Framing Fact-Checks as a “confirmation” increases engagement with corrections of misinformation: a four-country study

Natalia Aruguete[*] Flavia Batista [†] Ernesto Calvo [‡] Matias Guizzo Altube [§] Carlos Scartascini [¶] Tiago Ventura [‖]

August 22, 2023

Abstract

Previous research has extensively investigated why users spread misinformation online, while less attention has been given to the motivations behind sharing fact-checks. This article reports a four-country survey experiment assessing the influence of *confirmation* and *refutation* frames on engagement with online fact-checks. Respondents randomly received semantically identical content, either affirming accurate information (“It is TRUE that p”) or refuting misinformation (“It is FALSE that *not* p”). Despite semantic equivalence, confirmation frames elicit higher engagement rates than refutation frames. Additionally, confirmation frames reduce self-reported negative emotions related to polarization. These findings are crucial for designing policy interventions aiming to amplify fact-check exposure and reduce affective polarization, particularly in critical areas such as health-related misinformation and harmful speech.

Significance Statement

This study unveils a crucial insight into online fact-checking dissemination by revealing that confirmation frames (”It is TRUE that p”) elicit higher engagement and reduce negative emotions associated with polarization compared to semantically equivalent refutation frames (”It is FALSE that not p”). The implications of this discovery are central to crafting effective policy interventions, as it suggests a more potent strategy to enhance the reach of fact-checks and mitigate the emotional divisions exacerbated by misinformation, especially in areas like health and harmful speech.

[*] Universidad Nacional de Quilmes, UNQ. Email: nataliaaruguete@gmail.com. Webpage: http://unq.academia.edu/nataliaaruguete.

[†] University of Maryland, Government and Politics, UMD. Address: 4118 Chiconteague, College Park, MD 20742, USA. Email: fbatista@umd.edu.

[‡] University of Maryland, Government and Politics, UMD. Address: 3140 Tydings Hall, College Park, MD 20742, USA. Email: ecalvo@umd.edu. Webpage: http://gvptsites.umd.edu/calvo/

[§] IADB. 1300 New York Avenue, N.W., Washington, DC 20577, USA. Email: matiasgu@iadb.org .

[¶] IADB. 1300 New York Avenue, N.W., Washington, DC 20577, USA. Email: carlossc@iadb.org. Webpage: https://www.cscartascini.org

[‖] Center for Social Media and Politics, New York University. Address:14 East 4th Street, New York, NY 10012. tav5082@nyu.edu. Webpage: https://www.venturatiago.com/

Fact-checking is today the first line of defense against misinformation (Bode and Vraga, 2015; Del Vicario et al., 2016; Lazer et al., 2018; Van Der Linden et al., 2017). It is frequently 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 et al., 2020, p. 350). Research shows that fact checks successfully influence people's discernment of misinformation and *nudge* users to update their beliefs after correction, whether in survey experiments or field experiments, and across different cultural contexts (Arechar et al., 2022; Bode and Vraga, 2015; Clayton et al., 2020; Porter and Wood, 2021). The effect of fact-checking interventions extends over time, with minimal evidence of backfire effects from exposure to fact-checking corrections (Nyhan et al., 2020; Swire-Thompson et al., 2020).

To curb the spread of misinformation, fact-checkers can employ two distinct framing strategies: they can either publish **confirmation frames** that replace misinformation with accurate information, or they can publish **refutation frames** that warn social media users about content tagged as misinformation (Aruguete et al., 2023). Choosing *confirmations* provides users with factually accurate content they can share with peers. Opting for *refutations* allows fact-checkers to decrease the sharing of inaccurate, misleading, or false content. The effectiveness of increasing "good" content versus reducing "bad" content has not been experimentally tested. In this paper, we evaluate the impact of *confirmation* and *refutation* frames on the sharing behavior of social media users.

The lack of studies measuring the impact of *confirmation (TRUE)* and *refutation (FALSE)* frames is surprising, given the central role content labeling plays in fact-checking interventions. As noted by Shin and Thorson (2017), "[u]nlike traditional journalism, which emphasizes detached objectivity and adheres to the 'he said, she said' style of reporting, contemporary fact-checking directly engages in adjudicating factual disputes by publicly deciding whose claim is correct or incorrect" (Shin and Thorson, 2017, p.1). The decision to intervene using *confirmation* or *refutation* frames is an editorial choice that is independent of the source material (Vosoughi et al., 2018).

This paper presents experiments conducted in four different countries to assess the effect of *confirmation* and *refutation* frames on the sharing behavior of social media users. Our experiments expose nationally representative samples of Argentine, Brazilian, Chilean, and Colombian respondents to edited Facebook posts framed as a confirmation of accurate information or a refutation of misinformation. The experiment rotates the confirmation and refutation frames, the choice of labels (labeled vs. unlabeled), and the type of vaccine (Moderna, AstraZeneca, and Sputnik V).[1] The empirical analysis and robustness checks include several control variables (i.e., socio-demographic, attitudinal, and health status variables) and validation checks (i.e., processing time and pseudo-placebo treatment).

Our primary outcome measures the decision to engage (i.e., "like," "share," and "comment") with the fact check and the self-reported affective response to the fact-checking post. Our hypotheses, pre-registered at https://osf.io/ prior to the collection of the data, posit that respondents will engage more with confirmation frames than refutation frames [hypothesis 1 (H1)]. We propose this effect to be independent of other factors prompting engagement with a correction, such as cognitive congruence and partisan attachment. Our primary hypothesis stems from two theoretical mechanisms: the heavier cognitive burden of refutation frames and the positive valence charge associated with confirmation frames. We offer specific hypotheses and dedicated tests for each mechanism.

First, negation is known to impose a heavier cognitive load (Christensen, 2020). Research

[1] The variation in vaccines only takes place in Argentina.

in cognitive linguistics and cognitive psychology has documented differences in processing semantically equivalent positive or negative statements. Kaup et al. (2006) show that individuals are faster to process statements such as "the umbrella was open" compared to its semantically equivalent "the umbrella was not closed." Subjects also display faster response times for "the umbrella was closed" than for "the umbrella was not open." Indeed, this cognitive effort is not the result of the state of the umbrella (i.e., *open* or *closed*), but the result of how we process negation statements. In social networks, a higher cognitive burden could conceivably deter a swift, automatic, and affective response (Aruguete and Calvo, 2018; Kahneman, 2011), leading to more evaluative sharing behavior. We hypothesize that refutation frames exert a higher cognitive burden on respondents, thus resulting in longer reading times [hypothesis 2 (H2)] that curtail sharing.

Second, we expect that the confirmation of pro-attitudinal beliefs will carry a positive valence charge compared to the refutation of a counter-attitudinal belief. A standard sentiment analysis using state-of-the-art RoBERTa (Loureiro et al., 2022) shows that "It is true that vaccines are effective" is classified as *Positive* (i.e., Cardiff scores are Positive: 0.782, *Neutral*: 0.209, *Negative*: 0.009). Meanwhile, "It is false that vaccines are not effective" is classified as *negative* (i.e., Cardiff scores are *Positive*: 0.024, *Neutral*: 0.278, *Negative*: 0.698). This is because the words "true" and "false" function not only as Boolean operators but also convey positive and negative connotations in social conversation.

Confirmation statements such as "it is TRUE that p" convey that the content is socially acceptable and less likely to expose users to public scrutiny and criticism. Tetlock (2002) coins the term "intuitive politician" to describe the behavior of risk-averse subjects who seek to preserve their reputation by aligning themselves with socially accepted positions. "People behave like intuitive politicians when they seek to maintain a positive reputation or fulfill the social duties for which they are accountable" (Margolin et al., 2018). Therefore, confirmation frames communicate greater social acceptability and widespread consensus with published content.

Refutation frames, in contrast, suggest that there are dissenting opinions and raise the potential for conflict. That is, refutation frames suggest that there are at least some individuals or groups with competing beliefs (Tetlock, 2002). Therefore, statements framed as confirmations will have a positive valence charge that is independent of the pro- or counter-attitudinal preferences for the denoted content in the message. We hypothesize that confirmation frames will elicit positive emotional reactions and refutations will elicit negative ones [hypothesis 3 (H3)].

To sum up our pre-registered hypotheses, we anticipate the statement "it is TRUE that p" to enhance engagement compared to "it is FALSE that not p" [hypothesis 1 (H1)], because the former is both cognitively simpler to process [hypothesis 2 (H2)], and because *TRUE* carries an inherent positive valence charge [hypothesis 3 (H3)]. Conversely, refutation statements are cognitively challenging, and sharing refutation messages aligns one with an in-group social media user at odds with an out-group user's beliefs.

From Theory to Design

The two-arm design exposes respondents to a Facebook post that randomly confirms a clinically correct statement or refutes a clinically incorrect statement. Crucially, the experiment did not spread misinformation to participants; both the *confirmation* and the *refutation* frames communicated that vaccines are effective against the Omicron variant. In the Argentine version of the experiment, three different vaccines (Sputnik V, Moderna, and AstraZeneca) were taken into account to test for differences in perceived vaccine quality. For each vaccine, participants were exposed to *confirmation* and the *refutation* frames. In Chile, Brazil, and Colombia, two distinct designs were employed, presenting *confirmation* and *refutation* treatments either with explicit labels or without labels (see Figure 1 for the treatments implemented in Colombia - the complete set of treatments is provided in the Supplementary Information File (SIF)).[2]

Additionally, we introduced confirmation and refutation frames unrelated to our health correction and devoid of any correlation with political preferences. This pseudo-placebo treatment measures the independent valence charge associated with the use of the words "true" and "false" in a post about dogs.

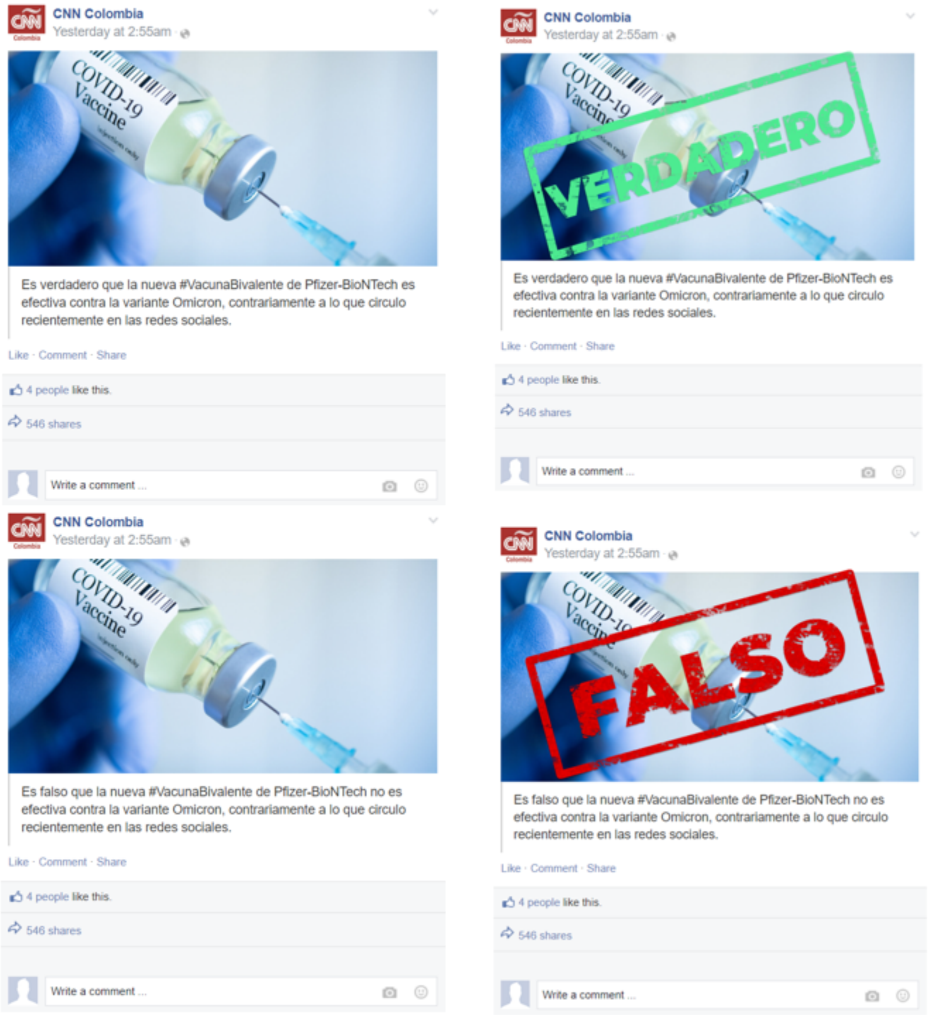
In all four countries, simple randomization was implemented, with respondents having equal chances of being assigned to each treatment (*confirmation* or *refutation* of the vaccines or dog treatments) and to each of the design alternatives (label, no-label, and, in the case of Argentina, vaccine type).[3]

After exposure to the treatments, respondents were asked to indicate whether they would "like," "share," and/or "comment" on the Facebook post. The response format allowed for multiple selections, with an explicit "ignore" option that was exclusive if chosen. Additionally, participants were asked to self-report their emotional response to the post, choosing from a list that included Ekman's six basic emotion categories: *fear*, *anger*, *joy*, *sadness*, *disgust*, and *surprise*, as well as an additional positive category, *optimism*. Multiple responses were allowed, except for the alternative *indifferent*, which was exclusive if selected.

The sequence of presentation (the Facebook treatments, the sharing behavior, and the emotional response) remained the same for all survey respondents. Additionally, the researchers recorded the time-to-read (the elapsed time spent viewing the post), the time-to-react (the elapsed time before responding to the behavior question), and the time-to-feel (the elapsed time

[2] See Figures S1, S2, S3, and S4.

[3] Table S1 through Table S4 present summary statistics and balance across the treatments.



CNN Colombia
Yesterday at 2:55am

Es verdadero que la nueva #VacunaBivalente de Pfizer-BioNTech es efectiva contra la variante Omicron, contrariamente a lo que circulo recientemente en las redes sociales.

Like · Comment · Share

4 people like this.

546 shares

Write a comment ...

CNN Colombia
Yesterday at 2:55am

Es verdadero que la nueva #VacunaBivalente de Pfizer-BioNTech es efectiva contra la variante Omicron, contrariamente a lo que circulo recientemente en las redes sociales.

Like · Comment · Share

4 people like this.

546 shares

Write a comment ...

CNN Colombia
Yesterday at 2:55am

Es falso que la nueva #VacunaBivalente de Pfizer-BioNTech no es efectiva contra la variante Omicron, contrariamente a lo que circulo recientemente en las redes sociales.

Like · Comment · Share

4 people like this.

546 shares

Write a comment ...

CNN Colombia
Yesterday at 2:55am

Es falso que la nueva #VacunaBivalente de Pfizer-BioNTech no es efectiva contra la variante Omicron, contrariamente a lo que circulo recientemente en las redes sociales.

Like · Comment · Share

4 people like this.

546 shares

Write a comment ...

Figure 1 Images of the *Confirmation* ("It is TRUE that p") and *Refutation* ("It is FALSE that not p") treatments used in Colombia. Left images without labels. Right images with labels. The confirmation and refutation frames are semantically equivalent but differ in their cognitive accessibility and their valence charge. All four treatments are factually correct and conform to the design used by our partner organization in Argentina, *Chequeado*. The designs for each country, and the placebo, are reported in the Supplemental Information File to this article.

before reporting an emotional reaction). The survey collected additional information to allow the inclusion of various demographic, political, and COVID-19 risk factors in the empirical analysis.

Results

Confirmation frames led to systematically higher engagement than *refutation* frames across the four countries involved in the study. Detailed results are shown in Figure 2, and full regression tables available in the Supplemental Information File (SIF).[4] Focusing on the *engagement* category (as represented by the first set of bars for each country in Figure 2), it was evident that the confirmation frame significantly increased overall engagement with the correction, affirming hypothesis 1 (H1). For example, in Argentina, engagement rose from 0.189 (or 18.9% of combined likes, shares, and comments) to 0.371 (or 37.1%), which is a significant positive difference of 18.2 percentage points (effectively a two-fold increase). Differences in engagement for Brazil, Chile and Colombia amounted to 13, 15, and 14 percentage points, respectively.

When examining the individual components of engagement, similar trends were observed. In all four countries, the "like" category showed the highest differences. For Argentina and Chile, confirmation frames produced a three-fold higher proportion of likes compared to refutation frames. For Brazil and Colombia, the difference was about twice as large.

The impact of the confirmation and refutation frames on reported emotion [hypothesis 3 (H3)] is consistent with our expectations. As illustrated in Figure 3, individuals who were exposed to the confirmation frame reported significantly more "joyful" and "optimistic" responses, significant at the $p < 0.01$ level. These differences are quite pronounced, ranging from a more than two-fold increase in reported optimism and joy in Brazil, to more than a five-fold increase in Argentina.

[4] Full models and robustness checks for all four countries are reported in Tables S5, S6, S7, and S8 of the SIF accompanying this article.

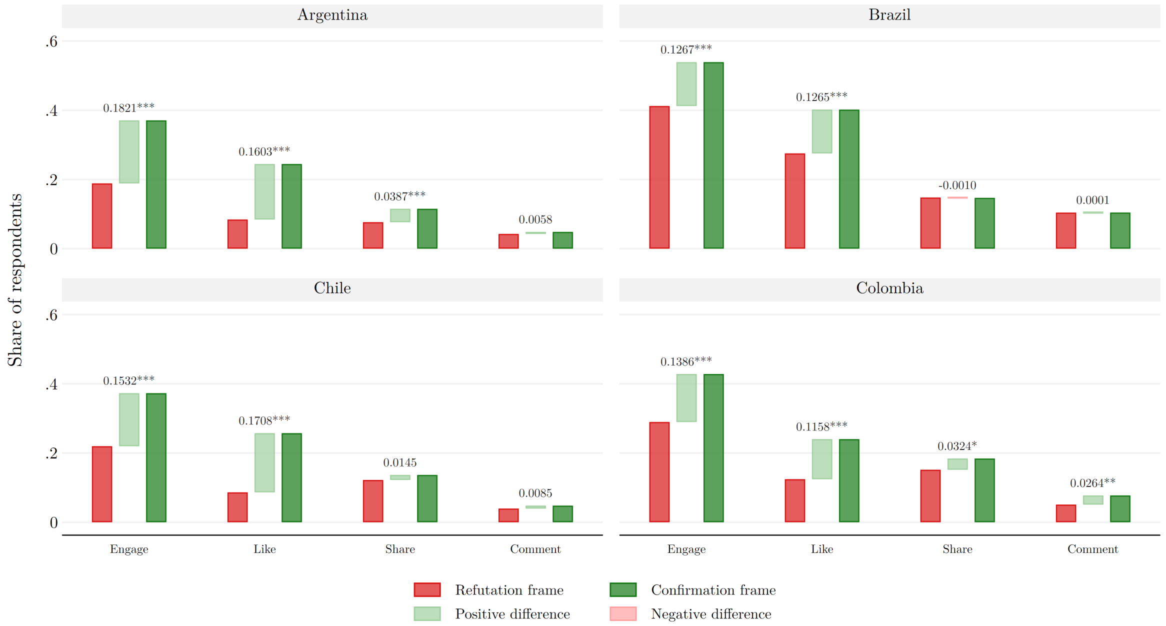


Figure 2 Regression results on “engagement,” “like,” “share,” and “comment” for the four countries. The first red bar for each dependent variable corresponds to the refutation frame. The third green bar corresponds to the confirmation frame. The middle bar shows the difference between the refutation and confirmation frames. When the difference is positive (confirmation frames generate more engagement than refutation) the bar is light green. Light red when the difference is negative. Full regression results are provided in the Supplemental Information File to this article. The summary of the differences in engagement between frames for the four countries is shown in Table 1.

In contrast, the refutation frame was primarily linked with negative emotions, such as “anger,” “disgust,” and “stress.” For instance, in Argentina, the refutation frame was four times more likely to induce anger than the confirmation frames. Similarly, the refutation frame was at least twice as likely to elicit anger in the other countries included in the study.

Table 1 presents the results of the differences –depicted as the middle bars in the figures– for all four countries. It emphasizes the variances between the *refutation* and *confirmation* frames, taking into account various control measures.[5] The results presented in Table 1 underline the consistency of the proposed framing effects, suggesting a notable influence of the confirmation

[5] The complete set of results can be found in Table S5 through Table S8 in the Supplemental Information File (SIF). The SIF also includes extensive robustness checks, alternative estimates with and without controls, estimates with heterogeneous effects for socio-demographic and attitudinal questions, and heterogeneous effects by party.

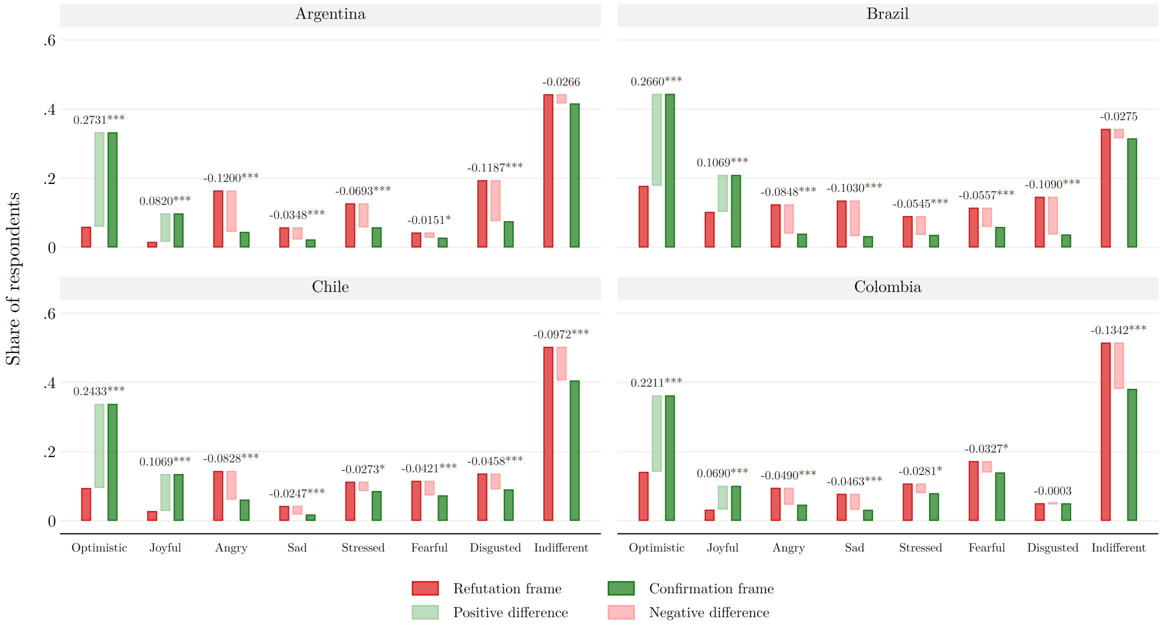


Figure 3 Regression results for the reported emotions for the four countries. The first red bar for each dependent variable corresponds to the refutation frame. The third green bar corresponds to the confirmation frame. The middle bar shows the difference between the refutation and confirmation frames. When the difference is positive (confirmation frames generate more engagement than refutation), the bar is light green. It is light red when the difference is negative. Full regression results are provided in the Supplemental Information File to this article. The summary of the differences in engagement between frames for the four countries is shown in Table 1.

and refutation frames on the outcomes observed.

Across all four countries, Table 1 consistently indicates that the emotional responses to the treatments are more positive for the confirmation frames and more negative for the refutation frames. The confirmation frame elicits feelings of optimism and joyfulness, according to the self-reported emotions of respondents. On the other hand, the refutation frame consistently provokes more negative emotions, such as anger, sadness, stress, fear, and disgust.

Interestingly, even though the refutation frame communicates the same core information as the confirmation frame, it consistently evokes stronger negative emotional reactions. This finding suggests that the manner in which information is presented or framed plays a crucial role in determining its emotional impact on the recipient. The refutation frame seems to incite affective polarization, a phenomenon that is well-documented in the existing literature.

Figure 4 illustrates the findings from the Argentine survey, showcasing the average rates of engagement (i.e., the sum of "like," "share," and "comment" rates) for the *confirmation* frame (TRUE label) and the *refutation* frame (FALSE label). These results validate the expectation that the effect of framing is consistent regardless of the specific vaccine brands. The results from testing with the AstraZeneca, Sputnik V, and Moderna vaccines were statistically indistinguishable. This is especially noteworthy given that different vaccines were associated with political decisions, and thereby became ideologically charged, and were believed to have varying degrees of effectiveness. Such consistency emphasizes the robustness of the framing effect.The Supplemental Information File provides a comprehensive description of the findings for each behavior separately (e.g., "like," "share," and "comment").

Figure 5 visualizes the engagement results from Brazil, Chile, and Colombia, contrasting the effects of treatments with explicit labels to those without labels.[6] In Brazil and Colombia, the

[6] Labels refer to the large banners placed over the picture, as depicted in Figure 1.

Table 1 Difference of Means between the *confirmation* and *refutation* frames

Variable	Argentina	Brazil	Chile	Colombia
Reactions				
Engage	0.188*** (0.018)	0.131*** (0.024)	0.152*** (0.023)	0.147*** (0.024)
Like	0.163*** (0.015)	0.127*** (0.022)	0.171*** (0.018)	0.120*** (0.020)
Share	0.042*** (0.012)	0.001 (0.018)	0.010 (0.017)	0.035* (0.019)
Comment	0.008 (0.009)	-0.002 (0.015)	0.006 (0.010)	0.026** (0.012)
Emotions				
Optimistic	0.278*** (0.015)	0.258*** (0.022)	0.246*** (0.020)	0.226*** (0.022)
Joyful	0.087*** (0.009)	0.106*** (0.018)	0.105*** (0.014)	0.071*** (0.013)
Angry	-0.121*** (0.012)	-0.083*** (0.014)	-0.081*** (0.015)	-0.049*** (0.013)
Sad	-0.035*** (0.008)	-0.102*** (0.014)	-0.023*** (0.009)	-0.049*** (0.011)
Stressed	-0.072*** (0.012)	-0.051*** (0.012)	-0.025* (0.015)	-0.028* (0.015)
Fearful	-0.015* (0.008)	-0.051*** (0.014)	-0.042*** (0.015)	-0.037** (0.019)
Disgusted	-0.121*** (0.014)	-0.112*** (0.014)	-0.046*** (0.015)	0.001 (0.011)
Indifferent	-0.028 (0.020)	-0.029 (0.023)	-0.102*** (0.025)	-0.133*** (0.025)

Note: Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Each cell corresponds to a different regression using as sample the survey of the country indicated in the header. Coefficients represent the effect of the confirmation frame on the reaction or emotion indicated in the first column compared against the refutation frame. All regressions control for age, sex, educational attainment, employment status, partisan attachment, having had COVID-19, number of doses administered of COVID-19 vaccine, and time spent reading the post. Full set of models in the SIF file to this article.

presence of labels produces larger differences in engagement, ranging from 1.5 to 6 percentage points. In Chile, however, higher differences are noted when there are no labels. Intriguingly, across all three countries, individuals are more inclined to engage with negative frames when they lack a label.[7]

[7] Separate estimates for the components of engagement can be found in the supplemental information file.

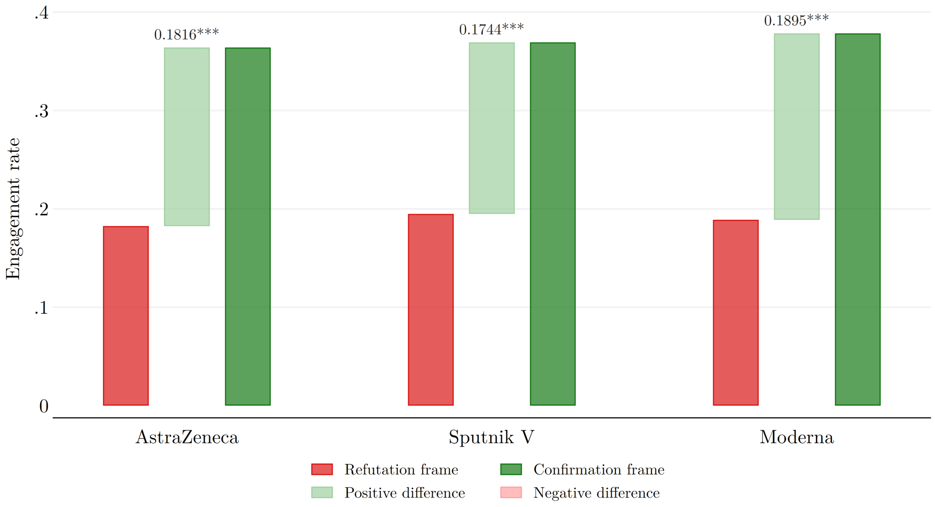


Figure 4 Argentine experiment: Overall engagement (like+share+comment) using the confirmation and refutation frames, TRUE or FALSE alternatively. Separate means are presented for each vaccine brand: AstraZeneca, Sputnik V, and Moderna. The TRUE and FALSE statements are semantically identical but differ in their cognitive accessibility and their valence charge. Both the TRUE and FALSE adjudications are factually correct.

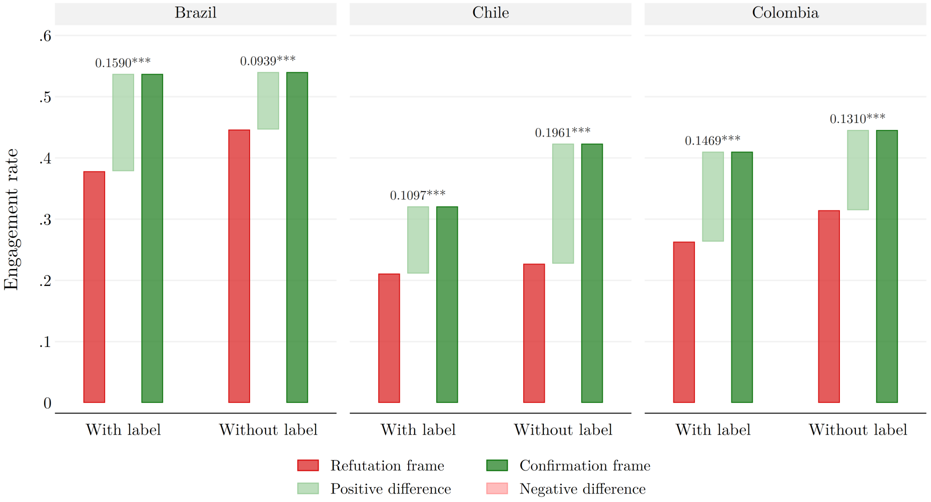
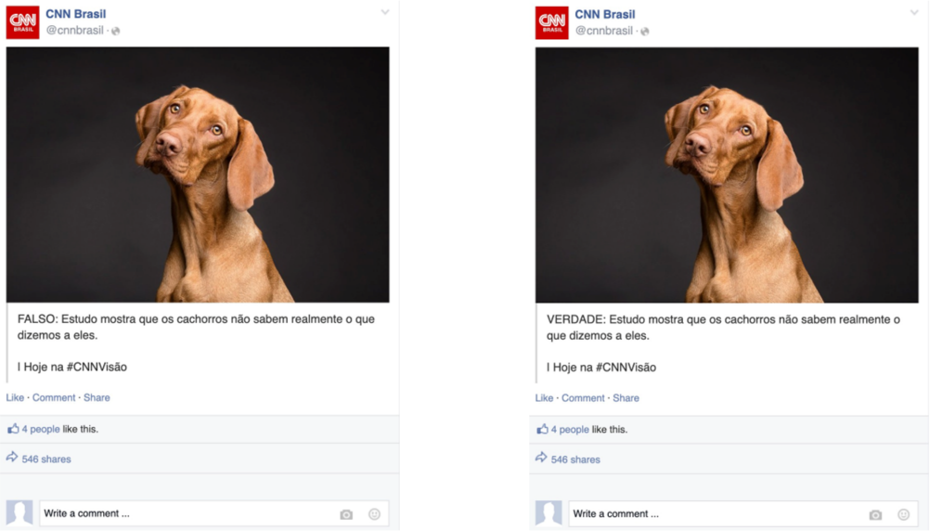


Figure 5 "Engagement" rate using the confirmation and refutation frames, TRUE or FALSE alternatively. Separate means are presented for the treatments with and without explicit labels. The TRUE and FALSE statements are semantically identical but differ in their cognitive accessibility and their valence charge. Both the TRUE and FALSE adjudications are factually correct.

Beyond Vaccines: Dogs do not Understand What We Say to Them

In order to examine if the observed effects were particularly influenced by the context of vaccines, a topic that has been heavily politicized in numerous countries, we conducted a distinct exercise to test the robustness of our findings. This experiment involved presenting survey participants with a minimally modified CNN social media post framed either as a confirmation ("True: Dogs do not really understand what we say to them") or as a refutation ("False: Dogs do not really understand what we say to them"). As can be seen in Figure 6, which uses the treatments in Brazil as an example, the only difference between the two treatments lies in the inclusion of the words "True:" or "False:". This ensures that the same content is conveyed while maintaining consistency in the semantic meaning and cognitive complexity of the message.



CNN Brasil
@cnnbrasil ·

FALSO: Estudo mostra que os cachorros não sabem realmente o que dizemos a eles.

I Hoje na #CNNVisão

Like · Comment · Share

4 people like this.

546 shares

Write a comment ...

CNN Brasil
@cnnbrasil ·

VERDADE: Estudo mostra que os cachorros não sabem realmente o que dizemos a eles.

I Hoje na #CNNVisão

Like · Comment · Share

4 people like this.

546 shares

Write a comment ...

Figure 6 Images of the *confirmation* ("It is TRUE that p") and *refutation* ("It is False that not p") pseudo-placebo treatments used in Brazil. The confirmation and refutation frames are semantically equivalent and intended to be equivalent in their cognitive accessibility and their valence charge, changing only the word "True" for "False." Both treatments conform to the design used by our partner organization in Argentina, *Chequeado*. The pseudo-placebo designs for each country are reported in the Supplemental Information File of this article.

This treatment offers an opportunity to investigate the direct and unconditioned impact of

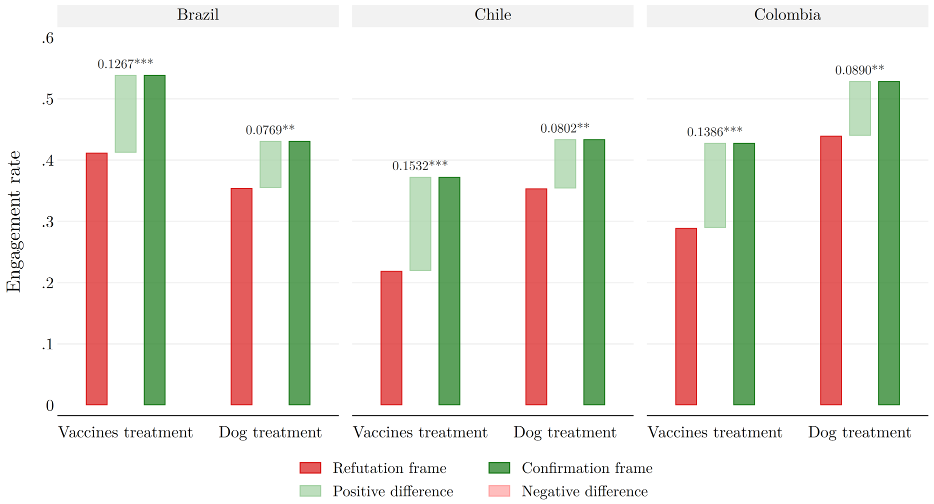


Figure 7 'Engagement' rate using the confirmation and refutation frames, TRUE or FALSE alternatively. Separate means are presented for the dog treatment (pseudo-placebo) and the vaccine treatments (pooling labeled and unlabeled treatments). The TRUE and FALSE statements are semantically identical but differ in their cognitive accessibility and their valence charge. Both the TRUE and FALSE adjudications are factually correct.

the words "True" and "False." Figure 7 includes the estimates of the "dog" treatment for Brazil, Chile and Colombia.[8] The results point to positive and statistically significant effects of the confirmation frame on "engagement." While these differences are smaller in percentage terms compared to the previous analysis, they still present differences that range between 8 and 9 percentage points, equating to roughly a 20 percent increase in engagement across all three countries. Furthermore, the effects on the "Like" behavior in Brazil and Chile are also positive and statistically significant. This simple exercise shows the power of the "TRUE" and "FALSE" labels on any type of post, indicating the substantial influence on engagement levels.

A Rejection of the Cognitive Difficulty Hypothesis, H2

Our findings provide no evidence to suggest a higher cognitive burden associated with the FALSE frame, as hypothesized in Hypothesis 2 (H2). Two major observations support this

[8] The full results can be found in Table S28 through Table S30 of the Supplemental Information File.

conclusion. First, there is no consistent relationship between education level and engagement with the confirmation and refutation frames, indicated by the absence of significant patterns across countries. This can be seen in Table S15 of the Supplemental Information File (SIF). The influence of the confirmation frame across varying education levels, both on the propensity to react (Table S24) and the self-reported emotion (Table S24), does not show that more educated individuals are less susceptible to the frames.

Second, we notice no significant decrease in the impact of the confirmation versus refutation frames attributable to the time respondents spent reading the treatments. Contrary to our expectations, longer reading times did not lessen the behavioral and emotional differences between the confirmation and refutation frames. In fact, in Argentina and Brazil, an increase in reading time correlated with a statistically significant rise in reported "likes" for the confirmation frame ($p < 0.05$). The influence of reading time on the overall engagement behavior is illustrated in Figure 8 for respondents from all countries in the vaccines treatments.

This impact of extended exposure time is significant: prolonged exposure to the TRUE frame amplifies the differences in "likes" between the confirmation and refutation frames. Thus, a more thorough reading of the post increases the probability that the confirmation frame will attract a higher "like" rate than the refutation frame. Similar results are reported in Tables S26 and S27 of the Appendix for all countries, with Brazil demonstrating results analogous to those of Argentina, while Chile and Colombia show a more modest positive correlation. In none of the four cases is there a statistically significant decline in the behavioral response gap between the confirmation and refutation frames. Consequently, the increased propensity to share the confirmation frame can be exclusively attributed to its positive valence charge, as per Hypothesis 3 (H3), rather than the cognitive difficulty associated with the refutation frame, as per Hypothesis 2 (H2).

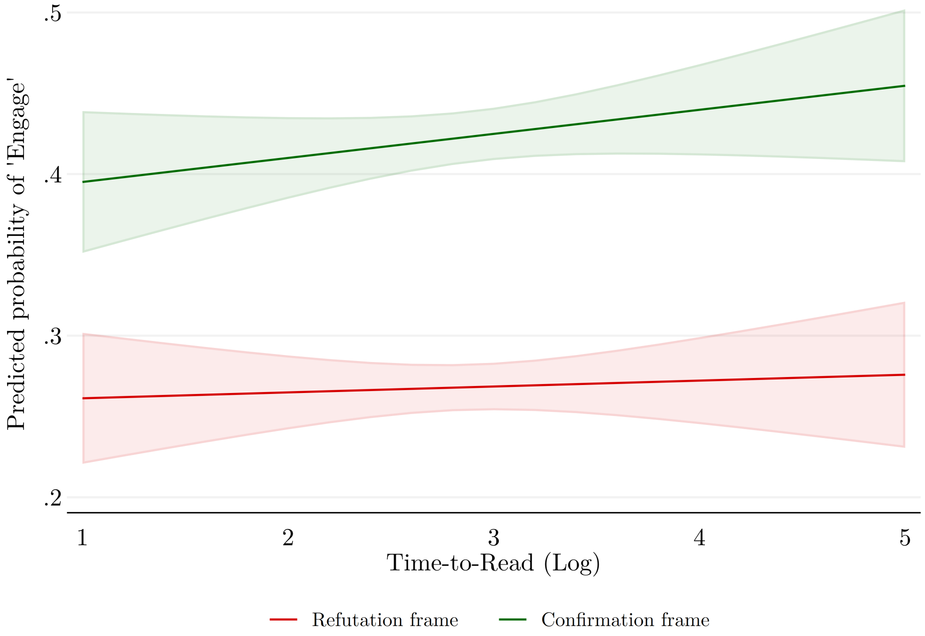


Figure 8 "Engagement" rate and Time-to-Read the Facebook Post in the vaccines treatments. Longer reading times are associated with larger differences in the response to the confirmation and refutation frames. The results refute the cognitive difficulty hypothesis, as higher attention does not reduce the differences between the *confirmation* and *refutation* frames. Probability estimates are obtained from a linear probability model, controlling for socio-demographic characteristics. Shaded area corresponds to the 95% confidence interval.

Other Results: Partisanship, Vaccination Status, and Other Sources of Heterogeneity

In addition to heterogeneity in education and reading time, the Supplementary Information File presents additional exercises where we look at differences according to political affiliation, vaccination status, and other socio-demographic indicators. The differences in engagement remain for individuals in each of these different group categories. Confirmation frames about the vaccines tend to elicit more relative engagement and positive emotions among government supporters in Argentina, where the incumbent government aggressively pursued quarantine and mask mandates. On the other hand, the opposite is true for Brazil and Chile, where anti-COVID policies were more divisive and weakly enforced.[9]

Across the four countries, the differences in engagement and positive emotions tend to be higher for those who were vaccinated (twice or more) than those who were not.[10] These results suggest that while the effects appear to be fairly universal, variations do exist among different groups, in line with expectations. Therefore, the impact of different fact-checking strategies will not be uniform across all individuals. This indicates that tailoring the framing and the message to suit specific demographics could still be desirable for maximizing the efficacy of the message.

Discussion

The results from the four survey experiments lend support to a higher intention to "engage" and "like" fact-checks framed as confirmations compared to the semantically equivalent refutation [hypothesis 1 (H1)]. All four surveys provide evidence that the findings remain robust across a variety of experimental designs, including different brands of the COVID-19 vaccine, with or without the use of labels, and across a diverse range of socio-demographic categories.

[9] See Tables S16, S17, S18, and S19 in the SIF.

[10] See Table S20 and S21 in the SIF.

Moreover, the observed emotional responses and the absence of an effect related to cognitive effort suggest that this discrepancy arises from distinct interpretations of the confirmation and refutation frames [hypothesis 3 (H3)]. We speculate that, despite their semantic equivalence, confirmation frames draw the reader's attention towards the health benefits of the vaccine, while refutation frames draw attention to the misinformation event itself.

The rejection of the cognitive burden hypothesis [hypothesis 2 (H2)] further bolsters a valence-driven interpretation of the results. We find no evidence suggesting that rates of liking or sharing stem from difficulties in comprehending the confirmation and refutation frames. Nor is there a significant difference in the mean processing time for each frame. Intriguingly, we observe an increase in "likes" and "shares" for the confirmation frame with prolonged reading times. Given that the reading time is similar for both the confirmation and refutation frames, yet longer reading times increase the probability of liking and sharing the confirmation frame, the only plausible explanation is that a deeper understanding enhances the positive valence of the confirmation frame.

The results of our experiments have significant policy implications. Fact-checkers aiming to expand their posts' reach would likely benefit from more frequent use of the confirmation frame. Our analysis of TRUE versus FALSE frames usage among 22 fact-checkers in Latin America revealed that refutation frames are four times more likely to be used. Some fact-checkers exclusively use refutation frames, thereby potentially reducing their corrections' exposure and likely increasing the stock of negative valence content on social media.

The findings in this paper also indicate that the effect of confirmation and refutation frames operates independently of other demographic, partisan, and health-associated moderators of fact-check sharing. The often emphasized negative partisan effects of misinformation can overshadow the fact that negative and positive valence charges in health messages are not solely a result of our partisan predispositions. Fact-checkers can choose different editorial strategies to frame a correction either as a contribution to the overall amount of correct information present on social networks or as a contribution to the overall stock of polarized content. The standard use of the label "FALSE" can be seen not only as a warning about toxic content but also as a reminder to readers that social media is highly polarized. This may divert attention away from crucial health issues and towards the partisan conflict underlying them.

Methods

Survey Information

The survey experiments were conducted in Argentina in February of 2022, Chile in November of 2022, Brazil in December of 2022, and Colombia in March of 2023. The surveys were designed by the Interdisciplinary Lab for Computational Social Science (iLCSS) at the University of Maryland, College Park, in collaboration with the Fact-Checking Agency Chequeado.

All four surveys were administered online by the polling firm Netquest.[11]. The sample comprised 12,000 adult respondents from Argentina, Brazil, Chile, and Colombia, with stratification based on gender, age, and education in accordance with current census data. The survey took a median time of 22 minutes to complete. In addition to the experiment, it included a battery of socio-demographic, attitudinal, and political questions.

[11] Netquest is a reputable survey company with large global panels of respondents. Netquest panels opt-in respondents, using quota sampling to achieve a nationally representative sample on key demographics, such as age, gender, population, and income. An Independent assessment of the quality of Netquest panels compared with a probabilistic sample was recently published by (Castorena et al., 2023), finding very small deviations from optimal sampling

Design Information

The "vaccine" experiments use a two-arm design that exposes respondents to one of two equivalent statements that *confirm* the efficacy of the vaccine or refute their inefficacy. The design randomly prompts respondents to read either the confirmation statement "*It is TRUE that the new VacunaBivalente is effective against the Omicron variant*" or the refutation of the corresponding misinformation "*It is FALSE that the new VacunaBivalente is not effective against the Omicron variant.*" In Argentina, the brand of the vaccine (Sputnik V, Moderna, and AstraZeneca) is rotated. In Brazil, Chile, and Colombia, the use of labels is rotated, and a pseudo-placebo treatment about dogs is included.[12]

The flow of the experiment is as follows. First, respondents are exposed to either a *confirmation* or *refutation* frame. The time respondents spend reading the statement (time-to-read) is measured, beginning with the image loading and ending when the respondent progresses to the next page of the online survey. The second page asks respondents if they would "like," "share," "comment," or "ignore" the Facebook post. The time-to-respond is again measured. Finally, on the third page, respondents are asked to self-report their emotional reaction to the question.

The statistical models utilized in the paper employ simple two-tailed mean tests. The conditional effects of time variables and other socio-economic indicators are further assessed using general linear regression models as well as ordinary least-square models.

[12] The treated individuals are well-balanced as shown in Table S1 to S4 in the SIF.

Variable Definitions

Dependent Variables

- **Engagement.** After seeing the Facebook Post, respondents are asked to “like,” “share,” “comment,” or “ignore” it. Each reaction is treated as a dependent variable. In addition, there is an indicator variable (engage) for the selection of at least one active reaction (like, retweet, or reply) by the respondent.

- **Emotions.** After seeing the Facebook Post, respondents are asked if the publication elicited any of the following emotions: Anger, contempt, disgust, optimism, stress, sadness, fear, or indifference. Respondents can mark more than one option. Each emotion is associated with one indicator variable and is treated as a single dependent variable.

Treatment Variables

- **True/False framing.** Binary variable indicating if the statement is a *confirmation* or *refutation* frame (“It is TRUE that p”) or refutation framework (“It is FALSE that not p”).

- **Argentina: Brand of the vaccine.** Set of indicators for the vaccine brand mentioned in the vignette: AstraZeneca, Moderna, or Sputnik V.

- **Brazil, Chile, and Colombia: Explicit label used in the treatment.** A categorical variable that indicates if the label was included or not.

Control Variables

- **Time to read.** Time in milliseconds (log) spent by the respondent.

- **Partisan attachment.** Set of binary variables indicating vote intention in hypothetical presidential elections: *Frente de Todos* (center-left ruling party), *Juntos por el Cambio* (center-right opposition party), and *voto en blanco* (none of the above).

- **Age.** Set of staggered indicator variables for six age groups: 18 to 25 years old, 26 to 35 years old, 36 to 45 years old, 46 to 55 years old, 56 to 65 years old, and more than 65 years old.

- **Sex.** Binary variable indicating if the respondent is a woman.

- **Education.** Set of indicator variables for the highest level of education attained (completed or incomplete): Primary, secondary, university (undergraduate), or graduate level.

- **Employment status.** Binary variable indicating if the respondent is employed at the time of answering the questionnaire.

- **Vaccination status.** Set of indicator variables for the number of COVID-19 vaccine doses received: None, one, or two or more.

- **COVID-19 status.** Binary variable indicating if the respondent ever got COVID-19.

Ethics

Human Subjects and Ethics approval was granted by the University of Maryland Institutional Review Board prior to the implementation of surveys in each country. The project approvals are registered under the identification code IRB 1825785, beginning with IRB [1825785-1] "COVID-19, Trust, and Misinformation" approved on October 27, 2021. Further approvals for each survey are registered under the identification codes 1825785-2 through [1825785-8], with final approval on January 19, 2023. Decisions by the review board granted all four surveys expedited review category 7. Waiver of Consent Documentation, 45CFR46.117(c). Waiver of Consent

45CFR46.116(f)(3) (deception) was granted as we exposed respondents to treatments created by our research team. A disclaimer provided respondents with information on how to contact the researchers or IRB if needed.

Competing Interests

The authors declare that they have no competing interests related to this work.

Funding Statement

This research is part of the Inter-American Development Bank project: "Reforms, public policy, and polarization. The effect of social network interventions", BID/ITEC RG-T4022-P002.

References

Arechar, A. A., Allen, J. N. L., Cole, R., Epstein, Z., Garimella, K., Gully, A., Lu, J. G., Ross, R. M., Stagnaro, M., Zhang, J., et al. (2022). Understanding and reducing online misinformation across 16 countries on six continents.

Aruguete, N., Bachmann, I., Calvo, E., Valenzuela, S., and Ventura, T. (2023). Truth be told: How 'true' and 'false' labels influence user engagement with fact-checks. *New Media Society*.

Aruguete, N. and Calvo, E. (2018). Time to #Protest: Selective Exposure, Cascading Activation, and Framing in Social Media. *Journal of Communication*, 68(3):480–502.

Bode, L. and Vraga, E. K. (2015). In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, 65(4):619–638.

Castorena, O., Lupu, N., Schade, M., and Zechmeister, E. (2023). Online surveys in latin america. *PS: Political Science Politics*, 56(2):273–280.

Christensen, K. R. (2020). The neurology of negation: fmri, erp, and aphasia. In *The Oxford handbook of negation*, pages 725–739. Oxford University Press Oxford.

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., et al. (2020). Real solutions for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, 42(4):1073–1095.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., and Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559.

Kahneman, D. (2011). *Thinking, fast and slow.* Macmillan.

Kaup, B., Lüdtke, J., and Zwaan, R. A. (2006). Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics*, 38(7):1033–1050.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

Loureiro, D., Barbieri, F., Neves, L., Anke, L. E., and Camacho-Collados, J. (2022). Timelms: Diachronic language models from twitter. *arXiv preprint arXiv:2202.03829*.

Margolin, D. B., Hannak, A., and Weber, I. (2018). Political fact-checking on twitter: When do corrections have an effect? *Political Communication*, 35(2):196–219.

Nyhan, B., Porter, E., Reifler, J., and 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*, 42:939–960.

Porter, E. and Wood, T. J. (2021). The global effectiveness of fact-checking: Evidence from simultaneous experiments in argentina, nigeria, south africa, and the united kingdom. *Proceedings of the National Academy of Sciences*, 118(37).

Shin, J. and Thorson, K. (2017). Partisan selective sharing: The biased diffusion of fact-checking messages on social media. *Journal of Communication*, 67(2):233–255.

Swire-Thompson, B., DeGutis, J., and Lazer, D. (2020). Searching for the backfire effect: Measurement and design considerations. *Journal of applied research in memory and cognition*, 9(3):286–299.

Tetlock, P. E. (2002). Social functionalist frameworks for judgment and choice: intuitive politicians, theologians, and prosecutors. *Psychological review*, 109(3):451.

Van Der Linden, S., Maibach, E., Cook, J., Leiserowitz, A., and Lewandowsky, S. (2017). Inoculating against misinformation. *Science*, 358(6367):1141–1142.

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

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

Acknowledgements

We thank Laura Zommer, Olivia Sohr, Ariel Riera, and Leticia Small from *Chequeado* for their suggestions to the design and the materials used for this research project. We also thank Felipe Gonzalez Alzaga, Carolina Batista, and Carolina Verena Franco for excellent research assistance. Previous versions of this article benefited from comments from colleagues of Emory University, the University of Chicago, the University of Maryland, Facebook, and Twitter.

Author contributions statement

Must include all authors, identified by initials, for example: E.C. T.V. and N.A. conceived the experiments, E.C. and T.V. conducted the experiments, E.C., T.V., C.S. and M.G.A. analyzed the results. All authors reviewed the manuscript.