Trustful Voters, Trustworthy Politicians: A Survey Experiment on the Influence of Social Media in Politics*

Natalia Aruguete† Ernesto Calvo ‡ Carlos Scartascini § Tiago Ventura ¶

November 30, 2020

Number of words: 11,517

Abstract

Recent increases in uncivil dialogue, political polarization, and fake news in social media raise questions about the relationship between negative online messages and the decline in political trust. We implement a *trust game* in a survey experiment with 4,800 respondents in Brazil and Mexico. Our design models the effect of social media on *trust* and *trustworthiness*. Survey respondents alternate as agents (*politicians*) and principals (*voters*), with rewards contingent on their preferred "candidate" winning the election. We measure the extent to which voters place their *trust* in others and are themselves *trustworthy*, that is, willing to honor requests that may not benefit them. Results provide robust support for a negative effect of uncivil partisan discourse on trust behavior and null results on trustworthiness.

*This research is part of the *Inter-American Development Bank* project: "Transparency, Trust, and Social Media", 1300600-01-PEC. PI: Ernesto Calvo, 2019-2020. We thank Elizabeth Zechmeister, Noam Lupu, and Maita Schade from LAPOP, who coordinated the probabilistic selection of respondents in Brazil and Mexico. We also thank Julia Rubio, who contributed to the survey design, as well as the members of our Lab (iLCSS-UMD). The experimental design was preregistered at the OSF Framework and can be accessed at https://osf.io/hvjkt. The experiment has received approval from the university's Institutional Review Board, number 1552091-3.

†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.

¶University of Maryland, GVPT. 4118 Chincoteague, College Park, MD 20742, USA. venturat@umd.edu.

1 Trustful Voters and Trustworthy Politicians

Is uncivil discourse in social media reducing political trust among voters? Is social media engagement making voters less trustworthy? The increases in uncivil dialogue (Mason, 2016; Iyengar, Sood and Lelkes, 2012) and polarization (Banks et al., 2020; Bail et al., 2018), and the proliferation of fake news in social media (Lelkes, Sood and Iyengar, 2017; Lazer et al., 2018) raise questions about the relationship between online exposure and the recent decline in political trust. In this article we report results of an experiment that shows significant declines in trust behavior among users exposed to, and who engaged with, uncivil partisan messages in social media.

Understanding the effect of social media exposure on trust is substantively and theoretically important. Political trust is critical to citizens' commitment to the rule of law, norms and regulations, and democracy. Research shows that politicians perceived as trustworthy are associated with increases in political engagement, higher voter turnout, citizen support for existing policies, support for institutional reforms, compliance with political authorities, and reciprocity (Levi and Stoker, 2000). Unfortunately, measures of trust have shown steady declines across the world over the past decade (Keefer, Scartascini and Vlaicu, 2018; Murtin et al., 2018; Scartascini and Valle L., 2020). These declines are concurrent with the rise of social media as a dominant platform for interpersonal communication and for the delivery of political news.[1] However, there is little research that tests

[1] As noted by Deibert (2019)"The third painful truth is that the attention-grabbing algorithms underlying social media also propel authoritarian practices that aim to sow confusion, ignorance, prejudice, and chaos, thereby facilitating manipulation and undermining accountability" (Deibert, 2019, 1).

for the relationship between social media exposure, social media engagement, trust, and trustworthiness.[2]

In this paper, we describe a survey experiment that implements a variant of the well known *trust game* in an electoral context.[3] We test whether respondents *trust* others to act on their behalf and whether they are *trustworthy* with respect to the resources entrusted to them. Trusting behavior increases the potential rewards perceived by the participants (more votes) but may carry large costs if the players' trust is betrayed (that is, if their candidate loses the election). By providing prizes associated with winning the election and raffle tickets for each vote they contribute to the win, we ensure that incentives are aligned as in the traditional trust game. After an initial round of the game, we randomly treat a subset of respondents to negative and positive tweets from incumbent and opposition politicians and measure changes in *trust* and *trustworthy* behaviors. We expect negative tweets from out-group politicians to activate partisan identities, even if those identities are orthogonal to the actual game being played. We implement our survey experiments using two large, randomized panels of 2,400 Brazilian and Mexican respondents each.[4]

Results from our experiment are consistent with traditional trust games, both in the total numbers of votes entrusted to other players as well as in trust behavior across rounds

[2] See Witmer and Håkansson (2015) for an overview of this discussion.

[3] *Trust games* are a well-established methodological strategy for studying economic exchanges in low-information environments (Berg, Dickhaut and McCabe, 1995). Individuals benefit from engaging in economic exchanges, but they may obtain extra benefits from providing a lower service or a lower payment than promised. Lower levels of trust, therefore, increase transaction costs and reduce economic activity. *Trust games* seek to reproduce behavioral trust responses rather than attitudinal ones.

[4] An agreement between LAPOP-Vanderbilt and Netquest enabled us to ensure a probabilistic selection of survey respondents.

of play. Regarding the experimental treatments, we find robust and statistically significant declines in trust among voters exposed to negative messages from out-group political figures (*dissonant messages*). Findings also support an activation of partisan identities and higher memory availability (Kahneman, 2011) of polarization frames.[5] Results are more modest when only the negative tone of the social media post (*uncivil discourse*) is considered. Finally, we do not find a statistically significant effect on trustworthiness. Agents cast entrusted votes at the same rate, regardless of the treatments they received. Overall, results provide support for a negative effect of social media exposure on trust but no support for a decline in trustworthiness.

Further testing of our findings indicates that "dosage" matters. We find that incidental exposure to social media has modest effects on trust. Results are pronounced and statistically significant at higher levels of engagement with tweets. Differences in trust between the control and treatment group are great when they "do" Twitter (like, retweet, reply) as opposed to when they "read" it (no engagement). These findings are new and important, as they point to differences between social media platforms and more traditional news outlets.[6] Our findings support "anger" as an important mediator in reducing trust, in line

[5] As shown by Bail et al. (2018) and Nyhan and Reifler (2010), exposure to counter-attitudinal arguments may create a "backfire" effect that increases political polarization and induces motivated reasoning on the users' side. Our study builds on these findings to show similar negative effects on interpersonal trust as a consequence of exposure to social media messages from a misaligned high-level politician. Partisan frames also increase recall through higher memory "availability" (Kahneman, 2011), which also explains lower latency for treated responses.

[6] Research in political communication describes several other important differences, including changes in editorial gatekeeping, in the practices and routines of journalists, and in exposure to different news frames propagated by peers (Shoemaker and Reese, 2013; Tandoc, 2014; Aruguete and Calvo, 2018). Showing that engagement is an important mediator in reducing trust also contributes to current research

with recent research on negative emotions and polarization (Mason, 2016; Banks, 2014; Banks, White and McKenzie, 2019).

In all, our work offers three novel contributions to scholars interested in the study of social media, trust, and democratic governance. First, we find that social media exposure leads to declines in trust behavior rather than a mere change in attitudes.[7] The decline in trust behavior is self-interested and cannot be explained by the desire of the respondents to interpret the intent of the survey instruments.

Second, we show conclusively that social media engagement magnifies the effect of the experimental treatment. There is a larger decline in trust among respondents who shared the content of the treatment, compared with those who were simply exposed to partisan messages. This is critically relevant for the burgeoning literature on incidental exposure to news (Boczkowski, Mitchelstein and Matassi, 2018; Fletcher and Nielsen, 2018). Our two-way design, comparing engaged and nonengaged users in the randomized experiment, supports the view that sharing behavior increases the negative effect of social media treatments.

Third, we contribute methodologically to the study of *trust games*, presenting a survey design that replicates important behavioral responses from in-person lab experiments. Although the use of online survey experiments reduces the number of measurements taken from each treated individual, it can be rapidly scaled up. More important, the use of random samples of respondents ensures that results have higher external validity. Addi-

on incidental consumption of news (Boczkowski, Mitchelstein and Matassi, 2018), indicating that the consequences of incidental exposure may be more modest than previously thought.

[7]See Tucker et al. (2018) for an overview of the recent literature on social media

tionally, our design brings the trust game from the more traditional "investment" setting to an electoral political scenario, which could prove useful for understanding the role of trust in voter-politician interactions.

The organization of this paper is as follows. First, we describe the substantive importance of testing for the relationship between social media exposure, 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 cognitive dissonance, negative tone of the content, and sharing behavior. Fourth, we describe extensions of our results that describe the mediating effect of negative emotions on trust. We conclude with a discussion of possible further extensions of our work.

2 Trust and Trustworthiness

Beginning with the work of Adam Smith, trust and trustworthiness have been recognized as key factors in promoting cooperation and exchange (Smith, 1937). 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). Trust and trustworthiness have positive effects on the ability of people to make transactions and on the ability of governments to function (Arrow, 1974; Knack and Keefer, 1997; Gambetta, 1988; Jacobsen, 1999; Zak and Knack, 2001; Algan and Cahuc, 2014*a*; Bjørnskov and Méon, 2015; Algan et al., 2017). High trust correlates with higher growth,

social progress, and democratic stability (Algan and Cahuc, 2010, 2014*b*; Aghion et al., 2010; Keefer et al., 2020). Importantly, if citizens do not trust their governments, they will not demand public goods or policies whose benefits materialize only in the long run (Keefer, Scartascini and Vlaicu, 2018; Keefer et al., 2020; Scartascini and Valle L., 2020).

Studying trust has become ubiquitous in recent years. Most studies use well-known survey questions that measure *trust attitudes* rather than *trust behavior*.[8] 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. This point is forcefully made by the literature on transaction costs, which establishes the positive effects of individuals who place their trust in others, irrespective of whether they agree that "most individuals can be trusted" (Bloom et al., 2012).

In the last two decades, *trust games* have revolutionized the field of behavioral economics and political science, generating data on trust and trustworthy behavior rather than reports on attitudes, which are frequently vitiated by misreporting and desirability biases (Berg, Dickhaut and McCabe, 1995). In the traditional *trust game*, individual trust behavior is studied in group settings where cooperation leaves everyone better off and self-interested behavior can make everyone worse off. In these laboratory games, one individual has an initial endowment that she can retain or pass to a second individual. The amount she passes ("invests") is multiplied (usually by 3) by the time it reaches the second individual. A second individual can keep all the receipts or reciprocate by send-

[8]Examples include agreement questions such as "Most people can be trusted" as well as scale questions of reported trust in family, friends, and neighbors.

ing back part or all. The amount passed by the sender is said to capture *trust*, and the amount returned to the trustor by the trustee to capture *trustworthiness* (Camerer and Loewenstein, 2003). As such, "investing" in a common endeavor makes sense only if the individual believes that other players will reciprocate. Johnson and Mislin (2011) offer a very thorough overview and a meta-analysis of 162 trust games.

Democratic representation is a particular type of trust game in which a *principal* (the voter) sends her vote to an *agent* (the politician), who will then act on her behalf. The *principal-agent* relationship is a difficult one, 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 be principled and 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).[9]

2.1 Nuts and Bolts of the Electoral Trust Game

In the political trust game that we embedded in our survey experiment, each respondent selects one of two fictional candidates. Respondents are informed that they must collect votes for their candidate of choice throughout the survey. At the end of the survey, those who chose the candidate who won the most votes are allowed to enter a raffle for two new

[9]For a general discussion of trust and trustworthiness, see Hardin (2002). In Hardin, trustworthiness is described as an instrumental trait, where a politician seeks to build a reputation over repeated interactions. Similar descriptions of trustworthiness as reciprocity are found in in Croson and Buchan (1999) and Fehr and Gächter (2000)

iPads. Respondents are also informed that their number of raffle entries will be equal to the number of votes they contributed to their candidate. Therefore, collecting as many votes as possible is incentive-compatible: it increases the chances that their candidate will win *and* it increases their chances of winning an iPad.

In this survey-based political trust game, respondents act the parts of a politician (agent) and a voter (principal). When answering questions as the agent (*politician*), respondents are asked to cast or discard votes entrusted to them (trustworthiness) by a universal player. When answering in the role of the principal (*voters*), respondents must decide how many votes to cast directly (one vote cast equals one raffle entry) and how many to entrust to another player (one entrusted vote that is eventually cast equals two votes and two raffle entries). However, they are warned that the agent (politician) may discard those votes; these are the same choices they themselves have when playing the other role. To ensure that there is no deception, our team serves as a universal player that honors all votes—those cast directly as well as those entrusted to another player.

Following Mazar and Ariely (2006), and in order to approximate better the relationship between representatives and voters, the role of the agent is reinforced by an initial pledge: "If other players delegate (entrust) their votes to me, I agree to follow their preferences and to use them to support the candidate of his or her choosing." After reading the initial statement, they are given five votes in support of their candidate of choice. The pledge is not binding. Representatives are required to read it but need not promise to comply with it; nor are there any sanctions for defaulting. Next, respondents are given the option of winning another three votes for their candidate of choice if they correctly

answer "2+2/2".[10]

In the first round, all voters first play the role of the agent (trustworthiness), to ensure that they are aware of their own decision as trustee before they decide how many of their own votes they will entrust to another player. All votes cast by the agents count toward the candidate favored by the trustee. After playing as agents, respondents decide how many of their own votes they will cast directly or entrust to others. The number of votes entrusted is doubled by the time it reaches the second individual.

For the voter, placing her trust in another player (the agent) offers the possibility of doubling the votes her candidate receives (as well as the raffle tickets that she herself earns). However, the other player may decide not to cast those votes, something that is made clear at the outset. Therefore, placing her trust in another player hinges on her belief that others will comply with her request. The fraction of votes that are not entrusted to another respondent will still benefit her candidate, but those that are entrusted to others count twice as much. All votes cast (single votes) or entrusted (double votes) are accepted by our universal respondent. Respondents are not given the option to change candidates; they have no information about whether entrusted vote are cast, and, just as important, they have no information on the relative standing of their candidates.

While *trust* is measured by the number of delegated votes, *trustworthiness* is the act of casting delegated votes, even if those votes serve the candidate whom the respondent does not support (thereby lowering the likelihood of winning the raffle). In our survey, trust

[10]In our sample, almost 74% of Mexican respondents and 64% of Brazilian respondents gave the correct response of 3, with most of the respondents who erred mistakenly selecting 2. We used this as one of our validation questions for attention and math skills.

is a behavioral response by principals (voters), whereas trustworthiness is the response from agents (politicians). As we will show, there are good theoretical reasons for the two behaviors to be weakly correlated, as they express different types of cognitive beliefs about oneself (*belief-based guilt*) and about the other players (*first-order beliefs*).

2.2 *Belief-Based Guilt*, *First-Order Belief*, and Framing

In our experiment, both *trust* and *trustworthiness* are repeated independent behavioral responses to the *potential for realizing gains* by entrusting one's votes to others or of *incurring losses* by casting other respondents' votes. There are no expected future interactions and no gains in reputation. As in Cox (2004), we isolate *trustworthiness* from other preferences such as reciprocity or altruism, given that respondents have no information about the individual who entrusted votes to them or about the individual to whom votes will be entrusted.[11] But why would anyone cast votes if doing so reduced their chances of winning the election and the raffle? Given that there is no accountability in the model, or in many countries, one of the potential explanations for why agents fulfill the mandate entrusted to them is *guilt aversion*, a concept to which we now turn.

Trustworthiness and Guilt Aversion

Battigalli and Dufwenberg (2007) coin the term *guilt aversion* to describe the psychological cost of letting other people down: "Player i's guilt may depend on how much he

[11] To avoid deception, our survey experiment institutes a universal respondent who carries out all requests received from respondents. Therefore, all votes entrusted by respondents were doubled and counted toward the respective candidates' total tally. As in Cox (2004), we effectively create a triad where respondents' decisions to cast or entrust votes are independent from one another.

lets j down. Player i's guilt may also depend on how much j believes i believes he lets j down."(Battigalli and Dufwenberg, 2007, pg. 170). In the absence of information about player j, *simple guilt* emerges, where the decision to cast the entrusted votes is a function of the perceived rewards, m_i, of defaulting on the request made by others, and the general guilt sensitivity, $\theta_i G_{ij}$, when no other information exists about the principal, j.

$$Pr(TWY_{ij}) = \phi(m_i - \theta_i G_{ij}) \tag{1}$$

In Equation 1, the probability that we will be *trustworthy* is affected by the subjective value of the reward, m_i, and by the guilt sensitivity, θ_i, of *agent* i for a generic *principal*, j. We are agnostic about the social, political, or psychological origins of "guilt" and consider the guilt parameter, θ_i, as a placeholder for a simple aversion to default on a mandate. Therefore, simple guilt describes the individual's propensity to act on a generic request. This θ_i parameter is sensitive to a number of exogenous shocks. Framing effects, we argue, are one possible mechanism that modulates the guilt sensitivity parameter, θ_i.

For example, consider the experiment proposed by Ariely and Jones (2012), wherein respondents are asked to "read a pledge" before being given the chance to cheat. We may think of this pledge as a heuristic device that increases the relative value of θ_i, the cost of *"letting other people down"*.[12] In our experiment, we expect negative messages from out-

[12]In our experiment, individuals are offered a "pledge" and the opportunity to click on the option, "I read the pledge." As in Ariely and Jones (2012), respondents who selected this option were considerably more likely to cast the entrusted votes, even though the question only asked respondents to *read* rather than *sign* the pledge. Further, even if respondents believed they were signing rather than reading the pledge, there were no sanctions for defaulting on it. Therefore, the only cost of defaulting accommodates Battigalli and Dufwenberg (2007)'s definition of *simple guilt*.

group politicians to increase negative feelings toward others (Mason, 2016; Banks, 2014), thereby reducing the value of θ_i (and reducing trustworthy responses). This expectation follows from the literature on generic or procedural frames, where the way in which a problem is presented alters the perceived legitimacy of an actor (Entman, 1993) or event (Iyengar, 1990).

Our approach to trustworthiness differs from the classification proposed by Ashraf, Bohnet and Piankov (2006), who distinguish between "unconditional kindness", "expectations of reciprocity", and "[instrumental] reciprocity." In our case, guilt is the result of defaulting on a request from another respondent. There is no "kindness" in casting entrusted votes; there is no reciprocity expected nor information collected about the individual who entrusts votes to be cast; and, finally, there are no instrumental benefits to be gained from being trustworthy.

Trust and First-Order Belief

Unlike *belief-based guilt*, where we pay the cost or reap the benefit of our decision instantaneously, trust is a cognitive belief about the future behavior of others. When we entrust our votes to others, we do not know if they will comply with our request. If we were to find out that the *principal* failed to cast our votes we would feel betrayed rather than guilty. Our decision to trust another person depends on how we evaluate the behavior of other respondents, which may or may not be related to our own guilt sensitivity. We expect others to be less trustful as potential gains from deception increase, m_j^*. We also expect others to be less likely to fulfill their promises if they have been remorseless or dishonest in the past. Trust, therefore, is a cognitive belief about other people's behavior, where

the subjective gains from m_j^* and the subjective losses from guilt θ_j^* remain unobserved.

Given that *agent* j decides to cast entrusted votes following Equation 1, an action that is unobserved by *principal* i, the share of delegated votes depends on our belief that $m_j^* - \theta_j^* G_{ij} > 0$, a belief that is unrelated to and not informed by our own guilt aversion parameter θ_i. Notice that not even the likelihood of betrayal depends on $\theta_i - \theta_j^* < 0$, given that others defaulting on their promise to be trustworthy is unrelated to how trustworthy we are. We may feel no remorse when defaulting on the mandate we received; at the same time, we may still be outraged by the failure of others to do so. Therefore, as shown in Equation 2, our decision to entrust others or to cast votes ourselves depends on unobserved values of how attractive to the other player is the unobserved prize, m_j^*, and how costly the unobserved guilt, θ_j^*.

$$T_{ij} = \alpha V_i + (1 - \alpha) V_i \phi(m_j^* - \theta_j^* G_{ij}) \tag{2}$$

If we assume an empathetic respondent who will do for others what she expects others to do for her, trustworthiness and trust would show a weak positive correlation, $cor(\theta_i, \theta_j^*) > 0$. Notice that θ_j^* remains unobserved by i and does not reflect actual information about the principal. Therefore, the value of m_j^* and the simple guilt parameter, θ_j^*, represent expectations of the respondent and not actual behavioral traits of another player.[13]

[13]In our experiment, the baseline round offers respondents the possibility of duplicating a fraction of the votes entrusted to others. We then treat two-thirds of the respondents to social media frames with the remaining third as controls. Given that we provide no information about the principal, j, and that reward m_i is held constant, we measure the changes in the expected sensitivity of the guilt parameter

As described by Buskens, Frey and Raub (2018), a "[trust] problem can arise if a voter decides on casting a vote for a representative who might later choose to initiate or support policies other than those preferred by the voter. The representative can benefit from changing his position opportunistically and the voter may regret having voted for this representative" (p. 2). In politics, we expect representatives to fulfill their "mandate" just as in our experiment we expect them to cast the entrusted votes. As is also the case for a politician, however, there is no obligation to fulfill the mandate and, in our experiment, no sanction for defaulting on it. In what follows, we describe an experiment designed to evaluate the effect of social media on trustworthiness (*guilt aversion*) and trust (*first-order belief*).

3 Survey Experiment and Hypotheses

3.1 Game Sequence and Trust/Trustworthiness interventions

To capture the role of social media on changes in trust and trustworthiness, we embed a survey experiment in a political trust game and expose respondents to contextually appropriate tweets from government or opposition political figures in Mexico and Brazil[14] As already noted, the opening question of the survey invites respondents to select one of two fictional cartoon candidates and informs them that they will be able to collect votes for their candidate throughout the survey. Once all respondents answer the survey, those

only in the second round, $\theta^*_{j,R1} - \theta^*_{j,R2}$.

[14] A different experimental design was used in Argentina, with negative messages evaluated by exposure time (*dosage*). Results from that experiment are reported in a separate article.

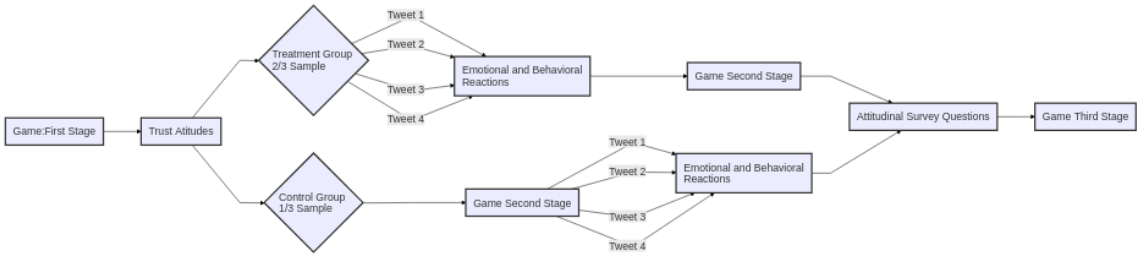
who supported the candidate who wins most votes are entered into a raffle for two new iPads. Respondents are informed that the number of their entries in the raffle will be equal to the number of votes they personally contribute to their candidate. Therefore, collecting as many votes as possible is incentive compatible: by making sure their candidate wins they become eligible to participate in the raffle, and by collecting more votes for their candidate they increase their chances of winning an iPad. At the time of the survey, the local price of an iPad was approximately 1.5 times times the median monthly salary in Brazil and half the median salary in Mexico.[15]

When answering questions as the agent (*politician*), respondents are asked to cast or discard votes entrusted to them (trustworthiness). When answering in the role of the principal (*voters*), respondents decide how many votes to cast directly and how many to entrust. In the first round, all voters first play the role of the agent (trustworthiness). As in our theory, we are interested in setting participants' guilt sensitivity parameter, θ, before they entrust votes to others. All votes cast by the agents count toward the candidate favored by the trustee. After playing as agents, respondents decide how many votes to cast directly or to entrust to others. All votes cast (single votes) or entrusted (double votes) are accepted by our "universal" respondent.[16]

[15]A total of four iPads were distributed in the raffles in Mexico and Brazil, making the odds/price ratio very attractive.

[16]The rounds of casting and entrusting are repeated three times, with some respondents assigned to social media treatments and others to the control group. In this paper, we focus on the first treatment sets, which compare the baseline stage (first round) to the first treatment stage (second round). Comparing the first and second rounds of our two identical experiments in Brazil and Mexico minimizes cross-effects from survey questions that could account for changes in levels of *trust* or *trustworthiness*. Separate analyses model the effect of frame elements on trust and trustworthiness in a third round.

Figure 1 Rounds of the *Trust Game* Embedded in the Survey



Note: All survey respondents take part in the first round. After we measure trust attitudes, a third of the respondents are assigned to the control group; the remaining two-thirds are treated with positive and negative political messages from in-group and out-group politicians (four different designs). Emotional and behavioral reactions include propensity to "retweet", "like" or "reply" as well as questions about affective responses ("anger", "joy", "disgust", etc.). The design allows us to isolate the effects of political dissonance and uncivil discourse. Further, in the extensions we provide evidence of behavioral and emotional triggers that reduce trust in respondents in the treatment group but not among the control group.

Rounds

In the appendix, we present a full description of the trust game and the embedded framing experiment, including graphics of the survey flow and the two rounds of the game. Here we present a summary. Figure 1 is a graphic representation of the experimental design and survey flow. We analyze the trustworthy-trust behavior twice, in a baseline round and in a second round. By survey design, respondents' odds of winning a reward are conditional on their candidate winning the overall election and the number of "votes" they contribute to that victory. Fulfilling a request to cast votes for another candidate,

therefore, reduces the likelihood that respondents will personally benefit once the survey is completed.

By comparing the baseline round and the second round, we can measure changes in *trust* and *trustworthiness*, which depend on the simple guilt parameters, $\theta_j^* G_{ij}$ and $\theta_i G_{ij}$, respectively. Given that the value of the reward (m^*) remains constant and that no other information is provided about likely partners, changes from the first to the second round can occur only through the guilt parameter. Even though θ_j^* remains unobserved, we can measure changes in trustworthiness between rounds 1 and 2, $\theta_{j,2}^* - \theta_{j,1}^*$, because everything else in Equation 1 remains constant. Therefore, our model is experimentally distinguished by the use of two rounds administered to treatment and control groups.

After the baseline round, one-third of the respondents are set as controls, while two-thirds are treated with positive and negative political messages (tweets) from in-group and out-group politicians.[17] Negative messages from out-group politicians are expected to decrease *trustworthiness* and *trust*, with social media frames altering the perceived value of the guilt parameters, $[\theta_i, \theta_j^*]$. Our experimental design allow us to compare the baseline trustworthiness and trust of respondents assigned to the treatment and control conditions. We expect "positive" messages to increase baseline trustworthiness and trust. We expect "negative" messages to decrease trustworthiness and baseline trust.

[17]The opposition politicians are Fernando Haddad (Brazil) or Francisco Calderon (Mexico), while the incumbent politicians are Eduardo Bolsonaro (Brazil) or Marti Bartres (Mexico). Respondents are informed upon conclusion of the survey that the tweets were edited by our team. Nothing in the tweets constitutes information that, if believed, could harm the respondents. Edited tweets ensure that the treatments are similar in all respects except the endorsement figures and the negative or positive frames. In the appendix, we present the images and text of the treatments.

3.2 Main Hypotheses: Trust and Trustworthiness

The experimental treatments in both Brazil and Mexico present tweets dealing with the COVID-19 crisis. The content of the tweets attributes blame (negative tweet) or signals cooperation (positive tweet). We randomly treat one-third of our sample to negative messages (attribution of responsibility) and one-third to positive messages (interparty cooperation). The control group—the remaining third—receives no social media messages. Among the two arms of the treatment, we randomly rotate the author of the tweets using two high-level politicians from different (opposing) parties. All the hypothesis were preregistered previously with the OSF/EGAP Framework [18]

Positive messages report to voters the willingness of political elites to cooperate with rivals to fight the COVID-19 pandemic. The messages signal to respondents the importance of unity and cooperation to manage the crisis. Negative messages blame political opponents for sowing conflict and weakening the needed response to the crisis. These negative tweets activate partisan identities and 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 positive social media messages will increase trustworthiness and trust, and negative messages will reduce both.

HT_0A**: Positive social media messages increase compliance by agents and trust among principals.**

HT_0B**: Negative social media messages decrease compliance by agents and trust**

[18]The preregistration can be accessed here https://osf.io/hvjkt.

among principals.

Because positive or negative political messages may be endorsed by politicians who align or not with the preferences of the respondents, we test for the effect of political congruence (in-group) or dissonance (out-group) on *trust* and *trustworthiness*. A broad literature in political behavior shows that partisanship 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). However, less is known about the effect of partisanship on trust behavior. Informed by the literature on partisan identities, we expect the endorsement of out-group politicians to augment the effect of positive and negative messages on trust and trustworthiness:

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

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

Considerable research suggests that individuals perceive social media platforms as conduits for increased polarization and mistrust. Therefore, we expect that the mean levels of trustworthiness and trust in individuals in the treatment group will be lower than among 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. Recent scholarship in both political science and psychology suggests that the amount of time spent on a survey question works as a measure of respondent effort. Those who are more cognitively engaged with the treatment receive stronger doses (Berinsky, Margolis and Sances, 2014; Wise and Kong, 2005; Malhotra, 2008). Similar effects for latency, measured as the time spent reading tweets, have been shown to increase the effects of social media framing on polarization (Banks et al., 2020). To capture the effects of attention, we capture the time the respondents spend reading each treatment condition. We expect:

HT_3**: Higher engagement, such as lower latency (more time spent reading the tweets) and active responses to tweets (retweet, like, and reply), will increase the effects of the treatments.**

4 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.[19] 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

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

did not agree to cast votes, 22% and 15%, respectively, agreed to do so in the second round.

It is worth highlighting that, although casting votes reduces the chances of winning one of the iPads, a majority of 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)

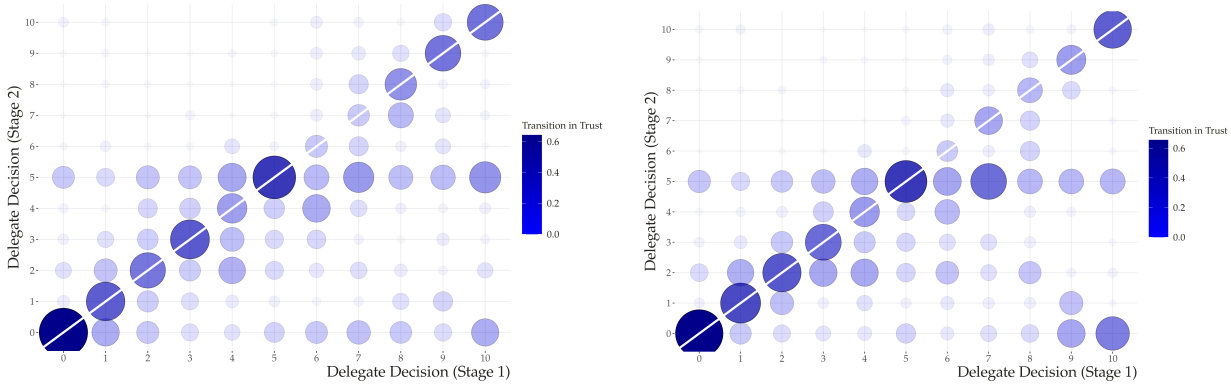
Descriptive Results for Trust

Figure 2 presents descriptive results on the number of votes [0,10] delegated (entrusted) to others, with the first round shown on the horizontal axis and the second on the vertical.

The circles in Figure 2 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 delegated 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, respectively.

Overall, we observe a decline in trust among respondents in our Mexican and Brazilian samples. Between the first and the second rounds of the game, respondents consistently reduced the number of votes entrusted to other players, as the reader can easily see from the more populated lower triangle in Figure 2. Therefore, from the first to the second round we observe that most respondents reduced the votes entrusted to other players and retained for themselves a larger number of votes.

Figure 2 Trust: First and Second Rounds of the Game, Compared



a) Brazil b) Mexico

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

5 Experimental Results

Descriptive evidence in the previous section shows that between the first and second rounds, fewer respondents agreed to cast the votes entrusted to them (lower trustworthiness) and smaller quantities were delegated to other respondents (lower trust). In Brazil, rates of agreement to cast entrusted votes (trustworthiness) declined from 64% to 59%, and in Mexico from 64% to 56%. Similarly, entrusted votes (trust) in Brazil declined from 3.4/10 votes 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 in the round. In the next two subsections, we show that social media exposure had no effect on the decline in trustworthiness but a significant effect on trust.

5.1 The Null Effect of Social Media Exposure on Trustworthiness

Table 3 on page 27 presents our findings on the effect of social media exposure 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 the treatments 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 social media messages. We then separate by treatment condition (negative/positive and in-group/out-group) and control for the first-round decision to cast votes.

While findings are suggestive and point in the right direction, estimates do not reject

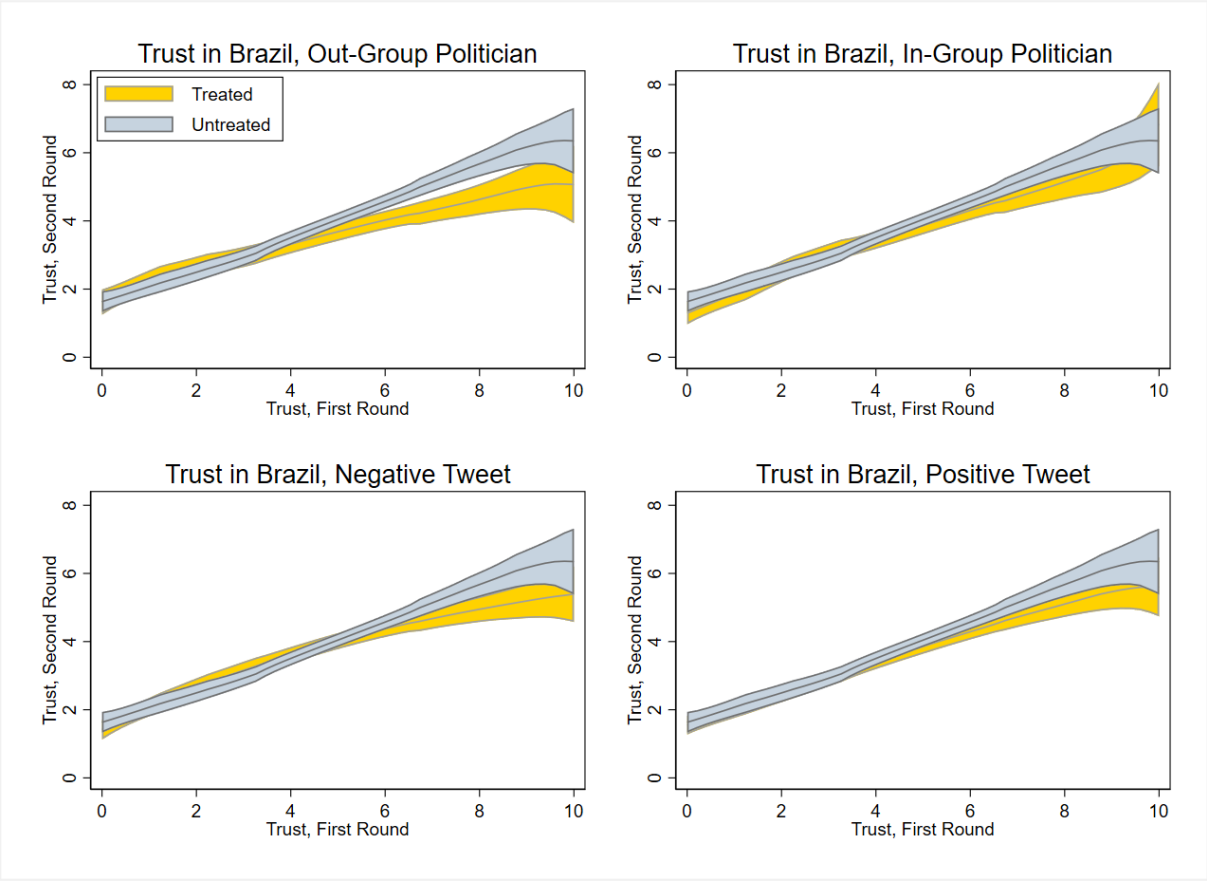
the null hypothesis that $\theta_{i,2} - \theta_{i,1} \neq 0$. Accordingly, we report null findings for the trustworthiness (agent) hypotheses, HT_1A and HT_1B. Only hypothesis HT_2 holds, showing a decline in trustworthiness in later rounds, consistent with most in-person implementations of the *trust game*. This decline, however, is not explained by social media exposure. Therefore, contrary to our expectations, exposure to social media messages, varying the endorsement and framing of the messages, has no effect on the trustworthiness of respondents.

5.2 The Negative Effect of Social Media Exposure on Trust

Unlike the case for trustworthiness, our model results show that social media exposure reduces overall trust. We begin by presenting very conservative estimates of the effect of our experiment on trust, separating dissonant messages (out-group politician) and uncivil messages (negative content) using nonparametric visual information. Then, we present statistical models and estimate the marginal effects of the treatments. The following section then describes an identification strategy, with very robust results, which allows us to distinguish the factors that mediate the decline in trust.

Figures 3 and 4 separate the results of our experiment by out-group/in-group politicians and by the negative/positive conditions. 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 (*uncivil discourse*), as they are significant for Brazil but not for Mexico.

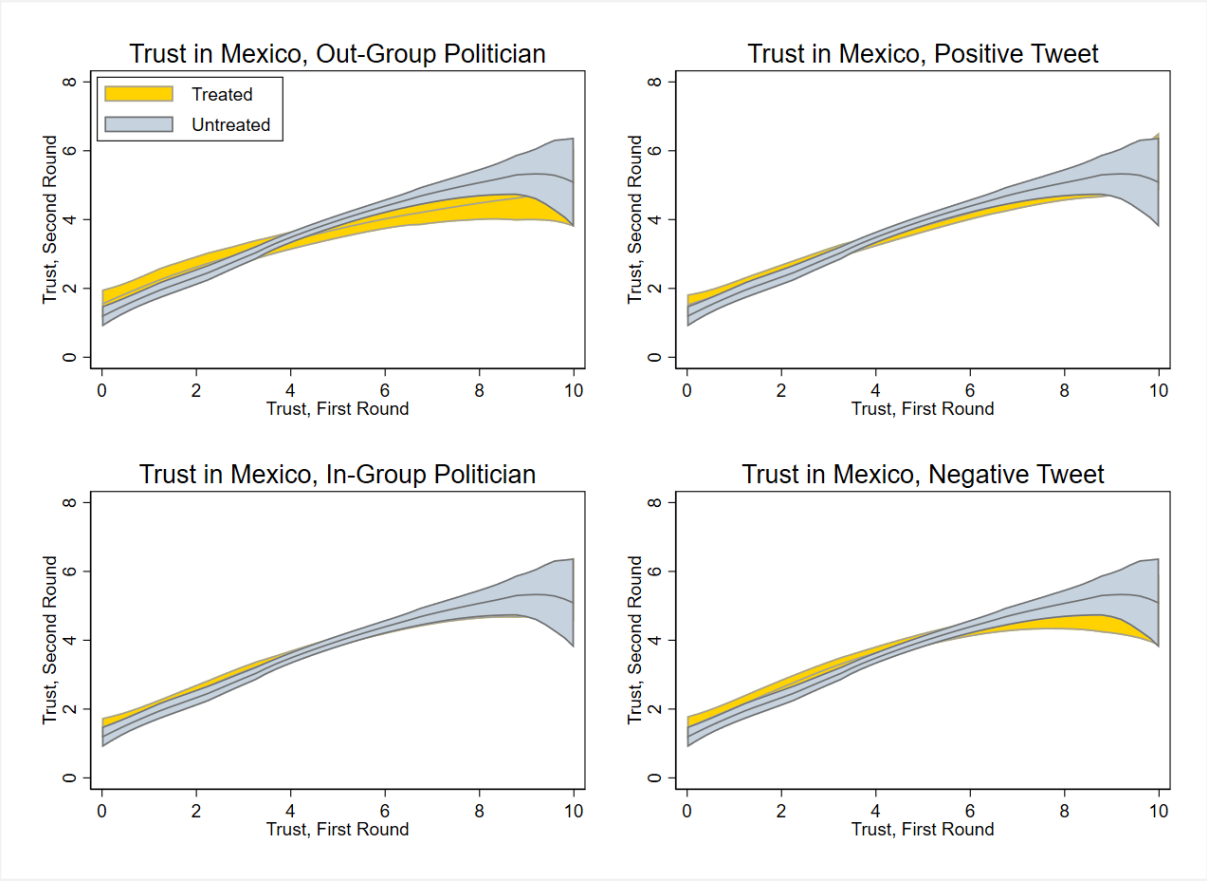
Figure 3 Changes in Trust among Treated and Untreated Respondents in Brazil



Note: Local polynomial lines with confidence intervals. Plots compare changes in trust (votes delegated) between the first and second rounds of the game in Brazil. Four treatment conditions are compared with the control group: dissonant tweets from an out-group politicians, congruent tweets from an in-group politician, negative tweets (*responsibility deflection*), and positive tweets (*cross-the-aisle*). The joint effect of out-group and negative tone is not evaluated in this figure.

The upper left plot in Figure 3 provides visual confirmation of 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 4). By contrast, exposing respondents to tweets from politicians they support yields small effects in Brazil and null results in Mexico.

Figure 4 Changes in Trust among Treated and Untreated Respondents in Mexico



Note: Local polynomial lines with confidence intervals. Plots compare changes in trust (votes delegated) between the first and second stages of the game in Mexico. Four treatment conditions are compared with the control group: dissonant tweets from an out-group politician, congruent tweets from an in-group politician, negative tweets (*responsibility deflection*), and positive tweets (*cross-the-aisle*). The joint effect of out-group and negative is not evaluated in this figure.

The lower left plots in Figures 3 and 4 show that, compared with the control group, negative 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 negative tweet, the results reported in this section are very conservative.

In Table 4, we present the results from benchmark ordinary least squares (OLS) analysis

to capture the effect of the treatments on declines in trust in the second round of the game. Because changes in trust are heterogeneous, as shown in Figures 3 and 4, we use an interactive linear model between the treatments and the decision to entrust votes at the first round of the game. Columns 1 to 3 present the results for Brazil of each different set of specifications, and columns 4 to 6 for Mexico. [20]

The first models for each country (1 and 4) estimate the treatment effect of the content of the tweets. If we do not take into account the first-round decision to entrust votes, we cannot reject the null hypothesis that after respondents are exposed to the treatment $\theta^*_{j,2} - \theta^*_{j,1} \neq 0$. As in Figures 3, 4, and 5, the effect of the treatment has the expected negative effect once the first-round decision is taken into account.

In models 2 and 5, we estimate the endorsement effects of 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.[21]

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 treatment with a tweet from a misaligned politician, respondents decrease the number of votes they entrust to other players. The effect is larger for higher levels of trust in

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

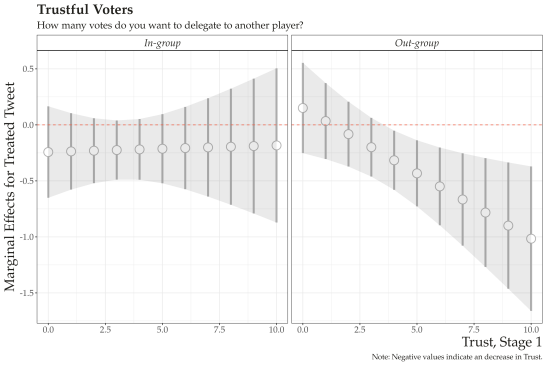
[21] As an example, consider the case of a Brazilian respondent who votes for Bolsonaro and is exposed to the tweet from Haddad. We score this treatment as an out-group message. The same rule is used for MORENA and PAN voters in Mexico. In-group coding, on the other hand, describes a voter exposed to a tweet from a politician she supports.

the first round, as reported in Figure 3. 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 effect on trust.

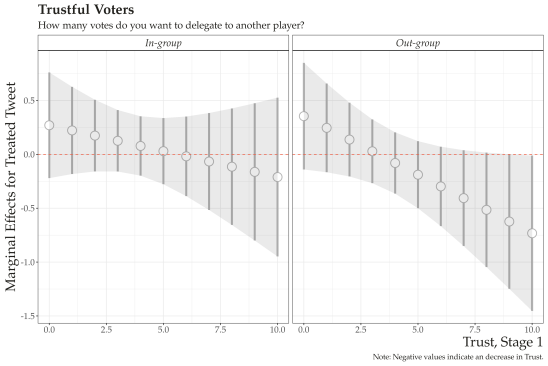
Finally, models 3 and 6 evaluate hypotheses $H1_A$ and $H1_B$. with respondents playing the role of principals (voters). We estimate the effects of being exposed to a negative message from an out-group politician. Results in both countries show statistically significant declines in trust after respondents are exposed to uncivil/negative social media messages from political opponents.

Results are fully described in Figure 5, 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 5 presents the effects of reading a tweet from a misaligned politician, no matter the tone of the message. Figure 6 separates the out-group treatment according to the positive and negative framing. The figures provide a clear visualization of how out-group messaging, in particular with a negative tone, has a detrimental effect on interpersonal trust. For both cases, we see that reading a negative dissonant message reduces by almost 10% the votes delegated to other players between the first and second stages of the trust game—and marginal effects are statistically different from zero on respondents who in the early stage of the game exhibited higher levels of trust. The effect is substantively significant and, more important, describes a low-dosage treatment (one tweet) compared with the large number of tweets that users are exposed to on a daily basis.

Figure 5 Marginal Effects of Cognitive Dissonance on Trust

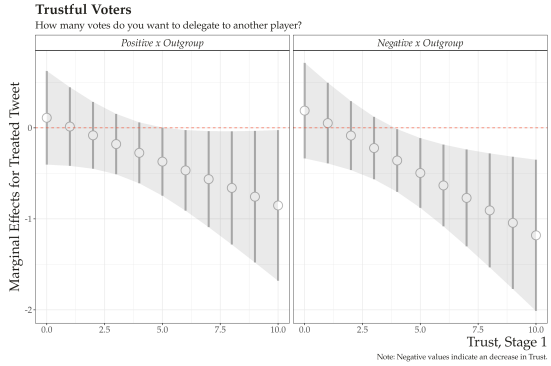


a) Brazil

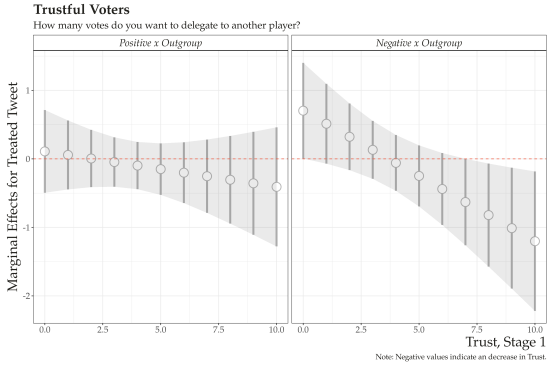


b) Mexico

Figure 6 Marginal Effects of Negative Treatment from a Misaligned Politician



a) Brazil



b) Mexico

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: Negative	0.035			−0.027		
	(0.036)			(0.034)		
Framing: Positive	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)	
Negative Out-group			−0.032			−0.011
			(0.052)			(0.063)
Positive Out-group			−0.024			0.005
			(0.050)			(0.052)
Negative x Trustworthiness (Round 1)	−0.021			0.013		
	(0.045)			(0.042)		
Positive 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)	
Negative Out-group x Trustworthiness (Round 1)			0.015			0.009
			(0.065)			(0.078)
Positive 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. Models 1, 2, and 3 report results for Brazil; Models 4, 5, and 6 for Mexico. The dependent variable uses the decision to cast votes entrusted by other players, thus measuring subjects' levels of trustworthiness. 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

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)
Framing: Negative	−0.052			0.299		
	(0.194)			(0.209)		
Framing: Positive	−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)	
Negative Out-group			0.083			0.676*
			(0.283)			(0.366)
Positive Out-group			0.110			0.015
			(0.274)			(0.324)
Negative x Trust (Round 1)	−0.032			−0.050		
	(0.043)			(0.046)		
Positive 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)	
Negative Out-group x Trust (Round 1)			−0.126**			−0.164**
			(0.062)			(0.079)
Positive 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. Models 1, 2, and 3 report results for Brazil; Models 4, 5, and 6 for Mexico. 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

6 Engagement with Social Media: The Role of Attention and Anger

Results in the previous section show a decline in trust among voters exposed to negative political messages from out-group politicians. Separating the treatment conditions, we find larger negative effects from exposure to out-group politicians. Treatment effects are more robust in our sample of Brazilian respondents. While results confirm the hypothesized effect of 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. As we will show next, our survey included validation checks to evaluate whether respondents properly interpreted the partisan leaning of the social media frames and, more importantly, questions about respondents' emotional response to the four partisan treatments in each country.

In this section we analyze these results in greater detail, introducing an identification strategy that helps to show that the detected effects on trust are the result of exposure to social media content. Our identification assumption applies a difference-in-difference design to survey data by manipulating when respondents play the trust game. The identification relies on a double-identification assumption:

$$Y_i(t, m), M(t) \perp T_i \tag{3}$$

Where $t \in T$ represents all the values for the treatment, $m \in M$ the mediator values,

and Y the responses to the trust game. This assumes (i) that the potential outcomes for Y between given treatment rounds can be ignored (which is achieved by randomization in the survey) and (ii) that the potential outcomes for the mediator as a function of the treatment assignment can be ignored. In our design these assumptions require us to rule out any effect of respondents answering **before** reading the tweet. As we discuss below with an example, we consider this a plausible assumption.

First, we 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 that this is a reasonable assumption, one that allows us to identify the heterogeneity of the treatment effects conditional on behavioral reactions to the 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 7 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 7, the left plot compares the effect of the *treated-*

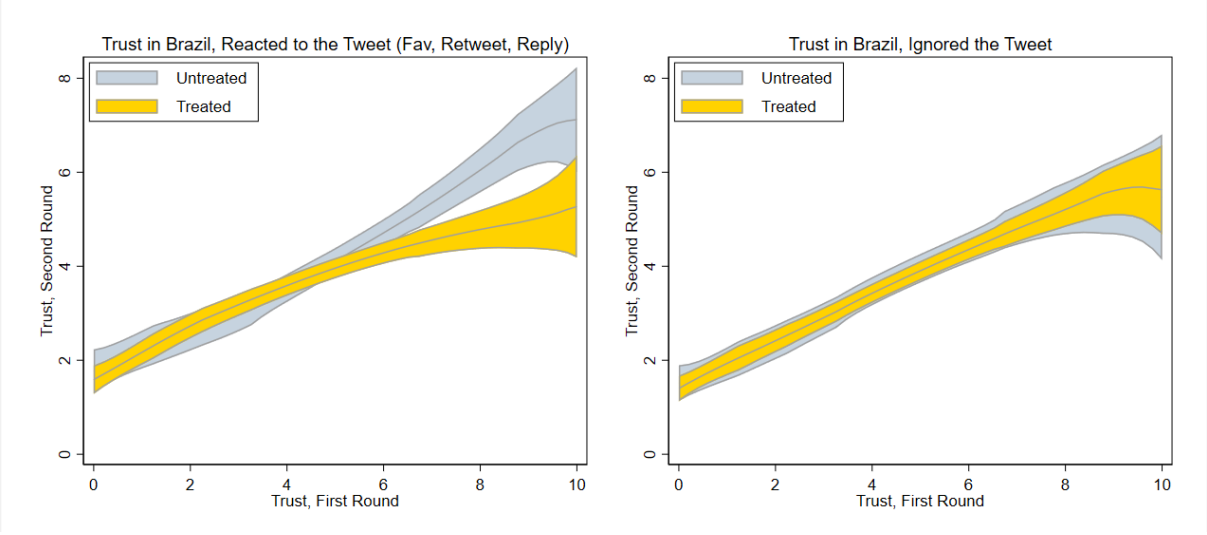
engaged group (like, retweet, reply) against the *control-engaged* group. Meanwhile, the right plot describes the *treatment/ignore* group against the *control/ignore* group. Notable is the significant decline in trust among respondents who like, retweet, or reply to a tweet in the treatment group compared with respondents in the control group who were equally engaged with the tweet. As important is the fact 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.

Described in terms of our theory, our test results show that $(\theta^*_{j,2} - \theta^*_{j,1}|E) < (\theta^*_{j,2} - \theta^*_{j,1}|\neg E)$, given that $(\theta^*_{j,2} - \theta^*_{j,1}|E) < 0$, while $(\theta^*_{j,2} - \theta^*_{j,1}|\neg E) = 0$. 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 8 depicts similar two-way comparisons, focusing on messages from out-group politicians (dissonant trait). Among those who like, retweet, or reply to the message (left plot), we see larger treatment effects. By contrast, incidental exposure (Boczkowski, Mitchelstein and Matassi, 2018) to the tweet, as shown in the plots to the right of Figure 8, has modest effects in Brazil and a null effect in Mexico. Indeed, conditioning on both treatment and attention provides the strongest evidence yet of the effect of social media

Figure 7 Changes in Trust When Respondents Engage with the Tweet (left) or Ignore It (right)



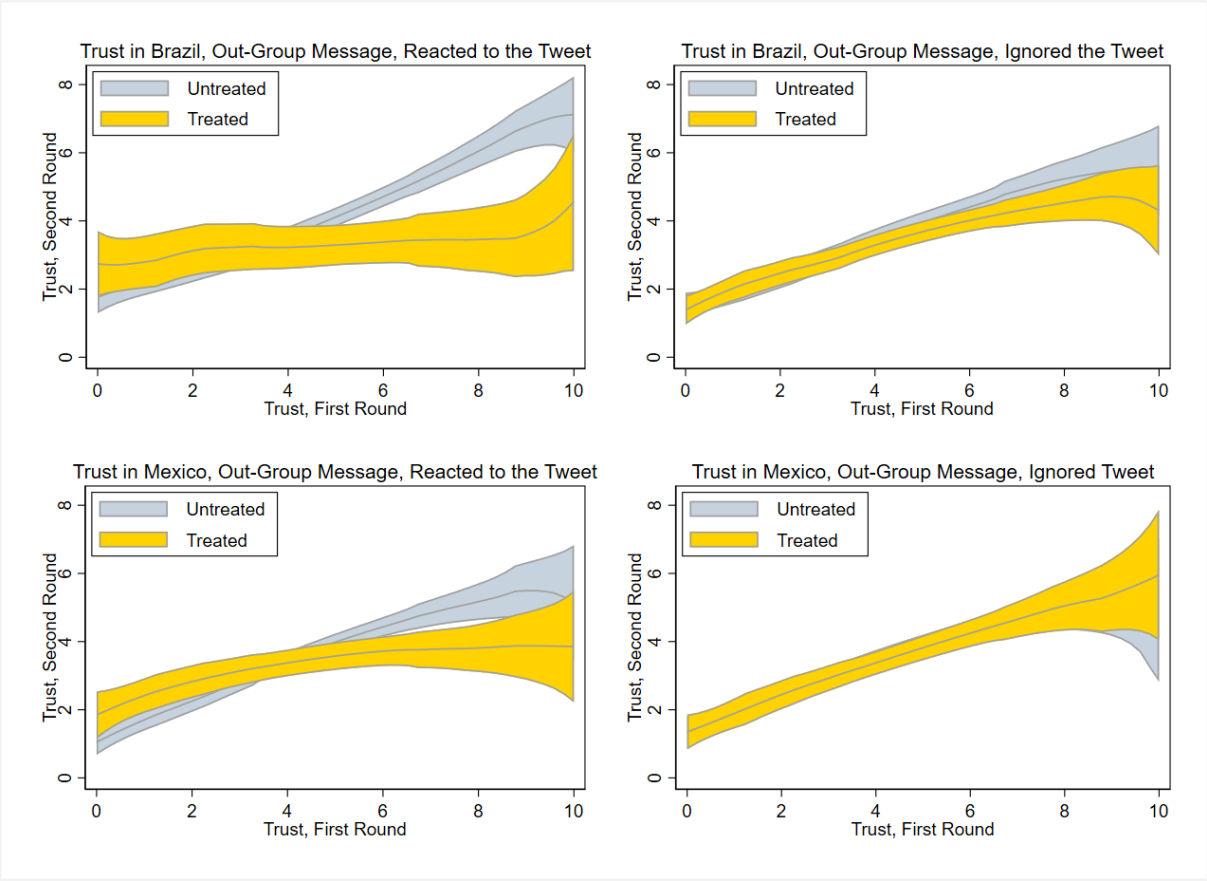
Note: The left plot estimates the treatment effect among voters who engaged with the tweet (like, retweet, or reply). The right plot estimates the treatment effect among those who did not engage (ignore). Results show a decline in trust only among respondents who saw and engaged with the tweet *before*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

on trust .[22]

An extensive political science literature argues that polarization is partially explained by the affective response of voters toward out-group peers and out-group politicians, with particular attention to the role of emotions such as anger and disgust (Mason, 2016; Banks, 2014; Banks, White and McKenzie, 2019; Iyengar, Sood and Lelkes, 2012; Iyengar and Westwood, 2015). Our design allows us to assess how similar dynamics might explain changes in trust due to incidental exposure to social media. The two-way identification strategy again shows that being angry has a large mediating role in explaining trust behavior among Brazilian respondents.

[22] Appendix D presents the results using a linear parameterization of the treatment effects using OLS. The results are similar.

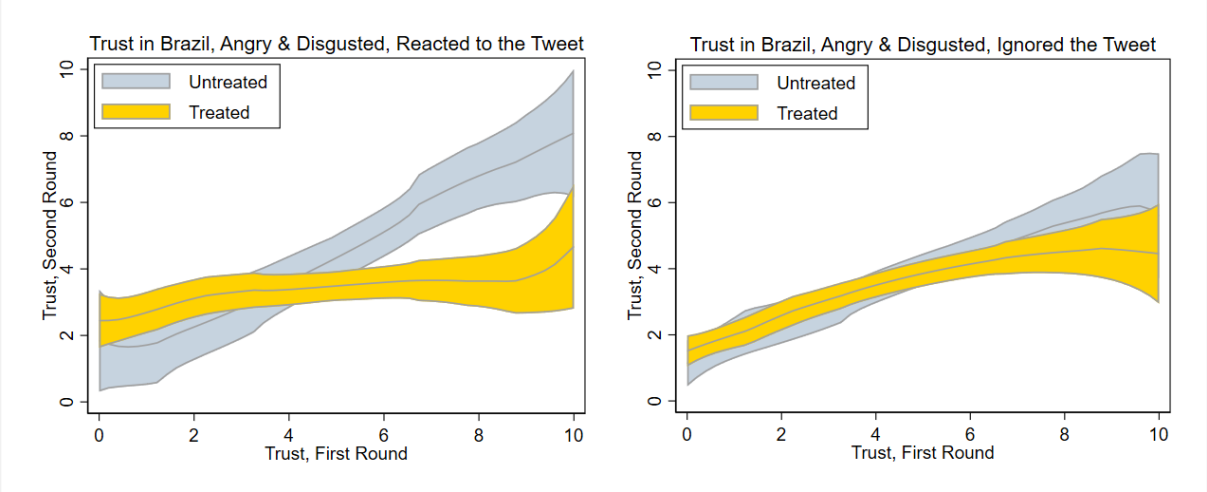
Figure 8 Changes in Trust When Respondents Engage with Dissonant Tweets



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 for partisan dissonance in the control group and no difference in the treatment and control respondents who ignored the tweet.

It is worth noting that while 33% of respondents in Brazil described their feelings toward the tweet as "anger" or "disgust", only 11% did so in Mexico. The difference in reported anger allows us to account for the larger effects detected in Brazil than in Mexico. Indeed, the double-identification approach allows us to see that much of the difference between the countries is explain by respondents' sensitivity to the frames we use.

Figure 9 Changes in Trust When Respondents Engage and Report Anger or Disgust in Connection with the Tweet



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

The comparisons between treatment and control groups with respondents who reacted or ignored our social media treatments is illuminating. Given that all other parameters are held constant, we can confidently state that $\theta^*_{j,R1} - \theta^*_{j,R2} < 0$. Similarly, we can confirm the importance of anger as a mediator: $(\theta^*_{j,2} - \theta^*_{j,1}|(A,E)) < (\theta^*_{j,2} - \theta^*_{j,1}|\neg(A,E))$, given that $(\theta^*_{j,2} - \theta^*_{j,1}|(A,E)) < 0$, whereas $(\theta^*_{j,2} - \theta^*_{j,1}|\neg(A,E)) = 0$. By splitting the sample between those who were angry and engaged with the tweet (both treatment and control) and those who were not angry and did not engage (both treatment and control), we are able to test for the different mechanisms that explain the decline in trust.

7 Concluding Remarks

Are polarization and uncivil dialogue reducing trust and trustworthiness? Results from our survey experiment provide compelling evidence of a negative effect of social media on trust behavior. Negative messages from out-group politicians reduce the propensity of survey respondents to entrust votes to their peers. As such, social media could be affecting political trust by activating partisan identities, even though these partisan identities are orthogonal to the game being played (candidates in the game are fictional and respondents do not know the partisanship of other respondents).

The negative effect on trust is considerably greater among randomly treated respondents who engage with social media messages through likes, retweets, and replies. By contrast, respondents in the control group who were equally engaged show no change in trust behavior. Together, the double-identification strategy discussed in Section 6 provides robust evidence of a social media effect on trust behavior. This effect comes not from the existence of information—which has been available for many decades from traditional mass media—but from the engagement with that information that social media allows. Moreover, the sentiments, particularly anger, that the tweets generate matter, too. Our identification strategy also allows us to test for a number of mediators discussed in the literature. We show that self-reported anger after exposure to tweets (Mason, 2016; Banks, 2014) is associated with steeper declines in trust behavior among respondents in the treatment group. This effect may be more prevalent in social media than in traditional media because of fewer gatekeepers, anonymity, larger quantities of information, and the

greater probability of finding extreme positions.

We also find no evidence of an effect on trustworthiness. That is, after exposure to negative social media messages, respondents are less likely to entrust resources to others but no less likely to cast votes entrusted to them. This is a particularly relevant finding given the disenchantment with their representatives common among Latin Americans. One possibility is that the result is affected by our design. Politicians (agents) have only the choice of casting the votes or not; they cannot decide how many to cast, as is usually the case in traditional trust games. Also, the pledge they were obliged to read before playing may have raised guilt levels. This is an interesting result in itself, as guilt seems to be replacing accountability (which is not part of our design) in the relationship between voters and politicians. It remains to be explored whether differences across countries in the quality of democracy can be explained not only by accountability but also by guilt and whether social media may attenuate guilt by showcasing the infractions of others.

Interesting extensions of the proposed model could be deployed to understand why social media exposure decreases trust but has negligible effects on trustworthiness. Indeed, results are consistent with social media reducing the association between θ_j^* and θ_i after exposure. Recent research by Corbacho et al. (2016) has shown that individuals who perceive others as corrupt are also more likely to engage in corruption themselves. By contrast, our experiment finds no equivalent association between perceiving others to be deceitful and behaving deceitfully. The dissociation between trust and trustworthiness in the treatment group, therefore, raises new questions about the connections between θ_j^* and θ_i.

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

Evidence has mounted that trust is important for thriving democracies and economies. Latin America and the world have seen large drops in trust over the past few decades. Failures by governments to deal with several economic crises (and a pandemic), growing inequality, and unfulfilled expectations may be the main drivers. Still, the quantity of information and how it is distributed matters as well. Social media was expected to bring additional transparency, higher accountability, and, hence, higher political trust. Unfortunately, the evidence does not seem to bear out these expectations.

References

Aghion, Philippe, Yann Algan, Pierre Cahuc and Andrei Shleifer. 2010. "Regulation and distrust." *The Quarterly Journal of Economics* 125(3):1015–1049.

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

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

Algan, Yann and Pierre Cahuc. 2014*b*. Trust, growth, and well-being: New evidence and policy implications. In *Handbook of economic growth*. Vol. 2 Elsevier pp. 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.

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

502.

URL: *http://dx.doi.org/10.1093/joc/jqy007*

Ashraf, Nava, Iris Bohnet and Nikita Piankov. 2006. "Decomposing trust and trustworthiness." *Experimental economics* 9(3):193–208.

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.

Banks, Antoine J, Ismail K White and Brian D McKenzie. 2019. "Black politics: How anger influences the political actions Blacks pursue to reduce racial inequality." *Political behavior* 41(4):917–943.

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

Berg, Joyce, John Dickhaut and Kevin McCabe. 1995. "Trust, reciprocity, and social history." *Games and economic behavior* 10(1):122–142.

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*

Bloom, N., C. Genakos, R. Sadun and J Van Reenen. 2012. "The Organization of Firms across Countries." *Quarterly Journal of Economics* 124(4):1663—-1705.

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.

Buskens, Vincent, Vincenz Frey and Werner Raub. 2018. "Trust Games." *The Oxford Handbook of Social and Political Trust* p. 305.

Camerer, Colin F and George Loewenstein. 2003. "Behavioral economics: Past, present, future.".

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.

Deibert, Ronald J. 2019. "The road to digital unfreedom: Three painful truths about social media." *Journal of Democracy* 30(1):25–39.

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.

Gambetta, Diego. 1988. "Trust: Making and breaking cooperative relations.".

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, 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*

Johnson, Noel D and Alexandra A Mislin. 2011. "Trust games: A meta-analysis." *Journal of Economic Psychology* 32(5):865–889.

Kahneman, Daniel. 2011. *Thinking, fast and slow*. Macmillan.

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* .

Knack, Stephen and Philip Keefer. 1997. "Does social capital have an economic payoff? A cross-country investigation." *The Quarterly journal of economics* 112(4):1251–1288.

Lazer, David MJ, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild et al. 2018. "The science of fake news." *Science* 359(6380):1094–1096.

Lelkes, Yphtach, Gaurav Sood and Shanto Iyengar. 2017. "The hostile audience: The effect of access to broadband internet on partisan affect." *American Journal of Political Science* 61(1):5–20.

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.

Mazar, Nina and Dan Ariely. 2006. "Dishonesty in everyday life and its policy implications." *Journal of public policy & Marketing* 25(1):117–126.

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.

Nyhan, Brendan and Jason Reifler. 2010. "When corrections fail: The persistence of political misperceptions." *Political Behavior* 32(2):303–330.

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.

Shoemaker, Pamela J and Stephen D Reese. 2013. *Mediating the message in the 21st century : a media Sociology perspective*. Third edit ed. New York: Routledge/Taylor & Francis Group.

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

Smith, Adam. 1937. "The wealth of nations [1776].".

Tandoc, Edson C. 2014. "Journalism is twerking? How web analytics is changing the process of gatekeeping." *New media & society* 16(4):559–575.

Tucker, Joshua A, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal and Brendan Nyhan. 2018. "Social media, political polarization, and political disinformation: A review of the scientific literature." *Political Polarization, and Political Disinformation: A Review of the Scientific Literature (March 19, 2018)* .

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.

Witmer, Hope and Peter Håkansson. 2015. "Social media and trust: A systematic literature review." *Journal of Business and Economics; 3* 6.

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.