The use of *Confirmation* and *Refutation* frames in Fact-Checking War Related Misinformations

Carolina Batista * Ernesto Calvo † Shibley Telhami ‡

February 3, 2024

Words: 2,942

Abstract

We implement a survey experiment measuring the effect of 'confirmation' versus 'refutation' frames in fact-checking wartime corrections. Respondents were presented with semantically equivalent content either confirming accurate information (e.g., "it is TRUE that President Zelensky REMAINED in Ukraine") or refuting its inaccurate version (e.g., "it is FALSE that President Zelensky LEFT Ukraine"). Even though both statements are semantically equivalent, our pre-registered experiment shows that 'confirmation' frames elicit greater engagement than 'refutation' frames. Refutation frames are also associated with negative sentiments like anger and disgust. In contrast, the equivalent confirmation frame elicits positive sentiments such as happiness. The experimental design mimics a Facebook post, employing four randomized treatments that vary in framing (confirmation vs. refutation) and news source (New York Times vs. Fox News). The survey was administered to 2,091 US adults in May 2022.

*University of Maryland, Government and Politics, UMD. Address: 4118 Chiconteague, College Park, MD 20742, USA. Email: batistac@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/

‡University of Maryland, Government and Politics, UMD. Address: 3140 Tydings Hall, College Park, MD 20742, USA. Email: sadat@umd.edu.

Confirmations and Refutations in Fact Checking

Misinformation is a defining characteristic of highly polarized political environments, with negative effects on public debate, policy implementation, and democratic governance (Lazer et al., 2018; Swire-Thompson and Lazer, 2019; West and Bergstrom, 2021). To counter misinformation, independent fact-checkers, journalists, and social media platforms invest significant time and resources in detecting misinformation and labeling content as factually correct or incorrect. However, reducing the prevalence of misinformation also depends on reaching target audiences and maximizing social media sharing of factual corrections. Indeed, the visibility of fact-checking messages rests on "the extent to which individuals share primarily attitude-consistent content with their social networks" (Shin and Thorson, 2017, p.). Therefore, understanding which corrective content is more broadly shared by social media users and the affective response it elicits is central to the fact checker's mission.

As discussed in Aruguete et al. (2023a), to curve misinformation, fact-checkers may label corrections in two very different ways: they can publish confirmation frames that communicate to readers factually correct content ("it is true that p") or publish refutation frames that identify content as misinformation ("it is false that not p"). While these statements are both viable interventions that address misinformation, they are not likely to be shared equally by users, nor will they elicit the same affective responses.

An unresolved problem in the study of confirmation and refutation frames is that the statement "it is true that p" is cognitively easier to process than its equivalent refutation "it is false that not p." In this research note, we address this issue explicitly, using a variation of the refutation frame that is cognitively similar to its confirmation. To test for the effect of cognitively similar and semantically equivalent statements, we compare "It is true that p" and "It is false

that q", where q is the antonym of p, eliminating one important confounding factor resulting from different cognitive difficulties in processing negation statements. Second, different from (Aruguete et al., 2023a), we explicitly evaluate the media and partisan effects of the confirmation and refutation statements, considering different endorsements of the same message by the New York Times or Fox News.

Confirmation frames ("it is true that p") and refutation frames ("it is false that antonym(p)") are important moderators of fact-checking corrections that have only recently been studied in depth. As described in Aruguete et al. (2023b),"[T]he 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." We show that TRUE and FALSE frames are important moderators of content sharing, distinct from other mechanisms such as cognitive congruence, affective activation, and partisanship.

The Experiment

Our experiment exposes a nationally representative sample of respondents to a Facebook post framed as a confirmation of accurate information ("it is TRUE that President Volodymir Zelensky remains in Ukraine despite social media claims to the contrary. 'I need weapons, not a ride' he said") or as a refutation of misinformation ("it is FALSE that President Volodymir Zelensky left Ukraine despite social media claims to the contrary. 'I need weapons, not a ride' he said"). The treatments differ in the the use of the words "TRUE" and "False" and in the use of "remains" and "left". This change holds the semantic content identical but changes the valence charge from neutral to negative. Sentiment analysis of the confirmation frame, using *Cardiffnlp*(Loureiro et al., 2022), classifies the statement as 'neutral' (0.808) rather than negative (0.118) or positive (0.073). By contrast, the refutation frame is classified as negative

(0.692) rather than neutral (0.297) or positive (0.011). However, the extent to which differences in sentiment classification are associated with engagement and emotional differences is often untested. The experiment also rotates two alternative sources for the message, the New York Times and Fox News.

We expect respondents to engage with confirmation frames more frequently than refutation frames. This effect is independent of other pro-attitudinal and counter-attitudinal preferences for engaging, liking, and sharing a correction, such as the effect of cognitive congruence, affective activation, and partisan attachment. Previous research shows that negative frames reduce engagement, likes, and shares due to two different mechanisms. First, negation frames carry a heavier cognitive burden (Christensen, 2020). Cognitive effort is higher when processing statements such as "the umbrella is not closed", compared to the similar statement "the umbrella is open" Kaup et al. (2006). Cognitive effort is also associated with lower social media engagement (Meinert and Krämer, 2022). Our experiment minimizes the difference in cognitive effort between the confirmation and refutation frames, using q, the antonym of p, when comparing sharing behavior.

Second, the confirmation of pro-attitudinal beliefs carries a higher positive valence than the refutation of a counter-attitudinal belief. More important, as discussed by Tetlock (2002), individuals behave socially as 'intuitive politicians,' weary of communicating content that may affect their reputation compared to socially accepted views (Margolin et al., 2018). A more extensive discussion of these mechanisms is described in (Aruguete et al., 2023b).

The Experimental Design

Our two-arm design with the treatments shown in Figure 1 exposes respondents to a Facebook post that randomly confirms a factually correct statement or refutes an incorrect one. Both the

confirmation and refutation of Facebook posts reproduce a fact-checking correction published by *FactCheck.org* on March 21, 2022, days after the invasion of Ukraine (https://www.factcheck.org/2022/03/zelens remains-in-ukraine-despite-false-claims-on-social-media/). Our post uses identical language for the refutation and its adaptation as a confirmation.

After exposure, we asked respondents whether they would share, like, or comment on the Facebook post (behavior), with an explicit "ignore" alternative that was exclusive if chosen. Second, we ask them to report their emotional response to the post using Ekman's six basic emotion categories (Ekman and Friesen, 1971): fear, anger, joy, sadness, disgust, and surprise. Participants can pick only one emotion from this list or choose "indifferent" as an alternative.

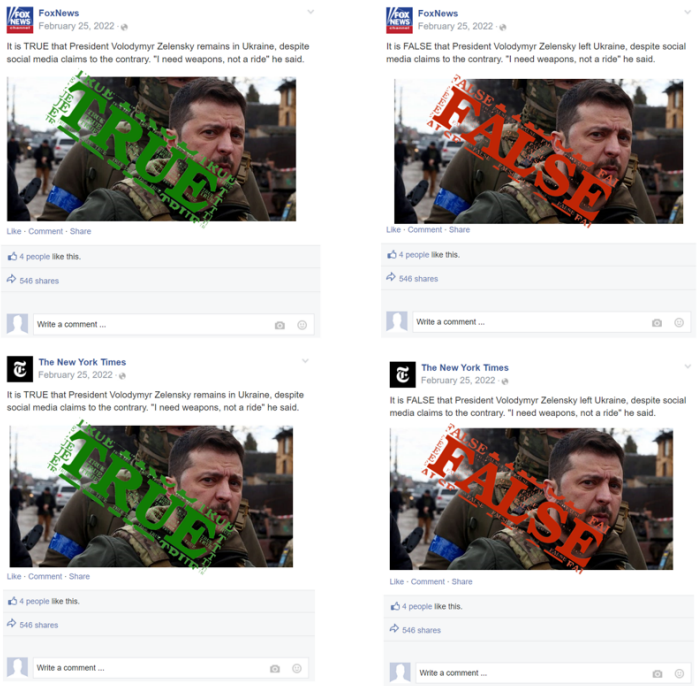

Figure 1 Flow: Respondents are randomized into a TRUE or FALSE group and a news source (four possible treatments). We measure Time-to-Read the assigned Facebook post. The survey the presents respondents with a behavior instrument: "would you 'like', 'share', 'reply', or 'ignore' the Facebook post". We measure Time-to-React. Finally, we ask how the post made them feel (affective response).

Main Hypotheses

Social media engagement and content sharing constitute a complex phenomenon that connects multiple independent cognitive and valence dispositions. We expect individuals to have distinct "click rates," in this instance, indicating their inclinations to engage with social media posts. Their overall "click rate" is moderated by cognitive and affective responses to statements to pro- and counter-attitudinal statements.

The first hypotheses of our study expect that pro-attitudinal confirmation ("it is TRUE that p") will be more likely to elicit engagement ("like", "share", "comment") than pro-attitudinal refutations ("it is FALSE that antonym(p)"), because negation statements impose a higher cognitive burden (Christensen, 2020). Using q, the antonym of p, instead of the double negation 'false that not p' we expect to minimize any cognitive difference between the confirmation and negation statement. Still, we test for the hypotheses in Aruguete et al. (2023a) that, all else equal, negation frames are associated with longer reading time:

HT_1: We expect refutation frames to increase the overall time-to-read of the treatment (higher cognitive demand).

The second hypothesis of our study expects confirmation frames to be liked, shared, and commented on at higher rates than refutation frames, due to its higher valance charge:

HT_2: Confirmation frames are 'liked,' 'shared,' and 'commented' at higher rates than pro-attitudinal refutations.

The third set of hypotheses considers the heterogeneous effects of Facebook posts on Democratic and Republican partisans exposed to news organizations with progressive or conservative stances. We expect that Democratic participants treated with content from the New York Times will display greater engagement compared to Republican participants. Conversely, we expect

reduced engagement among Democrats when exposed to Fox News compared to their Republican counterparts. All else equal, we expect:

HT_3a: Higher pro-attitudinal engagement by Democratic respondents exposed to the New York Times treatment.

HT_3b: Higher pro-attitudinal engagement by Republican respondents exposed to the Fox News treatment.

Model Estimation

To test the proposed hypotheses, we estimate a model wherein the dependent variable Y_{it} assumes a value of 1 if the respondent engages by liking, sharing, or commenting on the Facebook post and 0 otherwise. Our analysis reports quantities of interests: the average marginal component effects and marginal means.

Covariates

We include several covariates in our analysis. These consist of an indicator variable taking the value of 1 for the confirmation frame ("It is true that p") and 0 for the refutation frame ('it is false that q"); an indicator variable taking the value of 1 if the NYT is posting the correction and 0 if FoxNews is posting the correction; the time-to-read for the treatments; the time-to-react for the behavioral response; and the respondents' partisan preference, indicated by their voting choices. Additionally, we estimate models that incorporate partisanship, age, gender, race, and other socio-demographic variables.

Results

The null effect of cognitive effort

Figure 1 shows the time-to-read either the confirmation or refutation posts. We find no statistically significant difference between the confirmation and refutation conditions and no support for the first hypothesis. Therefore, the use q instead of the double negation "*false* that *not p*", eliminates this important confounding effect which affected previous studies.

We also find no support for a higher cognitive effort for the refutation when we compare differences in engagement as respondents spend more time reading the treatments. Figure 3 shows that confirmation frames are 'liked,' 'shared,' and 'commented' at higher rates than the refutation frame, with the constant difference between the frames across exposure times. Consequently, we do not find that the refutation frames elicit longer time-to-read by respondents, and we do not find that the differences between frames diminish as time-to-read increases.

Positive valence by media and treatment condition

Figure 4 shows higher engagement with confirmation frames compared to its refutation version. Higher engagement is confirmed for the subsample of respondents exposed to the New York Times post and Fox News post, although engagement is higher for the New York Times post. Figure 5 also shows higher engagement with the confirmation frame controlling for vote choice. Democratic and Republican participants are 1.75 and 2 times more inclined to engage with the confirmation frames than the refutation frames. Independent voters are nearly three times as likely to like, share, or comment on the confirmation frame.

Figure 6 shows that Republican respondents are marginally likelier to engage with the post

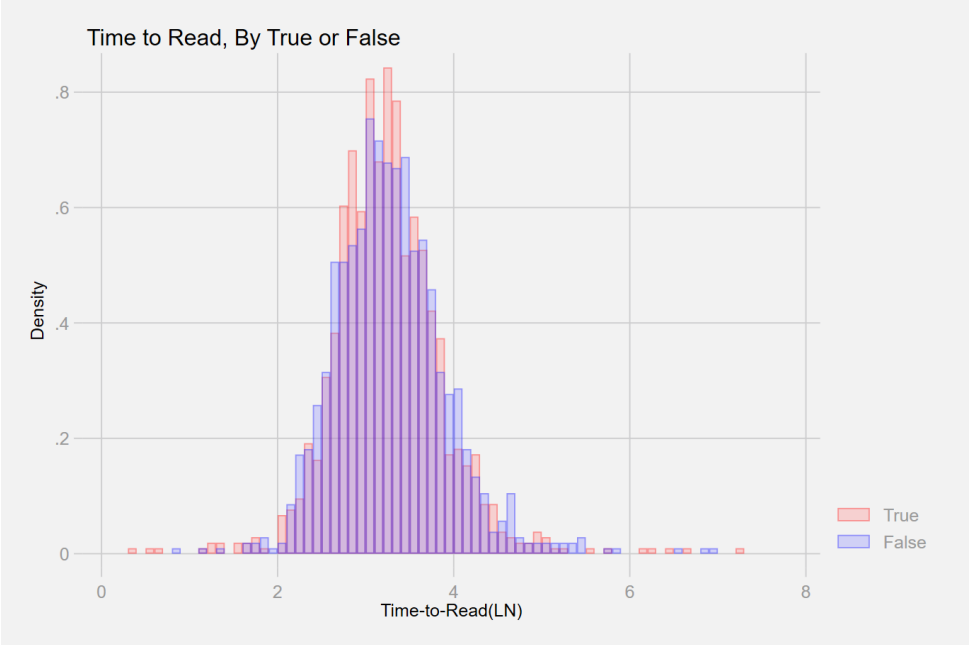

Figure 2 Confirmation frames ("It is true that p") and refutation frames ("It is false that antonym(p)") by news organization. The confirmation and refutation statements are semantically identical but differ in cognitive accessibility and valence charge. Both the TRUE and FALSE adjudications are factually correct.

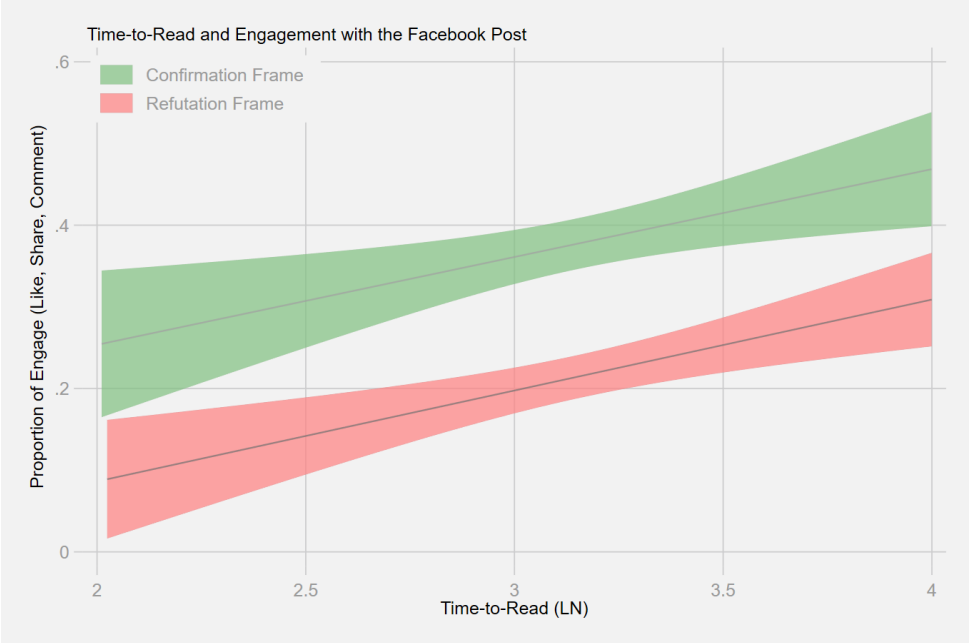

Figure 3 Confirmation (TRUE) and refutation (FALSE) by news organization. The confirmation and refutation statements are semantically identical but differ in cognitive accessibility and valence charge. Both the TRUE and FALSE adjudications are factually correct.

when the source is Fox News, while Democratic respondents are likelier to engage with the post when the source is the New York Times. However, the difference between Republican and Democratic respondents is not statistically significant.

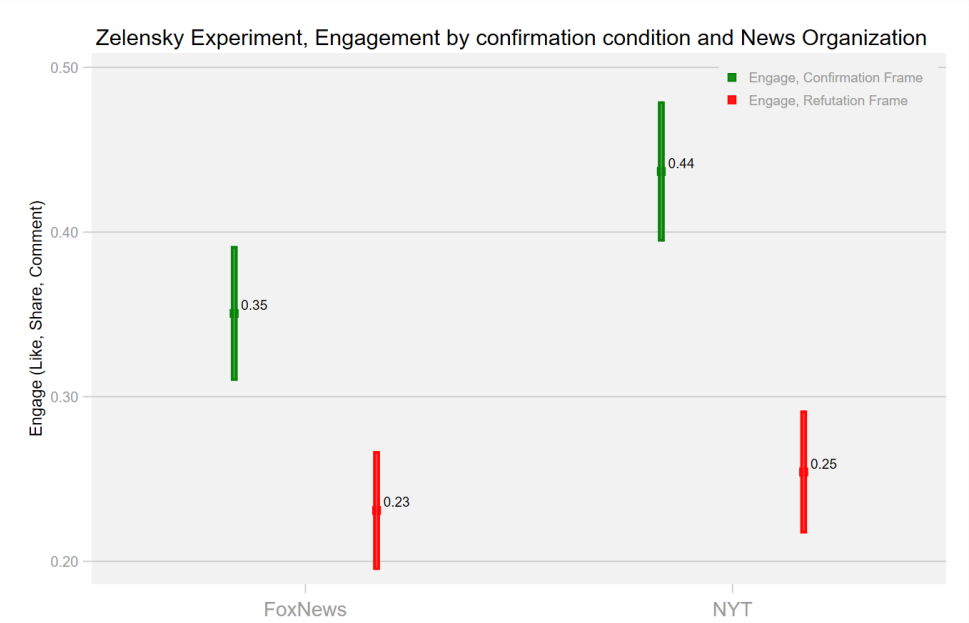

Figure 4 Confirmation (TRUE) and refutation (FALSE) by news organization. The confirmation and refutation statements are semantically identical but differ in cognitive accessibility and valence charge. Both the TRUE and FALSE adjudications are factually correct.

Considering that cognitive effort doesn't yield any measurable effect on engagement, the effect is driven by changes in valence. Table 1 reinforces this, as we are able to observe that the refutation frame elicits considerably more negative emotions, such as anger and disgust. Meanwhile, participants treated with the confirmation self-reported happiness and optimism after exposure to the treatment.

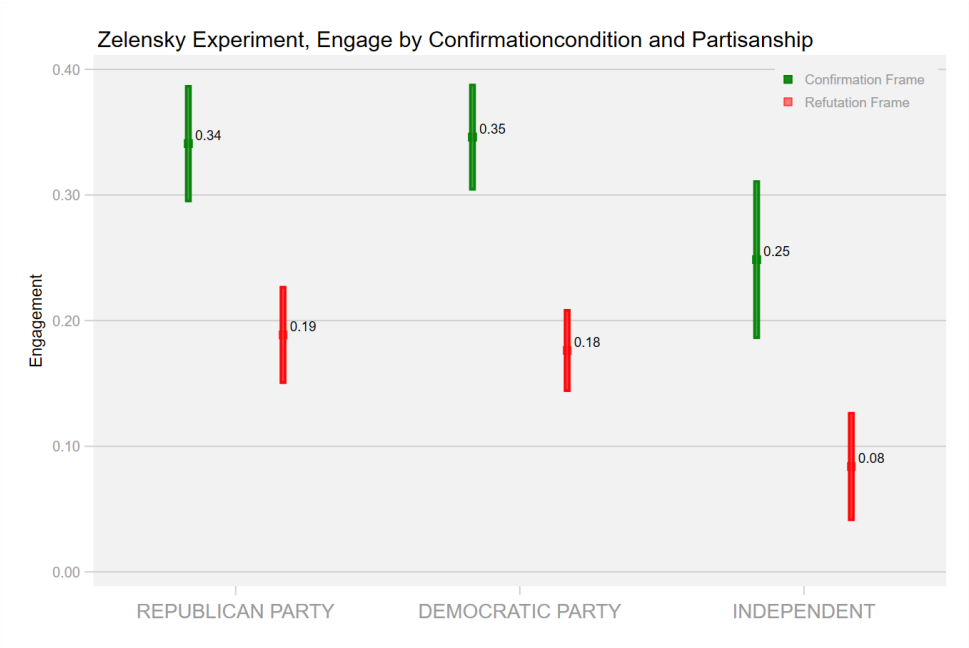

Figure 5 Confirmation (TRUE) and refutation (FALSE) by news organization. The confirmation and refutation statements are semantically identical but differ in cognitive accessibility and valence charge. Both the TRUE and FALSE adjudications are factually correct.

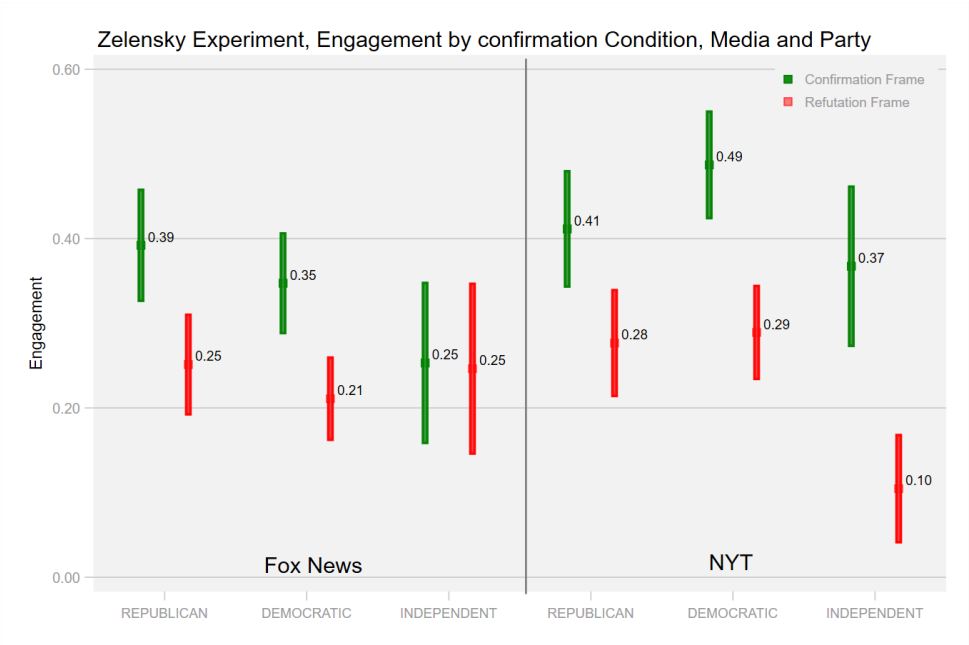

Figure 6 Confirmation (TRUE) and refutation (FALSE) by news organization. The confirmation and refutation statements are semantically identical but differ in cognitive accessibility and valence charge. Both the TRUE and FALSE adjudications are factually correct.

Table 1 Difference of Means between the *Confirmation* and *Refutation* Frames

Variable	Label "False"	Label "True"	Diff	P-value
Reactions				
Engage	0.243	0.394	0.151***	0.000
Like	0.167	0.328	0.160***	0.000
Share	0.046	0.061	0.015	0.117
Comment	0.058	0.056	-0.003	0.790
Ignore	0.757	0.606	-0.151***	0.000
Emotions				
Angry	0.131	0.089	-0.042***	0.003
Happy	0.049	0.118	0.069***	0.000
Disgusted	0.191	0.121	-0.070***	0.000
Fearful	0.026	0.050	0.023***	0.006
Sad	0.131	0.189	0.058***	0.000
Stressed	0.083	0.123	0.040***	0.003
Indifferent	0.389	0.310	-0.078***	0.000

Note: Robust standard errors in parentheses .
Full set of models in the SIF file to this article.
*** $p < 0.01$, ** $p < 0.05$, $p < 0.1$

Limitations of this Study

Some limitations of this study should be noted. First, while our study is internally valid, exposure and attention to the treatments is higher in surveys than in corrections that organically circulate on Facebook. Therefore, the external validity of the experiment cannot be assessed. Our study describes the behavioral mechanism for sharing *confirmations* and *refutations*, but the baseline rate for sharing the treatment would likely be lower in the wild.

A second limitation of this study is that we measure the behavioral response *after* the treatment rather than jointly with the treatment. This is done to retrieve measures of time-to-read and time-to-react. The experiment is internally valid because responses are collected following identical protocols. However, the behavior we measure is not identical to the unsolicited liking, sharing, and commenting on posts published on Facebook. As noted above, rates of "likes" are,

on average, higher than expected when multiple posts compete for the respondents' attention.

A third limitation of this study is that the affective response is self-reported. This is the standard instrument to measure affective responses in nationally representative surveys, but it adds the possibility that individuals report "desirable" emotions that may differ from the affective response they experience. Again, we expect findings to be internally valid, given that respondents are only presented with one of the possible treatments and, therefore, the self-report emotions are consistent within treatments.

Concluding remarks

Results from this experiment support a higher intent to "Engage" and "like" fact checks that use the confirmation frame rather than the refutation frame. Results converge with a previous study on COVID-19 vaccination. Consistent with our preregistered hypotheses, the effect of the confirmation frame on engagement is positive and statistically significant at $p < 0.01$. Our study extends prior research by showing that refutation frames with no additional cognitive effort will still elicit higher engagement.

The rejection of the cognitive burden hypothesis supports a valence-driven interpretation of confirmation frames in fact-checking. We find no evidence that difficulties in understanding the confirmation and refutation frames explain engagement, liking, or sharing rates. There is no significant difference in the mean processing time for each frame.

The results of our experiment have important policy consequences. Studies on the partisan effects of misinformation oftentimes obscure how fact-checkers contribute to polarization when framing corrections using refutation frames. Fact-checkers interested in reducing polarization and increasing exposure to their corrections will sometimes benefit from using confirmation

frames. This is not currently the norm, and fact-checkers concentrate on publishing refutation statements, which is warranted when dealing with disinformation campaigns but may be unnecessary for public service announcements. Our findings also differ from previous studies by showing that confirmation and refutation frames are robust to partisan media and voting preferences.

References

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

Aruguete, N., Batista, F., Calvo, E., Altube, M. G., Scartascini, C., and Ventura, T. (2023b). Reducing misinformation: The role of confirmation frames in fact-checking interventions.

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.

Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124.

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.

Meinert, J. and Krämer, N. C. (2022). How the expertise heuristic accelerates decision-making and credibility judgments in social media by means of effort reduction. *Plos one*, 17(3):e0264428.

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. and Lazer, D. (2019). Public health and online misinformation: challenges and recommendations. *Annual review of public health*, 41:433–451.

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

West, J. D. and Bergstrom, C. T. (2021). Misinformation in and about science. *Proceedings of the National Academy of Sciences*, 118(15):e1912444117.