a regression discontinuity design. We conclude with a discussion of theoretical and practical implications, and directions for future research.

Motivated reasoning and selective sharing of fact-checks

Considerable work exists on what motivates people to share political misinformation (e.g., Chadwick, Vaccari, & O'Loughlin, 2018; Guess, Nagler, & Tucker, 2019; Pennycook & Rand, 2019; Wagner & Boczkowski, 2019). In contrast, fewer studies address the motivations for sharing corrections to political misinformation. Still, prior research suggests that it is a behavior motivated by partisan goals (Amazeen, Vargo, & Hopp, 2019; Mattes & Redlawsk, 2020; Shin & Thorson, 2017). Just as people selectively prefer to *consume* ideologically congenial information (Garrett, 2009; Mummolo, 2016), they also selectively prefer to *share* ideologically congenial information (Ekstrom & Lai, 2020; Lewendosky et al., 2012).

The literature on motivated reasoning (Kunda, 1990) provides an explanation for partisan selective sharing. As a process by which people acquire, evaluate, and form related judgments about new information, motivated reasoning focuses on two primary goals: accuracy and directional motivations (Bolsen & Palm, 2019). An accuracy goal is defined by information-processing that seeks to form a precise, unbiased picture of the world. A directional goal, in contrast, prompts individuals to process information that supports or protects their pre-existing beliefs and identities. As Pietryka (2016, p. 369) noted, 'The guiding principle for work in motivated reasoning is that all reasoning is motivated, and, for political decisions, accuracy motivations tend to be weak while directional motivations are the norm.'

Prior research demonstrates the prevalence of directional goals over accuracy goals in the political realm. In an oft-cited experiment conducted in the United States, Taber and Lodge (2006) measured participants' opinion on gun control and affirmative action before and after being exposed to several pro- and contra-arguments. They found that participants were more likely to: (a) select arguments confirming their beliefs (i.e., a confirmation bias), (b) perceive congenial arguments as stronger (i.e., a prior attitude effect), and (c) invest more time counter-arguing against uncongenial arguments (i.e., a disconfirmation bias; see also Strickland, Taber, & Lodge, 2011; Hameleers et al., 2021). As Kunda (1990) concluded, when confronted with contrary evidence, people often become 'motivated skeptics.' Indeed, when corrective messages serve to undermine attitudes or values, individuals respond defensively to preserve their existing viewpoints (Lewandowsky et al., 2012). Conversely, there is evidence that political engagement and attitude strength can motivate sharing misinformation (Petersen et al., 2018; Valenzuela, Halpern, Katz & Miranda, 2019).

Prior research has linked the diffusion of political fact-checking to directional goals. Shin and Thorson (2017) predicted that pro-attitudinal fact checks are more likely to be shared than counter-attitudinal fact-checks. Their analysis of Twitter data collected during the 2012 US presidential election generally support that the sharing of political fact-checks suffers from confirmation bias (also Amazeen et al., 2019; Freiling et al., 2021). At the same time, there is little evidence that fact-checking activates a disconfirmation bias by which some individuals exposed to uncongenial corrections double down on their inaccurate beliefs (Guess & Coppock, 2018; Wood & Porter, 2019). Based on this body of work, our first hypothesis states that:

H1: Attitude-consistent fact-checks are more likely to be shared than counter-attitudinal fact-checks.

Although this hypothesis is rather confirmatory, it is necessary to delve into the moderators of partisan selective sharing (Li, 2020). Furthermore, there is limited work on whether partisan selective sharing of fact-checking messages results from motivated reasoning or some other process (e.g., cognitive dissonance, negativity bias, judgements of content quality, and so forth; see Stroud, 2011). Hence, we will follow Taber and Lodge's (2006) model of motivated reasoning and—unlike prior studies—explore the response time (i.e., time-to-retweet) associated with political fact-checks. Shall we find that the sharing of congenial fact-checks has a shorter response time compared to uncongenial fact-checks, we will have clearer evidence that partisan sharing of fact-checks is a process regulated by motivated reasoning (also Lodge & Taber, 2005). This, in turn, will enable us to test the boundary conditions for this process, which we explain next.

Differential effects of confirmations and refutations

Taber and Lodge (2006) noted that directional goals and subsequent selective information processing are driven by automatic affective processes. For most voters, political messages are 'hot,' that is, upon exposure to any political stimulus, associated attitudes come to mind automatically, even prior to semantic information. This primacy of affect means that upon being confronted with a message related to a presidential candidate, for instance, voters' feelings (i.e., likes or dislikes) for the candidate are aroused before conscious awareness of other associations (e.g., that the candidate is honest or dishonest, conservative or liberal, and so forth). These 'hot cognitions,' in turn, establish the direction and strength of biases in information processing as well as the impact of these biases on subsequent behavior.

The existence of 'hot' cognitions implies the existence of 'cold' cognitions as well. While the former result from processes that are spontaneous, fast, and below conscious awareness, the latter result from processes that are deliberative, slow, and self-aware.

Attendance to identity-relevant media sources may increase belief in—and sharing of—messages that bolster partisan attitudes and negative perceptions of out-group media (Shin & Thorson, 2017; Robertson et al., 2020; Slater, 2007). Analytic thinking and actively open-minded thinking, in contrast, can result in greater acceptance of counter-attitudinal corrective messages (Martel et al., 2021; see also Hameleers et al., 2021). Thus, one would expect that partisan selective sharing of fact-checks increases with hot cognition and decreases with cold cognition.

In the context of fact-checking, it is more likely for hot cognition to arise from pro-attitudinal messages labeled 'true' than it is for pro-attitudinal messages labeled 'false.' To understand this argument, it is important to consider that the defining feature of fact-checking is that of adjudicating the level of truth (or falseness) of political messages. When a fact-checking organization determines that a content is 'true' or 'false,' it validates the beliefs or claims of some users over the beliefs or claims of other users, such that the adjudication stage can be perceived as a decision that benefits one party and injures another. As with judiciary decisions, adjudicating a message as 'true' means awarding the source of that message as a carrier of truth, while adjudicating a message as 'false' means inflicting harm on the source of misinformation as a carrier of deceit. This means that when exposed to a fact-check, users not only are exposed to a content that may be congenial (or not) to their beliefs, but they also receive a cognitive 'award' or 'punishment' as the specific adjudication informs them whether their beliefs regarding the initial message are correct or incorrect.

Importantly, this system of 'awards' and 'punishments' can regulate the process of sharing pro-attitudinal fact-checks by increasing automatic processing—typical of hot cognition—toward the former while decreasing it toward the latter. As Redlawsk (2002, p. 1023) explained, 'it requires no effort to accept what one already knows is true.' Thus, the validation of a politically congruent belief (e.g., 'You thought this was true and, yes, we

verified it as true') should increase automatic processing and behavioral intention towards that message compared to a refutation that is also politically congruent (e.g., 'You thought this was false and yes, we verified it as false'). Put another way, a pro-attitudinal confirmation of a claim will require less cognitive effort when it validates content that was already available in memory and cognitively congruent. Therefore, an adjudication of 'true' will fit neatly with the original pro-attitudinal message and increase sharing. Meanwhile, an adjudication of 'false' will force an evaluative response (i.e., cold cognition), which should reduce the likelihood of sharing. In hypothesis form, the expectation is that:

H2: Attitude-consistent fact-checks rated 'true' are more likely to be shared than attitude-consistent fact-checks rated 'false.'

With this hypothesis, we are positing a moderating effect of the adjudication made by the fact-checking organization on partisan selective sharing of verifications, as specified in H1. Still, both hypotheses are consistent with the notion that motivated reasoning explains the diffusion of political fact-checks.

Materials and methods

Experimental design

To test the hypotheses, we implemented a two-stage, two-arms experiment exposing respondents to a social media post and then randomly assigning them to a 'true' or 'false' adjudication.[1] The experiment was embedded in an online survey fielded between April 27 and May 5, 2020—five months after the national election—by the polling firm Netquest. The nationally representative sample included 2,041 adult respondents from the 24 Argentinean provinces, stratified by gender, age, and education to match current census data. The survey flow is summarized in Figure 1, with two different implementations: an anti-government tweet (hereafter referred as *Audífono*) and an anti-opposition tweet (hereafter referred as

Ofelia). The two implementations had identical designs but opposite political leanings, to ensure that government and opposition supporters would be tested with pro-attitudinal and counter-attitudinal messages.

[INSERT FIGURE 1]

After exposure to the initial tweet that was to be fact-checked later, respondents were asked (Q1) whether they would engage with the publication (i.e., 'like,' 'retweet,' 'reply,' or 'ignore') and (Q2) how it made them feel (i.e., 'angry,' 'joyful,' 'sad,' 'disgusted,' 'stressed'). Subsequently, respondents were shown a fact check exposing them to a 'true' or 'false' adjudication. This was followed by questions asking respondents (Q3) whether they would engage with the publication, (Q4) whether they believed the adjudication was credible (i.e., the initial tweet was 'surely true,' 'likely true,' 'likely false,' or 'surely false'), and (Q5) how it made them feel. We included measures of time-to-respond at the first and second stage, which were used to evaluate hot/cold cognition. Between exposure to the initial tweet and the fact check we asked a battery of attitudinal questions capturing the political preferences and social media behavior of the respondents.

Stimuli

We ran two versions of our experiment. The first one used an anti-government tweet accusing former Argentina president Mauricio Macri of using an earpiece (*audífono* in Spanish) during the televised debate of 2019. The second one used an anti-opposition tweet accusing the member of the Buenos Aires city legislature Ofelia Fernández of collecting a hefty salary despite not having completed a high school degree. To maximize partisan recognition, the tweets were attributed to the accounts of two leading journalists that during the campaign were widely recognized as partisan and aligned with either the government or the opposition (*@lanataenel13* and *@robnavarro*, respectively). The text and hashtag included replicated

those that were used by Chequeado. Subsequently, both treatment groups received either a 'true' or 'false' adjudication, formatted with the design of the leading Argentinean fact-checking organization, Chequeado.

Because the survey was conducted five months after the election, the probability of pre-treatment effects is limited to highly informed, partisan respondents who may had vague familiarity with the stimuli. However, it is unlikely that most respondents recall the stimuli: as shown in the analysis of the observational data, the original claims garnered a few thousand retweets only and the number of fact-checking interventions by *Chequeado* during the election campaign was above 300.

Variables

Our main dependent variable, sharing, was measured in binary fashion: it takes the value of 1 if the respondent indicates an intent to retweet the stimulus and 0 otherwise. As a robustness check, we also measured engagement, a variable that takes the value of 1 if the respondent retweets, likes, or replies to the tweets, and 0 otherwise.

Our first set of independent variables is the respondent's vote intention (e.g., 'If the second round of the presidential election were to take place next week, whom would you vote for?'). A total of 799 respondents expressed a likely vote for the current president Alberto Fernández (39.1%); 747 supported former president and opposition leader Mauricio Macri (36.6%); and 495 indicated that they would vote blank (24.2%). To produce separate estimates for pro- and counter-attitudinal alignments with Fernández and Macri, those voting blank (i.e., independents) are the reference category in the statistical analyses.

To capture the processes implied by the hot cognition literature, our second set of independent variables measures the time-to-respond to the sharing questions.[2] The median reaction time was 5.4 seconds for *Audífono* and 5.1 seconds for *Ofelia*, with 95% of the

observations in the range between 2 and 16 seconds. In both designs, we gauged the conditional effect of a fast response to the initial tweet on the likelihood of retweeting the true or false adjudications attributed to Chequeado on the second stage of the experiment.

A third set of independent variables, analyzed as controls, measure the propensity to share any adjudication. Studies of digital nudging show that respondents vary in their propensity to share content and that sharing may be incentivized (Mirsch et al. 2018). Heterogeneity in the initial propensity to share is an important source of noise and, accordingly, one for which we control. Because emotional arousal may increase sharing (Berger & Milkman, 2011), we also included emotions triggered by the fact-check. In addition, we measured age, gender, frequent use of Facebook, frequent use of Twitter, support for fact-checking, political engagement online, ideology, trust, and political knowledge, all of which may be potential confounders. Restricted models with only the key variables are reported in the main text; unrestricted models with all covariates associated with sharing are reported in the online supplementary file. The estimations, in turn, take the form of a general linear model with a binomial distribution and a logit link function:

$$Sharing \ (\pi)$$

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 Fernandez + \beta_2 Macri + \ldots + \beta_k Var_k$$

We estimated separate equations for the original messages, *Audífono* and *Ofelia*, and for 'true' and 'false' adjudications. Based on the hypotheses, we expect the findings detailed on Table 1.

[INSERT TABLE 1]

Results

To facilitate the interpretation of our results, let us first walk through one of the two versions

of the experiment. The leftmost image in Figure 2 presents the original tweet in the *Audífono* experiment by *@robnavarro*: 'He deceives us till the end: Macri has an #earbud inside his ear. Did we stay through the full *#DebateAr2019* for this?' The image reinforces the text and was retrieved from tweets that circulated during the presidential debate on October 20, 2019.

[INSERT FIGURE 2]

The plot to the right of the anti-Macri *Audífono* tweet shows the share of respondents that like, retweet, and reply the original message. We see that the proportion of Fernández voters who would retweet (28%) the original message is significantly higher than the proportion of Macri voters who would do the same (6%). In all, voters with pro-attitudinal affinity to the *Audífono* tweet are more likely to share it. More interesting, however, is the effect of the 'true' or 'false' rating of the fact-check on sharing behavior. The plots on the right of Figure 2 show that the pro-attitudinal, 'true' ('verdadero' in Spanish) adjudication displays a high retweet rate (37% among Fernández voters). The 'false' ('falso') adjudication, however, did not elicit the same intention to share among those respondents most inclined to agree with it (15% among Macri voters). Thus, while being proven right about a pro-attitudinal belief seems to increase sharing, being proven right about a counter-attitudinal belief does not seem to elicit an equally strong response. As we will show, this is not simply a feature of the *Audífono* treatment; similar results were obtained with the *Ofelia* treatment.

While the descriptive results are consistent with H1, we formally tested the first hypothesis by estimating a series of linear models. Table 2 presents the estimates of the likelihood of sharing the first tweet and the likelihood of sharing the fact-checks rated 'true' and 'false,' with vote choice dummies as independent variables. In line with H1, pro-attitudinal fact-checks are shared more widely than counter-attitudinal fact-checks. Specifically, Fernández voters are significantly more likely to share the pro-attitudinal fact-check confirming the veracity of the *Audífono* tweet and less likely to share the counter-

attitudinal fact-check refuting it. The exact opposite is true for Macri voters. Similar findings are shown for the *Ofelia* experiment, with Macri voters sharing more widely the confirmation than the refutation, and Fernández voters sharing more widely the refutation than the confirmation. Taken together, the evidence supports H1.

Readers may also see that the constant in the linear models is negative and larger for the 'false' adjudication. In both experiments, the 'true' adjudication is more widely shared than the 'false' adjudication, with larger differences among pro-attitudinal users. In the adjudication models we also control for the likelihood of sharing the initial tweet and for the overall propensity to share tweets. Results with the controls are in the online supplementary file and produce substantively similar results. Models (2), (3), (5), and (6) in Table 2 also show that individuals who shared the original tweet were more likely to also retweet both the 'true' and 'false' adjudication. Again, however, the conditional effect of sharing the original tweet on the 'true' adjudication is larger than for the 'false' adjudication. By controlling for the likelihood of sharing all initial tweets (i.e., the 'trigger finger' effect), we know that the propensity to retweet the correction because the first tweet was shared is different from the overall incentive to share content.

[INSERT TABLE 2]

Thus far, the results are consistent with H2. To make better sense of the linear models, we estimated the predicted probabilities of sharing based on the estimates reported in Table 2. As shown in Figure 3, pro-attitudinal confirmations garner significantly higher probabilities of sharing than pro-attitudinal refutations. It is remarkable how similar the results are across the two treatments, which all but eliminates the possibility that the asymmetry between pro-attitudinal validations and refutations can be explained by partisan features. Thus, the experimental data support H2.

[INSERT FIGURE 3]

While the findings thus far are consistent with our two hypotheses, they do not provide evidence that they are a consequence of motivated reasoning or something else. Thus, we now evaluate whether the asymmetry between pro-attitudinal confirmations and refutations (now in the constant) can be explained by hot cognition—the automatic response to the initial tweet and to the pro-attitudinal confirmation. We do that by measuring the conditional effect of a faster sharing time to the original tweet (i.e., time-to-share) on the fact-checking adjudication. Figure 4 reports the marginal effect of a 'true' adjudication conditional on the time-to-share of the initial tweet, with separate estimates by vote choice. Modeling the marginal effect of the 'true' adjudication allows us to better understand how it differs from the false adjudication. The upper plot models the marginal effect on retweeting the adjudication and the lower plots illustrates the marginal effect on engagement with the adjudication (a combination of likes, retweets, and replies).

Readers will readily note that retweeting the original tweet magnifies the effect of the 'true' treatment. Indeed, actively engaging with the tweet is different from simply reading it. In the experiment, interacting with the original tweet by retweeting, liking, or replying amplifies results, with larger inter-party differences. Consider for example the case of *Audífono* (upper-left plot). The marginal effect of a 'true' adjudication is largest among the Fernández voters who rapidly retweeted the original tweet against Macri. By contrast, the marginal effect of the 'true' adjudication is negative and statistically significant for Macri supporters who retweeted the initial tweet. The differences in the marginal effect of a 'true' adjudication are largest when comparing Fernández and Macri voters at lower reaction times. Also, overall engagement is at a maximal difference in the pro- and counter-attitudinal effect of 'true' adjudications. Thus, Figure 4 shows that fast and automatic responses to the initial tweet—the typical markers of hot cognition—are key to understanding the difference between

a ‘true’ and ‘false’ adjudication, as measured by the average marginal effect of ‘true’ conditional on time-to-retweet and time-to-engage.

[INSERT FIGURE 4]

Partial replication with observational data

In the previous section, we provided evidence of selective sharing of ‘true’ and ‘false’ fact-checks with two experimental implementations that favoured either the government or the opposition. We highlighted a robust moderating effect in the form of a behavioural preference for pro-attitudinal confirmations compared to pro-attitudinal refutations and demonstrated that this effect is explained by a process consistent with jumping from hot to cold cognition. To assess the robustness of experimental findings, in this section we introduce supporting evidence of partisan selective sharing as stated in H1 using real-world social media data, collected during the second presidential debate of the 2019 election in Argentina. Specifically, we analyzed a rumor that circulated in Twitter by supporters of candidate Alberto Fernández and a ‘false’ adjudication to that rumor by the fact-checking organization Chequeado. We then evaluated the effect of Chequeado’s fact-check by implementing a regression discontinuity design (RDD), which describes differences between the original content as well as the pro- and counter-attitudinal refutation by Macri and Fernández voters, respectively.

Data Collection

During the two weeks prior to the election, we collected a total of 3,813,298 tweets using the following search strings: ‘Macri’, ‘Fernández’, ‘Peronismo’, ‘Cambiemos’, ‘debate’, ‘elección’, ‘auricular’, and ‘audífono.’ To build our sample, we used the Python base program Twarc to access both Twitter streaming and RESTful APIs. Whereas the former lets users capture tweets in real time, the latter allows access to a temporary repository of tweets that includes a large sample of all tweets published during the week prior to the query.

Prior to the analysis, we filtered singletons (i.e., one-time users) and tweets posted in languages other than Spanish. Using the *cluster* function in *igraph* (Csardi, 2006), we identified the primary connected cluster, eliminating users connected in smaller networks as well as users with low activity (namely, *in-degree = 0* or *out-degree* ≤ 3). The primary connected clusters contained the main networks that were politically engaged during the debate. Last, we implemented a random-walk community detection algorithm to identify the main communities and proceeded to visually identify those connected to the two leading presidential candidates. Those communities closely correspond to the two political groups competing in the election, the pro Macri coalition (Juntos por el Cambio, JxC) and the pro Fernández coalition (Frente de Todos, FdT). The online supplementary file provides the list of the top 30 users in each of the communities, which were validated by the authors to ensure they included the leading authorities of the candidates' communities. The curated network contained 91,982 high activity users and 1,250,030 tweets-retweets.

For the regression discontinuity design, we restricted the estimation to those retweets that discuss the *Audífono* rumor, that is, that Macri was using a hearing aid during the October 20, 2019 televised debate. For this task, we simply searched mentions to the following words: 'audífono,' 'auricular,' 'oído,' and 'oreja.' These publications comprise a small subset of 3,600 tweets published within the 6-hour window of the presidential debate.

The first publication that mentions *Audífono* appeared at 10:36 PM, October 20, 2019, over an hour after the beginning of the debate, when House Representative Araceli Ferreyra published a tweet that displayed an image of Macri and accused the presidential candidate of using an earpiece. Chequeado published a fact-check on the *Audífono* rumor, adjudicating it as 'false,' on October 21, 2019, at 1:20 AM. In the 21 hours that followed the correction, a total of 1,376 users shared the correction, 70.6% of which were users identified with the Macri coalition in the primary connected network. Meanwhile, 28.4 % were published by

users identified with the Fernández community. As shown in Table 3, in the three hours prior to Chequeado's adjudication, most of the tweets spreading the rumor were shared by users in the Fernández network.

[INSERT TABLE 3]

Model

The simple descriptive results from Table 3 are a strong indicative of the effects of pro-attitudinal preferences (H1). However, these simple models do not allow us to estimate the effect of Chequeado's intervention. Thus, we used an interrupted time series analysis, a variety of regression discontinuity designs (RDD) in which the running variable is time (Morgan & Winship, 2005). Twitter data is ideal for our approach because of the granularity and high frequency of tweets. Our primary parameter of interest is the change in social media users' behavior upon the correction.

The dependent variable is time-to-retweet, which captures changes in reaction time on users' behavior before and after the correction. It was measured with the number of seconds elapsed from the time a tweet is posted by a user to the time it is retweeted by a second user. Our unit of analysis is therefore any retweet collected using the methods described above, from which we retrieved information about the time, the author of the tweet, and the user who retweeted the original message. Previous research has extensively used time-to-retweet to understand heterogeneity on content propagation, news sharing, and activation on Twitter (Aruguete & Calvo, 2018; Lee et al., 2015; Stieglitz & Dang-Xuan, 2013).

The exact time of the Chequeado correction is the cut-off of the discontinuity regression model. Indeed, because we know exactly when the correction was published, there is no measurement error on the cut-off. Regression discontinuity models assume that effects are continuous at the cutoff. When dealing with time as a running variable, the continuity

assumption requires that no omitted variable that systematically affects the outcome also changes after the intervention. Given that we have the precise minute when the adjudication was granted, and that we only consider data from a small window of hours around the cutoff, it is reasonable to assume that this assumption holds. To estimate the models, we followed the recommended setting of local polynomial, a triangular weighting function and data-driven bandwidth selection (Calonico et al., 2014). To ensure the results are robust to different modeling choices, a variety of model specifications are included in the online supplementary file.

Results

The upper plot on Figure 5 presents the evolution of the *Audífono* rumor on Twitter. The vertical axes report the log of the time-to-retweet, with lower values indicating that users are more engaged (lower reaction time), and the horizontal is centered at the moment when Chequeado published the fact-check with the correction, adjudicating it as 'false.' Each point represents multiple retweets, and the lines are from LOESS smoother models before and after the correction. Intuitively, the upper plot on Figure 5 shows two important patterns. As the *Audífono* rumor first appeared, Macri supporters exhibited lower time-to-share, reacting faster to messages discussing the issue online. This pattern changed, and in the hours after the debate ended (red line in the right plot), both communities were tweeting at roughly the same speed. After Chequeado published the fact-check on its website and on social media, both communities sped up their sharing behavior and decreased their time-to-retweet, as the discontinuity at the centered gray line shows.

[INSERT FIGURE 5]

Based on our theoretical framework, we would expect that Macri supporters, when exposed to a pro-attitudinal refutation, would show lower time-to-retweet when compared to

Fernández supporters, for whom the correction is counter-attitudinal. The bottom plot of Figure 5 presents this comparison using the regression discontinuity models. Across all models, we observe lower time-to-retweet (i.e., reaction time) at adjudication. However, just as we expected, the decline in time-to-retweet is far larger among the supporters of Macri. Meanwhile, the effect of Chequeado's adjudication among Fernández supporters is more modest and less robust —the point-estimates are not statistically different from zero on models with a larger bandwidth. Further supporting H1, the pro-attitudinal refutation reduced time-to-retweet and increased sharing by Macri supporters, while the counter-attitudinal refutation failed to do the same among Fernández supporters, who were confronted with an adjudication penalizing sharing in their own community.

Discussion

Past research on people's motivation for sharing fact-checks to political misinformation shows that partisan goals drive this behavior, though little is known about the exact mechanisms explaining this process. Building on the literature on partisan selective sharing and motivated reasoning, this paper aimed to shed light on some of cognitive factors that may have a significant effect on individuals' engagement with verifications about political issues in online settings. Consistent with research exploring the role of directional goals in the context of motivated reasoning, our experiments show that users favor content that is congenial to their beliefs, especially when fact-checking organizations 'award' such congruent beliefs by rating them as 'true' instead of 'false.' That is, a pro-attitudinal fact-check labeled 'true' (i.e., a validation) is more likely to be shared on social media than an equally congenial fact-check labeled 'false' (i.e., a refutation).

This has theoretical and practical implications. We demonstrate that most participants exhibit faster rection times (i.e., time-to-retweet) to validations compared to refutations, even

when both are politically congruent with participants' political identities. This result suggests that sharing fact-checking messages is regulated by the 'hot cognition' hypothesis, which posits that political affairs are affectively charged and that this affective charge is automatically activated upon exposure to the concept (Lodge & Taber, 2005; Taber & Lodge, 2006). It could be argued, then, that corrections to political misinformation—even when attributed to non-partisan, professional news outlets—will be processed by most individuals like a partisan message, that is, less as an objective, quality piece of information and more like a subjective, biased one. This is consistent with prior research showing that participants' preexisting beliefs, ideology, and knowledge regulate the ability of fact-checking to correct political misinformation (Walter et al., 2020). To borrow from persuasion research (O'Keefe, 2015), when it comes to the diffusion of political fact-checking, receiver factors seem to matter more than either communicator or message factors.

From a practical point of view, fact-checkers should consider presenting their work with a 'true' adjudication more often. Our findings suggest that social media users are more inclined to spread verifications than corrections debunking rumors and misinformation. The primacy of verification is consistent with extant research on proper debunking of misinformation (Lewandosky et al., 2020), which stresses the importance of stating the truth first instead of providing a simple retraction (e.g., 'this claim is not true'). This is because 'true' ratings provide a quick factual alternative to the causal 'gap' in explaining what happened if the misinformation is corrected. Having a causal alternative facilitates "switching out" the inaccurate information in an individual's initial understanding and replaces it with a new version of what happened.

As any research, there are several limitations that future research could tackle. Although we strived to increase the external validity of the survey experiment by using real sources of misinformation and fact-checking, there is always the possibility that the messages

read by our participants would not be selected in the real world. In fact, we used a forced exposure design. While the observed data used in the second part of the study partially replicated the experimental findings, thus alleviating external validity concerns, it was not a full replication. For that, additional fact-checks (some of them labeled 'false') should have been analyzed. Furthermore, in both studies we measured immediate effects and, thus, we do not know whether the longer-term effects of repeated exposure to fact-checks on user engagement. In the future, research on sharing fact-checking messages should examine emotional mechanisms as well as individual differences moderating the effects of fact-checking adjudications, such as trust on fact-checking organizations. Last, the results of this study derive from a single-country study, which leaves open the role played by contextual variables, such as political party systems, media systems, and cultural beliefs and values.

All in all, this article makes a modest but important contribution to current research on misinformation corrections. Considering the rising line of inquiry that addresses fact-checks sharing and diffusion, this paper constitutes a step toward that direction examining the cognitive moderators to further increase our understanding of how the reception and processing of political fact-checking messages on social media and their visibility.

Notes

References

Amazeen, A., Vargo, C. & Hopp, T. (2019). Reinforcing attitudes in a gatewatching news era: Individual-level antecedents to sharing fact-checks on social media. *Communication Monographs*, *86*(1), 112–132. https://doi.org/10.1080/03637751.2018.1521984

Aruguete, N., & Calvo, E. (2018). Time to# protest: Selective exposure, cascading activation, and framing in social media. *Journal of Communication*, *68*(3), 480–502. https://doi.org/10.1093/joc/jqy007

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research, 49*(2), 192-205. https://doi.org/10.1509/jmr.10.0353

Bolsen, T., & Palm, R. Motivated reasoning and political decision making. *Oxford Research Encyclopedia of Politics.*Retrieved from https://oxfordre.com/politics/politics/view/10.1093/acrefore/9780190228637.001.0001/acrefore-9780190228637-e-923

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, *82*(6), 2295–2326. https://doi.org/10.3982/ECTA11757

Chadwick, A., Vaccari, C., & O'Loughlin, B. (2018). Do tabloids poison the well of social media? Explaining democratically dysfunctional news sharing. *New Media & Society, 20*(11), 4255–4274. https://doi.org/10.1177/1461444818769689

Ekstrom, P. D., & Lai, C. K. (2021). The selective communication of political information. *Social Psychological and Personality Science*, *12*(5), 789–800. https://doi.org/10.1177/1948550620942365

Freiling, I., Krause, N. M., Scheufele, D. A., & Brossard, D. (2021). Believing and sharing misinformation, fact-checks, and accurate information on social media: The role of anxiety during COVID-19. *New Media & Society*. Advance online publication. https://doi.org/10.1177/14614448211011451

Garrett, R. K. (2009). Politically motivated reinforcement seeking: Reframing the selective exposure debate. *Journal of Communication*, *59*(4), 676–699. https://doi.org/10.1111/j.1460-2466.2009.01452.x

Graves, D. (2018). Understanding the promise and limits of automated fact-checking. Reuters Institute. Retrieved from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-02/graves_factsheet_180226%20FINAL.pdf

Guess, A., & Coppock, A. (2020). Does counter-attitudinal information cause backlash? Results from three large survey experiments. *British Journal of Political Science*, *50*(4), 1497–1515. https://doi.org/10.1017/S0007123418000327

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1), e4586. https://doi.org/10.1126/sciadv.aau4586

Hameleers, M. van der Meer, T. & Vliegenthart, R. (2021): Civilized truths, hateful lies? Incivility and hate speech in false information – evidence from fact-checked statements in the US. *Information, Communication & Society.* Advance online publication. https://doi.org/10.1080/1369118X.2021.1874038

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Lee, J., Agrawal, M., & Rao, H. (2015). Message diffusion through social network service: The case of rumor and non-rumor related tweets during Boston bombing 2013. *Information Systems Frontiers*, *17*(5), 997–1005. https://doi.org/10.1007/s10796-015-9568-z

Lewandowsky, S., Ecker, U. K. H., 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. https://doi.org/10.1177/1529100612451018

Lewandowsky, S., Cook, J., Ecker, U. K. H., Albarracín, D., Amazeen, M. A., Kendeou, P., Lombardi, D., Newman, E. J., Pennycook, G., Porter, E. Rand, D. G., Rapp, D. N.,

Reifler, J., Roozenbeek, J., Schmid, P., Seifert, C. M., Sinatra, G. M., Swire-Thompson, B., van der Linden, S., Vraga, E. K.,…, Zaragoza, M. S. (2020). *The debunking handbook 2020*. Available at https://sks.to/db2020. https://doi.org/10.17910/b7.1182

Li, J. (2020) Toward a research agenda on political misinformation and corrective information. *Political Communication*, *37*(1), 125-135. https://doi.org/10.1080/10584609.2020.1716499

Lodge, M., & Taber, C.S. (2005). The automaticity of affect for political leaders, groups, and issues: An experimental test of the hot cognition hypothesis. *Political Psychology, 26*(3), 455-482. https://doi.org/10.1111/j.1467-9221.2005.00426.x

Martel, C. Mosleh, M., & Rand, D. (2021). You're definitely wrong, maybe: correction style has minimal effect on corrections of misinformation online. *Media and Communication*, *9*(1), 120–133.https://doi.org/10.17645/mac.v9i1.3519

Mattes, K. & Redlawks, D. (2020). Voluntary exposure to political fact checks. *Journalism & Mass Communication Quarterly*, *97*(4), 913–935. https://doi.org/10.1177/1077699020923603

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge University Press.

Mummolo, J. (2016). News from the other side: How topic relevance limits the prevalence of partisan selective exposure. *Journal of Politics, 78*(3): 763–773. https://doi.org/10.1086/685584

O'Keefe, D. J. (2015). *Persuasion: Theory and research*. Sage Publications.

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Petersen, M.B., Osmundsen, M., Arcenaux, K. (2018). A "Need for Chaos" and the Sharing of Hostile Political Rumors in Advanced Democracies. Paper presented at the American Political Science Association, Boston, MA. 30 August-2 September.

Pietryka, M. T. (2016). Accuracy motivations, predispositions, and social information in political discussion networks. *Political Psychology*, *37*(3), 367–386. https://doi.org/10.1111/pops.12255

Redlawsk, D. P. (2002). Hot cognition or cool consideration? Testing the effects of motivated reasoning on political decision making. *Journal of Politics*, *64*(4), 1021–1044. https://doi.org/10.1111/1468-2508.00161

Robertson, C. T., Mourão, R. R., & Thorson, E. (2020). Who uses fact-checking sites? The impact of demographics, political antecedents, and media use on fact-checking site awareness, attitudes, and behavior. *The International Journal of Press/Politics*, *25*(2), 217–237. https://doi.org/10.1177/1940161219898055

Shin, J. & Thorson, K. (2017), Partisan selective sharing: The biased diffusion of fact-checking messages on social media. *Journal of Communication*, *67*, 233–255. https://doi.org/10.1111/jcom.12284

Slater, M. D. (2007). Reinforcing spirals: The mutual influence of media selectivity and media effects and their impact on individual behavior and social identity. *Communication Theory*, *17*(3), 281–303. https://doi.org/10.1111/j.1468-2885.2007.00296.x

Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of Management Information Systems*, *29*(4), 217–248. https://doi.org/10.2753/MIS0742-1222290408

Strickland, A. A., Taber, C. S., & Lodge, M. (2011). Motivated reasoning and public opinion. *Journal of Health Politics, Policy and Law*, *36*(6), 935–944. https://doi.org/10.1215/03616878-1460524

Stroud, N. J. (2011). *Niche news: The politics of news choice*. Oxford University Press.

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*(3), 755–769. https://doi.org/10.1111/j.1540-5907.2006.00214.x

Valenzuela, S., Halpern, D., Katz, J. E. & Miranda, J.P. (2019). The paradox of participation versus misinformation: Social media, political engagement, and the spread of misinformation, *Digital Journalism*, *7*(6), 802–823. https://doi.org/10.1080/21670811.2019.1623701

Valenzuela, S., Piña, M., & Ramírez, J. (2017). Behavioral effects of framing on social media users: How conflict, economic, human interest, and morality frames drive news sharing. *Journal of Communication, 67*, 803–826. https://doi.org/10.1111/jcom.12325

Wagner, M. C., & Boczkowski, P. J. (2019). The reception of fake news: The interpretations and practices that shape the consumption of perceived misinformation. *Digital Journalism*, *7*(7), 870–885. https://doi.org/10.1080/21670811.2019.1653208

Walter, N., Cohen, J., Holbert, L., & Morag, Y. (2020) Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, *37*(3), 350–375. https://doi.org/10.1080/10584609.2019.1668894

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, *41*(1), 135–163. https://doi.org/10.1007/s11109-018-9443-y

Table 1. Experimental design and hypotheses

	Stimuli:					
	Audífono			*Ofelia*		
Vote choice:	Initial tweet	Fact-checking adjudication: True	Fact-checking adjudication: False	Initial tweet	Fact-checking adjudication: True	Fact-checking adjudication: False
Fernández	Pro-attitudinal message **H1 (+)**	Pro-attitudinal confirmation **H2 (++)**	Counter-attitudinal refutation	Counter-attitudinal message	Counter-attitudinal confirmation	Pro-attitudinal refutation **H2 (+)**
Macri	Counter-attitudinal message	Counter-attitudinal confirmation	Pro-attitudinal refutation **H2 (+)**	Pro-attitudinal message **H1 (+)**	Pro-attitudinal confirmation **H2 (++)**	Counter-attitudinal refutation

Note: We expect sharing to increase for pro-attitudinal messages (+). We expect sharing spikes with pro-attitudinal confirmations (++). Shaded cells indicate higher sharing rates, with green for pro-attitudinal confirmations and red for pro-attitudinal refutations.

Table 2. Partisanship, cognitive congruence, and social media sharing

	Audífono			*Ofelia*		
	(1)	(2)	(3)	(4)	(5)	(6)
	Audífono, Original	Correction, True	Correction, False	Ofelia, Original	Correction, True	Correction, False
Fernández voter	0.98***	0.68***	0.12	-0.37***	0.19	1.19***
	(0.16)	(0.23)	(0.28)	(0.14)	(0.24)	(0.28)
Macri voter	-1.05***	-0.53*	0.71**	0.42***	0.53**	-0.14
	(0.22)	(0.27)	(0.28)	(0.13)	(0.23)	(0.31)
Retweeted original		2.61***	2.01***		2.98***	1.77***
		(0.26)	(0.29)		(0.20)	(0.25)
Overall sharing		0.49	0.65*		0.56*	0.84**
		(0.37)	(0.34)		(0.33)	(0.33)
Constant	-1.91***	-2.09***	-2.71***	-1.11***	-2.36***	-3.10***
	(0.13)	(0.20)	(0.24)	(0.10)	(0.21)	(0.27)
N	2,041	1,030	983	2,041	998	1,015
Pseudo R^2	.09	.28	.14	.02	.11	.13

Note: Standard errors are in parentheses. Shaded regions indicate pro-attitudinal match (cognitive congruence). Shaded gray for the original, shaded green for pro-attitudinal confirmation 'true,' and shaded red for pro-attitudinal refutation 'false.'

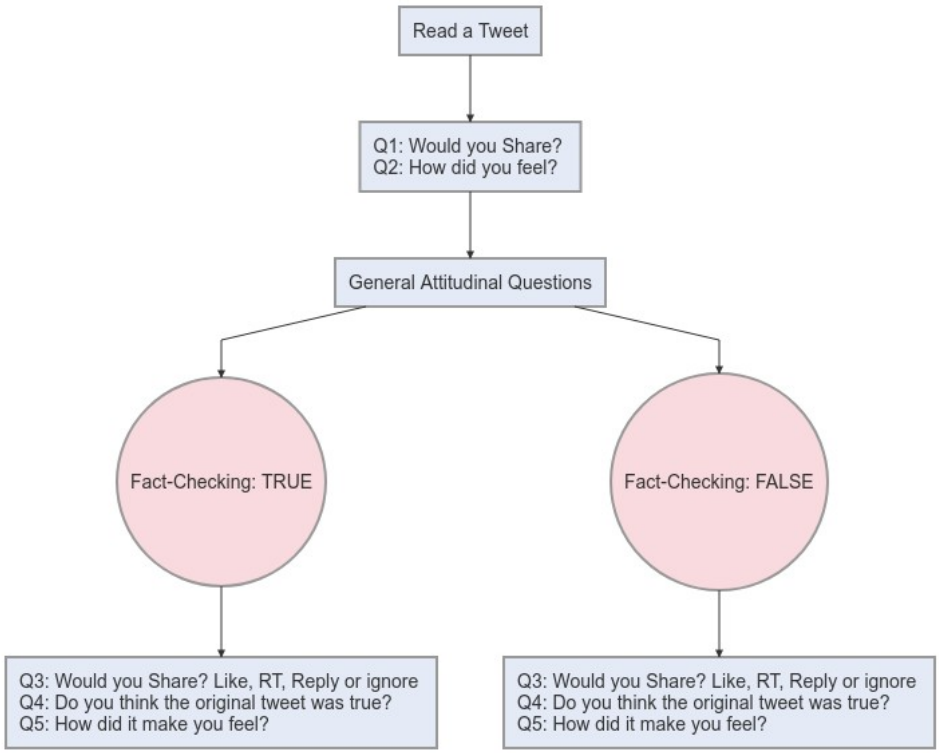
*** $p < .01$, ** $p < .05$, * $p < .10$

Table 3. Sharing of *Audífono* content before and after the adjudication by Chequeado (six-hour window)

	Before (3 Hours)	After (3 hours)
Fernández supporters	69.63%	7.16%
Macri supporters	3.58%	14.12%
Others	4.93%	0.58%
N	1,034	
	Before (3 Hours)	After (21 hours)
Fernández supporters	26.12%	20.06%
Macri supporters	1.34%	49.91%
Others	1.85%	0.73%
N	2,757	

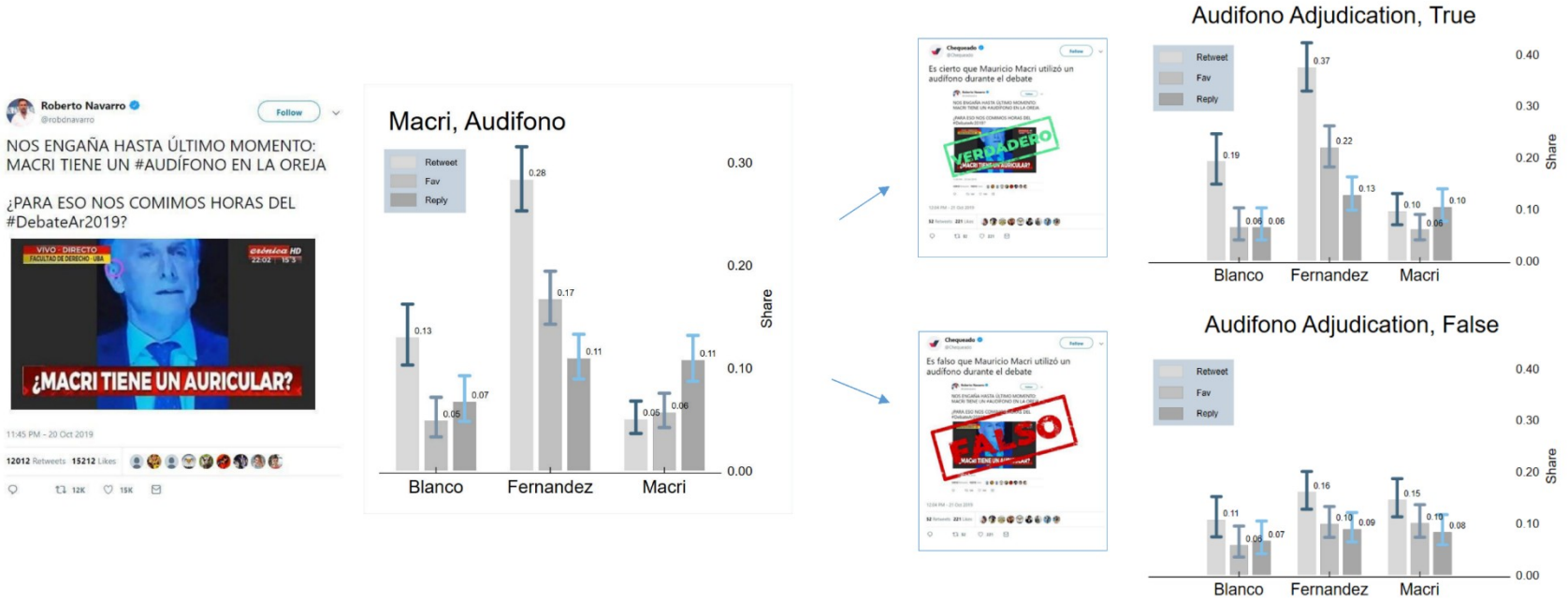
Note: Data collection described in the online supplementary file. Shaded regions indicate pro-attitudinal match (cognitive congruence). Shaded green for the original, pro-attitudinal message, and shaded red for the pro-attitudinal refutation 'false.'

Figure 1. Design of each of the two fact-checking experiments, *Audífono* and *Ofelia*



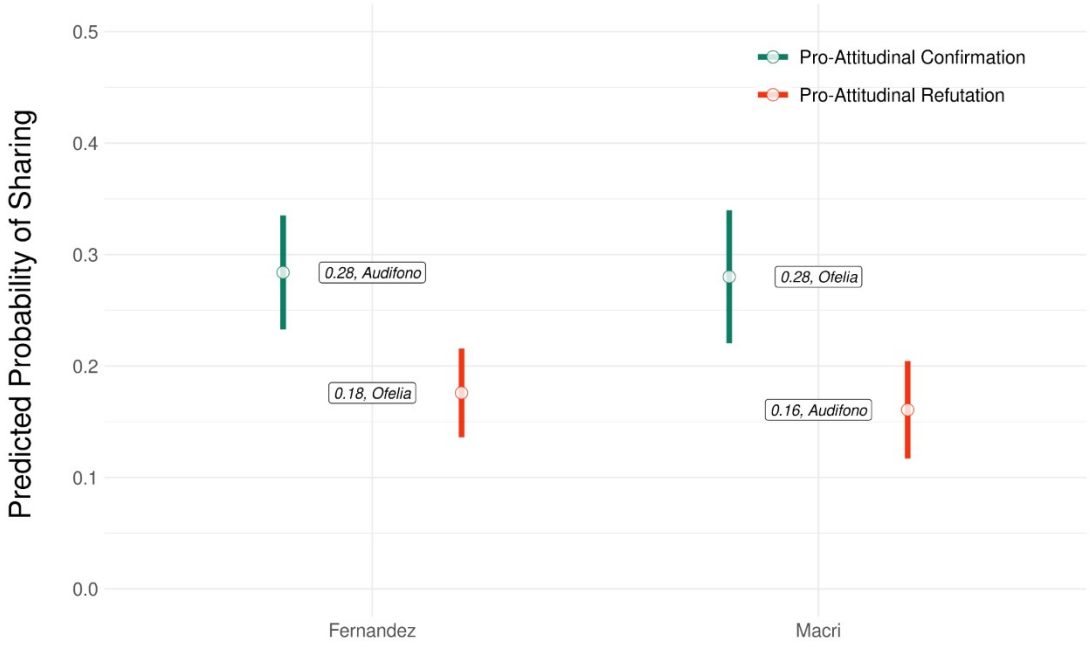
Note: Survey respondents were treated with the original tweet. After exposure they were distracted with attitudinal questions. Then they were randomly assigned to treatment and control groups exposed to 'true' or 'false' adjudications. Response time measures the time-to-respond for each of the instruments.

Figure 2. Experimental design with initial tweet and ‘true’ and ‘false’ adjudications



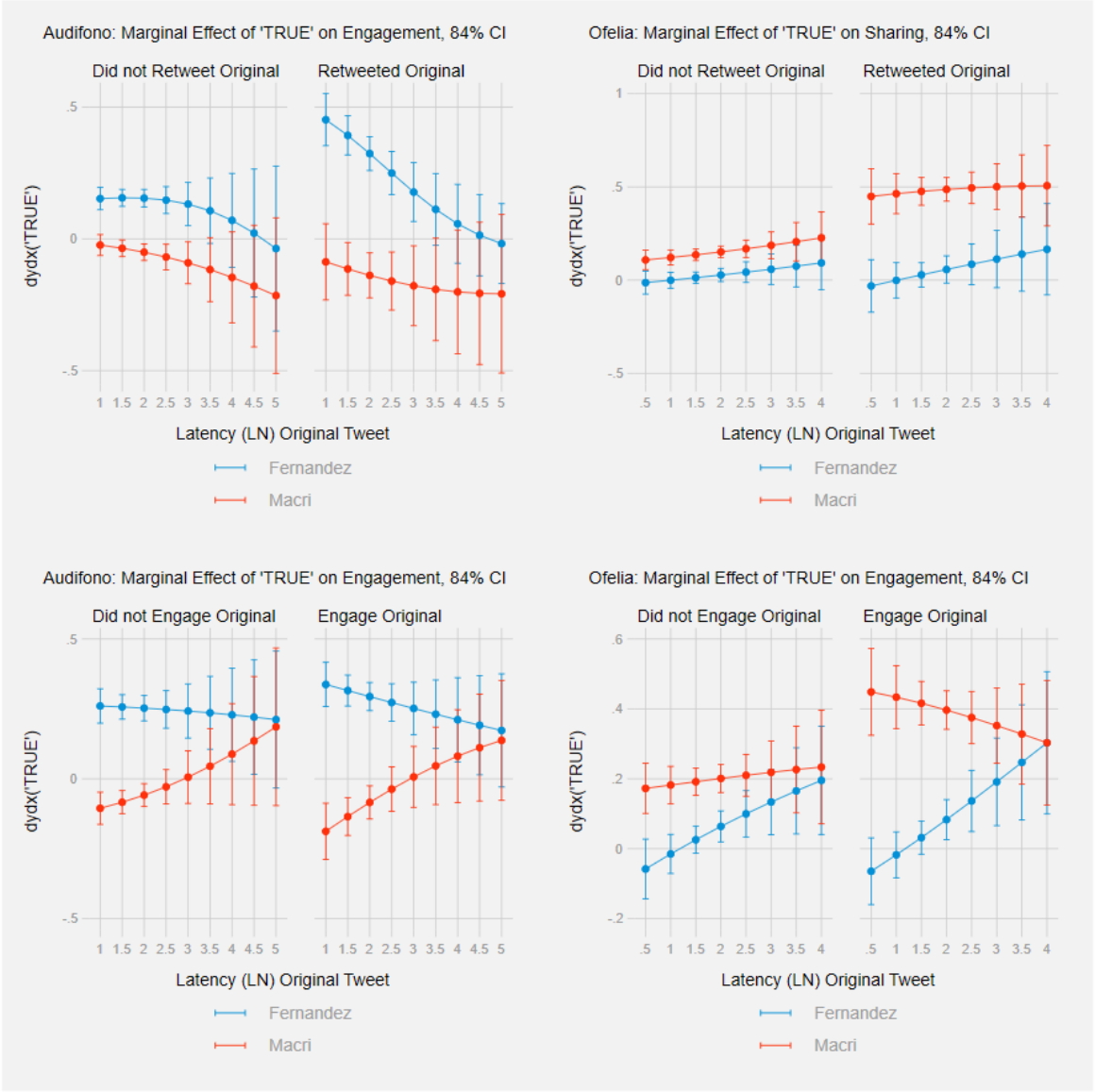
Note: Sharing the original tweet and sharing the ‘true’ or ‘false’ adjudications in the first experiment, *Audífono*.

Figure 3. Sharing of the original tweet and the ‘true’ and ‘false’ adjudication



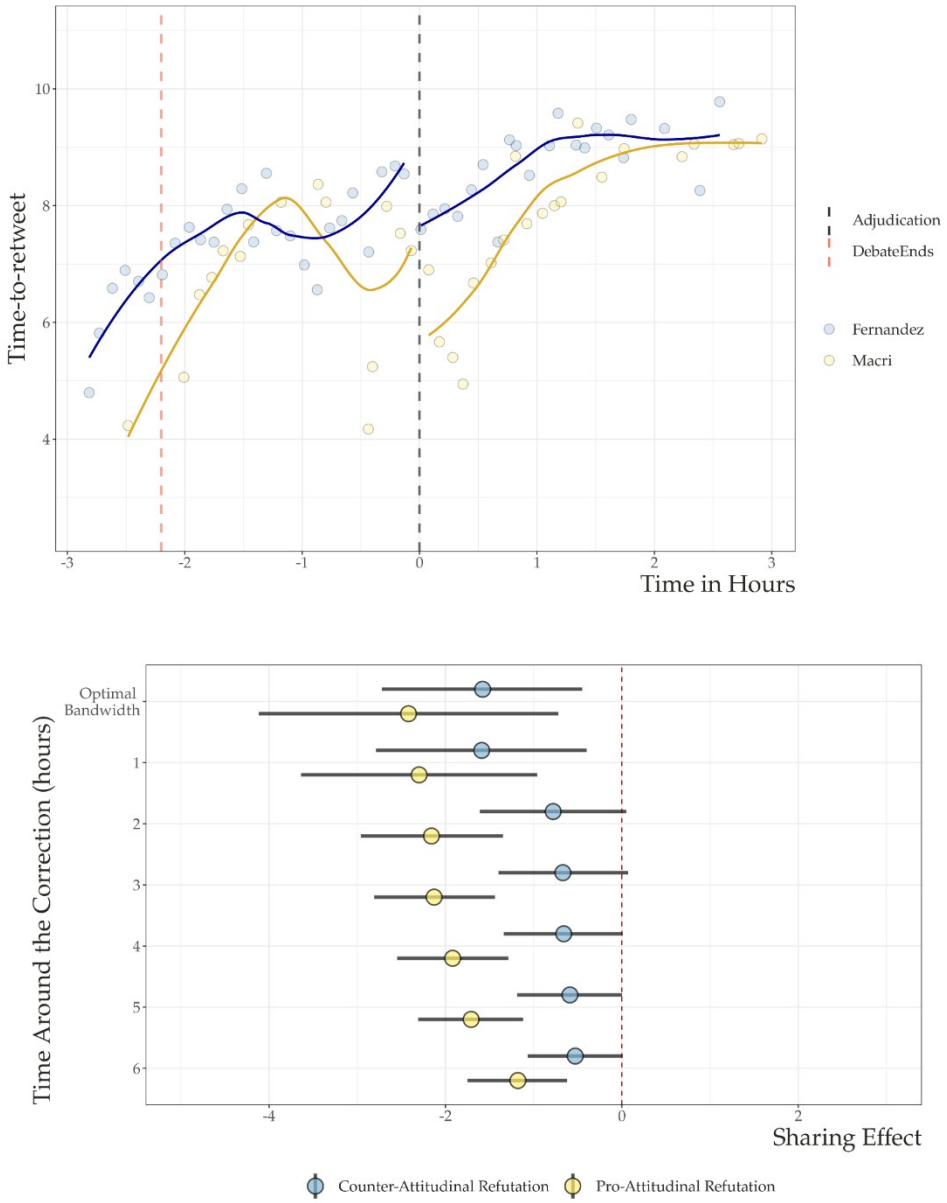
Note: Predicted probabilities estimated from models in Table 2.

Figure 4. Marginal effect of ‘true’ adjudication on sharing, conditional on the time-to-retweet



Note: Lines (vertical axis) describe the marginal effect of a ‘true’ adjudication on sharing, conditional on lower reaction time (i.e., ‘hot cognition’). Pro-attitudinal voters display higher marginal effects for the ‘true’ adjudication.

Figure 5. Regression discontinuity design (RDD) measuring the change in the time-to-retweet rate after Chequeado publishes its 'false' adjudication



Note: The upper plot presents visual evidence the effect of Chequeado's adjudication on sharing. The lower plot describes point-estimates for the discontinuity models, with Local Linear Estimates as in Calonico et al. (2014), with alternative bandwidths for robustness.

[1] The initial posts adapted two existing publications from the 2019 general election in Argentina that had been corrected by the leading fact-checking NGO, Chequeado. This was done with in coordination and with prior consent from Chequeado. Observational Twitter data was also collected during the election cycle, evaluated jointly between the researchers and the Chequeado team. The survey took place after the presidential election and the selection of cases was made to ensure that deceptive posts complied with IRB-Human Subjects regulations. These regulations include a selection of posts on issues that did not address medical or otherwise sensitive information that could harm the respondents and a disclaimer at the end of the survey on the use of edited Tweets. The IRB-Human Subjects compliance is included in the online supplementary file.

[2] We follow Taber and Lodge (2006) and separate exposure time from reaction time. Respondents are first exposed to the treatments (Tweets). Then, we ask each question as a separate event. This allows us to measure the reaction time, the time that elapses from the moment the question is presented to the moment the respondents answer this question. Latency, exposure time to the tweet, is also recorded but not manipulated experimentally, allowing the responding to decide the length of time they want to read this question.