Keep your promises, even when your peers do not:

A Survey Experiment on the Influence of Social Media on Trust

Natalia Aruguete[*] Ernesto Calvo [†] Carlos Scartascini [‡] Tiago Ventura [§]

March 16, 2025

Number of words: 8,995

Abstract

This study measures the effect of partisan and polarizing social media messages on political trust and trustworthiness in Brazil and Mexico. We implemented two survey experiments with approximately 2,300 respondents each, using a modified "trust game" to measure the effects of polarizing social media messages on two dimensions: trust (the belief that others will fulfill their pledges) and trustworthiness (fulfilling the pledges made to others). Among users exposed to polarizing partisan messages, findings show a statistically significant decline in trust (i.e., we perceive others will not keep their promises) and a null effect on trustworthiness (i.e., we keep the promises made to others). The decline in trust is larger if respondents actively '*like*,' '*share*,' or '*comment*' on the message. These findings underscore the role of active engagement with polarizing social media content as a mediator in diminishing trust.

JEL Codes: D72, D83, D91

Keywords: Trust, Trustworthiness, Social media, Political polarization

[*]Universidad Nacional de Quilmes, UNQ. Castro Barros 981, CABA, Argentina. nataliaaruguete@gmail.com

[†]University of Maryland, GVPT. 3140 Tydings Hall, College Park, MD 20742, USA. ecalvo@umd.edu.

[‡]IADB. 1300 New York Avenue, N.W., Washington, DC 20577, USA. CARLOSSC@iadb.org.

[§]Georgetown University, McCourt School of Public Policy, 37th Street NW O Street NW, Old North 100, Washington, DC 20057. Corresponding Author, email: tv186@georgetown.edu

1 Introduction

Understanding the effect of polarizing social media messages on trust and trustworthiness is substantively and theoretically important (Banks et al., 2020; Bail et al., 2018). Political trust is critical to citizens' commitment to the rule of law, norms, regulations, and democracy (Keefer, Scartascini and Vlaicu, 2018; Murtin et al., 2018; Scartascini and Valle L., 2020). Research shows that trust in governments increases political participation, voter turnout, support for institutional reforms, and improved compliance with political mandates (Levi and Stoker, 2000). In contrast, mistrust is associated with higher disaffection and lower support for long-term policies with broad-based benefits, such as investments in education. It also reduces support for policies whose benefits are difficult to observe, like bureaucratic reform. Mistrust increases support for policies whose effects are immediate and tangible, even if such policies do not foster long-term sustainable and inclusive growth (Keefer, Scartascini and Vlaicu, 2018).

In this article, we report the results of two pre-registered experiments measuring the effect of polarizing social media messages on trust (the expectation that peers will comply with pledges made to us) and trustworthiness (complying with pledges we make to peers). Our findings reveal a decline in trust among users exposed to polarizing social media messages but no decline in trustworthiness. Users perceive others as less likely to fulfill their pledges after reading polarizing social media messages, and the decline is larger if respondents actively *like*, *share*, or *comment* on the social media message. However, respondents abide by their pledges to others at unchanged rates, with trustworthiness largely unaffected. Therefore, when exposed to polarizing messages, respondents are less likely to *trust* others but remain equally *trustworthy*.

Our work offers three novel contributions to scholars studying social media, trust, and democratic governance. First, we find that partisan social media messages reduce trust behavior. Our experimental design allows us to show that this decline in trust behavior is self-interested and cannot be explained by the desire of the respondents to comply with the researcher or the instrument.

Second, we show that social media engagement magnifies the effect of the experimental treatment. This is a crucial contribution, showing that "doing" social media differs from "reading" social media. Engagement matters, and it affects political behavior. The finding is particularly relevant for the burgeoning literature on incidental exposure to news (Boczkowski, Mitchelstein and Matassi, 2018; Fletcher and

Nielsen, 2018; Settle, 2018; Weeks et al., 2017; Anspach, 2017). There is a larger decline in trust among respondents who actively *like* and *share* the treatment than in the control group. Our two-way design, comparing engaged and non-engaged users in the treated and control groups, shows that the decision to interact with a social media post increases the negative effect of polarizing messages.

Third, we contribute methodologically to the study of *trust games*, presenting a survey design that replicates important behavioral responses from in-person lab experiments. Our online survey design rapidly scales to large-N samples, increasing the study's external validity.

The organization of this paper is as follows. First, we describe the substantive importance of testing for the relationship between polarizing social media messages, trust, and trustworthiness. Second, we present our experimental design and its implementation in Mexico and Brazil. Third, we present our general experimental results, with estimates that distinguish between partisan source identity and the polarizing tone of the content. Fourth, we describe extensions of our results that show the mediating effect of sharing behavior on the effects of exposure on interpersonal trust. We conclude with a discussion of possible further extensions of our work.

2 Trust, trustworthiness, and social media framing

Trust and trustworthiness are fundamental forces that shape societies and institutions (formal and informal) and co-evolve with them (Arrow, 1974; Guiso, Sapienza and Zingales, 2004). Research shows that trust and trustworthiness have positive effects on the ability of people to make transactions and on the ability of governments to function (Arrow, 1974; Jacobsen, 1999; Zak and Knack, 2001; Algan and Cahuc, 2014; Bjørnskov and Méon, 2015; Algan et al., 2017). High trust correlates with higher growth, social progress, and democratic stability (Algan and Cahuc, 2010, 2014; Keefer et al., 2020). More importantly for democratic governance, recent research shows that declining trust makes citizens less likely to reach consensual policy decisions (Ryan et al., 2020).

Studying trust has become ubiquitous across many different fields. Most studies use well-known survey questions that measure *trust attitudes* rather than *trust behavior*. Examples include agreement questions such as "Most people can be trusted" and scale questions of reported trust in family, friends, and neighbors. This is problematic, as there is consistent evidence that trust attitudes and trust behavior are weakly correlated (Wilson, 2017). Importantly, the analytical connection between the social benefits

of trust and trustworthiness makes sense in terms of behaviors rather than attitudes.

In the last two decades, *trust games* have revolutionized economics and political science, generating data on trust and trustworthy behavior rather than descriptions of individual-level attitudes. Democratic representation is a particular type of trust game in which a voter (*the principal*) supports a politician (the *agent*) to act on her behalf. The *principal-agent* relationship is difficult, with decisions made by a politician often hidden from the public's view. This raises the specter of abuse by officeholders, who are expected to fulfill their mandates even if these do not align with their preferences or interests.

We expect politicians to be *worthy of our trust*, although they frequently deceive us (Hardin, 2002). We also consider ourselves to be *worthy of the trust of others*, although we are often willing to explain away why we default on our promises (Ariely and Jones, 2012). Our paper seeks to clarify the relationship between trust and trustworthiness in democratic representation, using a "trust game" that models how exposure to polarizing messages induces changes in interpersonal trust and trustworthiness behavior.[1]

Our experiment rotates four different social media posts to measure the effect of framing on trust and trustworthiness. Canonical work on framing effects (Entman, 1993; Iyengar, 1990; Arugute, Calvo and Ventura, 2023) shows that distinct frames alter the perceived legitimacy of an actor or event, with more polarizing frames activating "us vs them" identities and increasing negative feelings toward others (Mason, 2016; Banks, 2014). In the Supplemental Information Files (SIF), Section B, we provide a theoretical model based on a simple guilty game with latent parameters to describe the mechanism behind changes in trust and trustworthiness. Although evidence for the effects of social media usage on polarization has been widely investigated (Bail et al., 2018; Banks et al., 2020; Allcott et al., 2020; Asimovic et al., 2021), relatively little is known about similar effects on a broader set of citizens' behavior, such as interpersonal trust and trustworthiness. Of particular relevance is distinguishing the effects of incidental exposure and active engagement (Boczkowski, Mitchelstein and Matassi, 2018; Fletcher and Nielsen, 2018; Anspach, 2017; Stroud, 2010). Indeed, changes in trust behavior are modest when consumption is incidental and much more substantive when respondents actively engage with the treatments.

[1] For a general discussion of trust and trustworthiness, see Hardin (2002), Croson and Buchan (1999), and Fehr and Gächter (2000).

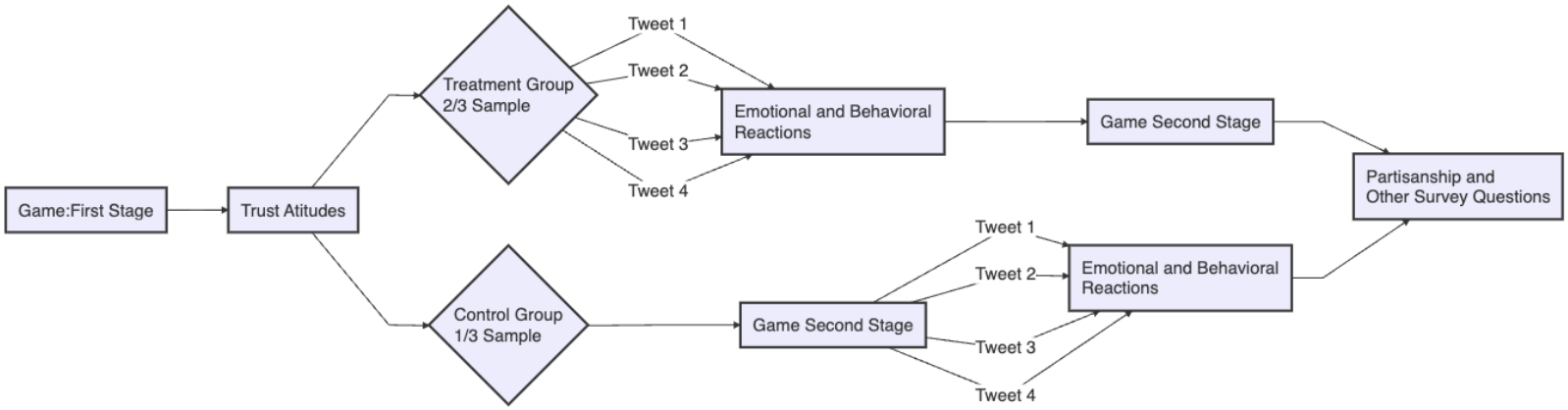
3 Trust Game: Nuts and Bolts

Game Sequence and Trust/Trustworthiness Interventions

To measure the effect of polarizing social media messages on trust and trustworthiness, we embed a political *trust game* in a survey experiment. In our game sequence, respondents select one of two fictional cartoon candidates who they are willing to support. We incentivize respondents to collect votes for their candidate with raffle tickets that allow them to win an iPad, conditional on their candidate winning the election. Following Cox (2004), the respondents' decisions to cast or entrust votes are independent of one another.

Respondents collect tickets to the raffle by earning votes for their candidate. Therefore, collecting as many votes as possible is incentive compatible: making sure their candidate wins makes them eligible to participate in the raffle, and collecting more votes increases their chances of winning the prize. Four iPads were distributed in Mexico and Brazil, making the odds/price ratio very attractive. At the time of the survey, the local price of an iPad was approximately 1.5 times the median monthly salary in Brazil and half the median salary in Mexico.

Figure 1 Survey Diagram



Respondents play two distinct roles: first, as an *agent*, respondents pledge to cast votes entrusted to them by other players (trustworthiness). Second, as voters, respondents cast votes directly (each vote collects one raffle ticket) or entrust votes to others (one vote collects two raffle tickets). In each round, respondents play firstly as agents, depositing votes entrusted to them (trustworthiness). Secondly they play as voters, depositing their votes or entrusting them to others (trust).

Each vote cast directly counts as a single raffle ticket (single vote). Every vote entrusted to others counts double (two raffle tickets), but only if deposited. Unbeknownst to our respondents, every entrusted vote is deposited by our "universal" respondent, but they do not receive this information. We repeat this procedure and analyze changes between the first and the second round. Figure 1 illustrates the experimental design and survey flow.

First Round: Setting a Baseline

Respondents select one of two candidates, Laura or Juan,[2] who have no distinctive markers other than their gender. When selecting a candidate, respondents are informed that they will have multiple opportunities to increase the votes they allocate to their candidate of choice and, more importantly, supporters of the overall survey winner (Laura or Juan) are eligible to participate in a raffle for one of two new iPads.

Respondents are then informed of how to increase the votes for their candidates and earn more raffle tickets. First, they win votes by reading a commitment pledge to act as agents for other respondents, to reinforce how votes are collected and to ensure that the importance of depositing votes is conveyed.[3]

[2]We used the respective translations for the names in Portuguese in the Brazilian survey.

[3]The instruction tells respondents that they can win five more votes for their candidate if they read the pledge: "*If other players delegate their votes to me, I agree to follow their preferences and cast their votes for the candidate they choose.*" To win the five votes, players are asked to answer either, "*I read the pledge*" or "*I did not read the pledge.*" We do not require them to sign the pledge, only to read it.

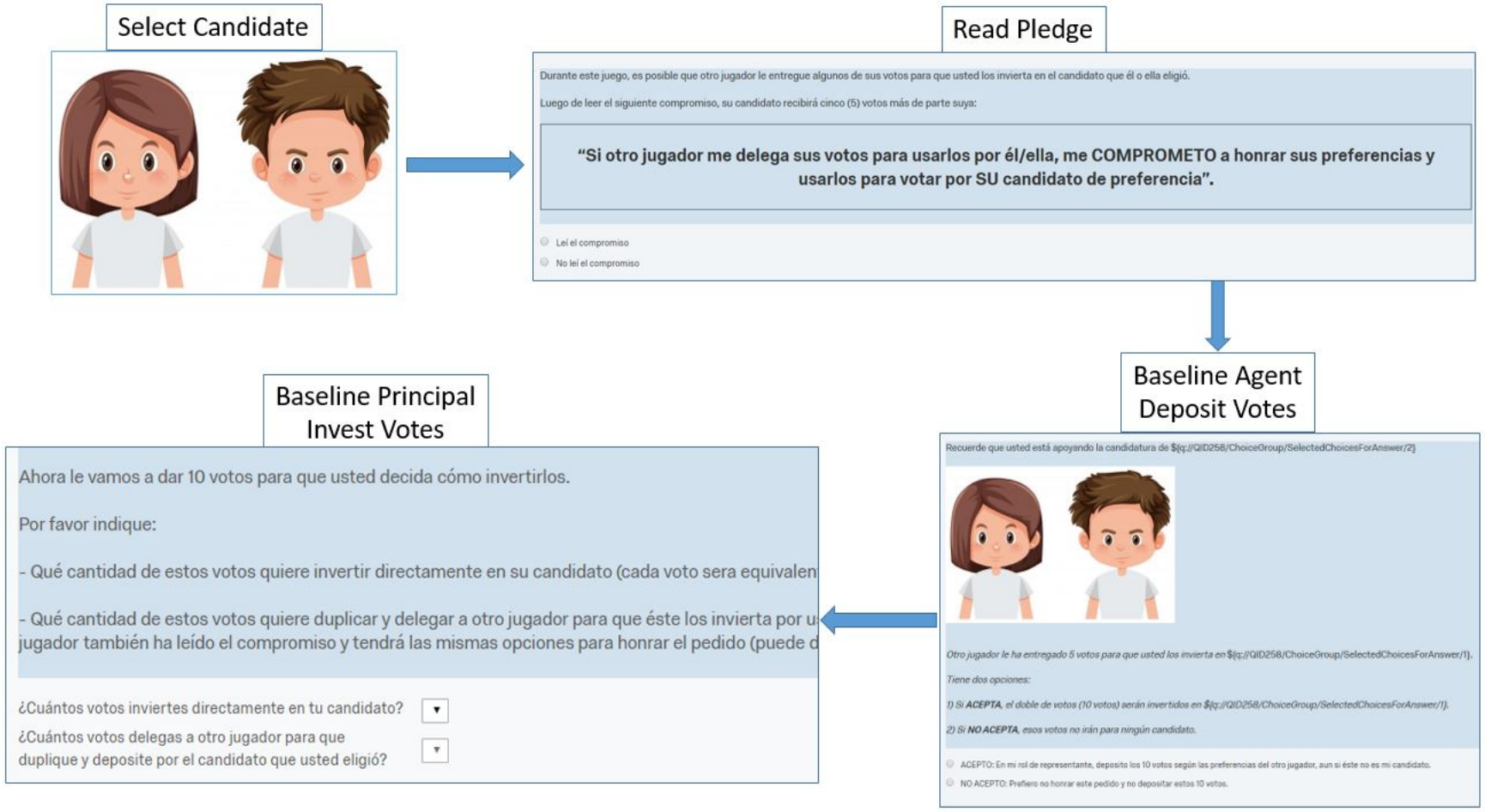


Figure 2 Experimental design: Respondents are asked to select a candidate to support throughout the survey. Respondents accumulate votes for their candidates; each vote counts as a ticket for a raffle among the winning candidate's supporters. Respondents may cast votes directly (one vote = one raffle ticket) or entrust votes to another player to cast. Entrusted votes count double, increasing the odds that their preferred candidate will win and the number of raffle tickets to win the iPad.

In this first round, we set a trustworthiness baseline, asking respondents to cast votes entrusted to them by another respondent. We then offered respondents 10 votes they could use to increase their raffle tickets. They may use these votes in two possible ways: (i) they can cast votes directly to support their candidate, or (ii) they can entrust votes to be cast on their behalf, with the provision that entrusted votes count double if deposited by a peer. All these steps are summarized in Figure 2.

Second Round: The Tweet Treatment

Once the experiment's baseline is set, we distract the respondent by asking various attitudinal, behavioral, and socio-demographic questions. These include questions about standard economic and political attitudes, measures of political knowledge, and perceptions of personal trust and trust in institutions.

We randomly select two-thirds of respondents to be exposed to tweets before playing the second round, with the remaining third serving as a control group (see Figure 1). We also collect the time they spend reading the Tweet, a standard approach serving as validation check and a measure of attention to the frames (Iyengar, 2011). After exposing respondents to the tweets, we ask if they would *'like'*, *'retweet'*, *'reply'*, or *'ignore'* the tweet they just read. We follow up with a question that asks how the tweet made them feel ("angry", "sad", "hopeful", etc.). Finally, we measure *trustworthiness* and *trust* for the treated group. They are then asked to invest in their candidates and to cast votes.

The treatment group reads the tweets before playing the second round of the game. The control group plays the second round and only then reads the tweet, ensuring we register a behavioral response to the tweets for the entire sample.

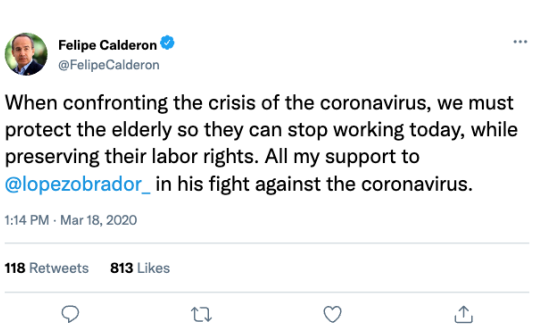
Social Media Frames

The experimental treatments in Brazil and Mexico present respondents with COVID-19 messages by prominent politicians. In March and April of 2020, COVID-19 was a salient and partisan issue in both countries (Calvo and Ventura, 2021; Aruguete et al., 2021). The message of the tweets varied in two dimensions: first, the party of the tweet's author (incumbent or opposition politician), and second, the tone of the messages (positive or negative). The experiment randomly rotated the author of the tweets between two high-level political figures from either the government or the opposition. On the other hand, the tweet's tone attributed blame to the out-group politician (polarizing tweet) or signaled a

willingness to cooperate (non-polarizing tweet) during the crisis.The complete wording of the treatment is presented in Section A of the SIF.

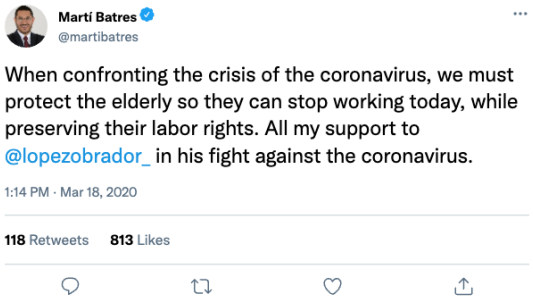
For the author, we use two prominent political figures in each country. In Brazil, we use Eduardo Bolsonaro, a member of the legislature and son of President Jair Bolsonaro, and Fernando Haddad, the leading candidate of the Workers' Party in the 2018 national election. For the Mexican case, we use Martí Batres, current senator from the ruling party, the National Regeneration Movement (MORENA); and Felipe Calderón, Mexico's president from 2006 to 2012, a leader of the opposition to the current government.

Figure 3 Tweets for the Treatment Conditions



a) Felipe Calderón x Non-polarizing Tweet (T1)

b) Felipe Calderón x Polarizing Tweet (T2)



c) Marti Batres x Non-polarizing Tweet (T3)

d) Marti Batres x Polarizing Tweet (T4)

To vary the tone of the message, we use a non-polarizing and a polarizing tone related to the COVID-19 crisis. In both countries, we use the same wording for the non-polarizing message, varying the polarizing message to increase congruence between the content and the political context in each case.

Non-polarizing messages frame the crisis as a moment of national union in which the president should lead the country; in the polarizing message, the author avoids blame for the crises and shifts responsibility to the opponent. The *Supplemental Information File* shows the complete set of treatments in each country. Here, we illustrate our framing design in Figure 3, with translated versions of the tweets used in Mexico. [4]

As described, one-third of the respondents were randomly assigned to the control group and two-thirds to the treatment group (see Figure 1). Participants in the treatment group read the social media message before the second round of the trust game, while those in the control group read the message after the second round. Therefore, when entrusted with or entrusting votes to others, the control group is unaffected by the Tweet frame.

4 The Hypotheses: Trust and Trustworthiness

Non-polarizing messages report to voters the willingness of political elites to cooperate with rivals to fight the COVID-19 pandemic. The messages signal respondents the importance of unity and cooperation in managing the crisis. Polarizing partisan messages blame political opponents for sowing conflict and weakening the needed response to the crisis. These polarizing tweets frame the COVID-19 response as an "us vs. them" problem (Iyengar, Sood and Lelkes, 2012; Iyengar and Westwood, 2015; Mason, 2016). The initial hypotheses of the experiment, as stated, reflect the expectation that non-polarizing social media messages will increase trustworthiness and trust while polarizing messages will reduce both.

HT_0A**: Non-polarizing social media messages increase compliance by agents and trust among principals.**[5]

HT_0B**: Polarizing social media messages decrease compliance by agents and trust among principals.**

Because the tweets may be endorsed by politicians who are aligned or misaligned with the preferences of the respondent, we test for the effect of partisan alignment (in-group) or partisan misalignment

[4] Full wording of the Treatment Tweets for both countries in original language and in English provides in Table 1 in the Supplemental Information Files

[5] The Pre-Approved plan used the term "partisan" instead of "polarizing," which was correctly flagged by reviewers as confusing. The wording of the original hypotheses was modified accordingly.

(out-group) on *trust* and *trustworthiness*.[6] A broad literature in political behavior shows that partisan alignment is central to attitude formation in areas as distinctive as candidate evaluation, economic perceptions, support for democracy and authoritarianism, and policy preferences (Green, Palmquist and Schickler, 2004; Arceneaux, 2008; Slothuus and De Vreese, 2010; Evans and Andersen, 2006; Zaller, 1992). Informed by the literature on partisan identities, we expect the endorsement of out-group politicians to augment the effect of non-partisan and partisan messages on trust and trustworthiness:

HT_1A**: Non-polarizing social media messages from misaligned politicians result in larger gains in trustworthiness among agents and in trust among principals.**

HT_1B**: Polarizing social media messages from misaligned politicians result in larger declines in trustworthiness among agents and in trust among principals.**

Research suggests that individuals perceive social media platforms as conduits for increased polarization. Therefore, we expect the mean levels of trustworthiness and trust in individuals in the treatment group will be lower than in the control group. This leads to our third set of hypotheses.

HT_2**: On average, trustworthiness and trust will decline in later rounds of questioning, compared with the baseline measures.**

We also expect attention to the treatment conditions to moderate the effects of framing and cognitive dissonance. Research in political science and psychology suggests that the time spent answering a survey question is a valid measure of the respondent's cognitive effort (Berinsky, Margolis and Sances, 2014; Wise and Kong, 2005; Malhotra, 2008). The time spent reading tweets has been shown to increase the effect of the social media treatments (Banks et al., 2020). As we collect the time they spend reading the Tweet as a measure of attention, we can then test for the relationship between attention and engagement (Iyengar, 2011). We expect:

HT_3**: Higher engagement, such as lower latency (more time spent reading the tweets) and active engagement with the tweets ('likes,' 'retweets,' and 'replies'), will increase the effects of the treatments.**

[6]Throughout the paper we use the concepts of in-group/alignment and out-group/misalignment between voters and politicians interchangeably.

5 Descriptive Evidence: Trustworthiness and Trust

Descriptive Results for Trustworthiness

Tables 1 and 2 present descriptive information on the decision to cast the five entrusted votes (i.e., our measure of *trustworthiness*). In the first round, a total of 64% of Mexican and Brazilian respondents cast the entrusted votes, which, as noted earlier, reduced their chances of participating in the raffle. In the second round, casting rates declined to 59% and 51%, respectively.[7] Among those who agreed to cast entrusted votes in the first round, 20% in Brazil and 19% in Mexico defected in the second round. Among those who did not agree to cast votes, 22% and 15%, respectively, agreed to do so in the second round. Although casting votes reduces the chances of winning one of the prizes, most respondents still accepted their role of trustee and cast the votes of their peers as requested.

Table 1 Trustworthy, Transition Matrix (Brazil)

	Second Round		
First Round	Agree	Don't Agree	Total
Agree	51% (1213)	12% (295)	64% (1508)
Don't Agree	8% (189)	28% (666)	36% (855)
Total	59% (1402)	41% (961)	100% (2363)

Table 2 Trustworthy, Transition Matrix (Mexico)

	Second Round		
First Round	Agree	Don't Agree	Total
Agree	51% (1188)	13% (307)	64% (1495)
Don't Agree	5% (129)	31% (722)	36% (851)
Total	56% (1317)	44% (1029)	100% (2346)

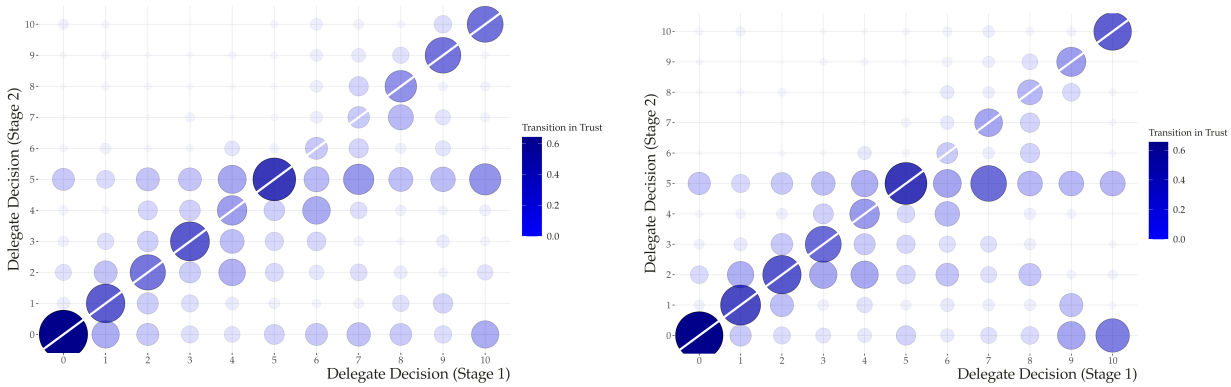
[7] We do not analyze the third round of the game here. However, trustworthiness in both countries remained almost unchanged in the third round.

Descriptive Results for Trust

Figure 4 presents descriptive results on the number of votes [0,10] entrusted to others, with the first round shown on the horizontal axis and the second on the vertical. The circles in Figure 4 describe the share of votes entrusted in the second round conditional on the respondent's decision in the first round. For example, the circles plotted on the diagonal of each figure represent respondents who entrusted the same amount of votes in the first and second rounds of the game. By contrast, the upper and lower triangles indicate an increase or decrease in trust.

Overall, we observe a decline in trust among respondents in our Mexican and Brazilian samples. Between the first and second rounds of the game, respondents consistently reduced the number of votes entrusted to other players and retained for themselves a larger number of votes, as can be readily inferred from the more populated lower triangle in Figure 4.

Figure 4 Trust: First and Second Rounds of the Game, Compared



a) Brazil b) Mexico

Note: The plots present changes in trust (votes entrusted) between the first and second rounds of the game in Brazil and Mexico. The upper triangle in each figure indicates the share of respondents who entrusted more votes in the second round (increase in trust), whereas the lower triangle indicates the share of subjects who entrusted fewer votes (decrease in trust).

6 Experimental Results

Descriptive evidence in the previous section shows that fewer respondents agreed to cast the votes entrusted to them (lower trustworthiness) between the first and second rounds, and smaller quantities were entrusted to other respondents (lower trust). In Brazil, deposits of entrusted votes (trustworthi-

ness) declined from 64% to 59%, and in Mexico from 64% to 56%. Similarly, entrusted votes to others (trust) in Brazil declined from 3.4/10 in the first round to 3.17/10 in the second, and in Mexico from 3.75/10 in the first to 3.24/10.

It is worth emphasizing that almost two-thirds of the respondents agreed to deposit the requested votes, while only a third of the resources were entrusted to others. The asymmetry between being trustworthy and distrusting others is very relevant and underscores different mechanisms for explaining these two behaviors.

The following two subsections show that, among the treated respondents (2/3 of respondents), polarizing social media messages had no measurable effect on trustworthiness but a statistically significant effect on trust compared to the control group (1/3 of respondents).

The Null Effect of Partisan and Polarizing Messages on Trustworthiness

Table 3 presents our findings on the effect of partisan and polarizing social media messages on trustworthiness. We estimate benchmark linear probability models to capture the effect of exposure to social media messages on the binary decision to cast votes entrusted by another player in the second round of the game. In the second round, our models interact exposure to the treatment with the subjects' first-round decision. Columns 1 to 3 present the results for Brazil, while columns 4 to 6 present those for Mexico. The baseline condition includes respondents who played the second round of the game without being exposed to polarizing social media messages. We then separate by treatment condition (partisan/non-partisan and in-group/out-group) and control for the first-round decision to cast votes.

In all, the estimates do not reject the null hypotheses in HT_1A and HT_1B. Only hypothesis HT_2 holds, showing a decline in trustworthiness in later rounds, consistent with most in-person implementations of the *trust game*. This decline, however, is not explained by exposure to the post. Therefore, contrary to our expectations, exposure to polarizing social media messages and varying the endorsement and framing of the message does not affect the trustworthiness of respondents. SIF Section D presents results without interaction with the subjects' first-round decision and reports null findings.

While the decision to trust another person depends on how we evaluate their behavior (i.e., their likelihood of not complying with our request), the decision to be trustworthy affects our self-image. It is governed by the tension between potential gains and our guilt sensitivity (Battigalli and Dufwenberg,

2007). We expect others to be less trustful as gains from deception increase, but we may consider the psychological cost of deceiving others too high. Therefore, we may expect others to be less likely to fulfill their promises even if we are not less likely to fulfill ours when tempted by rewards. The lack of a negative effect of polarizing partisan messages on trustworthiness is promising, as it may indicate that individuals could fear that others are not to be trusted without themselves being less trustworthy. The concluding section discusses potential venues to further test for this possibility.

The Negative Effect of Partisan and Polarizing Messages on Trust

Unlike the case for trustworthiness, our model results show that polarizing social media Tweets by out-group politicians reduce trust. We begin by presenting conservative estimates of the effect of our experiment on trust, separating dissonant messages (out-group politicians) and polarizing messages (negative tone). Then, we present statistical models and estimate the marginal effects of the treatments.

Figures 5 and 6 separate the results of our experiment by out-group/in-group condition (partisan alignment) and by the polarizing/non-polarizing conditions (tone of the tweet). Separating the two treatment conditions, we find robust and statistically significant results when respondents are exposed to messages by out-group politicians (*dissonant messages*). Results are inconclusive when considering only the negative tone of the social media post (*polarizing partisan message*), as they are significant for Brazil but not for Mexico.

The upper left plot in Figure 5 visually confirms a statistically significant difference between respondents in the treatment and control groups exposed to messages from out-group politicians. The negative effect of the tweet is larger for respondents who entrusted more than four votes in the first round. Results are substantively similar but less robust in the case of Mexico (Figure 6).

By contrast, exposing respondents to tweets from politicians they support yields small effects in Brazil and null results in Mexico. In addition, the lower left plots in Figures 5 and 6 show that, compared with the control group, polarizing political messages produce a modest decline in trust in Brazil but have no significant effect in Mexico. Given that we are not considering the joint effect of an out-group politician posting a partisan tweet, the results reported in this section are very conservative.

In Table 4, we present the results using benchmark ordinary least squares (OLS) analysis to capture the treatments' effect on trust in the second round of the game. Because changes in trust are heteroge-

Table 3 Regression Models: Treatment Effects of Framing and Endorsement on Trustworthiness

	Brazil			Mexico		
	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	0.344***	0.416***	0.462***	0.240***	0.303***	0.207*
	(0.074)	(0.082)	(0.097)	(0.076)	(0.095)	(0.107)
Trustworthiness (Round 1)	0.589***	0.591***	0.594***	0.634***	0.632***	0.634***
	(0.032)	(0.032)	(0.032)	(0.029)	(0.029)	(0.029)
Framing: Polarizing	0.035			−0.027		
	(0.036)			(0.034)		
Framing: Non-Polarizing	0.005			−0.019		
	(0.036)			(0.033)		
Out-group		−0.028			0.0005	
		(0.040)			(0.043)	
In-group		0.019			−0.030	
		(0.041)			(0.042)	
Polarizing Out-group			−0.032			−0.011
			(0.052)			(0.063)
Non-Polarizing Out-group			−0.024			0.005
			(0.050)			(0.052)
Polarizing x Trustworthiness (Round 1)	−0.021			0.013		
	(0.045)			(0.042)		
Non-Polarizing x Trustworthiness (Round 1)	−0.010			0.012		
	(0.045)			(0.042)		
Out-group x Trustworthiness (Round 1)		0.028			0.002	
		(0.050)			(0.054)	
In-group x Trustworthiness (Round 1)		−0.007			0.039	
		(0.050)			(0.052)	
Polarizing Out-group x Trustworthiness (Round 1)			0.015			0.009
			(0.065)			(0.078)
Non-Polarizing Out-group x Trustworthiness (Round 1)			0.038			−0.001
			(0.062)			(0.066)
N	2,128	1,607	1,156	2,219	1,426	1,084
Adjusted R^2	0.331	0.347	0.346	0.391	0.395	0.379

Notes: The models use benchmark OLS estimation. The dependent variable uses the decision to cast votes entrusted by other players, thus measuring subjects' levels of trustworthiness. A battery of individual-level pre-treatment controls—such as, age, income, employment, education, gender, and individual level of trust—are controlled for in all six estimations. *p<0.1; **p<0.05; ***p<0.01

Figure 5 Changes in Trust among Treated and Untreated Respondents in Brazil

Note: Plots compare changes in trust (votes entrusted to others) between the first and second rounds of the game in Brazil. Four treatment conditions are compared with the control group: the effect of a message from an out-group politician, the effect of an in-group politician, the effect of a polarizing tweet attacking opponents, and the effect of a non-polarizing tweet calling for unity. This figure does not evaluate the joint effect of out-group and partisan tone. Local polynomial lines with confidence intervals.

neous, as shown in Figures 5 and 6, we use an interactive linear model between the treatments and the decision to entrust votes in the first round of the game. Columns 1 to 3 present the results for Brazil of each set of specifications, and columns 4 to 6 for Mexico.[8]

SIF Section D presents similar linear models without modeling the heterogeneity of the results conditional on the first-stage levels of trust. In the models that do not consider the first-round decision to entrust votes, we cannot reject the null hypothesis that their trust behavior changes after respondents are exposed to the treatment. However, when heterogeneity is fully modeled, we detect a significant reduction in trust among participants who were more trustful in the first stage of the game. We argue that modeling the first-round decision deals correctly with the mechanical effects of our measurement

[8]The control group for all models consists of respondents who played the second round of the game without reading the social media message.

Figure 6 Changes in Trust among Treated and Untreated Respondents in Mexico

Note: Plots compare changes in trust (votes entrusted to others) between the first and second rounds of the game in Mexico. Four treatment conditions are compared with the control group: the effect of a message from an out-group politician, the effect of an in-group politician, the effect of a polarizing tweet attacking opponents, and the effect of a non-polarizing tweet calling for unity. This figure does not evaluate the joint effect of out-group and partisan tone. Local polynomial lines with confidence intervals.

choice. For example, a participant who donated a small amount in the first round, even with a negative shock in their trust behavior has little space to signal their decrease in trust in the second round. On the other hand, a participant who expressed high levels of trust in the first round had a higher budget to signal changes in their willingness to trust other players after being exposed to our treatment. As in Figures 5, 6, and 7, the effect of the treatment has the expected negative effect once the first-round decision is taken into account in the interactive models presented in Table 4.

In models 1 and 4, we model the effects of the polarizing framing of the tweets for Brazil and Mexico. In models 2 and 5, we estimate the effects of reading a message from an out-group politician. We consider the vote intention of the respondent, "if elections were to take place next week," and the author of the tweet to distinguish the effect of a message posted by an in-group or out-group politician.

While we cannot reject the null hypothesis for an unmediated effect of polarizing framing on trust

behavior (models 1 and 4), we find that exposure to a tweet from an out-group politician, independent of the content of the message, yields a statistically significant decrease in trust among respondents in Brazil. After being treated with a tweet from a misaligned politician, respondents decrease the votes they entrust to other players. The effect is larger for higher levels of trust in the first round, as reported in Figure 5. Although the results are substantially similar in Mexico, the magnitude of the effects is smaller. Although the interaction term is not statistically distinct from zero, even for the Mexican case, reading a tweet from a misaligned politician has a negative marginal effect on trust.

Finally, models 3 and 6 evaluate hypotheses $H1_A$ and $H1_B$, with respondents playing the role of principals (voters). We estimate the effects of being exposed to a partisan message sent from an out-group politician. Results in both countries show statistically significant declines in trust after respondents are exposed to polarizing partisan social media messages from political opponents. These interactive effects help us understand the null results in models 1 and 4. A polarizing framing only reduces trust behavior when associated with an out-group politician; in-group polarizing content has a null effect on participants' later trust decisions.

Results are fully described in Figure 7, with marginal effects for two of our treatment conditions from models 3 and 6. Results describe the marginal change in the number of votes [0,10] entrusted in the second round as a function of trust in the first round. Figure 7 presents the effects of reading a tweet from a misaligned politician (models 2 and 4) and Figure 8 separates the out-group treatment according to the non-polarizing and polarizing tone of the tweet (models 3 and 5). The figures clearly show how out-group polarizing messages reduce interpersonal trust behavior. For both cases, we see that reading a polarizing message from an out-group politician reduces by almost 10% the votes entrusted to other players between the first and second stages of the trust game on respondents who, in the early stage of the game, exhibited higher levels of trust. These effects become statistically significant compared to the control group as trust in the first stage increases. The effect is substantively significant and, more importantly, describes a low-dosage treatment (one tweet) compared with the large number of tweets that users are exposed to on a daily basis.

However, it is also possible to see that the effect of the treatment is larger in Brazil. An analysis of the respondent's affective response to the tweets in Brazil and Mexico shows similar angry reactions. Therefore, in the next section, we explore in greater detail the mechanisms that explain differences in

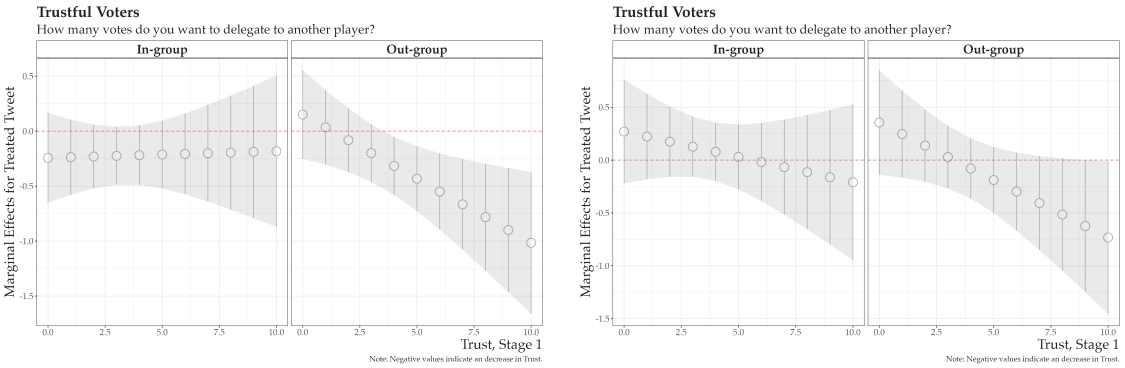
Table 4 Regression Models: Treatment Effects of Framing and Endorsement on Trust

	Brazil			Mexico		
	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	2.276***	2.037***	1.985***	2.514***	2.114***	2.322***
	(0.433)	(0.481)	(0.575)	(0.444)	(0.552)	(0.612)
Trust (Round 1)	0.460***	0.460***	0.460***	0.459***	0.462***	0.462***
	(0.031)	(0.031)	(0.031)	(0.032)	(0.032)	(0.032)
Tone: Polarizing	−0.052			0.299		
	(0.194)			(0.209)		
Tone: Non-Polarizing	−0.006			0.150		
	(0.195)			(0.204)		
Out-group		0.101			0.300	
		(0.216)			(0.264)	
In-group		−0.315			0.266	
		(0.217)			(0.261)	
Polarizing Out-group			0.083			0.676*
			(0.283)			(0.366)
Non-Polarizing Out-group			0.110			0.015
			(0.274)			(0.324)
Polarizing x Trust (Round 1)	−0.032			−0.050		
	(0.043)			(0.046)		
Non-Polarizing x Trust (Round 1)	−0.033			−0.039		
	(0.044)			(0.045)		
Out-group x Trust (Round 1)		−0.104**			−0.081	
		(0.048)			(0.057)	
In-group x Trust (Round 1)		0.022			−0.041	
		(0.050)			(0.058)	
Polarizing Out-group x Trust (Round 1)			−0.126**			−0.164**
			(0.062)			(0.079)
Non-Polarizing Out-group x Trust (Round 1)			−0.081			−0.018
			(0.062)			(0.069)
N	2,092	1,583	1,140	2,216	1,425	1,083
Adjusted R^2	0.232	0.234	0.218	0.200	0.196	0.202

Notes: The models use benchmark OLS estimation. The dependent variable uses the number of votes subjects (principals) entrusted in round 2 to another player to be doubled and cast for the principal's candidate. A battery of individual-level pretreatment controls—such as, age, income, employment, education, gender, and individual level of trust—are controlled for in all six estimations. *p<0.1; **p<0.05; ***p<0.01

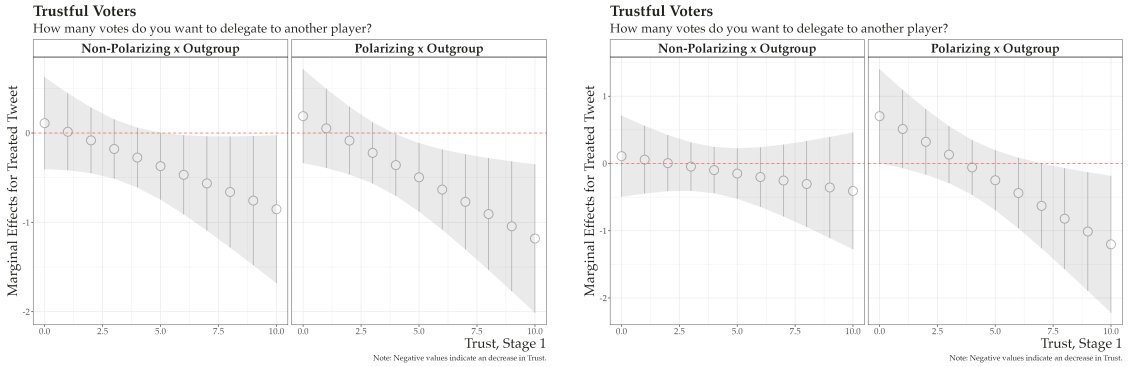
the decline in trust across both countries.

Figure 7 Marginal effects of source cue from Tweets on trust



a) Brazil b) Mexico

Figure 8 Marginal effects of the partisan treatment from a misaligned politician



a) Brazil b) Mexico

7 Mechanisms: The Role of Attention and Engagement

While results from the previous sections confirm the hypothesized effect of partisan and polarizing social media frames on trust, they provide limited information about the mechanisms that underlie our results or about the differences observed between Brazil and Mexico. Our survey, however, included validation checks to evaluate whether respondents correctly interpreted the partisan leaning of the social media frames and, more importantly, questions about respondents' engagement with the partisan treatments. In this section, we analyze these results in greater detail, introducing a double-identification strategy that isolates the effect of attention to social media on declines in trust.

Consider the effect of the treatment among respondents who engaged with the political tweets (by retweeting, liking, or replying) ***before*** answering our trust question (treatment group), compared with

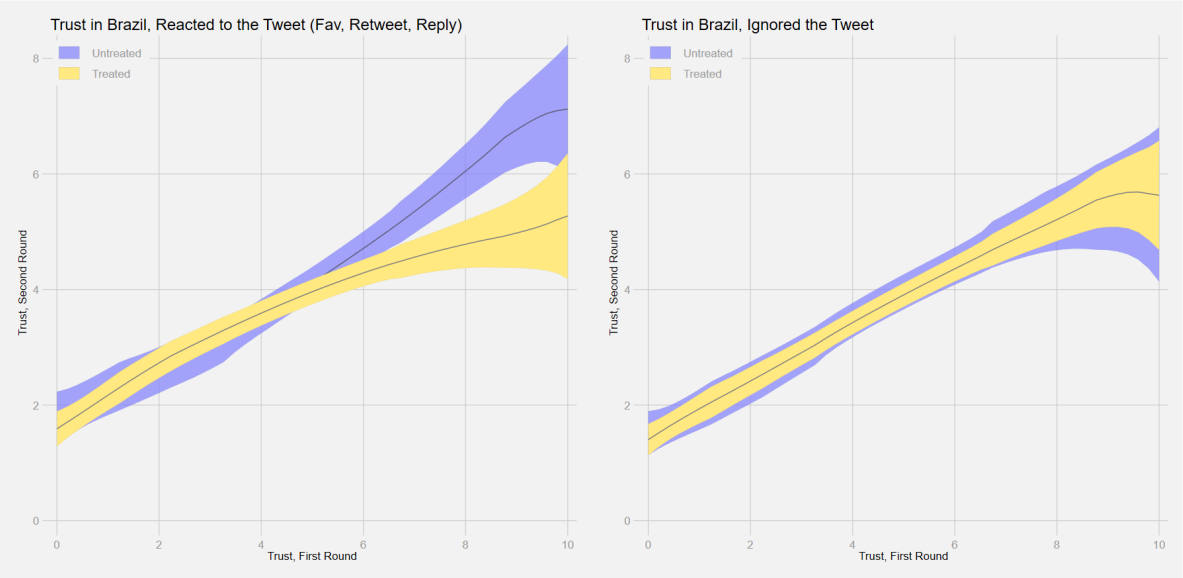
those in the control group who engaged with the tweet ***after*** answering the trust question. Given that the treatment consists exclusively of manipulating whether respondents play the trust game **before** or **after** reading the social media messages, our double-identification assumption only needs to assume that respondents assigned to the control group would have engaged with the tweet in the same way if they had been in the treatment group and not answered the trust question before engaging. We believe this is a reasonable assumption that allows us to identify the heterogeneity of the treatment effects conditional on behavioral reactions to the partisan and polarizing social media message.

Throughout this section, we repeat the same double-identification strategy (engaged treatment/engaged control, ignore treatment/ignore control) to isolate the mechanisms that explain a decline in trust. Consider Figure 9, which, as in the previous section, plots the trust decision in the second round (vertical axis) against the decision in the first round (horizontal axis). In Figure 9, the left plot compares the effect of the *treated-engaged* group (like, retweet, reply) against the *control-engaged* group. Meanwhile, the right plot describes the *treatment/ignore* group against the *control/ignore* group. Notable is the significant decline in trust among respondents who like, retweet, or reply to a tweet in the treatment group compared with respondents in the control group who were equally engaged with the tweet. Equally important is that those who ignore the tweet are almost indistinguishable between groups.

Results are revealing, showing a significant decline in trust only among respondents who engaged with the tweet ***before*** the second round (treatment), and null effects for respondents who engaged with the tweet but did so ***after*** the second round (control). In other words, if we consider only respondents who felt strongly about the tweet, the effect is large and significant only for the treatment group. By splitting the sample between those who engage with the tweet (treatment and control) and those who did not (treatment and control), we prove hypothesis HT_3 and are able also to test for the different mechanisms that explain the decline in trust.

Figure 10 depicts similar two-way comparisons, focusing on messages from out-group politicians (dissonant trait). We see larger treatment effects among those who like, retweet, or reply to the message (left plot). By contrast, incidental exposure (Boczkowski, Mitchelstein and Matassi, 2018) to the tweet, as shown in the plots to the right of Figure 10, has modest effects in Brazil and a null effect in Mexico. Indeed, conditioning on treatment and attention provides the strongest evidence yet of the effect of partisan and polarizing social media messages on trust. These differences are statistically significant at

Figure 9 Changes in Trust When Respondents Engage with the Tweet (left) or Ignore It (right)

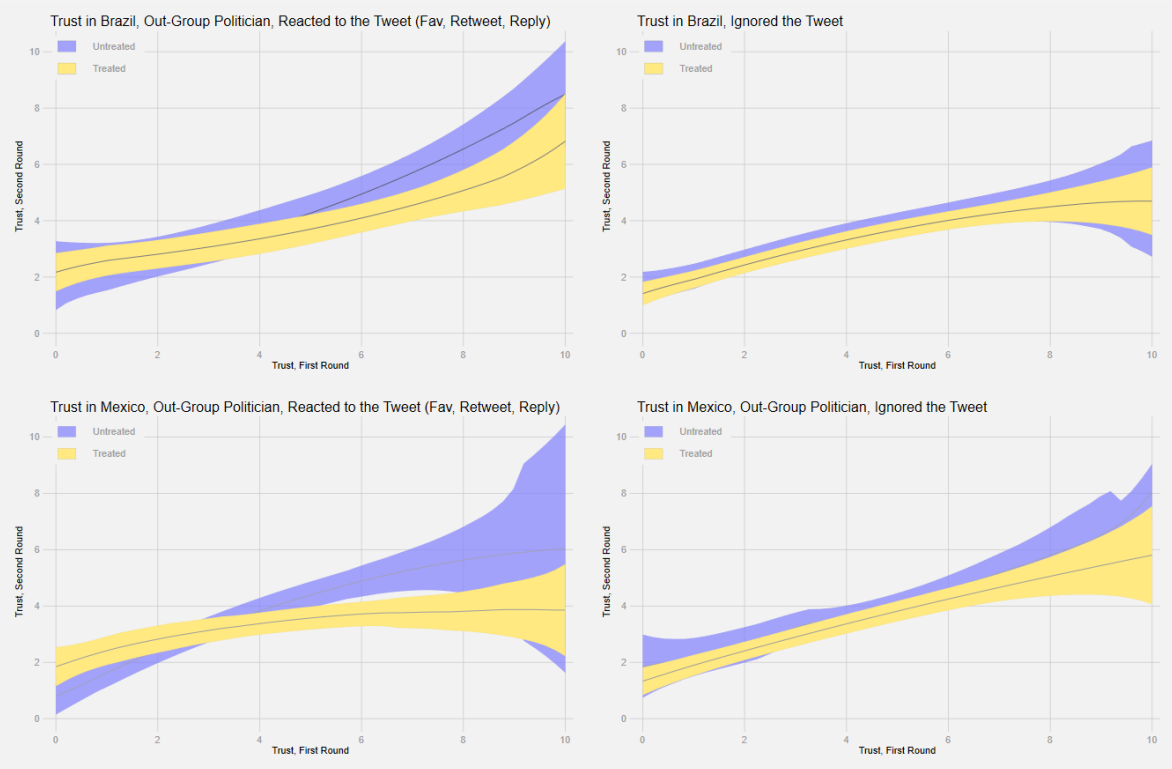


Note: The left plot estimates the treatment effect among voters who engaged with the tweet (like, retweet, or reply). The right plot estimates the treatment effect among those who did not engage (ignore). Results show a decline in trust only among respondents who saw and engaged with the tweet *before* the second round of the experiment. Those who engaged with the tweet *after* the experiment showed no decline in trust. We also found no effect among treatment and control respondents who ignored the tweet.

conventional levels, as shown in SIF Section D. Using a linear parameterization of the treatment effects, the effects of exposure to an out-group message and to our treatments overall only affect trust behavior of users who reacted to the tweets, as presented in Figures 9 and 10. As shown in SIF Section D, Tables 5 and 6, these differences are statistically significant at conventional levels. The comparisons between treatment and control groups with respondents who reacted or ignored our treatments are illuminating of how engagement, a crucial feature of the social media era, magnifies the declines of trust.[9]

[9] Again using our formal model in SIF Section B, given that all other parameters are held constant, we can confidently state that $\theta^*_{j,R1} - \theta^*_{j,R2} < 0$.

Figure 10 Changes in Trust When Respondents Engage with Dissonant Tweets



Note: The left plot estimates the treatment effect among Brazilian and Mexican voters who engaged with the tweet (like, retweet, or reply). The right plots estimate the treatment effect among those who ignore the tweet. Results show large declines in trust for dissonant tweets only among treated respondents who engaged with the tweet *before* the second round of the experiment. There is no effect of partisan dissonance in the control group and no difference in the treatment and control respondents who ignored the tweet.

8 Concluding Remarks: The Trust-Trustworthiness Gap

As political polarization increases, are we more likely to default on our promises to others? Are we less likely to trust others? This article shows that voters keep their promises to others even if they expect their peers not to do the same. Most of our respondents were willing to deposit votes entrusted to them, even if this decision reduced their chances of winning a reward. However, respondents did not trust their peers to be equally principled. The gap between trustworthiness and trust increased after exposing respondents to polarizing political messages. After exposure, respondents were equally trustworthy but reduced their trust in others. This is a sign of the times we live in. Polarization has little effect on our decisions to behave as good citizens, but we suspect others not to behave in the same way.

The absence of a negative effect on trustworthiness is a welcome finding that warrants emphasis. Almost two-thirds of the respondents agreed to deposit the entrusted votes as requested, a decision that, while contrary to their personal interest, is socially desirable. The high number of individuals willing to act upon their peers' requests contrasts sharply with the number of individuals entrusting votes to others. The asymmetry between trustworthiness and distrusting others is significant, especially considering that a lack of trust may persist even if politicians prove themselves trustworthy representatives of their voters. More principled politicians might be unable to overcome the increasing distrust from a toxic public sphere.

Our findings also show the negative effects of polarizing messages on trust are greater when respondents actively engage with social media messages through likes, retweets, and replies. Active engagement, therefore, increases the trust-trustworthiness gap. That is, as political interest and civic engagement increase, we are equally likely to be trustworthy but less likely to trust our peers. This result speaks to the challenges of addressing uncivil partisan messages through civic education and active participation. Overall, the double-identification strategy discussed in Section 7 provides robust evidence that the negative effect of polarizing partisan messages on trust behavior is greatest among those who are more keen to act on their political beliefs.

We should note that we measure engagement by asking participants to indicate their behavioral reaction to the message rather than providing them with the option to do so as in a "real" Twitter environment. Even though recent research shows that survey-based methods measuring sharing behavior in social media correlate with behavioral data (Mosleh, Pennycook and Rand, 2020), our measure of engagement lacks ecological validity. We expect researchers to improve the design in future studies by adopting more realistic environments to measure behavioral reactions using survey experiments.

A positive implication of our findings is that voters do not consider partisan political messages as a valid reason to transgress their principles. Recent research by Corbacho et al. (2016) shows that individuals who perceive others as corrupt are more likely to engage in corruption themselves. By contrast, our experiment finds no equivalent association between perceiving others as deceitful and behaving deceitfully. Therefore, the dissociation between trust and trustworthiness in the treatment group raises new questions to be explored in future work. Why do voters sometimes perceive that misdeeds by others are a valid reason to behave badly while at other times stick to their principles.

We believe implementing the proposed trust game as a survey experiment in two countries was a success. We find consistent estimates of trust behavior that are readily comparable across the two cases, with a design that allows us to determine the quality of the treatment (i.e., stronger in Brazil than in Mexico) and the importance of the mediating factors involved (i.e., engagement). We believe the survey design can be easily replicated and used to explore differences within and across countries, as with the laboratory version of the traditional trust game.

In 2020, as the world faced a serious health crisis, responsible social media platforms implemented measures to reduce toxic speech and address the increase in partisan political messages that delivered misinformation. With the purchase of X/Twitter by Elon Musk and the dismantling of the content moderation units at META, many of these safeguards have been lifted. As the public sphere becomes un-moderated, the implications of this study become more relevant. If trust is important for thriving democracies and economies, the increase in the trust-trustworthiness gap can be a second-order source of instability. More research is needed to assess how an increase in the presence of unmoderated toxic speech will alter the willingness of voters to behave ethically and the expectations that others will do the same.

How to address this trust-trustworthiness gap is an important problem that needs to be addressed by future research. It is common to demand reassurance from our politicians, asking them to prove they can be trusted and that they will not default on their promises. However, reassuring voters might be a tall order if uncivil partisan messages create perceptions that our behavior and that of our peers are increasingly dissimilar.

References

Algan, Yann and Pierre Cahuc. 2010. "Inherited trust and growth." *American Economic Review* 100(5):2060–92.

Algan, Yann and Pierre Cahuc. 2014. "Trust, Growth, and Well-Being: New Evidence and Policy Implications." *Handbook of Economic Growth* 2:49–120.

Algan, Yann, S Guriev, E Papaioannou and E Passari. 2017. "The European Trust Crisis and the Rise of Populism." *Brookings Papers on Economic Activity* 2:309–400.

Allcott, Hunt, Levi Boxell, Jacob Conway, Matthew Gentzkow, Michael Thaler and David Y Yang. 2020. "Polarization and Public Health: Partisan Differences in Social Distancing during COVID-19." *Available at SSRN 3570274* .

Anspach, Nicolas M. 2017. "The New Personal Influence: How Our Facebook Friends Influence the News We Read." *Political Communication* 34(4):590–606.
URL: *https://doi.org/10.1080/10584609.2017.1316329*

Arceneaux, Kevin. 2008. "Can partisan cues diminish democratic accountability?" *Political Behavior* 30(2):139–160.

Ariely, Dan and Simon Jones. 2012. *The (honest) truth about dishonesty.* Harper Collins Publishers New York, NY.

Arrow, Kenneth J. 1974. *The limits of organization.* WW Norton & Company.

Aruguete, Natalia, Ernesto Calvo, Francisco Cantú, Sandra Ley, Carlos Scartascini and Tiago Ventura. 2021. "Partisan cues and perceived risks: The effect of partisan social media frames during the covid-19 crisis in Mexico." *Journal of Elections, Public Opinion and Parties* 31(sup1):82–95.

Arugute, Natalia, Ernesto Calvo and Tiago Ventura. 2023. "Network activated frames: content sharing and perceived polarization in social media." *Journal of Communication* 73(1):14–24.

Asimovic, Nejla, Jonathan Nagler, Richard Bonneau and Joshua A Tucker. 2021. "Testing the effects of

Facebook usage in an ethnically polarized setting." *Proceedings of the National Academy of Sciences* 118(25).

Bail, Christopher A, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout and Alexander Volfovsky. 2018. "Exposure to opposing views on social media can increase political polarization." *Proceedings of the National Academy of Sciences* 115(37):9216–9221.

Banks, Antoine, David Karol, Ernesto Calvo and Shibley Telhami. 2020. "#Polarized Feeds: Two experiments on polarization, framing, and social media." *International Journal of Press/Politics* .

Banks, Antoine J. 2014. "The public's anger: White racial attitudes and opinions toward health care reform." *Political Behavior* 36(3):493–514.

Battigalli, Pierpaolo and Martin Dufwenberg. 2007. "Guilt in games." *American Economic Review* 97(2):170–176.

Berinsky, Adam J., Michele F. Margolis and Michael W. Sances. 2014. "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys." *American Journal of Political Science* 58(3):739–753.
URL: *http://dx.doi.org/10.1111/ajps.12081*

Bjørnskov, Christian and Pierre-Guillaume Méon. 2015. "The Productivity of Trust." *World Development* 70:317 – 331.
URL: *http://www.sciencedirect.com/science/article/pii/S0305750X15000169*

Boczkowski, Pablo J, Eugenia Mitchelstein and Mora Matassi. 2018. ""News comes across when I'm in a moment of leisure": Understanding the practices of incidental news consumption on social media." *New Media and Society* 20(10):3523–3539.

Calvo, Ernesto and Tiago Ventura. 2021. "Will I get COVID-19? Partisanship, social media frames, and perceptions of health risk in Brazil." *Latin American politics and society* 63(1):1–26.

Corbacho, Ana, Daniel W Gingerich, Virginia Oliveros and Mauricio Ruiz-Vega. 2016. "Corruption as a self-fulfilling prophecy: evidence from a survey experiment in Costa Rica." *American Journal of Political Science* 60(4):1077–1092.

Cox, James C. 2004. "How to identify trust and reciprocity." *Games and economic behavior* 46(2):260–281.

Croson, Rachel and Nancy Buchan. 1999. "Gender and culture: International experimental evidence from trust games." *American Economic Review* 89(2):386–391.

Entman, Robert M. 1993. "Framing: Toward clarification of a fractured paradigm." *Journal of communication* 43(4):51–58.

Evans, Geoffrey and Robert Andersen. 2006. "The political conditioning of economic perceptions." *The Journal of Politics* 68(1):194–207.

Fehr, Ernst and Simon Gächter. 2000. "Fairness and retaliation: The economics of reciprocity." *Journal of economic perspectives* 14(3):159–181.

Fletcher, Richard and Rasmus Kleis Nielsen. 2018. "Are people incidentally exposed to news on social media? A comparative analysis." *New media & society* 20(7):2450–2468.

Green, Donald P, Bradley Palmquist and Eric Schickler. 2004. *Partisan hearts and minds: Political parties and the social identities of voters*. Yale University Press.

Guiso, Luigi, Paola Sapienza and Luigi Zingales. 2004. "The role of social capital in financial development." *American economic review* 94(3):526–556.

Hardin, Russell. 2002. *Trust and trustworthiness*. Russell Sage Foundation.

Iyengar, Shanto. 1990. "Framing responsibility for political issues: The case of poverty." *Political behavior* 12(1):19–40.

Iyengar, Shanto. 2011. "Laboratory experiments in political science." *Cambridge handbook of experimental political science* pp. 73–88.

Iyengar, Shanto, Gaurav Sood and Yphtach Lelkes. 2012. "Affect, not ideologya social identity perspective on polarization." *Public opinion quarterly* 76(3):405–431.

Iyengar, Shanto and Sean J Westwood. 2015. "Fear and loathing across party lines: New evidence on group polarization." *American Journal of Political Science* 59(3):690–707.

Jacobsen, Dag Ingvar. 1999. "Trust in Political-Administrative Relations: The Case of Local Authorities in Norway and Tanzania." *World Development* 27(5):839 – 853.
URL: *http://www.sciencedirect.com/science/article/pii/S0305750X99000327*

Keefer, Phil, Ana María Rojas M, Carlos Scartascini and Joanna Valle L. 2020. Trust to Advance Inclusive Growth. In *Inclusion in Times of Covid-19*, ed. Victoria Nuguer and Andrew Powell. Inter-American Development Bank pp. 41–52.

Keefer, Philip, Carlos Scartascini and Razvan Vlaicu. 2018. "Shortchanging the Future: The Short-Term Bias of Politics." *Better Spending for Better Lives. How Latin America and the Caribbean Can Do More with Less. Development in the Americas report. Washington, DC, United States: Inter-American Development Bank* .

Levi, Margaret and Laura Stoker. 2000. "Political trust and trustworthiness." *Annual review of political science* 3(1):475–507.

Malhotra, Neil. 2008. "Completion time and response order effects in web surveys." *Public opinion quarterly* 72(5):914–934.

Mason, Lilliana. 2016. "A cross-cutting calm: How social sorting drives affective polarization." *Public Opinion Quarterly* 80(S1):351–377.

Mosleh, Mohsen, Gordon Pennycook and David G Rand. 2020. "Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter." *Plos one* 15(2):e0228882.

Murtin, Fabrice, Lara Fleischer, Vincent Siegerink, Arnstein Aassve, Yann Algan, Romina Boarini, Santiago González, Zsuzsanna Lonti, Gianluca Grimalda, Rafael Hortala Vallve et al. 2018. "Trust and its determinants: Evidence from the Trustlab experiment." *OECD Statistics Working Papers* 2018(2):0_1–74.

Ryan, John Barry, Talbot M Andrews, Tracy Goodwin and Yanna Krupnikov. 2020. "When Trust Matters: The Case of Gun Control." *Political Behavior* pp. 1–24.

Scartascini, Carlos and Joanna Valle L. 2020. Whom do we trust? The role of inequality and perceptions. In *The Inequality Crisis: Latin America and the Caribbean at the Crossroads.* Inter-American Development Bank pp. 329–351.

Settle, Jaime E. 2018. *Frenemies: How social media polarizes America.* Cambridge University Press.

Slothuus, Rune and Claes H De Vreese. 2010. "Political parties, motivated reasoning, and issue framing effects." *The Journal of Politics* 72(3):630–645.

Stroud, Natalie Jomini. 2010. "Polarization and Partisan Selective Exposure." *Journal of Communication* 60(3):556–576.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-2466.2010.01497.x*

Weeks, Brian E., Daniel S. Lane, Dam Hee Kim, Slgi S. Lee and Nojin Kwak. 2017. "Incidental Exposure, Selective Exposure, and Political Information Sharing: Integrating Online Exposure Patterns and Expression on Social Media." *Journal of Computer-Mediated Communication* 22(6):363–379.
URL: *https://doi.org/10.1111/jcc4.12199*

Wilson, Rick K. 2017. Trust Experiments, Trust Games, and Surveys. In *The Oxford Handbook of Social and Political Trust*, ed. Eric M. Uslaner. Oxford University Press.

Wise, Steven L and Xiaojing Kong. 2005. "Response time effort: A new measure of examinee motivation in computer-based tests." *Applied Measurement in Education* 18(2):163–183.

Zak, Paul J and Stephen Knack. 2001. "Trust and growth." *The economic journal* 111(470):295–321.

Zaller, John. 1992. *The nature and origins of mass opinion.* Cambridge university press.