

NBER WORKING PAPER SERIES

DOES SOCIAL MEDIA EXACERBATE POLITICAL POLARIZATION?
EXPERIMENTAL EVIDENCE FROM VARYING ACCESS TO TWITTER/X DURING A
PRESIDENTIAL DEBATE

Rafael Di Tella
Ramiro H. Gálvez
Ernesto Schargrodsky

Working Paper 29458
<http://www.nber.org/papers/w29458>

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2021

All authors contributed equally to this work. An earlier version of this paper circulated under the title “Does Social Media Cause Polarization? Evidence from Access to Twitter Echo Chambers during the 2019 Argentine Presidential Debate.” We thank Hunt Allcott, Levi Boxell, James Druckman, Matthew Gentzkow, Gary King, Jesse Shapiro, Ro’ee Levy, seminar participants at Facebook Core Data Science and especially Vincent Pons, for very helpful suggestions. We thank Julián Regatky, Lucía Freira, Lucas Soulès and Agustín Garassino for exceptional research assistance. This work was approved by the Ethics Committee of INECO (<https://www.ineco.org.ar/>), under the protocol titled “Political campaigns, political debates, and social networks” (PI Ernesto Schargrodsky, approved on 09/2019). This work was pre-registered under AEARCTR-0004850. This material is based on work supported by a Facebook Unrestricted Research Gift (“Content Policy Research on Social Media Platforms”). The views expressed here are strictly those of the authors and do not reflect the views of the Inter-American Development Bank, its Board of Directors, the countries they represent, or the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2021 by Rafael Di Tella, Ramiro H. Gálvez, and Ernesto Schargrodsky. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Does Social Media Exacerbate Political Polarization? Experimental Evidence from Varying
Access to Twitter/X during a Presidential Debate

Rafael Di Tella, Ramiro H. Gálvez, and Ernesto Schargrodsky

NBER Working Paper No. 29458

November 2021

JEL No. D72, L82, L86, O33, P16, Z13

ABSTRACT

We study how variation in social media access (Twitter/X) during a salient political event affects political polarization and stress, and how these exposure effects relate to broader patterns of selection into social media. We conducted a pre-registered laboratory experiment during the 2019 Argentine presidential debate, recruiting both Twitter users and non-users. Twitter users were randomly assigned to one of four conditions: watching the debate with their usual phone access, watching the debate with their phones switched off, watching the debate while interacting with an experimenter-run Twitter account that shared a balanced mix of partisan content, or watching a politically neutral documentary without phones. Non-Twitter participants were randomly assigned to watch the debate or the documentary without phones. Selection comparisons indicate that recruited Twitter users are more polarized than recruited non-users. Causal exposure comparisons within the Twitter-user sample show that two strategies often proposed to reduce political polarization on social media, removing access to the platform and exposing users to counter-attitudinal content, do not reduce polarization and may instead produce mild increases. The latter treatment also significantly increases physiological stress (as measured by salivary cortisol). Further exploratory analyses reveal substantial heterogeneity tied to users' online political environments. Politically segregated individuals — that is, users embedded in echo chambers — exhibit clear backfire effects when access is removed or cross-ideological interactions are encouraged. In particular, both treatments increase polarization among these users. In addition, when cross-ideological interaction is encouraged, segregated users concentrate a larger share of their Twitter engagement within like-minded networks and show higher cortisol levels. By contrast, non-segregated users show little to no response and closely resemble non-users. Moreover, the selection differences observed between recruited Twitter users and non-users in our sample are concentrated among politically segregated users.

Rafael Di Tella
Harvard University
Harvard Business School
and NBER
rditella@hbs.edu

Ernesto Schargrodsky
Universidad Torcuato Di Tella
and CAF and also CONICET
eschargr@utdt.edu

Ramiro H. Gálvez
Universidad Torcuato Di Tella
rgalvez@utdt.edu

1. Introduction

Political polarization has surged in recent years within Western democracies, with social media frequently cited as an important contributing factor (see, for example, Lelkes, Sood, and Iyengar 2017; Guriev, Melnikov, and Zhuravskaya 2021; Boxell, Gentzkow, and Shapiro 2022). A prominent explanation emphasizes the rise of social media “echo chambers” as a key driver of this trend (Sunstein, 2018 and Eady et al., 2019).¹ According to this view, echo chambers limit users’ exposure to diverse perspectives by providing content tailored to their partisan preferences. Recent evidence is consistent with this concern, showing that individuals embedded in political echo chambers receive disproportionately large amounts of positive content about their ingroup and substantial amounts of negative content targeting their outgroup (see, for example, Rathje, Van Bavel, and van der Linden, 2021).

In light of these concerns, a growing body of research seeks to understand and mitigate the impact of social media on political polarization, with two broad strategies receiving particular attention: i) restricting individuals’ access to social media and ii) exposing users to challenging, “counter-attitudinal” content. The evidence on the effectiveness of these strategies, however, remains mixed. Three recent examples of the first strategy are Allcott et al. (2020), Mosquera et al. (2020), and Allcott et al. (2024).² Allcott et al. (2020) investigate the impact of Facebook deactivation in the four weeks leading up to the 2018 U.S. midterm elections, finding that the intervention increased subjective well-being while decreasing both factual news knowledge and political polarization. Mosquera et al. (2020) report similar effects of deactivating Facebook for one week. Allcott et al. (2024) explore the impact of Facebook and Instagram deactivation on political beliefs and behavior. Participants were randomly assigned to deactivate their accounts for six weeks leading up to the 2020 U.S. election. The intervention reduced online political participation and had mixed effects on news knowledge and

¹ Echo chambers refer to online spaces where users’ existing opinions and beliefs are reinforced through repeated interactions with like-minded peers (see Cinelli et al. 2021; and Robbett, Colón, and Matthews 2023).

² Another study closely related to this literature is Beknazar-Yuzbashev and Stalinski (2022), who exploit Facebook’s introduction of a filter that allows users to hide ads from their feeds to examine the effects of political advertising during the 2020 U.S. elections. They find no significant overall effect on turnout, but do uncover suggestive heterogeneous effects across party lines. Guess et al. (2023b) take an intermediate approach by assigning a random sample of consenting U.S.-based Facebook users to feeds that excluded all reshares for a three-month period. This intervention substantially reduced exposure to political news, including content from untrustworthy sources, and decreased overall engagement, as measured by clicks, reactions, and partisan news interactions. However, the treatment did not produce measurable changes in political polarization or in any individual-level political attitudes. More recently, Arceneaux et al. (2026) study a randomized field experiment during the 2024 U.S. presidential election in which participants were incentivized to deactivate Facebook for the two weeks before the election, while a separate treatment encouraged engagement with ideologically balanced news. Consistent with the broader deactivation literature, they find mostly null or very small effects on political and social outcomes, reinforcing the view that Facebook’s average effects in real-world electoral settings are limited.

misinformation beliefs, while showing no significant impact on affective or issue polarization, perceptions of election legitimacy, candidate favorability, or voter turnout.³

Regarding the second strategy, which involves exposing users to counter-attitudinal content, the evidence is far from conclusive. Some studies suggest that such exposure can help challenge stereotypes and reduce intergroup prejudice (see Pettigrew and Tropp, 2006). For example, Levy (2021) offers subjects subscriptions to liberal or conservative news outlets on Facebook and finds that exposure to counter-attitudinal news decreases negative attitudes toward the “opposing political party,” with no effect on political opinions. Other researchers point toward null or even detrimental effects of these interventions. Nyhan et al. (2023) examine the prevalence and effects of exposure to politically like-minded content on Facebook. They find that although most of the content users encounter comes from like-minded sources, experimentally reducing this exposure by roughly one-third over several months produces no detectable changes in political attitudes or polarization. Bail et al. (2018) find that exposing Twitter (now X) users to bots that retweet counter-attitudinal content increases political polarization. Bail (2021) further argues that social media provides online extremists with a sense of identity and community, while exposure to outgroup content sharpens “us versus them” distinctions.

Against this backdrop of mixed and sometimes contradictory findings, our paper presents results from a pre-registered experiment that tests these two ameliorating strategies in the context of a presidential debate and, by additionally benchmarking outcomes against a sample of non-Twitter users, offers a broader perspective on how social media use relates to political polarization. To implement the experiment, we invited participants to attend a neutral venue on the day of the first presidential debate of Argentina’s 2019 election (October 13, 2019), where they took part in an academic study in exchange for monetary compensation. Participants came from two populations: active Twitter users and non-Twitter users. Twitter users were randomly assigned to one of four treatment arms that varied their access to social media. The first group watched the debate with no restrictions on cellphone use (Twitter-Allowed) and serves as the benchmark for all comparisons within Twitter users. The second group, designed to mimic the deactivation strategy, watched the live debate without access to their cellphones (Debate-Only). The third group, intended to mirror the counter-attitudinal exposure strategy, watched the debate while interacting with a Twitter account managed by our team that shared content spanning a wide ideological spectrum (Twitter-Interact). The fourth group, which serves as the benchmark for comparing polarization between Twitter and non-Twitter users, was asked to turn off their cellphones and watch a documentary on Argentine

³ In line with previous studies, Ventura et al. (2023) conducted a multimedia deactivation experiment to assess the impact of reduced exposure to potential misinformation sources on WhatsApp during the weeks preceding Brazil’s 2022 presidential election. The intervention significantly decreased participants’ exposure to false rumors. However, it did not yield significant changes in belief accuracy, political polarization, or well-being.

history instead of the live debate (Documentary). Non-Twitter users were assigned to either the Debate-Only or the Documentary condition. Across all groups, we measured several pre-registered outcome variables, including survey-based polarization indices and a biological marker of stress (salivary cortisol).

Having collected these outcomes, our design allows us to examine two related questions that correspond to the ideas of *exposure* and *selection* (Braghieri et al., 2024). Exposure refers to how varying access to social media causally affects individuals’ attitudes or behaviors, whereas selection refers to differences between social media users and non-users that emerge from individuals’ decisions to use social media in the first place. To study exposure, we compare post-debate outcomes for Twitter users in the Debate-Only and Twitter-Interact groups with those in the Twitter-Allowed group. This comparison allows us to assess how different levels and types of access to social media influence polarization indices and physiological stress, while also evaluating the effectiveness of the two ameliorative strategies proposed in the literature. To study selection, we compare outcomes for recruited Twitter users and non-Twitter users assigned to otherwise similar conditions. This analysis shows how polarization differs between the two groups and provides suggestive evidence on whether the observed patterns are consistent with selection into social media use.

Our results reveal a consistent set of findings. When we manipulate exposure to social media within the Twitter-user sample, the two strategies commonly proposed to mitigate polarization do not appear to reduce it. Instead, both the Debate-Only and Twitter-Interact treatments tend to be associated with higher values of our survey-based polarization measures relative to the Twitter-Allowed benchmark, although most of these differences are not statistically significant. We also find that the Twitter-Interact treatment produces a statistically significant increase in physiological stress, as captured by changes in salivary cortisol. Taken together, these patterns suggest that, while the evidence of increased polarization remains limited, interventions intended to reduce polarization may nonetheless generate short-run backfire effects in the immediate aftermath of a politically salient event. Turning to selection, we find that, under comparable conditions, Twitter participants display higher polarization than non-Twitter participants, while no significant differences emerge in physiological stress. This is consistent with selection into Twitter use being associated with higher underlying levels of polarization.

To deepen our interpretation of these patterns, we complement our pre-registered analysis with an additional exploration of online segregation within the Twitter user sample. This extension is motivated by the view that echo chambers are often cited as a driver of social-media-induced polarization, even though active social media use does not necessarily place users within one. To study this dimension, we construct a pre-treatment indicator of the political homogeneity of each participant’s Twitter environment, based on the ideological composition of the accounts they follow and the share of those accounts that are political in nature. When we divide Twitter users into

politically segregated and non-segregated groups, we find that the exposure and selection patterns documented in the full sample are driven almost entirely by the users embedded in politically segregated environments. Among non-segregated Twitter participants, polarization levels are statistically indistinguishable from those of non-Twitter participants — consistent with the absence of meaningful selection differences, while exposure to treatments produces essentially null effects. Among segregated participants, by contrast, the patterns shift sharply. The selection differences observed in the full sample appear to be concentrated in this group, who display substantially higher polarization than non-Twitter participants. With respect to exposure, both the Debate-Only and Twitter-Interact treatments generate clear backfire effects, yielding substantially higher levels of polarization than the corresponding Twitter-Allowed benchmark. The Twitter-Interact treatment also raises cortisol only among segregated participants, suggesting that the physiological response to interacting with partisan content is concentrated among individuals already embedded in politically homogeneous environments. Additional outcomes capturing Twitter activity during and after the debate reinforce this interpretation, since segregated users in the Twitter-Interact group concentrate their engagement even more within like-minded networks, whereas non-segregated users do not.

Our findings add a note of caution to the mixed evidence in the existing literature by identifying the contexts in which the two main strategies proposed to reduce polarization can produce unintended backfire effects. They also show that these backfire effects are heterogeneous and concentrated among users embedded in politically segregated online environments, while individuals with more diverse online exposure exhibit little or no response.

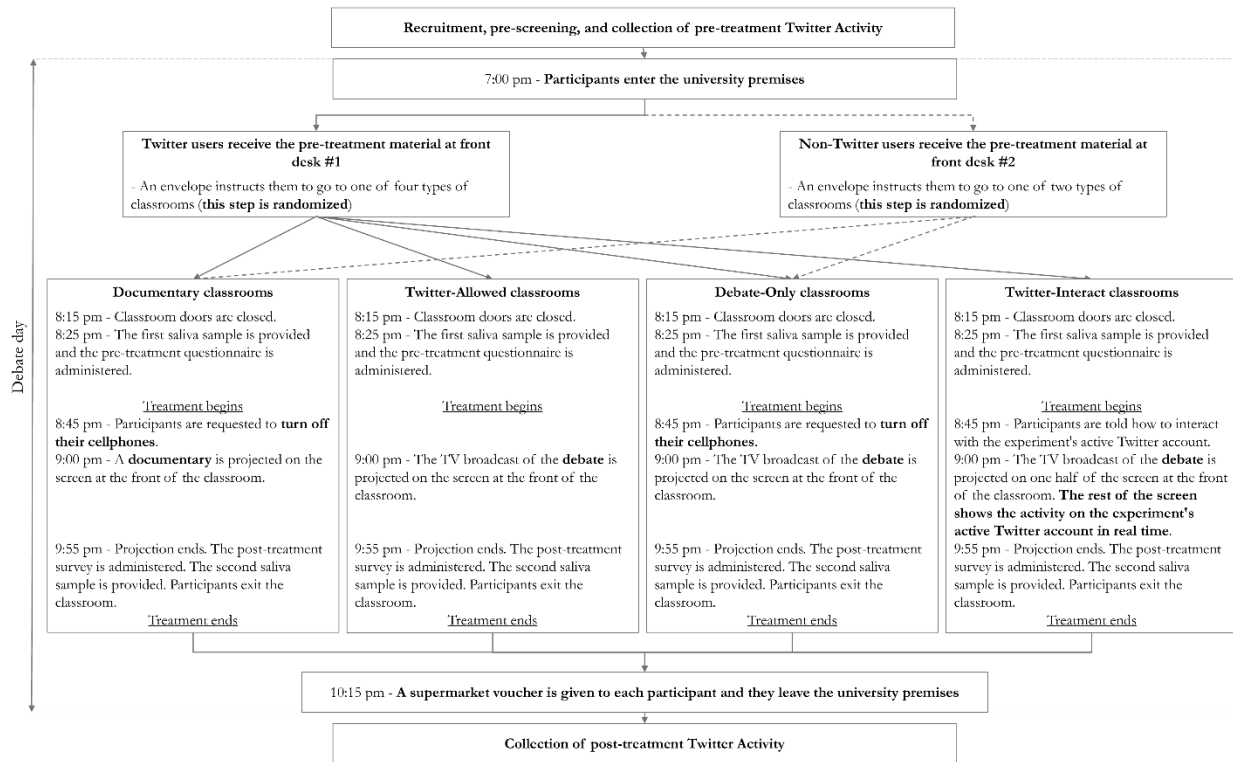
The rest of the paper is organized as follows: Section 2 describes the context of our experiment, as well as our data and methods. Section 3 presents our results. Section 4 discusses plausible interpretations for our findings and concludes.

2. Materials and methods

2.a. Experimental design

This section describes the experimental design. We conducted a controlled randomized experiment combining survey responses, biological markers, and Twitter activity data from active users. Twitter and non-Twitter participants were randomly assigned either to watch the presidential debate under different conditions of social media access or to watch a politically neutral documentary. Figure 1 summarizes the study’s structure and sequence of procedures. The remainder of this section presents the timing of the experiment, recruitment procedures, randomization strategy, treatments, and post-treatment data collection.

Figure 1. Experiment overview



2.a.1. Timing

The experiment was scheduled to coincide with the first of the two 2019 Argentine presidential debates (October 13, 2019), which marked the first time in Argentine history that candidates' participation in the debate was legally mandatory. Six presidential candidates participated in the debate, which was divided into two halves. The first half covered international affairs and economics, while the second covered education, health, human rights, and gender-related topics. Each candidate was allowed to speak on each topic for two minutes. Once each candidate had spoken, each was given half a minute to expand on or respond to their rivals' statements.⁴

At the time, Argentine society exhibited a pronounced degree of political polarization, and the country was undergoing an acute economic crisis under an International Monetary Fund stabilization program. Against this backdrop, the 2019 presidential debate emerged as a highly salient national event, attracting widespread public attention and extensive media coverage. The key figures in this

⁴ A complete transcription of the debate can be found in Infobae (2019).

event were the incumbent president, Mauricio Macri of *Juntos por el Cambio* (JxC), a coalition typically described as center-right, and *Alberto Fernández* of *Frente de Todos* (FdT), a populist coalition rooted in the Peronist movement whose vice-presidential candidate was two-time former president Cristina Fernández de Kirchner. The debate attracted immense public interest, with television ratings nearly matching those of the *Superclásico* between Boca Juniors and River Plate held the following week — the country’s most widely watched sporting rivalry.⁵

2.a.2. Recruitment

Participants were invited to attend a university campus located in the City of Buenos Aires on Sunday, October 13, 2019. While this in-person format was more costly than a remote setup, it enabled the collection of additional data, including salivary cortisol measurements, and avoided attrition and non-compliance. Recruitment targeted both Twitter and non-Twitter users and took place between October 4 and October 12, 2019 (the day before the presidential debate). All participants were required to be between 18 and 70 years old and eligible to vote in the 2019 presidential election. Additional criteria applied to Twitter users, who were required to have a personal Twitter account created before June 20, 2019, and to have demonstrated activity by posting or retweeting at least three messages between June 20 and September 20, 2019.

The recruitment strategy involved two complementary approaches. First, we created a dedicated Twitter account exclusively for recruitment and pinned to its profile a tweet inviting users to participate in an academic study on electoral preferences (without mentioning the presidential debate — see Figure A1 in Appendix C). The tweet, which offered a supermarket voucher as compensation and included a registration link, was retweeted by several prominent local academics. We refer to participants recruited through this channel as the *Twitter-recruited* sample ($n = 126$).

Second, we hired a professional recruitment agency that implemented two parallel recruitment strategies. The first involved recruiting participants through its regular online panel ($n = 220$, including 106 Twitter users and 114 non-Twitter users), which we refer to as the *recruitment-agency online* sample. The second involved in-person recruitment by field agents targeting individuals not included in the online panel ($n = 342$, including 215 Twitter users and 127 non-Twitter users), which we refer to as the *recruitment-agency offline* sample.

⁵ In line with Hernández et al. (2021) and Sood and Iyengar (2016), we anticipated that such a politically charged event would provide a suitable context for examining reactions to social media influences, particularly in the short term. For an analysis of the relationship between users’ political ideologies and the news they share during Argentina’s 2019 election period, see del Pozo et al. (2024).

During recruitment, all prospective participants were pre-screened for eligibility. Twitter users were asked to follow the experiment’s recruitment account, which allowed the collection of their public Twitter data, including the list of accounts they followed. Table A1 reports pre-treatment characteristics for both the Twitter and non-Twitter samples.

2.a.3. Randomization strategy

Upon arrival on campus, Twitter and non-Twitter participants were directed to separate front desks, where each participant received an individual sealed envelope containing the experimental materials.⁶ The outside of each envelope displayed instructions assigning participants to one of twenty-four classrooms. The classrooms were divided into four types, each corresponding to a distinct treatment arm (see Section 2.a.4), and each was equipped with a large-screen projector and an audio system. Participants were not informed that multiple classroom types existed. The envelopes distributed at check-in differed by participant type (Twitter or non-Twitter user). Twitter users could be assigned to classrooms corresponding to any of the four treatments (Twitter-Allowed, Debate-Only, Twitter-Interact, and Documentary), whereas non-Twitter users were assigned only to classrooms corresponding to the Debate-Only and Documentary treatments.

Within each front desk, sealed envelopes were used to ensure random assignment to classrooms and, consequently, to treatment conditions. For Twitter users, we first used a computer to generate a random ordering of the four treatment arms. Experimenters then followed this ordering as a fixed cycle, distributing envelopes sequentially and repeatedly cycling through the four conditions so that each arriving Twitter participant was assigned to the next treatment in the sequence. For non-Twitter participants, we likewise used a computer to randomize whether the first arrival would be assigned to the Debate-Only or Documentary condition, after which subsequent arrivals were assigned alternately between the two treatments. This procedure guaranteed random assignment within each sample and reduced the likelihood that friends or relatives arriving together would be assigned to the same group — thereby limiting opportunities for discussion or mutual influence during the session. Table A2 and Table A3 report pre-treatment characteristics across treatment arms for Twitter and non-Twitter users, respectively.⁷

⁶ Questionnaires can be found at <https://www.socialscisearch.org/trials/4850>.

⁷ Overall, the pre-treatment characteristics are balanced across treatment arms, suggesting successful randomization. In Table A2, differences in nine out of forty-four variables are statistically significant at 5%. In Table A3, differences in two of the thirty-six variables are statistically significant at the 5% level. Nevertheless, no clear pattern is observed, as these differences cannot be attributed to any single group (in other words, the treatment arm that differs from the rest varies across variables). Our main regressions control for many of these variables. We additionally checked the robustness of our results using multiple sets of controls. In all cases, our results remained robust to these alternative specifications.

Once all participants had reached their designated classrooms, the doors were closed. Participants were then provided with an informed consent form, which they read and signed — with no participant declining to sign it. Following this, participants were instructed to provide an initial saliva sample and to complete a pre-treatment questionnaire covering sociodemographic characteristics, political ideology, and patterns of technology use.

2.a.4. *Treatments*

We implemented four experimental treatments (Twitter-Allowed, Twitter-Interact, Debate-Only, and Documentary) for two purposes. First, within the Twitter sample, the treatments vary participants' access to Twitter during the debate, allowing us to study exposure effects with a focus on the two mitigation strategies discussed in the literature. Second, by including a comparison group of non-Twitter participants in the Debate-Only and Documentary conditions, the design enables us to characterize selection differences between Twitter users and non-Twitter users. We describe each treatment in detail below.

Twitter-Allowed

The Twitter-Allowed treatment, applied exclusively to participants from the Twitter sample, involved watching a live broadcast of the first half of the presidential debate while having unrestricted access to their cellphones, including their personal Twitter accounts. Participants were explicitly informed that they could browse and post freely on social media during the session, and they were given access to the university's Wi-Fi network to support normal usage. The only restrictions were that participants were not permitted to make phone calls or watch videos, even with headphones, and were instructed to refrain from discussing or debating with others in the classroom. Experiment assistants assigned to each classroom explained these rules before the session began and monitored compliance throughout the session.

The treatment was designed to closely replicate how Twitter users typically experience a debate by watching it while engaging with social media. The rationale for including this treatment was to establish a realistic benchmark that captures users' ordinary online behavior during major public events. Accordingly, the Twitter-Allowed group serves as the baseline condition within the Twitter sample.

Twitter-Interact

The Twitter-Interact treatment, applied exclusively to participants from the Twitter sample, consisted of watching a live broadcast of the first half of the presidential debate while being instructed to follow and engage with a designated Twitter account created for the experiment, referred to as the “*active account*.” This account primarily retweeted debate-related content drawn from a diverse set of

Twitter accounts spanning different points on the political spectrum. This set included both relatively moderate sources offering factual or measured commentary and more opinionated partisan accounts, some of which adopted a sharper or more confrontational tone. The intention was not to privilege highly affective content but rather to reproduce a realistic cross-section of the political debate as it appeared on Twitter during the event.

This setup ensured that, for each participant, the active account's activity appearing in their personal Twitter feed reflected a mix of pro-attitudinal, neutral, and counter-attitudinal content. Specifically, during the debate, the active account retweeted twenty-one tweets from *FdT*-leaning accounts and twenty-three from *JxC*-leaning accounts.⁸ The account also posted a small number of neutral tweets (four) and messages from an NGO focused on fact-checking (three). Participants were asked to perform at least one action (liking, retweeting, or replying) on a minimum of three posts or retweets from the active account every ten minutes. This requirement was introduced to ensure they actually viewed and engaged with the curated stream, serving as a compliance check. Experiment assistants made clear that using Twitter during the session was required.

To increase the salience of the active account's activity, the classroom screen was divided into two halves. The live broadcast of the debate was shown on the left side, and on the right side an embedded timeline projected the active account's tweets and retweets in real time (see Figure A2 in Appendix C).⁹ Additionally, to increase the likelihood that participants would see content originating from the active account, they were instructed to switch their Twitter feed to a chronological timeline.¹⁰

The rationale for the Twitter-Interact group was to expose Twitter participants, particularly those embedded in echo chambers, to counter-attitudinal information. Comparing this group with the Twitter-Allowed group allows us to assess whether encouraging engagement with diverse viewpoints serves as an effective strategy for mitigating polarization on social media.

We acknowledge that the Twitter-Interact condition constitutes a bundled treatment, combining elements of both exposure and action, with the latter serving as a compliance device.

⁸ About half of these tweets (47.7%) were authored by well-known partisan Twitter users, identified as “seed accounts” in our Argentine political landscape network (see Section 2.b.1).

⁹ It could be argued that being exposed to segregated content on a shared screen could spark debate among segregated individuals in a classroom, increasing awareness of disagreement among participants. However, we did not observe this kind of interaction in any classroom.

¹⁰ Guess et al. (2023a) examine the effects of Facebook's and Instagram's feed algorithms during the 2020 U.S. election by randomly assigning users to chronologically ordered rather than algorithmically ranked feeds. They find that, although this intervention substantially altered users' on-platform experiences and content exposure, it did not significantly affect either issue polarization or affective polarization over the three-month study period.

However, we argue that its primary effects are more closely tied to the exposure component. As discussed in Section 3.c, participants in the Twitter-Allowed group substantially increased their interactions with political accounts during the debate relative to their pre-debate activity. This pattern suggests that participants in the Twitter-Interact group would likely have engaged with political content even without explicit instruction, although perhaps less intensively. That said, alternative interpretations remain plausible, and we acknowledge that the observed effects could reflect not only exposure, but also the consequences of explicitly encouraging users to engage with content from a broader ideological spectrum.

Debate-Only

For Twitter and non-Twitter participants assigned to the Debate-Only treatment, the session consisted of watching a live broadcast of the first half of the presidential debate with their cellphones turned off — a condition verified by the experiment assistant in each classroom. Within the Twitter sample, this treatment is designed to isolate the experience of watching the debate without any exposure to social media. Comparing the Debate-Only and Twitter-Allowed groups therefore allows us to estimate the effect of removing access to Twitter during a politically salient event and to assess whether restricting social media use mitigates polarization during major political broadcasts.

We acknowledge that turning off participants' phones also prevents them from accessing other apps in addition to Twitter.¹¹ One might therefore argue that differences relative to the Twitter-Allowed group could arise from the loss of those other apps. To assess the plausibility of this concern, we show that participants in the Twitter-Allowed condition exhibited a significant increase in Twitter use during the debate (see Section 3.c), confirming that Twitter was actively used during the live debate within our sample. Should the reader prefer a broader interpretation, the estimates from the Debate-Only group can alternatively be viewed as capturing the effect of completely restricting cellphone access while watching the debate.

Although our analysis of the Debate-Only treatment focuses primarily on Twitter users, a subset of non-Twitter participants was also assigned to this condition. Including them allows us to compare the Debate-Only group with the Documentary group (described below), creating a clean debate-versus-no-debate contrast for individuals who do not use Twitter. This structure mirrors earlier experimental designs that isolate the effects of watching a political debate independent of digital platform use (see, for example, Bidwell, Casey, and Glennerster, 2020). While this comparison is not central to our main research question, it provides useful contextual evidence on how exposure to

¹¹ While one alternative could have been to ask participants to uninstall the Twitter app or deactivate their Twitter accounts, this might have introduced non-compliance issues that would have been challenging to detect.

political content alone may influence attitudes and emotional responses for individuals who do not use social media.

Documentary

The Documentary treatment included both Twitter and non-Twitter participants, who were instructed to turn off their cellphones and watch “*Historia Argentina, DVD 3 (1838 – 1880)*” on the classroom’s main screen, with no presidential debate shown. The documentary focuses on Argentine history during the period of the country’s first Constitution in the 19th century.

The purpose of this treatment was to ensure that participants were exposed to neither the debate nor the social media reactions surrounding it, thereby isolating them from the politically charged context of the event. This condition therefore provides a politically neutral benchmark in which neither Twitter users nor non-Twitter users are exposed to debate-related arousal or to social media activity. Because this treatment places both groups in an equivalent, non-political environment, we use it as our primary setting for examining selection into Twitter use and assessing whether differences in polarization exist between users and non-users under a common low-arousal condition.¹² This comparison, however, should be interpreted with caution, since Twitter use is not randomly assigned. As a result, differences between the two groups in the Documentary condition should be understood as descriptive cross-group differences within our recruited sample, rather than as identified causal effects of selection into Twitter use or as evidence necessarily generalizable to the broader population of Twitter users and non-users.

2.a.5. Post-treatment procedures

When the debate reached its first break, 55 minutes into the projection, all classroom screens were turned off and participants completed a post-treatment questionnaire that included several polarization-related items. They then provided a second saliva sample and returned both samples along with the completed questionnaire before leaving the classroom. Upon exiting the university, participants received a supermarket gift voucher as compensation. In the days that followed, we continued to monitor the Twitter activity of users in the Twitter sample to observe any post-treatment behavioral changes.

¹² In the experiment pre-registration, we referred to the Documentary condition as the “control” group because it minimizes exposure to both the debate and social media. In this paper, however, we use Twitter-Allowed as the primary benchmark for estimating exposure effects among Twitter users, since it most closely reflects the status quo environment in which the mitigation strategies would operate. We use the Documentary condition primarily to study selection into Twitter use by comparing users and non-users in a common politically neutral setting. This change in benchmark constitutes a deviation from the pre-registration, but it is interpretative only and does not affect the underlying statistical analysis. All pairwise differences between treatment arms, as well as their statistical significance, remain unchanged.

2.b. Measuring participants’ political positions from pre-treatment data

Our analysis characterizes participants’ ideological orientation and, among Twitter users, the extent to which their social media environments are politically segregated. For Twitter users, we draw on social media data and online behavior, which provide valuable insights into ideological leanings and the structure of online networks (see, for example, Kosinski et al., 2013 and Barberá, 2015). We follow a two-step procedure. First, we use Twitter data to construct a network representing the Argentine political landscape. Second, we use this network to infer ideological orientation and the degree of segregation among Twitter users based on their observed interaction patterns. For non-Twitter participants, whose online behavior is not observed, ideological orientation is instead predicted from pre-treatment survey data. The following subsections describe these procedures in detail.¹³

2.b.1. *The Argentine political landscape network*

To start building our political network, we followed the approach of Barberá (2015) and Bail et al. (2018) and, before launching participant recruitment, created a curated list of partisan Twitter accounts — which we refer to as “*seed accounts*.” Then, we assigned to each seed account i an ideological score θ_i equal to 0 if account i was an *FdT*-leaning account, and θ_i equal to 1 if account i was a *JxC*-leaning account. These accounts include the main electoral candidates’ Twitter accounts (for the legislative and executive branches in the main districts), the main political parties’ Twitter accounts, accounts of journalists and media outlets often described as partisan, as well as accounts of youth activists, artists, academics, and scientists with clear partisan leanings. Additionally, the list includes some widely followed Twitter accounts with identifiable partisan orientations. Overall, we listed 95 seed accounts (48 *JxC*-leaning and 47 *FdT*-leaning).

Then, for each seed account, we identified all of its “friends” (i.e., the Twitter accounts it follows).¹⁴ Having collected this information, we built a directed network where each node represents a Twitter account (where the seed accounts are a subset) and each directed edge indicates that an account befriends another.¹⁵ The resulting network contains 132,598 nodes and 216,472 edges. Finally,

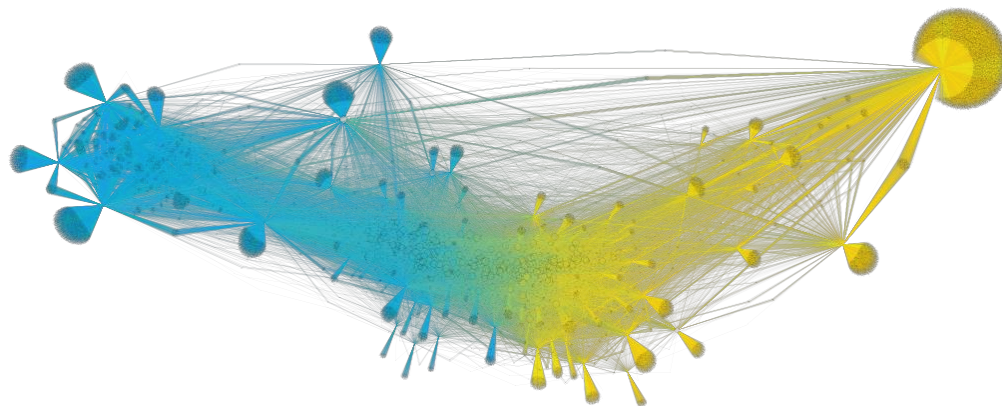
¹³ Importantly, all data and procedures used to construct the segregation measure, including the collection of Twitter followership data and the construction of the Argentine political landscape network, were completed prior to randomization and before any outcome data were observed.

¹⁴ We do this by means of the Twitter API. On average, each *FdT*-leaning seed account follows 2,382 other accounts, while each *JxC*-leaning seed account follows 2,178 other accounts.

¹⁵ We chose to build the network using outgoing edges for two reasons. First, some seed accounts have millions of followers, which would have made it unfeasible to download the data. Second, and more importantly, outgoing edges reflect active choices by the account holder, but incoming edges do not necessarily indicate this.

we made use of a simple label propagation model (see London and Getoor, 2013) to assign an ideological score to every non-seed account. Specifically, for every non-seed account i we assigned a value of θ_i equal to the average of the ideological scores of the nodes directing toward i (all of which are, by construction, seed accounts). Figure 2 illustrates the resulting network, which we refer to as the “*Argentine political landscape network*.”

Figure 2. Argentine political landscape network



Notes. The figure includes the 95 seed accounts and the 132,503 non-seed accounts they follow. A continuous color scale reflects users’ predicted political orientation. Light blue (yellow) indicates an entirely *FdT*-leaning (*JxC*-leaning) account. The size of each node reflects its in-degree.

Consistent with prior research (e.g., Adamic and Glance, 2005, and Halberstam and Knight, 2016), connections between accounts with opposing ideological positions are relatively rare, reflecting the presence of ideological echo chambers. Figure A3 shows that the distribution of ideological scores within the network is highly bimodal. Moreover, the network structure closely aligns with findings from previous studies on Argentine politics and social media (e.g., Aruguete and Calvo, 2018 and Cicchini et al., 2022).

2.b.2. *Measuring social media segregation among Twitter participants*

For clarity, we begin by describing how we measure social media segregation among Twitter users before turning to participants’ ideological orientations in Section 2.b.3. Social media segregation captures the extent to which users are exposed to politically homogeneous content within their online environments. On Twitter, exposure primarily occurs through the accounts a user follows. Thus, when a participant i follows another user j , they are exposed to j ’s activity (tweets, retweets, quoted retweets, and, to a lesser extent, likes). As a result, if user i mostly follows accounts that share their political leanings, they are effectively situated within a pro-attitudinal echo chamber.

Building on this idea, we construct a measure of online segregation for each Twitter participant based on the political composition of the accounts they follow. Specifically, for each user i , we calculate the average ideological score of the accounts they follow within the Argentine political landscape network, excluding all followed accounts outside the network. This average, denoted λ_i^f and referred to as the *follower-based ideological score*, behaves as follows: i) if the user predominantly follows *FdT*-leaning accounts, λ_i^f will be close to 0; ii) if the user primarily follows *JxC*-leaning accounts, λ_i^f will be close to 1; and iii) if the user follows a balanced mix of accounts, λ_i^f will be close to 0.5. We then define the *raw segregation score* for user i , denoted r_i , as $2 \cdot |\lambda_i^f - 0.5|$. This score ranges from 0 to 1, with higher values indicating greater ideological segregation.

A potential drawback of r_i is that it does not consider the degree to which a user follows political accounts.¹⁶ To correct for interest in politics, we define i 's *politicization score* (p_i) as the number of accounts included in the Argentine political landscape network that i follows divided by the total number of accounts i follows.¹⁷ Then, we calculate i 's *normalized segregation score* (s_i) as $r_i \cdot p_i$. This score also ranges from 0 to 1. Values close to 0 indicate that Twitter user i is not strongly embedded in a political echo chamber, either because they follow few political accounts or because the political orientations of those accounts are relatively balanced. By contrast, values close to 1 suggest that user i is deeply embedded in an environment characterized by a high degree of political-ideological homogeneity. Figure A4 plots the distribution of the normalized segregation score s_i and its components, r_i and p_i , for our Twitter-user sample.

Finally, we classify a participant i as “*segregated*” (i.e., in an echo chamber) if s_i is greater than the median value of the normalized segregation score calculated over our Twitter-user sample, and as “*non-segregated*” otherwise. This results in 223 participants being categorized as segregated and 224 as non-segregated. Table 1 reports the distribution of participants by segregation status and treatment

¹⁶ As an illustration, consider a user i , who follows 1,000 accounts, all of which are included in the political network and have an ideological score equal to 1, and a user j who follows 999 soccer-related accounts, but just a single account that is included in our political network and happens to have an ideological score equal to 1. In this case, both r_i and r_j will be equal to 1, even if i is more segregated than j . In fact, one could argue that j is not segregated at all, as j 's Twitter interests center on other topics.

¹⁷ An ex-post validation exercise indicates that participants who answered the maximum score in “interest in politics” in the pre-treatment questionnaire had a mean value of p equal to 0.447, while participants who answered a lower score had a mean value of p of 0.327. A two-sided t -test indicates that this difference is highly significant ($t = 5.57, p < 0.01$). 44.4% of all participants answered the maximum score in “interest in politics.”

arm. Pre-treatment characteristics by treatment arm are reported separately for segregated and non-segregated participants in Table A4 and Table A5.¹⁸

Table 1. Distribution of Twitter users by segregation status and treatment arm

Treatment arm	Segregation status		Total
	<i>Non-segregated</i>	<i>Segregated</i>	
Twitter-Allowed	50	60	110
Debate-Only	63	49	112
Twitter-Interact	62	50	112
Documentary	49	64	113
Total	224	223	447

2.b.3. *Measuring participants’ political orientation*

When calculating a subset of our outcome variables (see Section 2.c), we need to infer whether subjects lean toward the *FdT* or the *JxC* coalition. Because our sample includes both Twitter and non-Twitter users, we rely on two complementary approaches to estimate ideological orientation. For participants in the Twitter sample, we infer political preferences directly from their observed Twitter activity. For participants in the non-Twitter sample, we later apply a supervised machine learning classifier to predict ideological orientation based on pre-treatment survey data. Below, we provide details on each approach.

Predicting ideological orientation for Twitter users

For participants in the Twitter sample, we infer political orientation from their likes. We rely on likes rather than follower relationships because preliminary tests on an internal sample showed that follower networks were substantially less informative than likes for distinguishing participants’ political orientation. This is consistent with the nature of following: it is relatively infrequent, often accumulated over long periods, and may reflect interest, habit, or social ties rather than endorsement. Likes, by contrast, are more immediate, require less effort, and are more closely tied to users’ reactions to specific pieces of content.¹⁹

¹⁸ Randomization was carried out separately for Twitter and non-Twitter participants, but it was not stratified by the segregation indicator, which is constructed from pre-treatment Twitter followership data. Therefore, when we split the Twitter sample ex post into segregated and non-segregated participants, exact covariate balance across treatment arms within each subgroup is not guaranteed by design. However, the modest differences in pre-treatment covariates that we document within subgroups are consistent with proper randomization.

¹⁹ Likes also offer several advantages over other forms of engagement such as mentions or retweets. First, they are by far more common. Second, at the time, likes were less public than retweets and quoted retweets (they were not often shown in other users’ feeds), so users tended to be less cautious when liking tweets. Although likes are private today, at the time of our study they were publicly visible, which further supports their interpretation as genuine expressions of endorsement.

Concretely, for each Twitter user i , we estimate a *like-based party preference score* (λ_i^l), defined as the median ideological score of the accounts in our political landscape network whose tweets user i liked between June 20 and October 12, 2019.²⁰ If this median equals 0.5 (34 cases), we replace it with the mean of those ideological scores. If the mean is also 0.5 (2 of those 34 cases), or if the user did not like any tweets from accounts included in the political network during this period (24 cases), we assign λ_i^l the user's *follower-based ideological score* (λ_i^f). Accordingly, participants who primarily liked tweets from accounts with ideological scores close to 0 (*FdT*-leaning) have values of λ_i^l close to 0, whereas those who primarily liked tweets from accounts with ideological scores close to 1 (*JxC*-leaning) have values of λ_i^l close to 1. Finally, we classify participants as *FdT*-leaning if $\lambda_i^l < 0.5$ and as *JxC*-leaning otherwise. This procedure yields 232 *FdT*-leaning and 215 *JxC*-leaning participants.

Figure A5 displays the distributions of λ^l (the *like-based party preference score*) and λ^f (the *follower-based ideological score*) across three Twitter participant groups: i) those who reported voting for the *FdT* candidate in the primary elections (light blue), ii) those who reported voting for the *JxC* candidate (yellow), and iii) those who indicated voting for other candidates (gray).²¹ Overall, the patterns observed indicate that our metric exhibits strong discriminative performance.

Predicting ideological orientation for non-Twitter users

For participants who do not use Twitter, it is not possible to measure online behavior directly. However, because several of our outcome variables require identifying each participant's ideological orientation, we predict ideological orientation for non-Twitter users using a supervised machine learning model trained on the Twitter sample. Specifically, we employ a random forest model (Hastie, Tibshirani, and Friedman, 2009) trained on Twitter users' pre-treatment survey responses to infer whether a participant is *FdT*- or *JxC*-leaning.²² We then use this model to generate predictions for the non-Twitter sample, using the same set of pre-treatment predictors. Random forest was chosen for

Third, mentions and quoted retweets are often used to show opposition to tweets, whereas the vast majority of likes can be interpreted as endorsements.

²⁰ When a user j 's tweets are liked n times by i , θ_j is counted n times in the calculation of λ_i^l .

²¹ Although the survey question on primary vote intention was administered post-treatment, our interventions were strictly neutral with respect to candidates. Consistent with this, an ex-post analysis in the full sample provides no evidence that treatment affected reported primary vote preferences. Results are available upon request.

²² The predictors used in the model are the sociodemographic, ideology-related, and technology-use variables from the pre-treatment questionnaire, including items such as age, gender, education, social values, economic attitudes, views on human rights, political interest, and measures of phone and internet usage. Missing values were imputed using the sample mean for continuous variables and the mode for categorical variables, and for each variable with imputed observations we included a separate indicator identifying those observations.

its ability to capture nonlinear relationships and interactions, and for its robust performance in moderate samples (see Mullainathan and Spiess, 2017).²³ The model achieves an out-of-sample validation accuracy of 72%, based on five-fold cross-validation.

Figure A6 shows the distribution of the predicted probability of leaning toward JxC in the non-Twitter sample, plotted separately for three groups defined by their reported vote choice in the primary elections: i) those who voted for the *FdT* candidate (light blue), ii) those who voted for the JxC candidate (yellow), and iii) those who voted for other candidates (gray). As in Figure A5, the distributions display clear and systematic differences across groups, indicating that the predicted-probability measure has strong discriminatory power.

2.c. Outcome variables

We analyze two main sets of outcome variables. The first includes the primary outcomes, consisting of measures of political polarization and physiological stress derived from post-treatment survey and biological data. These outcomes were pre-registered in our analysis plan and constitute the core focus of the study. The second set includes additional measures related to Twitter activity, which were not pre-registered but provide complementary evidence on how exposure to political debate and social media influences online behavior.

2.c.1. *Primary outcome variables*

Polarization indexes

Our post-treatment survey included several questions designed to capture different dimensions of political polarization. Based on these responses, we constructed five indexes measuring participants' polarization across commonly studied dimensions. In all cases, higher values indicate higher levels of polarization. The five indexes are *Elite*, *Social Distance*, *Partisanship*, *Institutional*, and *Engagement*. Below, we describe how each index is constructed. Additional details are provided in Appendix A.²⁴

²³ See also Angrist et al. (2020) and Sohnesen and Stender (2017) for examples of random forest being used in the context of applied social science research.

²⁴ When calculating an index, and whenever an index or subcomponent is constructed by averaging or summing multiple items, each item is first normalized so that its distribution in the Twitter-Allowed group has mean 0 and standard deviation 1. If an item is missing for a given participant, it is excluded from the corresponding calculation.

We first focus on affective polarization, that is, the dislike or distrust citizens have toward those identified with other political parties.²⁵ Following Druckman and Levendusky (2019), we distinguish dislike toward party elites from dislike toward other partisans. Accordingly, the *Elite* index captures the dislike of outgroup party elites and is constructed as the average of two components: 1) the perception that the outgroup party’s presidential candidate is a threat to the well-being of Argentines, and 2) the difference between the normalized number of positive traits participants assign to the ingroup party leader and the normalized number of positive traits subjects assign to the outgroup leader.²⁶ To determine each participant’s ingroup and outgroup parties, we employ the procedure outlined in Section 2.b.3.

The *Social Distance* index captures affective polarization toward ordinary supporters of the opposing party. It is computed as the average of three components: 1) the participant’s discomfort with having an immediate family member marry someone with opposing political views, 2) their discomfort with socializing with someone who holds opposing views, and 3) their discomfort with working closely with someone from the opposite political camp.

The *Partisanship* index captures the degree to which participants are firmly committed to their political choice.²⁷ It is computed as the average of two components: 1) the extent to which participants report being certain about their intended vote, and 2) the concentration of their preferences across candidates. The latter is measured as the standard deviation of scores assigned in an exercise where each participant distributed 100 preference points among all presidential candidates.²⁸

²⁵ See Iyengar et al. (2019), Boxell, Gentzkow, and Shapiro (2020), and Rogowski and Sutherland (2016). Affective polarization has been shown to have risen in the last decade (Druckman et al., 2021) and it has been speculated to have a negative impact on democratic norms and accountability (see, for example, Graham and Svobik, 2020, and Gidron, Adams, and Horne, 2020).

²⁶ Our pre-analysis plan listed “differences in participants’ positive and negative perceptions of politicians across treatment arms” as a primary outcome. This pre-registered outcome corresponds directly to the two attitudinal components of the *Elite* index, which are constructed from the same survey items identified in the pre-analysis plan.

²⁷ Partisanship has been shown to correlate with political polarization (see Lupu, 2015) and has been shown to be heightened when elections are approaching (see Schwalbe, Cohen, and Ross, 2020). In Allcott et al. (2020) this dimension is referred to as *vote polarization*.

²⁸ Assigning points evenly across several candidates results in a low standard deviation, indicating weak partisanship, whereas allocating nearly all points to a single candidate yields a high standard deviation, reflecting strong partisan conviction. Since the answers of a few participants did not add up to 100, we normalized the scores for all participants to ensure that they add up to 1 before calculating the standard deviation. Our results remain consistent when we calculate dispersion scores based on participants’ assigned scores ranging from 1 to 10 for each presidential candidate. Additionally, the results remain robust when using the Herfindahl-Hirschman index to measure dispersion/concentration.

The *Institutional* index captures both the extent of participants' distrust in political institutions and the importance they place on loyalty over competence when evaluating executive appointments — a perspective reflecting tolerance for institutional weakness (see Edwards, 2001 and Egorov and Sonin, 2011). This measure relates to what the literature often terms issue or ideological polarization (see, for example, Allcott et al., 2020, Levy, 2021, and Bail et al., 2018). The index is computed as the average of three components: 1) the difference between how credible participants found allegations of electoral fraud by the outgroup party versus the ingroup party, 2) the difference between the level of disagreement with a highly publicized Supreme Court ruling against their own party versus one against the opposing party, and 3) the importance assigned to ministerial loyalty relative to technical competence in executive branch appointments.

The *Engagement* index captures the degree to which participants are willing to engage in the electoral process (e.g., voting, contributing money, working or volunteering for a campaign, attending a campaign event). High political engagement has been associated with ideological extremes, particularly in voters holding unfavorable views of the opposing party (Pew Research Center, 2014b). The index is calculated as the average of two components: 1) a score indicating the willingness to serve as an election observer at a polling place, volunteer, or donate funds to their preferred candidate, and 2) a score indicating the subject's willingness to persuade others to vote for their preferred candidate.

In addition to these five indexes, we construct a set of aggregate polarization measures. Our primary measure, referred to simply as the *Combined Index*, is the average of the five polarization indexes described above and serves as our main polarization outcome.²⁹ To assess robustness, we also compute three alternative versions. *Combined Index V2* averages the twelve individual components underlying the five indexes, thereby assigning equal weight to each component rather than to each broader dimension. Because political engagement is often correlated with polarization, although some studies suggest that intense polarization may also discourage engagement (Hetherington et al., 2008), we also construct two variants that exclude this dimension. *Combined Index V3* mirrors the main index but omits the Engagement index, while *Combined Index V4* averages the ten components underlying V3.

Cortisol variation

There is growing evidence that political events and polarized encounters can elicit physiological stress responses (Simons and Green, 2016). For instance, elections have been shown to increase cortisol levels among voters and to heighten anxiety and physiological stress among both youth and adults during highly charged electoral periods (Waismel-Manor, Ifergane, and Cohen, 2011,

²⁹ The individual indexes are only weakly correlated with one another. The largest estimated Pearson correlation between any pair is 0.366, for the *Partisanship* and *Engagement* indexes.

Suzuki et al., 2023, American Psychological Association, 2020). Against this background, we include cortisol variation as an outcome variable to assess whether our experimental interventions generate measurable physiological stress. Cortisol is a well-established biomarker of stress and can be measured accurately using non-invasive saliva tests (Kirschbaum and Hellhammer, 1989).³⁰

In our study, we measure salivary cortisol before and after treatment to capture participants’ stress response to the experimental intervention. We define $\Delta\text{Cortisol}$ as the difference between post- and pre-treatment measurements. Using $\Delta\text{Cortisol}$ rather than post-treatment levels alone is advantageous because it accounts for baseline individual differences in cortisol production, which may reflect long-term stress exposure, diurnal patterns, or personality traits.

2.c.2. Complementary outcome variables

Twitter actions during the debate

For participants assigned to the Twitter-Allowed and Twitter-Interact groups, we calculate two measures of Twitter behavior during the debate, capturing its quantity and composition, denoted as *Activity Level* and *Contact Segregation*, respectively.³¹

Activity Level captures the volume of online engagement and is defined as the total number of original tweets, replies, mentions, retweets, quoted retweets, and likes produced by a participant during the segment of the debate viewed in the experimental session.³²

Contact Segregation measures the degree to which participants’ Twitter interactions during the debate occurred within an ideologically homogeneous circle. We construct *Contact Segregation* following a four-step procedure (which mimics the procedure described in Section 2.b.2 to classify Twitter users as segregated or non-segregated). First, for a given account i , we compute its *activity politicization score* (p_i^{act}) as the proportion of its activity (replies, mentions, retweets, quoted retweets, and likes) performed during the experimental debate segment that was directed toward accounts included in the political landscape network. Second, we calculate the mean ideological score of the accounts with

³⁰ Cortisol has been studied in prior research such as Dickerson and Kemeny (2004), Young (2004), Wirtz et al. (2008), Ditzen et al. (2009), Carney and Yap (2010), Di Tella et al. (2019), and Haushofer et al. (2022).

³¹ The pre-registration focused exclusively on polarization and cortisol outcomes and did not include analyses of Twitter behavioral measures. The Twitter outcomes presented in this section were therefore not pre-registered and should be viewed as complementary exploratory evidence. Note that no additional Twitter-based variables were examined beyond those presented here.

³² To address potential issues arising from exceptionally high activity levels exhibited by some users during the analyzed period, we apply the inverse hyperbolic sine (IHS) transformation, defined as $\ln(x + \sqrt{x^2 + 1})$ (see Burbidge, Magee, and Robb, 1988).

which it interacted ($\bar{\theta}_i^{act}$).³³ Third, we determine i 's *raw activity segregation score* (r_i^{act}) as $2 \cdot |\bar{\theta}_i^{act} - 0.5|$. Finally, *Contact Segregation* for account i is obtained by multiplying its *activity politicization score* by its *raw activity segregation score* ($p_i^{act} \cdot r_i^{act}$).³⁴ Higher values of *Contact Segregation* indicate that a participant's debate-related activity was concentrated among ideologically aligned accounts, reflecting engagement within a politically homogeneous online environment.

Variation in Twitter actions

In addition to examining activity during the debate itself, we also assess whether the experimental interventions produced short-term changes in participants' subsequent online behavior. To do so, for each Twitter participant, we calculate *Activity Level* and *Contact Segregation* using all Twitter actions performed on the day after the debate. Because these measures can vary substantially across individuals, we adjust for baseline differences in Twitter use by subtracting the corresponding measures computed for the day before the debate. The resulting differences, denoted $\Delta Activity Level$ and $\Delta Contact Segregation$, capture short-term changes in both the volume and composition of participants' online engagement following the debate. Positive values of $\Delta Activity Level$ indicate an increase in overall posting and interaction relative to baseline, while positive values of $\Delta Contact Segregation$ indicate that participants' post-debate interactions became more ideologically homogeneous.

2.d. Empirical strategy

Regression specifications

In our analysis, we estimate the values of a given outcome Y across treatment arms at two levels. First, we examine effects in the full sample of participants. Later, we explore heterogeneous effects by online segregation status. In this section, we describe the regression specification used in each case.

When considering all participants as a whole, we estimate the following regression model:

$$Y_i = \sum_{k,j} \beta_{k,j} (Twitter_{i,j} \cdot T_{i,k}) + \gamma \cdot X_i + \epsilon_i, \quad (1)$$

where Y_i denotes the value of outcome Y for individual i , $Twitter_{i,j}$ is a set of dummy variables indicating whether participant i is a Twitter user or not ($j \in \{Twitter, non-Twitter\}$), $T_{i,k}$ represents dummy variables for treatment assignment ($k \in \{Twitter-Allowed, Debate-Only, Twitter-Interact,$

³³ If user i interacted with user j a total of n times, θ_j counts n times when calculating this average.

³⁴ If a user did not interact with any account in the political landscape network, we set its contact segregation equal to 0.

Documentary}),³⁵ X_i is a vector of pre-treatment covariates (age, age², years of education, female = 1, household head = 1, recruitment agency online = 1, and recruitment agency offline = 1), and ϵ_i is an error term. To account for the possibility of correlation in the error term within a classroom, we cluster the standard errors at the classroom level in all regressions (see Abadie et al., 2023).³⁶ Additionally, before estimating Eq. 1, we normalize Y and all the control variables, so that the distribution of the Twitter-Allowed group has a mean of 0 and a standard deviation of 1. Note that, because the model excludes a constant term and all variables are standardized, the coefficients $\beta_{k,j}$ should be interpreted as the mean standardized outcome for participants in treatment arm k and user type j . In the full-sample analysis, these coefficients provide the main inputs for our estimates of selection and exposure.

To examine heterogeneous effects by segregation status, we estimate a specification analogous to Eq. 1, but distinguish among three groups: non-segregated Twitter users, segregated Twitter users, and non-Twitter participants. Formally, we interact each treatment indicator with a set of three mutually exclusive dummy variables, denoted $S_{i,l}$, where each participant belongs to exactly one category. The model specification is:

$$Y_i = \sum_{k,l} \beta_{k,l} \cdot (S_{i,l} \cdot T_{i,k}) + \gamma \cdot X_i + \epsilon_i, \quad (2)$$

where $l \in \{\text{non-segregated}, \text{segregated}, \text{non-Twitter}\}$, and all other variables follow the definitions in Eq. 1. As in Eq. 1, the specification in Eq. 2 excludes a constant term, and all variables are standardized relative to the mean and standard deviation of the Twitter-Allowed group. Under this specification, the coefficients $\beta_{k,l}$ should be interpreted as the mean standardized outcome for individuals in treatment arm k and segregation category l . In the analysis by online segregation, these coefficients provide the main inputs for our estimates of selection and exposure.

Identification of selection and exposure

Our experimental design allows us to identify two objects of interest: exposure effects, which correspond to causal treatment effects among Twitter users, and selection patterns between Twitter

³⁵ Because only Twitter users were eligible for the Twitter-Allowed and Twitter-Interact treatments, these indicators are mechanically equal to zero for non-Twitter participants.

³⁶ By randomly assigning participants to classrooms, we ensure that no systematic differences in pre-treatment characteristics exist across groups. For this reason, clustering the errors at the classroom level accounts for potential intra-classroom dependencies that may arise during the experiment, in particular, shared discussions about the debate. However, as mentioned before, interaction between participants was actively discouraged, and experiment assistants assigned to each classroom ensured compliance with this rule. All our main results remain unaltered when classical or HC1/robust standard errors are used (for a discussion on the topic, see King and Roberts, 2015).

and non-Twitter participants placed in comparable environments. Under this design, both are estimated through linear hypothesis tests applied to the coefficients from Eq. 1 and Eq. 2.

To identify exposure effects within the Twitter sample, we compare each treatment group with the Twitter-Allowed condition, which approximates users' natural behavior of watching the debate while using Twitter. Formally, in the full-sample specification (Eq. 1), exposure is assessed through linear tests of the form $H_0: \beta_{Debate-Only, Twitter} = \beta_{Twitter-Allowed, Twitter}$ and $H_0: \beta_{Twitter-Interact, Twitter} = \beta_{Twitter-Allowed, Twitter}$. When incorporating segregation status (Eq. 2), these comparisons are repeated separately for non-segregated and segregated Twitter users in turn.

To assess selection, we compare Twitter and non-Twitter participants within the Documentary condition. In this condition, all individuals are placed in a politically neutral setting and watch a non-political documentary with their phones switched off, ensuring that neither group is exposed to the debate or to social media activity. Because both groups face the same low-arousal informational environment, differences in outcomes can be interpreted as selection comparisons between Twitter and non-Twitter participants under a common setting (though they should not be interpreted as identifying causal effects of selection into Twitter use). Although the Debate-Only condition could also be used for this purpose, we consider it less suitable. For Twitter users, watching the political debate without their usual online interaction represents a marked departure from their normal behavior, whereas for non-Twitter participants, passively watching the debate is closer to their usual experience. This creates an asymmetry in the comparison, since one group is removed from its status quo while the other is not. The Documentary condition mitigates this concern by removing both the debate and phone activity, thereby providing a common low-arousal, politically neutral setting for selection comparisons.³⁷

Formally, for a given outcome Y , we assess selection by comparing the estimated means of Twitter and non-Twitter participants within the Documentary condition. In the full-sample specification (Eq. 1), this corresponds to testing $H_0: \beta_{Documentary, Twitter} = \beta_{Documentary, non-Twitter}$. Under Eq. 2, we conduct analogous comparisons by testing $H_0: \beta_{Documentary, l} = \beta_{Documentary, non-Twitter}$ for $l \in \{\text{non-segregated, segregated}\}$.

³⁷ For completeness, we note that the selection estimates reported throughout the paper are qualitatively the same whether they are computed by comparing Twitter and non-Twitter participants within the Documentary condition or within the Debate-Only condition, which indicates that our conclusions about selection do not depend on which of these two common environments we use for the comparison.

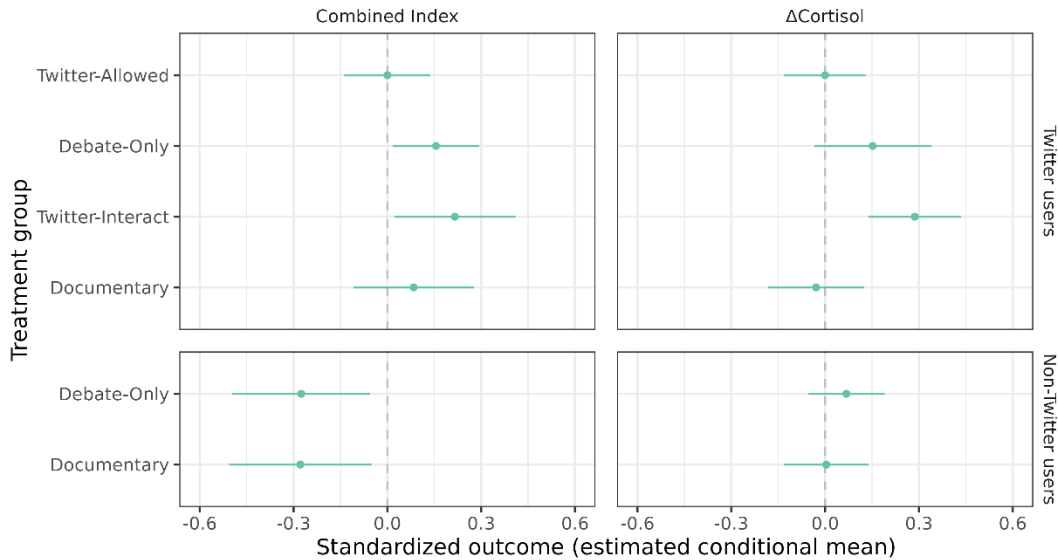
3. Results

This section presents our main findings. First, we report results for the full sample of participants. Second, Section 3.b explores treatment heterogeneity by incorporating online segregation into the analysis. Third, Section 3.c provides complementary evidence based on behavioral outcomes measured through Twitter activity. Finally, Section 3.d summarizes the main results before turning to the broader discussion in Section 4.

3.a. Results for the full sample

We begin by examining selection and exposure patterns using our full sample of participants.³⁸ Figure 3 reports estimates obtained from Eq. 1 based on the entire sample. Concretely, for each outcome variable and treatment group, the figure displays estimated conditional means and their 90% confidence intervals. Results are presented for the *Combined Index*, our main measure of political polarization, and for $\Delta\text{Cortisol}$.

Figure 3. Combined Index and $\Delta\text{Cortisol}$ by treatment group, full sample



Notes. The figure displays estimated conditional means for each treatment group and outcome variable, as obtained from Eq. 1. All outcome variables are standardized such that the distribution for the Twitter-Allowed group has a mean of zero and a standard deviation of one. Error bars represent 90% confidence intervals, with standard errors clustered at the classroom level. See Section 2.c for variable definitions and Table A6 for detailed estimates.

³⁸ Results presented in this section correspond to the analyses pre-registered in the AEA trial registry (AEARCTR-0004850).

We begin by examining selection patterns by comparing Twitter users and non-Twitter users within the Documentary condition (see Section 2.d). Visual inspection of Figure 3 shows that, within this condition, the *Combined Index* is higher among Twitter users than among non-Twitter users. Formally, a test comparing the estimated means across these two samples shows that this difference is statistically significant at the 5% level. In contrast, when focusing on $\Delta\text{Cortisol}$, the corresponding test reveals no statistically significant difference between the two groups. Taken together, these findings suggest that selection into Twitter use is associated with higher levels of political polarization, but not with differences in physiological stress as measured by cortisol. Detailed results are available in Table A7.³⁹

Focusing on exposure within the Twitter sample, visual inspection of Figure 3 indicates that the *Combined Index* is higher for participants in the Debate-Only and Twitter-Interact conditions compared with those in the Twitter-Allowed benchmark group. Although the estimated differences are not statistically significant, the direction of the effects is consistent with a backfire pattern, suggesting that interventions designed to reduce polarization may instead have increased it in the short run.⁴⁰ In contrast, results for $\Delta\text{Cortisol}$ indicate that the Twitter-Interact treatment significantly increases cortisol relative to the Twitter-Allowed benchmark, with the difference statistically significant at the 5% level. This finding suggests that requiring participants to actively engage with partisan content during the debate may have induced greater physiological stress. No differences are observed between the Documentary and Twitter-Allowed conditions for either the *Combined Index* or $\Delta\text{Cortisol}$. Detailed results are available in Table A8.⁴¹

While *Combined Index* provides a comprehensive summary of political polarization, it aggregates several distinct dimensions that may not respond uniformly to the experimental interventions. To gain deeper insight into the underlying dynamics, we disaggregate *Combined Index*

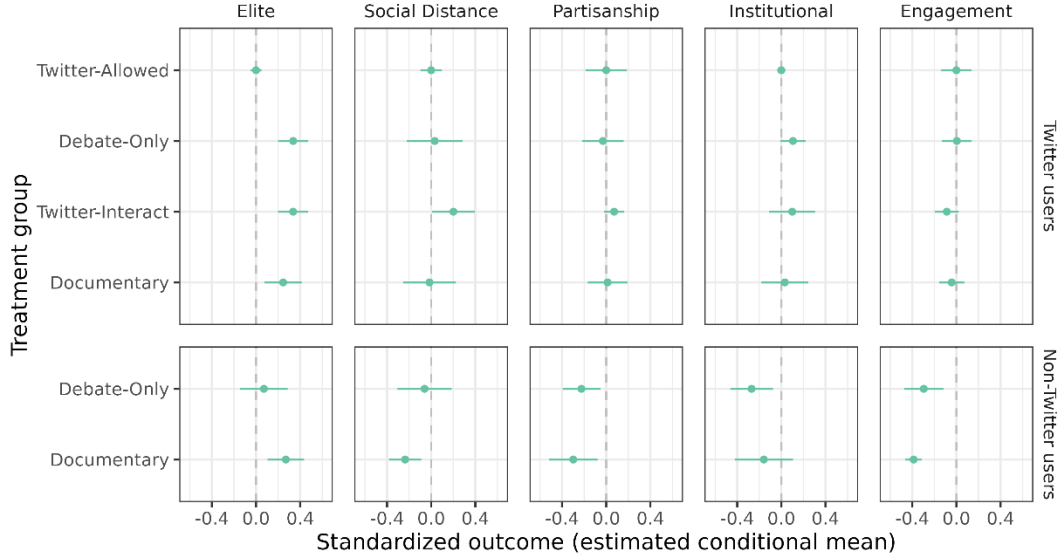
³⁹ The same patterns hold when the comparisons are conducted using the Debate-Only sample. See Table A7 for details.

⁴⁰ Results for the main *Combined Index* are broadly robust to alternative definitions of the index (see Section 2.c.1). The main exception is *Combined Index V3*, which excludes the *Engagement* index. Under this specification, both the Debate-Only and Twitter-Interact groups show statistically significant increases in polarization at the 10% level relative to the Twitter-Allowed benchmark. In addition, under *Combined Index V4*, which averages the ten components underlying V3, the Twitter-Interact group also shows a statistically significant increase at the 10% level, while the Debate-Only estimate remains statistically insignificant. See Table A8.

⁴¹ Within the non-Twitter-user sample, the differences between the Debate-Only and Documentary groups for both outcomes are negligible and not statistically significant, suggesting that, for this group, watching the debate versus not watching it had no discernible impact on either polarization or stress. Detailed results are available upon request.

into its five constituent measures and analyze them individually.⁴² Figure 4 reports the results from estimating Eq. 1 using each individual index as the outcome variable.

Figure 4. Individual polarization indexes by treatment group, full sample



Notes. The figure displays estimated conditional means for each treatment group and outcome variable, as obtained from Eq. 1. All outcome variables are standardized such that the distribution for the Twitter-Allowed group has a mean of zero and a standard deviation of one. Error bars represent 90% confidence intervals, with standard errors clustered at the classroom level. See Section 2.c for variable definitions and Table A9 for detailed estimates.

Regarding selection patterns, Figure 4 reveals a broadly consistent trend across the different dimensions of polarization. Among participants in the Documentary condition, differences between Twitter and non-Twitter users are positive for all indices except the *Elite* index, suggesting that Twitter users tend to display higher levels of polarization across most dimensions. Two of these differences are statistically significant, with the *Partisanship* index significant at the 10% level and the *Engagement* index significant at the 1% level. However, after adjusting for multiple comparisons across outcomes

⁴² Because the outcomes analyzed in this paper span distinct conceptual domains (e.g., survey-based polarization, cortisol levels, and Twitter activity), we do not apply a single multiple-testing correction across all regressions. However, we make an exception for the individual polarization indexes reported in Figure 4 and Figure 6, where we examine treatment differences across the five individual polarization indexes (*Elite*, *Social Distance*, *Partisanship*, *Institutional*, and *Engagement*). It could be argued that these indexes capture related dimensions of a broader underlying construct and are jointly informative and collected, so interpreting them together warrants accounting for multiple comparisons. For this reason, we report both unadjusted and Holm-adjusted significance levels for each set of polarization outcomes. Concretely, for each hypothesis test, whether pertaining to selection or exposure, we treat the five individual polarization indexes corresponding to a single treatment contrast as one family and apply Holm's correction within that five-test family.

using the Holm procedure, the *Partisanship* index loses its significance, while the *Engagement* index remains statistically significant at the 1% level. Detailed estimates are reported in Table A10.⁴³

When examining exposure effects within the Twitter sample, statistically significant differences emerge only for the *Elite* index. Participants in both the Debate-Only and Twitter-Interact conditions exhibit higher *Elite* index values than those in the benchmark Twitter-Allowed group, consistent with a backfire effect and significant at the 1% level. A similar but weaker pattern appears for the Documentary condition, where the difference is significant at the 5% level. For all other indices, no statistically significant differences are detected. After accounting for multiple comparisons across outcomes, the *Elite* index difference for the *Documentary* condition is no longer statistically significant, whereas the differences for the *Debate-Only* and *Twitter-Interact* conditions remain significant at the 1% level. Detailed estimates are reported in Table A11.

Taken together, the results based on the individual indexes align closely with those obtained for the *Combined Index*. Differences between Twitter and non-Twitter users point to greater polarization among the former, while interventions intended to mitigate polarization appear to have had limited effects overall, with several estimates pointing in a counterproductive direction.

3.b. Heterogeneity by online segregation

Up to this point, our findings reveal marked selection differences, with Twitter participants exhibiting systematically higher levels of polarization than non-Twitter participants. At the same time, exposure effects within the Twitter sample, though more limited in magnitude, are consistent with backfire dynamics, suggesting that interventions intended to reduce polarization may increase polarization in the short run.

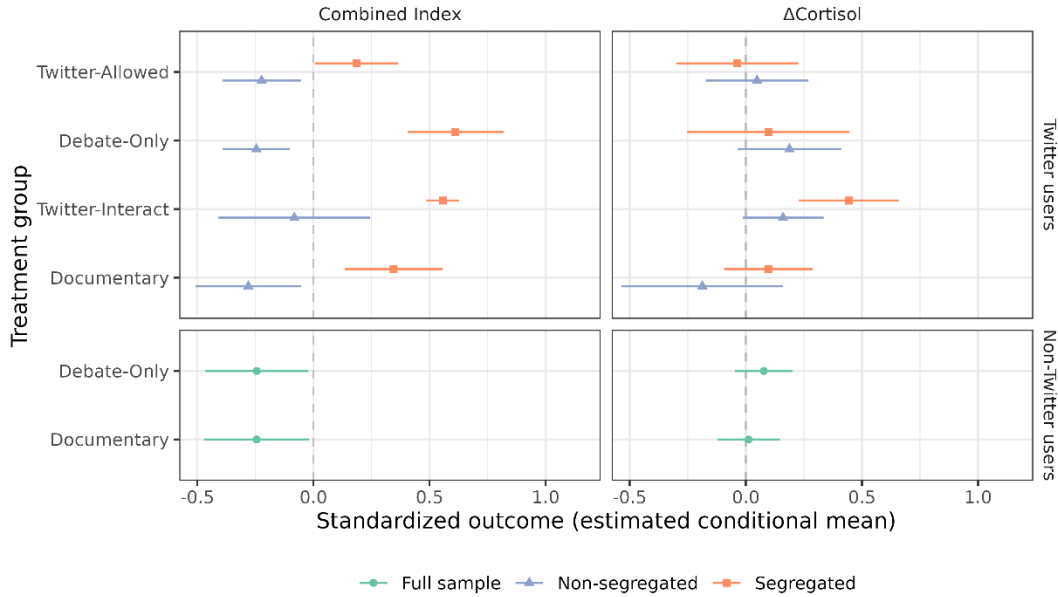
Motivated by these patterns, we next examine whether online segregation offers a more nuanced understanding of these effects. Following the procedure described in Section 2.b.2, we divide the Twitter sample into two groups: participants embedded in political echo chambers (segregated) and those who are not (non-segregated). We then revisit the analyses presented in Section 3.a, this time introducing online segregation as an additional dimension within the Twitter sample. Note that our pre-analysis plan registered heterogeneity analyses by sociodemographic characteristics, ideology, and beliefs, after hypothesizing that echo chambers distort beliefs and perceptions, potentially leading to polarized views. We consider online segregation as a closely related manifestation of ideological heterogeneity, as it reflects how individuals' online environments shape their exposure to differing

⁴³ A qualitatively similar pattern holds when the comparisons are conducted using the Debate-Only sample. See Table A10 for details.

viewpoints. However, because this dimension was not explicitly pre-registered, results regarding online segregation could be interpreted as suggestive evidence.

Figure 5 replicates Figure 3, but separates the Twitter sample into segregated and non-segregated participants. Conditional means for each treatment group and outcome variable are estimated using Eq. 2.

Figure 5. Combined Index and Δ Cortisol by treatment group and segregation status



Notes. The figure presents estimated conditional means for each treatment group and each outcome variable with the Twitter sample disaggregated into segregated and non-segregated participants (see Section 2.b.2), as obtained from Eq. 2. All variables are standardized so that the distribution for the Twitter-Allowed group has a mean of 0 and a standard deviation of 1. Error bars represent 90% confidence intervals, with standard errors clustered at the classroom level. See Section 2.c for variable definitions and Table A12 for detailed estimates.

Regarding selection, whereas Figure 3 suggested that Twitter users were more polarized than non-Twitter users, Figure 5 shows that this difference is driven almost entirely by participants classified as segregated. When comparing Documentary non-Twitter users with Documentary non-segregated Twitter users, the difference in the *Combined Index* is essentially zero and statistically insignificant. By contrast, comparing non-Twitter users with segregated Twitter users reveals substantially higher polarization among the latter, with the difference significant at the 1% level. As in Figure 3, no significant differences are observed for Δ Cortisol, either between non-Twitter users and

non-segregated Twitter users or between non-Twitter users and segregated Twitter users. Detailed estimates are presented in Table A13.⁴⁴

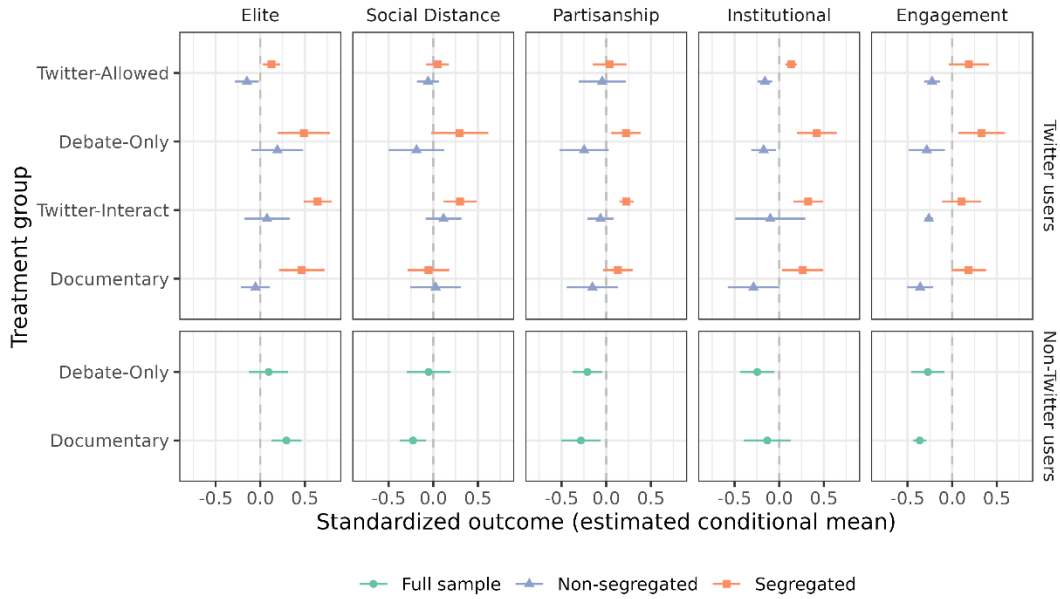
In terms of exposure, the distinction between segregated and non-segregated participants offers a clearer understanding of how online environments shape treatment effects. In particular, Figure 5 shows that, for the non-segregated Twitter sample, neither the Debate-Only group nor the Twitter-Interact group has *Combined Index* values that differ significantly from their Twitter-Allowed benchmark group. This pattern does not hold for the segregated sample. Here, both the Debate-Only and Twitter-Interact groups display positive, large, and statistically significant differences relative to their respective Twitter-Allowed benchmark (significant at the 5% level for Debate-Only and at the 1% level for Twitter-Interact).⁴⁵ For $\Delta Cortisol$, the difference between the Twitter-Interact and Twitter-Allowed groups (the only significant effect observed in the full sample) remains statistically significant at the 5% level, but only within the segregated sample. Note that the absence of an effect for $\Delta Cortisol$ in the non-segregated Twitter-Interact group makes it less likely that the higher cortisol levels are driven simply by greater social media activity, as non-segregated users were also required to interact with the experiment's active account. Taken together, these results suggest that backfire effects are confined to the segregated sample. Detailed estimates are presented in Table A14.

For completeness, Figure 6 replicates Figure 4, but splits the Twitter sample into segregated and non-segregated participants.

⁴⁴ The same patterns hold when the comparisons are conducted using the Debate-Only sample. See Table A13 for details.

⁴⁵ Approximately 68% of segregated participants assigned to the Twitter-Interact treatment followed at least one user retweeted by our Active Twitter account whom they would view as pro-attitudinal. In contrast, only about 12% of segregated participants followed retweeted accounts that could be considered counter-attitudinal (see Figure A7 for details). This pattern suggests that the effects observed for the Twitter-Interact segregated sample cannot be attributed to increased exposure to congruent, in-group content, since most of these participants were already following pro-attitudinal accounts prior to treatment.

Figure 6. Individual polarization indexes by treatment group and segregation status



Notes. The figure presents estimated conditional means for each treatment group and each outcome variable with the Twitter sample disaggregated into segregated and non-segregated participants (see Section 2.b.2), as obtained from Eq. 2. All variables are standardized so that the distribution for the Twitter-Allowed group has a mean of 0 and a standard deviation of 1. Error bars represent 90% confidence intervals, with standard errors clustered at the classroom level. See Section 2.c for variable definitions and Table A15 for detailed estimates.

Regarding selection, we find no significant differences between non-segregated Twitter users and non-Twitter participants when comparing the Documentary groups. The only exception is the *Elite* index, where the expected direction is reversed: non-segregated Twitter users score lower than their non-Twitter counterparts (a difference that is statistically significant at the 1% level). By contrast, all individual indexes for segregated Twitter users exceed those of the non-Twitter group, with three reaching statistical significance — *Partisanship* and *Institutional* at the 5% level, and *Engagement* at the 1% level. After accounting for multiple comparisons across outcomes using the Holm procedure, the results remain largely unchanged. The only differences introduced by this correction are that the decline in the *Elite* index among non-segregated users becomes significant at the 5% level, and the increase in the *Institutional* index for the segregated group is now significant at the 10% level. Overall, these findings indicate that the selection differences in polarization between recruited Twitter and non-Twitter users are concentrated among segregated participants, whereas non-segregated users exhibit polarization levels comparable to those of non-Twitter participants. Detailed estimates are reported in Table A16.⁴⁶

⁴⁶ The same patterns hold when the comparisons are conducted using the Debate-Only sample. See Table A16 for details.

Turning to exposure, the non-segregated sample again shows little evidence of systematic differences across treatment groups. The only exception is a marginally significant backfire effect in the *Elite* index for the Debate-Only condition, at the 10% level. By contrast, the segregated sample exhibits a stronger and more consistent pattern, with all statistically significant effects pointing in the backfire direction. Relative to the Twitter-Allowed benchmark, differences in the *Elite* index are highly significant, reaching the 1% level for Twitter-Interact and the 5% level for both Debate-Only and Documentary. Additional significant effects also emerge among segregated users: *Social Distance* is significant at the 10% level for Twitter-Interact, while *Institutional* is significant at the 5% level for Debate-Only and at the 10% level for Twitter-Interact.⁴⁷ After accounting for multiple comparisons, however, these results weaken substantially: only the backfire effect in the *Elite* index for the segregated Twitter-Interact group remains statistically significant at the 1% level. Detailed estimates are presented in Table A17.

Before concluding this section, we address two potential concerns regarding our measure of online segregation. First, recall that this measure is a composite index that combines a raw segregation component with a politicization component (Section 2.b.2). This raises the possibility that only one of these two dimensions is driving the results. In Appendix D.1, we show that when each component is considered separately, by constructing alternative segregation classifications based solely on raw segregation or solely on politicization, the heterogeneous effects documented above do not replicate. Under these alternative definitions, backfire patterns either disappear or arise among groups classified as non-segregated, indicating that none of the segregation components on its own captures the patterns observed with the composite index. These findings support the interpretation that the combination of ideological similarity and political interest captured by the composite measure is what drives the heterogeneity in our main results.

Second, a related concern is that the heterogeneous effects attributed to segregation may instead reflect other characteristics correlated with it. If so, the observed treatment heterogeneity could simply capture differences along those other dimensions rather than the role of segregation itself. To assess this possibility, we follow a two-step approach, described in detail in Appendix D.3. First, we train a random forest classifier to identify the pre-treatment variables most predictive of segregation status. We then re-estimate our main treatment-effect regressions by introducing, one at a time, interaction terms between each treatment indicator and the five variables most predictive of segregation (age, interest in politics, Instagram use, beliefs about the role of luck in poverty, and the importance attributed to the welfare state). Across all specifications, the backfire effects remain concentrated among segregated participants and largely absent among non-segregated participants. This evidence suggests that the patterns documented in this section are unlikely to be driven by

⁴⁷ Observing effects on *Social Distance* aligns with Voelkel et al. (2023), who report that affective polarization is indeed malleable to experimental treatments (in their case, a set of interventions aimed at reducing it).

demographic, attitudinal, or behavioral factors correlated with segregation, and instead point to online segregation itself as the main dimension shaping individuals' responses to the interventions.

3.c. Additional results on Twitter actions

So far, our results indicate that most of the observed effects are driven by segregated Twitter users, while non-segregated users display polarization levels comparable to non-Twitter participants and show no evidence of either depolarizing effects or backfire effects. To complement this analysis, we now take advantage of the fact that, within our Twitter sample, we collected data on participants' Twitter activity before, during, and after the debate. This allows us to examine how segregated and non-segregated participants across treatment arms used Twitter over time.⁴⁸

Twitter actions during the debate

A key feature of our design is its alignment with the timing of the presidential debate, given the expectation that such a political shock would increase the salience of Twitter. Data from the Twitter-Allowed group confirm this, as their activity rose sharply around the debate. In the three days prior to the debate, average active engagement during the debate-time window was 27.8% for segregated subjects in the Twitter-Allowed condition and 26.5% for their non-segregated counterparts. During the debate itself, engagement increased markedly, reaching 70% among segregated and 63.3% among non-segregated Twitter-Allowed participants. Beyond the intensity of participants' Twitter activity, the nature of their engagement also shifted toward politics. Interest in political content increased during the debate, as their Twitter interactions were disproportionately directed toward political accounts. In the three days prior to the debate, the likelihood that segregated subjects in the Twitter-Allowed condition had at least one interaction with actors included in the Argentine political landscape network was 21.7%, compared to 9.5% for non-segregated users. During the debate, these figures rose sharply, reaching 63.3% and 57.1% for segregated and non-segregated users, respectively.⁴⁹

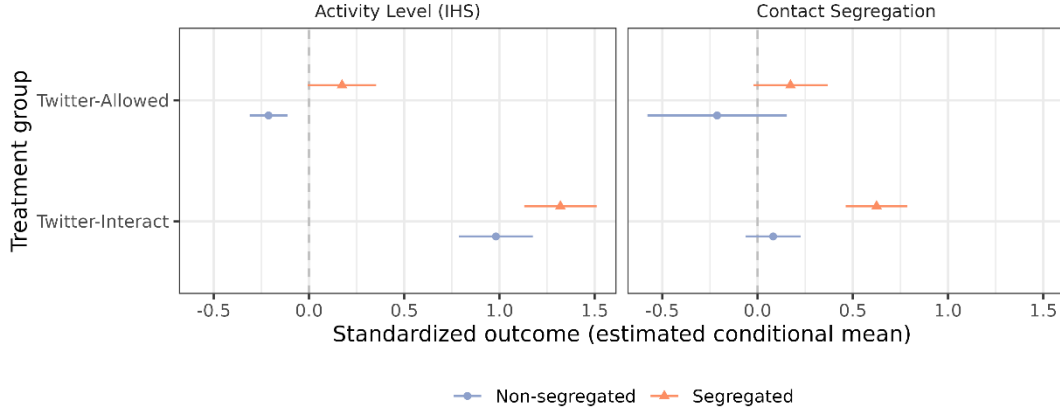
Building on this evidence of heightened engagement, we next compare the behavior of segregated and non-segregated participants across treatment arms to examine how they used Twitter

⁴⁸ For the sake of parsimony and exposition, in this section we concentrate on the analysis considering the segregated and non-segregated split, leaving out the analysis of the full sample of participants as a whole. Results for the full sample are available upon request. In addition, because non-Twitter users do not exhibit any activity on these outcomes, they cannot be considered for the analysis presented here. Finally, recall from Section 2.c that the Twitter-related outcomes reported here were not part of our original pre-registered set and, thus, should be interpreted as complementary suggestive evidence.

⁴⁹ These data also help address a relevant concern when introducing a treatment designed to mirror the natural use of Twitter, namely the possibility that participants would engage only minimally with Twitter and instead turn to other social media platforms such as WhatsApp, Facebook, or Instagram.

during the debate. Figure 7 reports, by segregation status, *Activity Level* and *Contact Segregation* for those groups allowed to use Twitter during the debate.⁵⁰

Figure 7. Twitter activity and contact segregation by treatment group and segregation status



Notes. The figure reports estimated conditional means for those groups allowed to use Twitter during the debate and each outcome variable, disaggregated by segregation status (see Section 2.b.2), as obtained from Eq. 2. All variables are standardized so that the distribution for the Twitter-Allowed group has a mean of 0 and a standard deviation of 1. Error bars represent 90% confidence intervals, with standard errors clustered at the classroom level. See Section 2.c for variable definitions and Table A18 for detailed estimates.

Figure 7 shows that, for both the non-segregated and segregated samples, *Activity Level* estimates are higher in the Twitter-Interact group compared to the Twitter-Allowed group. However, since the Twitter-Interact treatment required participants to interact during the debate with at least some content from our active Twitter account, these increases must be interpreted as compliance with the experimental design.

The pattern differs for *Contact Segregation*, where statistically significant effects appear only among segregated users. In this group, the Twitter-Interact treatment led participants to concentrate their interactions even more within politically homogeneous networks, whereas non-segregated users in the Twitter-Interact group did not exhibit such behavior. Overall, the evidence suggests that while both segregated and non-segregated Twitter-Interact participants complied with the treatment, only

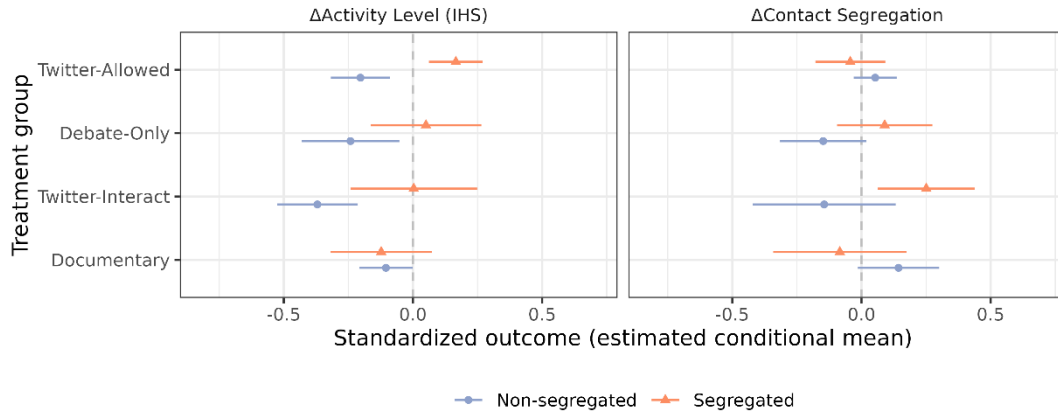
⁵⁰ Appendix D.2 further decomposes *Contact Segregation* into its two underlying components, *raw activity segregation* and *activity politicization*, and shows that the increase in *Contact Segregation* observed for segregated participants in the Twitter-Interact group is driven primarily by a rise in the ideological concentration of their interactions rather than solely by a greater share of activity directed toward political accounts. See Table A21 for details.

segregated users intensified their echo-chamber engagement. Detailed estimates are presented in Table A19.⁵¹

Post-debate Twitter dynamics

We now turn to examining whether the treatment affected how participants used Twitter after the experiment. To this end, Figure 8 presents results for Δ Activity Level and Δ Contact Segregation, which jointly capture quantitative and qualitative changes in Twitter behavior on the day following the debate relative to the preceding day.⁵²

Figure 8. Δ Activity Level and Δ Contact Segregation by treatment group and segregation status



Notes. The figure reports estimated conditional means for each treatment group and outcome variable, disaggregated by segregation status (see Section 2.b.2), as obtained from Eq. 2. Δ Activity Level and Δ Contact Segregation are defined as the value of each measure on the day after the debate minus its value on the day before the debate (following day minus preceding day). All variables are standardized so that the distribution for the Twitter-Allowed group has a mean of 0 and a standard deviation of 1. Error bars represent 90% confidence intervals, with standard errors clustered at the classroom level. See Section 2.c for variable definitions and Table A18 for detailed estimates.

⁵¹ The Twitter API provides the timestamp of when a liked tweet was originally posted, but not when the like itself occurred. This limitation makes it difficult to assign likes precisely to short observation windows, since some likes may have been recorded for tweets posted before the debate but actually generated afterward. As a robustness check, we re-estimated the results in this section excluding likes from the activity measures. The results still indicate higher contact segregation in the Twitter-Interact condition, although the difference between segregated and non-segregated users becomes less pronounced.

⁵² Appendix D.2 further examines Δ Raw Activity Segregation and Δ Activity Politicization, the two component changes underlying variation in Δ Contact Segregation. The results show that post-debate changes in Δ Contact Segregation are driven primarily by changes in the ideological concentration of participants' interactions (that is, Δ Raw Activity Segregation), rather than by changes in the overall share of activity directed toward political accounts.

Focusing on Δ *Activity Level*, we find that both segregated and non-segregated participants in the Debate-Only and Twitter-Interact groups exhibited lower post-debate activity relative to the corresponding Twitter-Allowed benchmark group, although these differences were not statistically significant. The only exception appears in the Documentary condition, where segregated participants exhibit a statistically significant decline in activity compared with their benchmark group.

Turning to Δ *Contact Segregation*, the results vary markedly across groups. In the Debate-Only condition, non-segregated participants show a statistically significant decline in Δ *Contact Segregation* relative to Twitter-Allowed participants — significant at the 5% level and consistent with a depolarizing effect. Among segregated participants, by contrast, the estimates move in the opposite direction, pointing to an increase, although the effect is not statistically significant. In the Twitter-Interact condition, segregated participants exhibit a statistically significant increase in Δ *Contact Segregation* relative to Twitter-Allowed participants, also significant at the 5% level, indicating a clear backfire effect. No similar pattern appears among non-segregated users.⁵³ Detailed estimates are presented in Table A19.

Overall, the results presented in this subsection suggest that behavioral backfire dynamics were concentrated among participants embedded in political echo chambers. Segregated users tended to reinforce their insular engagement on Twitter during and after the debate, while non-segregated users exhibited no evidence of backfire and even, in some cases, signs of depolarization — as reflected in the observed reduction in Δ *Contact Segregation*.

3.d. Summary of results

Our findings reveal a consistent pattern. The polarization differences between recruited Twitter users and non-users are concentrated among individuals in more segregated online environments. Moreover, the interventions designed to mitigate polarization often backfire among segregated users, increasing rather than reducing polarization. This pattern appears under both the

⁵³ One valid concern regarding this result is that part of the observed effect on Δ *Contact Segregation* may be driven by Twitter-Interact participants' interactions with accounts to which they were exposed during the debate through the experiment's active Twitter account. This raises the question of whether the effect is merely a mechanical consequence of such exposure, or whether users also began interacting differently with accounts they had already been following but that were not retweeted by our active account. To assess this possibility, we re-estimated these results after removing all post-treatment interactions from Twitter-Interact participants with accounts retweeted by the active account but not followed by them prior to the experiment, since these are the accounts they likely would not have encountered otherwise. All findings remain unchanged, suggesting that the reported effects are not merely a mechanical artifact. Detailed results are available upon request.

Debate-Only and Twitter-Interact treatments. In addition, the Twitter-Interact treatment increases stress among segregated users.

Table 2 summarizes our main findings, distinguishing between selection and exposure patterns, and between the results observed for segregated and non-segregated users.

Table 2. Summary of main findings on selection, exposure, and segregation

Grouping	Focus	Main finding(s)	Segregated users	Non-segregated users
Full sample	<i>Selection</i>	Selection comparisons indicate higher polarization among Twitter participants than among non-Twitter participants (higher <i>Combined Index</i>). No difference in Δ <i>Cortisol</i> .	—	—
	<i>Exposure</i>	Debate-Only and Twitter-Interact groups exhibit mild but statistically insignificant backfire effects in <i>Combined Index</i> . Positive effect on Δ <i>Cortisol</i> for Twitter-Interact.	—	—
Considering online segregation	<i>Selection</i>	Selection differences in polarization between Twitter and non-Twitter participants are concentrated among segregated participants.	Higher baseline <i>Combined Index</i> relative to non-Twitter users. No differences in Δ <i>Cortisol</i> .	Comparable baseline <i>Combined Index</i> relative to non-Twitter users. No differences in Δ <i>Cortisol</i> .
	<i>Exposure</i>	Backfire effects are concentrated among segregated users, with none among non-segregated ones.	Debate-Only and Twitter-Interact increase <i>Combined Index</i> . Twitter-Interact also raises Δ <i>Cortisol</i> .	No significant effects.
	<i>Twitter activity (during and after debate)</i>	Segregation shapes how exposure affects engagement during and after the debate.	During: Higher interaction levels in Twitter-Interact (reflecting compliance), with interactions becoming more concentrated in ideologically aligned networks. After: No clear change in overall activity in the main exposure treatments, but higher <i>Contact Segregation</i> in Twitter-Interact.	During: Higher interaction levels in Twitter-Interact (reflecting compliance), without a corresponding increase in ideological concentration. After: No clear change in overall activity, with some evidence of depolarizing adjustment in Debate-Only (lower <i>Contact Segregation</i>).

4. Discussion and conclusion

There is a growing body of work showing that many individuals inhabit online partisan bubbles, receiving limited exposure to opposing views. This has raised concerns that social media may foster polarization and has led researchers to examine two main policy responses: i) reducing social media use, and ii) increasing exposure to counter-attitudinal content. In this paper, we present experimental evidence that evaluates the effectiveness of both strategies. In addition, by recruiting

both Twitter and non-Twitter participants, we provide evidence on how self-selection into Twitter relates to observed polarization patterns.

Our pre-registered results show two main patterns. First, on the selection margin, Twitter users in our sample are more polarized than non-Twitter participants, while cortisol responses do not differ across the samples. This indicates that Twitter participants in our sample exhibit stronger partisan views than non-Twitter participants, with no apparent differences in underlying physiological stress. Second, on the exposure margin, neither withholding access to Twitter nor encouraging interaction with diverse content reduces polarization among Twitter users. If anything, both interventions increase average polarization relative to unrestricted use. In parallel, the cross-partisan exposure condition (Twitter-Interact) significantly raises physiological stress, suggesting that forced engagement with ideologically mixed content during a high-salience event may generate discomfort rather than moderation, a pattern consistent with short-run backfire effects.

Motivated by the fact that echo chambers are often blamed for social-media-driven polarization even though active social media use does not necessarily place a user within an echo chamber, we complemented the pre-registered analysis with an exploration of online segregation within the Twitter sample. We find that the reported patterns are driven almost entirely by users embedded in politically homogeneous environments (echo chambers). On the selection margin, segregated Twitter users in our sample are markedly more polarized than non-Twitter participants, while non-segregated Twitter users are statistically indistinguishable from them. On the exposure margin, both withholding access to Twitter and encouraging diverse interactions generate clear backfire effects among segregated users. Relative to unrestricted Twitter use, both interventions increase polarization, and the cross-partisan exposure condition also produces a significant rise in cortisol. Behavioral evidence from Twitter use reinforces this interpretation: during the debate, only segregated users assigned to the Twitter-Interact treatment become more contact-segregated, and after the debate they continue to concentrate their activity more strongly within like-minded networks, whereas non-segregated users do not.

Although identifying mechanisms is beyond the scope of this paper, we offer a plausible interpretation of these results. Recent work argues that social media can provide highly partisan users with a sense of purpose, community, and self-worth, which fosters status-seeking and gang-like behavior (Bail, 2021). Our interpretation is consistent with this view. If segregated users rely on their online environments for a steady stream of pro-attitudinal reinforcement, then altering that balance can unsettle their partisan identities. By altering the mix of pro-attitudinal and counter-attitudinal content that social media users typically receive, especially those in segregated environments, both the Debate-Only and Twitter-Interact treatments heighten the perceived divide between “us” and “them” relative to the natural Twitter-Allowed condition. This heightened contrast provides a plausible explanation for the backfire effects we observe.

In the Debate-Only treatment, shutting off participants' cellphones removes access to the positive, morally framed reinforcement about the ingroup and the negative, morally framed criticism of the outgroup that typically circulates within their echo chambers (Brady et al., 2017, and Rathje, Van Bavel, and van der Linden, 2021). Without these cues, segregated users cannot rely on their usual stream of biased interpretations and must instead evaluate the candidates at face value. Since the ingroup candidate typically does not perform as positively as portrayed within their partisan bubbles, and the outgroup candidate typically does not perform as negatively, this setting confronts segregated users with information that challenges their partisan identities, increasing polarization, particularly toward party elites — who become the primary source of counter-attitudinal content in the Debate-Only condition.

In the Twitter-Interact treatment, even though segregated participants continue to receive pro-attitudinal content from their echo chambers, the unexpected encounter with conflicting counter-attitudinal posts disrupts the narratives that ordinarily dominate their feeds. This rebalancing of pro- and counter-attitudinal content during the debate, delivered not only by party elites but also by other Twitter users, directly threatens their political identities. As before, this threat may induce them to defend those identities, resulting in higher polarization levels. Notably, unlike in the Debate-Only group, in the Twitter-Interact condition the challenge comes from peers as well, which may help explain the increase in social distance that we estimate for segregated users.⁵⁴

By contrast, and consistent with our interpretation, non-segregated users have not developed a strong sense of purpose or community around a political echo chamber. As a result, neither the Debate-Only nor the Twitter-Interact treatment is likely to place them in threatening conditions, and we therefore do not expect to observe polarizing effects for this group.

The size of the backfire effects we report aligns with those found in related studies such as Bail et al. (2018) and Yang, Qureshi, and Zaman (2020).⁵⁵ When compared to the depolarizing effects documented in Allcott et al. (2020) and Levy (2021), our estimates are nearly twice as large as those in the former, and our effects on social distance exceed those in the latter by an order of magnitude. It

⁵⁴ Our results suggest that providing counter-attitudinal content is more stressful for segregated users than simply muting their echo chambers. In the Twitter-Interact group we observe higher affective polarization, larger increases in cortisol, and more segregated contact patterns. An alternative interpretation is that the disparity between the two treatments reflects differences in the effective “dose” of counter-attitudinal information. The Debate-Only treatment already exposes segregated viewers to counter-attitudinal cues from the outgroup presidential candidate, so Twitter-Interact may be understood as qualitatively similar but substantially higher in intensity. This interpretation connects naturally to Lord et al. (1979), as both the Debate-Only and Twitter-Interact treatments confront participants with a stronger dose of conflicting information, creating greater scope for the kind of biased evidence evaluation that generated polarization in their classic study (see also Taber and Lodge, 2006).

⁵⁵ Our finding that both muting and diversifying social-media feeds exacerbate polarization among segregated users aligns with theoretical evidence that higher information precision can amplify ideological dispersion (Liu, 2024).

is important to emphasize that all of these papers, including our own, examine short-term impacts on political polarization. To provide context, Boxell, Gentzkow, and Shapiro (2020) document a long-run increase of 1.08 standard deviations in affective polarization in the United States over the last four decades. Against this benchmark, the effect sizes we estimate for segregated users, namely 0.426 standard deviations in the Debate-Only condition and 0.372 standard deviations in the Twitter-Interact condition, are substantial. That said, comparisons between short-run experimental effects and long-run historical trends should be made with caution. In particular, it remains unclear how interventions of the sort studied here would affect polarization over longer horizons, especially since our experiment took place during a period of exceptionally high electoral salience, which prior evidence suggests can intensify political polarization (Hernández et al., 2021). This suggests that the effects we estimate may be smaller in less politically charged settings. Even so, previous work shows that elevated polarization can lead voters to select candidates of lower average quality (Duell and Valasek, 2019), underscoring that even short-lived increases in polarization, particularly in the immediate run-up to an election, may carry important societal consequences.

In summary, political polarization is a complex and multifaceted phenomenon. Our results shed light on how social media environments shape short-run political attitudes in the context of a politically salient event. By showing that two widely discussed mitigation strategies can produce heterogeneous and sometimes counterproductive effects, this study contributes to the broader research agenda on the relationship between digital environments and political behavior.

References

- Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. 2023. “When Should You Adjust Standard Errors for Clustering?” *The Quarterly Journal of Economics* 138 (1): 1–35. <https://doi.org/10.1093/qje/qjac038>.
- Adamic, Lada A, and Natalie Glance. 2005. “The Political Blogosphere and the 2004 U.S. Election: Divided They Blog.” In *Proceedings of the 3rd International Workshop on Link Discovery*, 36–43. LinkKDD ’05. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/1134271.1134277>.
- Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. “The Welfare Effects of Social Media.” *American Economic Review* 110 (3): 629–76. <https://doi.org/10.1257/aer.20190658>.
- Allcott, Hunt, Matthew Gentzkow, Winter Mason, Arjun Wilkins, Pablo Barberá, Taylor Brown, Juan Carlos Cisneros, et al. 2024. “The Effects of Facebook and Instagram on the 2020 Election: A Deactivation Experiment.” *Proceedings of the National Academy of Sciences* 121 (21): e2321584121. <https://doi.org/10.1073/pnas.2321584121>.
- American Psychological Association. 2020. “2020 Presidential Election a Source of Significant Stress for More Americans than 2016 Presidential Race.”
- Angrist, Joshua, Pierre Azoulay, Glenn Ellison, Ryan Hill, and Susan Feng Lu. 2020. “Inside Job or Deep Impact? Extramural Citations and the Influence of Economic Scholarship.” *Journal of Economic Literature* 58 (1): 3–52. <https://doi.org/10.1257/jel.20181508>.
- Arceneaux, Kevin, Can Zengin, and Jonathan M Ladd. 2026. “Stop Blaming Social Media for Everything: The Minimal Effects of Facebook in the 2024 U.S. Presidential Election.” https://www.jonathanmladd.com/uploads/5/3/6/6/5366295/facebook_field_experiment_2024_u_s_.pdf.
- Aruguete, Natalia, and Ernesto Calvo. 2018. “Redes Sociais e Política Na Argentina.” *Revista Eletrônica de Estudos Integrados Em Discurso e Argumentação* 1 (16): 299–313. <https://doi.org/10.17648/eidea-16-2224>.
- Bail, Christopher A. 2021. *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing*. Princeton University Press. <https://press.princeton.edu/books/hardcover/9780691203423/breaking-the-social-media-prism>.
- Bail, Christopher A, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, M B 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): 9216–21. <https://doi.org/10.1073/pnas.1804840115>.
- Barberá, Pablo. 2015. “Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data.” *Political Analysis* 23 (1): 76–91. <https://doi.org/10.1093/pan/mpu011>.
- Beknazar-Yuzbashev, George, and Mateusz Stalinski. 2022. “Do Social Media Ads Matter for Political Behavior? A Field Experiment.” *Journal of Public Economics* 214: 104735. <https://doi.org/10.1016/j.jpubeco.2022.104735>.
- Bidwell, Kelly, Katherine Casey, and Rachel Glennerster. 2020. “Debates: Voting and Expenditure Responses to Political Communication.” *Journal of Political Economy* 128 (8): 2880–2924. <https://doi.org/10.1086/706862>.
- Boxell, Levi, Matthew Gentzkow, and Jesse M Shapiro. 2022. “Cross-Country Trends in Affective

- Polarization.” *The Review of Economics and Statistics*, January, 1–60. https://doi.org/10.1162/rest_a_01160.
- Brady, William J, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. 2017. “Emotion Shapes the Diffusion of Moralized Content in Social Networks.” *Proceedings of the National Academy of Sciences* 114 (28): 7313–18. <https://doi.org/10.1073/pnas.1618923114>.
- Braghieri, Luca, Sarah Eichmeyer, Ro’ee Levy, Markus M Mobius, Jacob Steinhardt, and Ruiqi Zhong. 2024. “Article-Level Slant and Polarization of News Consumption on Social Media.” *Available at SSRN*.
- Burbidge, John B, Lonnie Magee, and A Leslie Robb. 1988. “Alternative Transformations to Handle Extreme Values of the Dependent Variable.” *Journal of the American Statistical Association* 83 (401): 123–27. <https://doi.org/10.1080/01621459.1988.10478575>.
- Carney, Dana R, Amy J C Cuddy, and Andy J Yap. 2010. “Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance.” *Psychological Science* 21 (10): 1363–68. <https://doi.org/10.1177/0956797610383437>.
- Cicchini, Tomas, Sofia Morena del Pozo, Enzo Tagliazucchi, and Pablo Balenzuela. 2022. “News Sharing on Twitter Reveals Emergent Fragmentation of Media Agenda and Persistent Polarization.” *EPJ Data Science* 11 (1): 48. <https://doi.org/10.1140/epjds/s13688-022-00360-8>.
- Cinelli, Matteo, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. “The Echo Chamber Effect on Social Media.” *Proceedings of the National Academy of Sciences* 118 (9). <https://doi.org/10.1073/pnas.2023301118>.
- del Pozo, Sofia M, Sebastián Pinto, Matteo Serafino, Lucio Garcia, Hernán A Makse, and Pablo Balenzuela. 2024. “Analyzing User Ideologies and Shared News during the 2019 Argentinian Elections.” *EPJ Data Science* 13 (1): 54. <https://doi.org/10.1140/epjds/s13688-024-00493-y>.
- Di Tella, Rafael, Lucía Freira, Ramiro H. Gálvez, Ernesto Schargrodsky, Diego Shalom, and Mariano Sigman. 2019. “Crime and Violence: Desensitization in Victims to Watching Criminal Events.” *Journal of Economic Behavior and Organization* 159: 613–25. <https://doi.org/10.1016/j.jebo.2017.10.005>.
- Dickerson, Sally S, and Margaret E Kemeny. 2004. “Acute Stressors and Cortisol Responses: A Theoretical Integration and Synthesis of Laboratory Research.” *Psychological Bulletin*.
- Ditzen, Beate, Marcel Schaer, Barbara Gabriel, Guy Bodenmann, Ulrike Ehlert, and Markus Heinrichs. 2009. “Intranasal Oxytocin Increases Positive Communication and Reduces Cortisol Levels During Couple Conflict.” *Biological Psychiatry* 65 (9): 728–31. <https://doi.org/10.1016/j.biopsych.2008.10.011>.
- Druckman, James N, Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan. 2021. “Affective Polarization, Local Contexts and Public Opinion in America.” *Nature Human Behaviour* 5 (1): 28–38. <https://doi.org/10.1038/s41562-020-01012-5>.
- Druckman, James N, and Matthew S Levendusky. 2019. “What Do We Measure When We Measure Affective Polarization?” *Public Opinion Quarterly* 83 (1): 114–22. <https://doi.org/10.1093/poq/nfz003>.
- Duell, Dominik, and Justin Valasek. 2019. “Political Polarization and Selection in Representative Democracies.” *Journal of Economic Behavior & Organization* 168: 132–65. <https://doi.org/10.1016/j.jebo.2019.10.004>.
- Eady, Gregory, Jonathan Nagler, Andy Guess, Jan Zilinsky, and Joshua A Tucker. 2019. “How Many People Live in Political Bubbles on Social Media? Evidence From Linked Survey and Twitter Data.” *Sage Open* 9 (1): 2158244019832705. <https://doi.org/10.1177/2158244019832705>.
- Edwards, George C, III. 2001. “Why Not the Best? The Loyalty-Competence Trade-off in Presidential

- Appointments.” *The Brookings Review* 19 (2): 12–16. <https://doi.org/10.2307/20080969>.
- Egorov, Georgy, and Konstantin Sonin. 2011. “Dictators and Their Viziers: Endogenizing the Loyalty–Competence Trade-Off.” *Journal of the European Economic Association* 9 (5): 903–30. <https://doi.org/10.1111/j.1542-4774.2011.01033.x>.
- Gidron, Noam, James Adams, and Will Horne. 2020. *American Affective Polarization in Comparative Perspective*. Elements in American Politics. Cambridge University Press. <https://doi.org/10.1017/9781108914123>.
- Graham, Matthew H, and Milan W Svoblik. 2020. “Democracy in America? Partisanship, Polarization, and the Robustness of Support for Democracy in the United States.” *American Political Science Review* 114 (2): 392–409. <https://doi.org/10.1017/S0003055420000052>.
- Guess, Andrew M, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, et al. 2023a. “How Do Social Media Feed Algorithms Affect Attitudes and Behavior in an Election Campaign?” *Science* 381 (6656): 398–404. <https://doi.org/10.1126/science.abp9364>.
- . 2023b. “Reshares on Social Media Amplify Political News but Do Not Detectably Affect Beliefs or Opinions.” *Science* 381 (6656): 404–8. <https://doi.org/10.1126/science.add8424>.
- Guriey, Sergei, Nikita Melnikov, and Ekaterina Zhuravskaya. 2021. “3G Internet and Confidence in Government.” *The Quarterly Journal of Economics* 136 (4): 2533–2613. <https://doi.org/10.1093/qje/qjaa040>.
- Halberstam, Yosh, and Brian Knight. 2016. “Homophily, Group Size, and the Diffusion of Political Information in Social Networks: Evidence from Twitter.” *Journal of Public Economics* 143: 73–88. <https://doi.org/10.1016/j.jpubeco.2016.08.011>.
- Hastie, T, R Tibshirani, and J H Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer Series in Statistics. New York, NY, USA: Springer.
- Haushofer, Johannes, Sara Lowes, Abednego Musau, David Ndeti, Nathan Nunn, Moritz Poll, and Nancy Qian. 2022. “Stress, Ethnicity, and Prosocial Behavior.” *Journal of Political Economy Microeconomics* 1 (2): 225–69. <https://doi.org/10.1086/722367>.
- Hernández, Enrique, Eva Anduiza, and Guillem Rico. 2021. “Affective Polarization and the Salience of Elections.” *Electoral Studies* 69: 102203. <https://doi.org/10.1016/j.electstud.2020.102203>.
- Hetherington, Marc J, Deborah Jordan Brooks, John G Geer, and Martin P Wattenberg. 2008. “Turned Off or Turned On? How Polarization Affects Political Engagement.” In *Red and Blue Nation?*, edited by Pietro S Nivola and David W Brady, 1–54. Consequences and Correction of America’s Polarized Politics. Brookings Institution Press. <http://www.jstor.org/stable/10.7864/j.ctt6wpdq.4>.
- Infobae. 2019. “El Primer Debate Presidencial 2019 Completo.” *Infobae*, 2019. <https://www.infobae.com/politica/2019/10/14/el-primer-debate-presidencial-2019-completo/>.
- Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. 2019. “The Origins and Consequences of Affective Polarization in the United States.” *Annual Review of Political Science* 22 (1): 129–46. <https://doi.org/10.1146/annurev-polisci-051117-073034>.
- King, Gary, and Margaret E Roberts. 2015. “How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It.” *Political Analysis* 23 (2): 159–179. <https://doi.org/10.1093/pan/mpu015>.
- Kirschbaum, Clemens, and Dirk H Hellhammer. 1989. “Salivary Cortisol in Psychobiological Research: An Overview.” *Neuropsychobiology*. Switzerland: Karger. <https://doi.org/10.1159/000118611>.

- Kosinski, Michal, David Stillwell, and Thore Graepel. 2013. "Private Traits and Attributes Are Predictable from Digital Records of Human Behavior." *Proceedings of the National Academy of Sciences* 110 (15): 5802–5. <https://doi.org/10.1073/pnas.1218772110>.
- 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. <https://doi.org/10.1111/ajps.12237>.
- Levy, Ro'ee. 2021. "Social Media, News Consumption, and Polarization: Evidence from a Field Experiment." *American Economic Review* 111 (3): 831–70. <https://doi.org/10.1257/aer.20191777>.
- Liu, Zanhui. 2024. "Information and Polarization." *Journal of Economic Behavior & Organization* 226: 106686. <https://doi.org/10.1016/j.jebo.2024.106686>.
- London, Ben, and Lise Getoor. 2013. "Collective Classification of Network Data." In *Data Classification: Algorithms and Applications*, edited by Charu Aggarwal. CRC Press. <https://www.taylorfrancis.com/chapters/edit/10.1201/b17320-20/collective-classification-network-data-ben-london-lise-getoor>.
- Lord, Charles G, Lee Ross, and Mark R Lepper. 1979. "Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence." *Journal of Personality and Social Psychology* 37 (11): 2098–2109. <https://doi.org/10.1037/0022-3514.37.11.2098>.
- Lupu, Noam. 2015. "Party Polarization and Mass Partisanship: A Comparative Perspective." *Political Behavior* 37 (2): 331–56. <https://doi.org/10.1007/s11109-014-9279-z>.
- Mosquera, Roberto, Mofioluwademi Odunowo, Trent McNamara, Xiongfei Guo, and Ragan Petrie. 2020. "The Economic Effects of Facebook." *Experimental Economics* 23 (2): 575–602. <https://doi.org/10.1007/s10683-019-09625-y>.
- Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106. <https://doi.org/10.1257/jep.31.2.87>.
- Nyhan, Brendan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y Chen, Hunt Allcott, et al. 2023. "Like-Minded Sources on Facebook Are Prevalent but Not Polarizing." *Nature* 620 (7972): 137–44. <https://doi.org/10.1038/s41586-023-06297-w>.
- Pettigrew, Thomas F, and Linda R Tropp. 2006. "A Meta-Analytic Test of Intergroup Contact Theory." *Journal of Personality and Social Psychology*. American Psychological Association. <https://doi.org/10.1037/0022-3514.90.5.751>.
- Pew Research Center. 2014. *Political Polarization in the American Public. How Increasing Ideological Uniformity and Partisan Antipathy Affect Politics, Compromise and Everyday Life*. Washington, D.C.: Pew Research Center. <https://www.pewresearch.org/politics/2014/06/12/political-polarization-in-the-american-public/>.
- Rathje, Steve, Jay J Van Bavel, and Sander van der Linden. 2021. "Out-Group Animosity Drives Engagement on Social Media." *Proceedings of the National Academy of Sciences* 118 (26). <https://doi.org/10.1073/pnas.2024292118>.
- Robbett, Andrea, Lily Colón, and Peter Hans Matthews. 2023. "Partisan Political Beliefs and Social Learning." *Journal of Public Economics* 220: 104834. <https://doi.org/10.1016/j.jpubeco.2023.104834>.
- Rogowski, Jon C, and Joseph L Sutherland. 2016. "How Ideology Fuels Affective Polarization." *Political Behavior* 38 (2): 485–508. <https://doi.org/10.1007/s11109-015-9323-7>.
- Schwalbe, Michael C, Geoffrey L Cohen, and Lee D Ross. 2020. "The Objectivity Illusion and Voter

- Polarization in the 2016 Presidential Election.” *Proceedings of the National Academy of Sciences* 117 (35): 21218–29. <https://doi.org/10.1073/pnas.1912301117>.
- Simons, Joseph J P, and Melanie C Green. 2016. “Divisive Topics as Social Threats.” *Communication Research* 45 (2): 165–87. <https://doi.org/10.1177/0093650216644025>.
- Sohnesen, Thomas Pave, and Niels Stender. 2017. “Is Random Forest a Superior Methodology for Predicting Poverty? An Empirical Assessment.” *Poverty & Public Policy* 9 (1): 118–33. <https://doi.org/10.1002/pop4.169>.
- Sood, Gaurav, and Shanto Iyengar. 2016. “Coming to Dislike Your Opponents: The Polarizing Impact of Political Campaigns.” *Available at SSRN*, no. 2840225. <https://doi.org/10.2139/ssrn.2840225>.
- Sunstein, Cass R. 2018. *#Republic: Divided Democracy in the Age of Social Media*. NED-New. Princeton University Press. <http://www.jstor.org/stable/j.ctv8xnhtd>.
- Suzuki, Sara, Lindsay Till Hoyt, Neshat Yazdani, Mariah Kornbluh, Elan C Hope, Melissa J Hagan, Alison K Cohen, and Parissa J Ballard. 2023. “Trajectories of Sociopolitical Stress during the 2020 United States Presidential Election Season: Associations with Psychological Well-Being, Civic Action, and Social Identities.” *Comprehensive Psychoneuroendocrinology* 16: 100218. <https://doi.org/10.1016/j.cpnec.2023.100218>.
- Taber, Charles S, and Milton Lodge. 2006. “Motivated Skepticism in the Evaluation of Political Beliefs.” *American Journal of Political Science* 50 (3): 755–69. <https://doi.org/10.1111/j.1540-5907.2006.00214.x>.
- Ventura, Tiago, Rajeshwari Majumdar, Jonathan Nagler, and Joshua Aaron Tucker. 2023. “Misinformation Exposure Beyond Traditional Feeds: Evidence from a WhatsApp Deactivation Experiment in Brazil.” *Available at SSRN*, May, 48. <https://doi.org/10.2139/ssrn.4457400>.
- Voelkel, Jan G, James Chu, Michael N Stagnaro, Joseph S Mernyk, Chrystal Redekopp, Sophia L Pink, James N Druckman, David G Rand, and Robb Willer. 2023. “Interventions Reducing Affective Polarization Do Not Necessarily Improve Anti-Democratic Attitudes.” *Nature Human Behaviour* 7 (1): 55–64. <https://doi.org/10.1038/s41562-022-01466-9>.
- Waismel-Manor, Israel, Gal Ifergane, and Hagit Cohen. 2011. “When Endocrinology and Democracy Collide: Emotions, Cortisol and Voting at National Elections.” *European Neuropsychopharmacology* 21 (11): 789–95. <https://doi.org/10.1016/j.euroneuro.2011.03.003>.
- Wirtz, Petra H, Johannes Siegrist, Ulrike Rimmele, and Ulrike Ehlert. 2008. “Higher Overcommitment to Work Is Associated with Lower Norepinephrine Secretion before and after Acute Psychosocial Stress in Men.” *Psychoneuroendocrinology* 33 (1): 92–99. <https://doi.org/10.1016/j.psyneuen.2007.10.003>.
- Yang, Qi, Khizar Qureshi, and Tauhid Zaman. 2020. “Mitigating the Backfire Effect Using Pacing and Leading.” *arXiv preprint*. <https://arxiv.org/abs/2008.00049>.
- Young, Allan H. 2004. “Cortisol in Mood Disorders.” *Stress* 7 (4): 205–8. <https://doi.org/10.1080/10253890500069189>.

Appendix (Online)

A. Index construction details

Questionnaires and materials can be found at <https://www.socialscisearch.org/trials/4850>. The calculation of individual polarization indexes is as follows:

- a. Elite (average of a1 and a2)
 - a1. Threat (Q. 33)
 - a2. Ingroup leader positive attributes – Outgroup leader positive attributes (Q. 29) *
- b. Social Distance (average of b1, b2, and b3)
 - b1. Marriage (Q. 24a)
 - b2. Working (Q. 24b)
 - b3. Socializing (Q. 24c)
- c. Partisanship (average of c1 and c2)
 - c1. Dispersion of the preferences over the presidential candidates (Q. 32)
 - c2. Strength of conviction (Q. 30)
- d. Institutional (average of d1, d2, and d3)
 - d1. Outgroup party does fraud – Ingroup party does fraud (Q. 26) *
 - d2. Court against ingroup party – Court against outgroup party (Q. 28) *
 - d3. Loyalty importance – Technical competence importance (Q. 27)
- e. Engagement (average of e1 and e2)
 - e1. Political volunteering (Q. 25a)
 - e2. Persuade others (Q. 25b)

* To construct these variables, we use predicted party preference scores obtained from pre-treatment Twitter and questionnaire data.

The different combined indexes are calculated as follows:

- Combined Index (average of a, b, c, d, and e)
- Combined Index V2 (average of a1, a2, b1, b2, b3, c1, c2, d1, d2, d3, e1, and e2)
- Combined Index V3 (average of a, b, c, and d)
- Combined Index V4 (average of a1, a2, b1, b2, b3, c1, c2, d1, d2, and d3)

All indexes are calculated as the averages of the lower-level variables or indexes. During the averaging process, each constituent is normalized to ensure that the distribution of the Twitter-Allowed group has a mean of 0 and a standard deviation of 1.

Component a2 is constructed as follows. For each leader of the *JxC* and *FdT* coalitions, respondents selected traits from a list, from which we retain the five positive traits. Each positive trait is coded as an indicator equal to one if selected and zero otherwise. We then standardize these indicators using the mean and standard deviation of the Twitter-Allowed group and average them to obtain a leader-specific positive-trait score. Finally, we compute the difference between the ingroup leader's score and the outgroup leader's score.

B. Appendix tables

Table A1. Twitter and non-Twitter sample pre-treatment characteristics

	non-Twitter sample	Twitter sample	n	p-value
Sociodemographics				
Age	42.548 (0.870)	33.293 (0.527)	688	0.000
Female = 1	0.573 (0.041)	0.521 (0.024)	688	0.284
Education years	14.021 (0.142)	14.333 (0.084)	688	0.042
Household head = 1	0.585 (0.049)	0.380 (0.025)	688	0.000
Household head education years	14.117 (0.154)	14.768 (0.104)	684	0.000
Residence = City of Buenos Aires	0.556 (0.033)	0.659 (0.023)	687	0.002
Residence = Greater Buenos Aires North	0.199 (0.032)	0.139 (0.016)	687	0.101
Residence = Greater Buenos Aires West	0.145 (0.020)	0.096 (0.010)	687	0.036
Residence = Greater Buenos Aires South	0.100 (0.020)	0.101 (0.014)	687	0.959
Household size	3.100 (0.086)	3.007 (0.060)	686	0.337
Employed = 1	0.846 (0.022)	0.782 (0.019)	685	0.025
Household head employed = 1	0.958 (0.010)	0.946 (0.012)	679	0.481
Internet and social media use				
Internet at home = 1	0.958 (0.012)	0.993 (0.004)	686	0.009
Internet at work = 1	0.862 (0.025)	0.877 (0.024)	483	0.701
Has cellphone = 1	0.992 (0.006)	0.998 (0.002)	688	0.330
Times a day the cellphone is checked	35.010 (1.208)	42.724 (0.654)	685	0.000
Regular Facebook use = 1	0.563 (0.046)	0.549 (0.018)	681	0.798
Regular Twitter use = 1	0.000 (0.000)	0.858 (0.016)	686	0.000
Regular Instagram use = 1	0.462 (0.028)	0.819 (0.017)	667	0.000
Regular Whatsapp use = 1	0.979 (0.010)	0.991 (0.004)	685	0.292
Ideology and beliefs				
Messi vs. Maradona = Messi	0.337 (0.025)	0.474 (0.024)	685	0.001
Messi vs. Maradona = Maradona	0.400 (0.026)	0.258 (0.022)	685	0.000
Messi vs. Maradona = Equally good	0.262 (0.023)	0.267 (0.022)	685	0.880
Sentence length	0.436 (0.035)	0.343 (0.023)	687	0.048
Poverty: Importance of effort	6.238 (0.198)	5.648 (0.176)	677	0.028
Poverty: Importance of luck	4.774 (0.194)	4.745 (0.122)	678	0.900
Poverty: Importance of opportunities	7.909 (0.134)	7.987 (0.062)	682	0.632
Human rights violations: USA	7.595 (0.127)	8.130 (0.073)	684	0.000
Human rights violations: Venezuela	8.040 (0.166)	8.429 (0.111)	683	0.056
Welfare state importance	7.500 (0.169)	7.316 (0.093)	685	0.276
Interest in politics	7.950 (0.174)	8.578 (0.089)	685	0.002
Adverse to risk=1	0.817 (0.025)	0.776 (0.016)	687	0.150
Twitter profile when recruited				
Number of followers (log10)	-	2.222 (0.031)	447	-
Number of friends (log10)	-	2.513 (0.018)	447	-
Number of Tweets (log10)	-	1.913 (0.035)	447	-
Account age (years)	-	6.905 (0.142)	447	-
Recruitment method				
Twitter recruited	0.000 (0.000)	0.282 (0.021)	688	0.000
Recruitment agency - online panel	0.473 (0.028)	0.237 (0.021)	688	0.000
Recruitment agency - offline	0.527 (0.028)	0.481 (0.022)	688	0.201
Cortisol				
First sample cortisol level (µg/dL)	0.150 (0.005)	0.157 (0.005)	659	0.296
Political positions from pre-treatment data				
Normalized segregation score	-	0.116 (0.005)	447	-
Segregation status = segregated	-	0.499 (0.027)	447	-
Party preference score	-	0.484 (0.021)	447	-
Party preference = JxC	0.581 (0.037)	0.481 (0.028)	688	0.041

Notes. All variables are measured prior to treatment assignment. “Regular [platform] use = 1” indicates that the participant reports using the specified social media platform more than once per day or continuously throughout the day. Political position variables are calculated as described in Section 2.b. Reported *p*-value comes from a joint hypothesis *F*-test where equality of group means is tested. Standard errors (in parentheses) are clustered at the classroom level.

Table A2. Pre-treatment characteristics across treatment arms, Twitter sample

	Twitter Allowed	Debate	Twitter Interact	Documentary	n	p-value
Sociodemographics						
Age	33.936 (0.277)	31.714 (1.307)	33.500 (1.214)	34.027 (0.811)	447	0.409
Female = 1	0.482 (0.054)	0.509 (0.040)	0.482 (0.038)	0.611 (0.039)	447	0.074
Education years	14.464 (0.108)	14.009 (0.170)	14.286 (0.175)	14.575 (0.119)	447	0.044
Household head = 1	0.427 (0.067)	0.312 (0.047)	0.402 (0.030)	0.381 (0.037)	447	0.374
Household head education years	15.028 (0.168)	14.696 (0.198)	14.714 (0.145)	14.640 (0.256)	444	0.417
Residence = City of Buenos Aires	0.727 (0.019)	0.652 (0.040)	0.649 (0.045)	0.611 (0.054)	446	0.056
Residence = Greater Buenos Aires North	0.100 (0.027)	0.134 (0.027)	0.135 (0.029)	0.186 (0.033)	446	0.257
Residence = Greater Buenos Aires West	0.073 (0.023)	0.089 (0.011)	0.090 (0.020)	0.133 (0.020)	446	0.184
Residence = Greater Buenos Aires South	0.100 (0.036)	0.107 (0.026)	0.126 (0.016)	0.071 (0.026)	446	0.343
Household size	2.891 (0.055)	3.062 (0.096)	3.187 (0.088)	2.884 (0.155)	446	0.027
Employed = 1	0.778 (0.017)	0.687 (0.045)	0.821 (0.026)	0.841 (0.028)	445	0.015
Household head employed = 1	0.963 (0.013)	0.936 (0.027)	0.955 (0.008)	0.929 (0.032)	441	0.694
Internet and social media use						
Internet at home = 1	1.000 (0.000)	0.991 (0.009)	0.991 (0.008)	0.991 (0.008)	447	0.330
Internet at work = 1	0.867 (0.031)	0.829 (0.085)	0.897 (0.026)	0.907 (0.029)	309	0.685
Has cellphone = 1	1.000 (0.000)	0.991 (0.009)	1.000 (0.000)	1.000 (0.000)	447	†
Times a day the cellphone is checked	41.295 (0.970)	43.131 (1.806)	43.862 (0.920)	42.588 (0.993)	446	0.290
Regular Facebook use = 1	0.587 (0.037)	0.569 (0.014)	0.468 (0.027)	0.571 (0.032)	441	0.006
Regular Twitter use = 1	0.836 (0.019)	0.893 (0.029)	0.856 (0.014)	0.848 (0.051)	445	0.450
Regular Instagram use = 1	0.771 (0.036)	0.838 (0.039)	0.848 (0.024)	0.818 (0.023)	442	0.326
Regular Whatsapp use = 1	0.982 (0.009)	0.991 (0.009)	1.000 (0.000)	0.991 (0.009)	445	0.118
Ideology and beliefs						
Messi vs. Maradona = Messi	0.376 (0.017)	0.482 (0.036)	0.518 (0.053)	0.518 (0.044)	445	0.001
Messi vs. Maradona = Maradona	0.312 (0.031)	0.205 (0.042)	0.295 (0.040)	0.223 (0.039)	445	0.113
Messi vs. Maradona = Equally good	0.312 (0.018)	0.312 (0.043)	0.187 (0.020)	0.259 (0.049)	445	0.000
Sentence length	0.382 (0.021)	0.297 (0.055)	0.312 (0.042)	0.381 (0.048)	446	0.282
Poverty: Importance of effort	5.890 (0.156)	5.679 (0.302)	5.882 (0.494)	5.149 (0.272)	442	0.129
Poverty: Importance of luck	4.318 (0.208)	5.045 (0.214)	4.755 (0.198)	4.856 (0.223)	443	0.095
Poverty: Importance of opportunities	8.036 (0.056)	7.884 (0.147)	8.036 (0.122)	7.991 (0.140)	445	0.802
Human rights violations: USA	8.164 (0.061)	8.135 (0.152)	8.018 (0.175)	8.204 (0.152)	446	0.861
Human rights violations: Venezuela	8.373 (0.084)	8.432 (0.105)	8.273 (0.371)	8.634 (0.180)	443	0.594
Welfare state importance	7.264 (0.216)	7.369 (0.152)	7.241 (0.163)	7.389 (0.198)	446	0.912
Interest in politics	8.573 (0.227)	8.676 (0.154)	8.514 (0.186)	8.549 (0.128)	445	0.902
Adverse to risk=1	0.745 (0.030)	0.795 (0.035)	0.768 (0.027)	0.796 (0.032)	447	0.612
Twitter profile when recruited						
Number of followers (log10)	2.179 (0.061)	2.264 (0.038)	2.196 (0.084)	2.250 (0.046)	447	0.634
Number of friends (log10)	2.465 (0.026)	2.544 (0.031)	2.463 (0.014)	2.581 (0.036)	447	0.004
Number of Tweets (log10)	1.922 (0.068)	1.996 (0.080)	1.801 (0.034)	1.933 (0.058)	447	0.042
Account age (years)	6.635 (0.343)	6.587 (0.223)	7.241 (0.154)	7.152 (0.258)	447	0.063
Recruitment method						
Twitter recruited	0.291 (0.039)	0.312 (0.045)	0.250 (0.041)	0.274 (0.041)	447	0.762
Recruitment agency - online panel	0.245 (0.029)	0.241 (0.056)	0.205 (0.008)	0.257 (0.049)	447	0.380
Recruitment agency - offline	0.464 (0.054)	0.446 (0.028)	0.545 (0.040)	0.469 (0.039)	447	0.244
Cortisol						
First sample cortisol level (µg/dL)	0.175 (0.005)	0.149 (0.009)	0.142 (0.013)	0.164 (0.006)	429	0.019
Political positions from pre-treatment data						
Normalized segregation score	0.116 (0.005)	0.118 (0.013)	0.109 (0.013)	0.120 (0.010)	447	0.937
Segregation status = segregated	0.545 (0.047)	0.437 (0.051)	0.446 (0.049)	0.566 (0.046)	447	0.129
Party preference score	0.470 (0.014)	0.473 (0.030)	0.507 (0.070)	0.484 (0.035)	447	0.941
Party preference = JxC	0.436 (0.033)	0.509 (0.044)	0.464 (0.089)	0.513 (0.035)	447	0.374

Notes. All variables are measured prior to treatment assignment. “Regular [platform] use = 1” indicates that the participant reports using the specified social media platform more than once per day or continuously throughout the day. Political position variables are calculated as described in Section 2.b. Reported *p*-value comes from a joint hypothesis *F*-test where equality of group means is tested. Standard errors (in parentheses) are clustered at the classroom level.

Table A3. Pre-treatment characteristics across treatment arms, non-Twitter sample

	Debate	Documentary	N	p-value
Sociodemographics				
Age	42.811 (1.562)	42.277 (0.740)	241	0.758
Female = 1	0.574 (0.053)	0.571 (0.064)	241	0.977
Education years	14.016 (0.171)	14.025 (0.231)	241	0.976
Household head = 1	0.525 (0.038)	0.647 (0.087)	241	0.199
Household head education years	14.190 (0.169)	14.042 (0.257)	240	0.630
Residence = City of Buenos Aires	0.516 (0.041)	0.597 (0.048)	241	0.206
Residence = Greater Buenos Aires North	0.205 (0.044)	0.193 (0.048)	241	0.858
Residence = Greater Buenos Aires West	0.164 (0.026)	0.126 (0.028)	241	0.324
Residence = Greater Buenos Aires South	0.115 (0.031)	0.084 (0.024)	241	0.435
Household size	3.016 (0.064)	3.186 (0.156)	240	0.313
Employed = 1	0.844 (0.023)	0.847 (0.037)	240	0.942
Household head employed = 1	0.967 (0.012)	0.949 (0.016)	238	0.392
Internet and social media use				
Internet at home = 1	0.959 (0.020)	0.958 (0.012)	239	0.965
Internet at work = 1	0.849 (0.035)	0.875 (0.037)	174	0.607
Has cellphone = 1	0.984 (0.010)	1.000 (0.000)	241	0.109
Times a day the cellphone is checked	35.208 (2.159)	34.811 (1.108)	239	0.870
Regular Facebook use = 1	0.492 (0.040)	0.636 (0.076)	240	0.095
Regular Twitter use = 1	0.000 (0.000)	0.000 (0.000)	241	†
Regular Instagram use = 1	0.469 (0.039)	0.455 (0.041)	225	0.809
Regular Whatsapp use = 1	0.967 (0.017)	0.992 (0.008)	240	0.189
Ideology and beliefs				
Messi vs. Maradona = Messi	0.328 (0.037)	0.347 (0.035)	240	0.702
Messi vs. Maradona = Maradona	0.410 (0.027)	0.390 (0.046)	240	0.709
Messi vs. Maradona = Equally good	0.262 (0.030)	0.263 (0.037)	240	0.993
Sentence length	0.467 (0.040)	0.403 (0.056)	241	0.356
Poverty: Importance of effort	6.492 (0.239)	5.983 (0.291)	235	0.178
Poverty: Importance of luck	4.916 (0.219)	4.629 (0.320)	235	0.460
Poverty: Importance of opportunities	7.933 (0.210)	7.886 (0.169)	237	0.861
Human rights violations: USA	7.658 (0.181)	7.530 (0.180)	238	0.614
Human rights violations: Venezuela	7.942 (0.286)	8.139 (0.160)	240	0.550
Welfare state importance	7.371 (0.257)	7.630 (0.213)	239	0.438
Interest in politics	8.033 (0.261)	7.864 (0.227)	240	0.627
Adverse to risk=1	0.802 (0.027)	0.832 (0.041)	240	0.542
Recruitment method				
Recruitment agency - online panel	0.418 (0.028)	0.529 (0.041)	241	0.024
Recruitment agency - offline	0.582 (0.028)	0.471 (0.041)	241	0.024
Cortisol				
First sample cortisol level ($\mu\text{g/dL}$)	0.151 (0.008)	0.148 (0.007)	230	0.797
Political positions from pre-treatment data				
Party preference = JxC	0.590 (0.048)	0.571 (0.057)	241	0.801

Notes. All variables are measured prior to treatment assignment. “Regular [platform] use = 1” indicates that the participant reports using the specified social media platform more than once per day or continuously throughout the day. “Party preference = JxC” is calculated as described in Section 2.b.3. Reported p -value comes from a joint hypothesis F -test where equality of group means is tested. Standard errors (in parentheses) are clustered at the classroom level.

Table A4. Pre-treatment characteristics across treatment arms, non-segregated Twitter sample

	Twitter Allowed	Debate	Twitter Interact	Documentary	n	p-value
Sociodemographics						
Age	29.320 (0.571)	26.063 (0.993)	29.081 (1.195)	28.388 (1.199)	224	0.043
Female = 1	0.580 (0.055)	0.603 (0.052)	0.548 (0.007)	0.673 (0.061)	224	0.144
Education years	14.120 (0.177)	13.365 (0.224)	14.000 (0.097)	14.041 (0.358)	224	0.047
Household head = 1	0.380 (0.055)	0.159 (0.018)	0.339 (0.016)	0.327 (0.074)	224	0.000
Household head education years	14.720 (0.241)	14.683 (0.299)	14.629 (0.224)	14.521 (0.307)	223	0.964
Residence = City of Buenos Aires	0.740 (0.079)	0.619 (0.047)	0.656 (0.050)	0.551 (0.046)	223	0.174
Residence = Greater Buenos Aires North	0.060 (0.033)	0.175 (0.037)	0.131 (0.031)	0.224 (0.055)	223	0.033
Residence = Greater Buenos Aires West	0.100 (0.044)	0.079 (0.033)	0.066 (0.028)	0.163 (0.047)	223	0.341
Residence = Greater Buenos Aires South	0.100 (0.038)	0.111 (0.045)	0.148 (0.005)	0.061 (0.029)	223	0.016
Household size	3.020 (0.107)	3.365 (0.134)	3.323 (0.104)	2.898 (0.208)	224	0.052
Employed = 1	0.694 (0.044)	0.603 (0.052)	0.806 (0.027)	0.755 (0.056)	223	0.004
Household head employed = 1	0.959 (0.016)	0.934 (0.029)	0.951 (0.019)	0.918 (0.054)	220	0.815
Internet and social media use						
Internet at home = 1	1.000 (0.000)	0.984 (0.015)	1.000 (0.000)	1.000 (0.000)	224	†
Internet at work = 1	0.889 (0.055)	0.781 (0.093)	0.833 (0.048)	0.857 (0.069)	136	0.758
Has cellphone = 1	1.000 (0.000)	0.984 (0.014)	1.000 (0.000)	1.000 (0.000)	224	†
Times a day the cellphone is checked	42.000 (2.124)	42.540 (1.976)	44.355 (0.989)	42.857 (1.396)	223	0.639
Regular Facebook use = 1	0.633 (0.042)	0.483 (0.057)	0.443 (0.034)	0.592 (0.048)	219	0.002
Regular Twitter use = 1	0.840 (0.046)	0.857 (0.042)	0.869 (0.024)	0.776 (0.078)	223	0.688
Regular Instagram use = 1	0.920 (0.042)	0.952 (0.031)	0.919 (0.038)	0.875 (0.044)	222	0.557
Regular Whatsapp use = 1	1.000 (0.000)	0.984 (0.014)	1.000 (0.000)	0.980 (0.020)	224	†
Ideology and beliefs						
Messi vs. Maradona = Messi	0.408 (0.046)	0.587 (0.032)	0.516 (0.072)	0.592 (0.059)	223	0.011
Messi vs. Maradona = Maradona	0.306 (0.035)	0.143 (0.040)	0.258 (0.051)	0.163 (0.065)	223	0.014
Messi vs. Maradona = Equally good	0.286 (0.036)	0.270 (0.036)	0.226 (0.028)	0.245 (0.083)	223	0.578
Sentence length	0.420 (0.017)	0.302 (0.075)	0.323 (0.043)	0.286 (0.056)	224	0.017
Poverty: Importance of effort	5.551 (0.328)	5.159 (0.341)	5.754 (0.392)	4.968 (0.537)	220	0.538
Poverty: Importance of luck	4.660 (0.436)	5.270 (0.312)	4.600 (0.401)	5.574 (0.302)	220	0.157
Poverty: Importance of opportunities	8.240 (0.194)	8.222 (0.179)	7.885 (0.279)	8.354 (0.244)	222	0.630
Human rights violations: USA	8.280 (0.134)	8.161 (0.218)	7.742 (0.290)	8.367 (0.243)	223	0.342
Human rights violations: Venezuela	8.540 (0.098)	8.290 (0.106)	7.951 (0.414)	8.667 (0.259)	221	0.162
Welfare state importance	7.400 (0.543)	7.698 (0.172)	7.274 (0.233)	8.082 (0.279)	224	0.156
Interest in politics	8.120 (0.256)	8.190 (0.208)	7.902 (0.196)	8.306 (0.223)	223	0.561
Adverse to risk=1	0.760 (0.046)	0.857 (0.043)	0.790 (0.019)	0.735 (0.051)	224	0.257
Twitter profile when recruited						
Number of followers (log10)	2.291 (0.089)	2.311 (0.074)	2.259 (0.113)	2.357 (0.052)	224	0.827
Number of friends (log10)	2.453 (0.056)	2.551 (0.065)	2.361 (0.042)	2.564 (0.060)	224	0.017
Number of Tweets (log10)	1.900 (0.090)	1.966 (0.110)	1.761 (0.026)	1.852 (0.085)	224	0.127
Account age (years)	6.755 (0.492)	6.128 (0.279)	7.119 (0.238)	7.369 (0.250)	224	0.008
Recruitment method						
Twitter recruited	0.240 (0.078)	0.254 (0.028)	0.274 (0.053)	0.265 (0.058)	224	0.980
Recruitment agency - online panel	0.240 (0.072)	0.190 (0.066)	0.129 (0.031)	0.184 (0.059)	224	0.454
Recruitment agency - offline	0.520 (0.035)	0.556 (0.062)	0.597 (0.054)	0.551 (0.076)	224	0.693
Cortisol						
First sample cortisol level (µg/dL)	0.170 (0.005)	0.153 (0.008)	0.157 (0.015)	0.180 (0.014)	215	0.228
Political positions from pre-treatment data						
Normalized segregation score	0.026 (0.002)	0.024 (0.002)	0.023 (0.003)	0.022 (0.002)	224	0.756
Party preference score	0.412 (0.041)	0.450 (0.041)	0.497 (0.054)	0.399 (0.038)	224	0.458
Party preference = JxC	0.320 (0.084)	0.476 (0.070)	0.435 (0.076)	0.367 (0.055)	224	0.447

Notes. All variables are measured prior to treatment assignment. “Regular [platform] use = 1” indicates that the participant reports using the specified social media platform more than once per day or continuously throughout the day. Political position variables are calculated as described in Section 2.b. Reported *p*-value comes from a joint hypothesis *F*-test where equality of group means is tested. Standard errors (in parentheses) are clustered at the classroom level.

Table A5. Pre-treatment characteristics across treatment arms, segregated Twitter sample

	Twitter Allowed	Debate	Twitter Interact	Documentary	n	p-value
Sociodemographics						
Age	37.783 (1.254)	38.980 (2.143)	38.980 (1.515)	38.344 (1.237)	223	0.927
Female = 1	0.400 (0.064)	0.388 (0.053)	0.400 (0.070)	0.563 (0.059)	223	0.115
Education years	14.750 (0.134)	14.837 (0.222)	14.640 (0.317)	14.984 (0.110)	223	0.487
Household head = 1	0.467 (0.098)	0.510 (0.102)	0.480 (0.044)	0.422 (0.034)	223	0.678
Household head education years	15.288 (0.299)	14.714 (0.296)	14.820 (0.251)	14.730 (0.357)	221	0.494
Residence = City of Buenos Aires	0.717 (0.044)	0.694 (0.072)	0.640 (0.059)	0.656 (0.066)	223	0.728
Residence = Greater Buenos Aires North	0.133 (0.038)	0.082 (0.044)	0.140 (0.030)	0.156 (0.044)	223	0.646
Residence = Greater Buenos Aires West	0.050 (0.042)	0.102 (0.029)	0.120 (0.037)	0.109 (0.047)	223	0.624
Residence = Greater Buenos Aires South	0.100 (0.050)	0.102 (0.029)	0.100 (0.036)	0.078 (0.035)	223	0.957
Household size	2.783 (0.164)	2.673 (0.201)	3.020 (0.106)	2.873 (0.216)	222	0.383
Employed = 1	0.847 (0.014)	0.796 (0.040)	0.840 (0.049)	0.906 (0.027)	222	0.108
Household head employed = 1	0.966 (0.018)	0.939 (0.040)	0.960 (0.024)	0.938 (0.023)	221	0.776
Internet and social media use						
Internet at home = 1	1.000 (0.000)	1.000 (0.000)	0.980 (0.018)	0.984 (0.015)	223	†
Internet at work = 1	0.854 (0.038)	0.868 (0.094)	0.972 (0.029)	0.941 (0.035)	173	0.086
Has cellphone = 1	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	223	†
Times a day the cellphone is checked	40.708 (1.228)	43.878 (1.962)	43.250 (1.090)	42.383 (1.495)	223	0.383
Regular Facebook use = 1	0.550 (0.069)	0.673 (0.059)	0.500 (0.088)	0.556 (0.049)	222	0.294
Regular Twitter use = 1	0.833 (0.022)	0.939 (0.026)	0.840 (0.042)	0.905 (0.052)	222	0.017
Regular Instagram use = 1	0.644 (0.092)	0.694 (0.103)	0.760 (0.040)	0.774 (0.047)	220	0.585
Regular Whatsapp use = 1	0.966 (0.018)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	221	†
Ideology and beliefs						
Messi vs. Maradona = Messi	0.350 (0.066)	0.347 (0.066)	0.520 (0.031)	0.460 (0.050)	222	0.028
Messi vs. Maradona = Maradona	0.317 (0.070)	0.286 (0.047)	0.340 (0.046)	0.270 (0.060)	222	0.774
Messi vs. Maradona = Equally good	0.333 (0.043)	0.367 (0.080)	0.140 (0.032)	0.270 (0.041)	222	0.001
Sentence length	0.350 (0.052)	0.292 (0.062)	0.300 (0.045)	0.453 (0.059)	222	0.172
Poverty: Importance of effort	6.167 (0.085)	6.347 (0.303)	6.041 (0.652)	5.281 (0.223)	222	0.003
Poverty: Importance of luck	4.033 (0.239)	4.755 (0.233)	4.940 (0.321)	4.328 (0.287)	223	0.068
Poverty: Importance of opportunities	7.867 (0.201)	7.449 (0.165)	8.220 (0.133)	7.719 (0.322)	223	0.004
Human rights violations: USA	8.067 (0.167)	8.102 (0.179)	8.360 (0.209)	8.078 (0.240)	223	0.702
Human rights violations: Venezuela	8.233 (0.116)	8.612 (0.189)	8.673 (0.359)	8.609 (0.312)	222	0.232
Welfare state importance	7.150 (0.064)	6.938 (0.270)	7.200 (0.207)	6.859 (0.448)	222	0.787
Interest in politics	8.950 (0.212)	9.312 (0.151)	9.260 (0.108)	8.734 (0.197)	222	0.059
Adverse to risk=1	0.733 (0.036)	0.714 (0.043)	0.740 (0.042)	0.844 (0.048)	223	0.198
Twitter profile when recruited						
Number of followers (log10)	2.087 (0.055)	2.204 (0.085)	2.117 (0.085)	2.168 (0.053)	223	0.605
Number of friends (log10)	2.474 (0.014)	2.535 (0.087)	2.588 (0.030)	2.594 (0.046)	223	0.001
Number of Tweets (log10)	1.941 (0.059)	2.034 (0.079)	1.851 (0.092)	1.995 (0.074)	223	0.457
Account age (years)	6.535 (0.265)	7.178 (0.338)	7.392 (0.283)	6.986 (0.303)	223	0.155
Recruitment method						
Twitter recruited	0.333 (0.029)	0.388 (0.085)	0.220 (0.043)	0.281 (0.045)	223	0.118
Recruitment agency - online panel	0.250 (0.062)	0.306 (0.056)	0.300 (0.014)	0.312 (0.062)	223	0.872
Recruitment agency - offline	0.417 (0.075)	0.306 (0.037)	0.480 (0.053)	0.406 (0.047)	223	0.047
Cortisol						
First sample cortisol level (µg/dL)	0.178 (0.010)	0.144 (0.015)	0.122 (0.010)	0.151 (0.012)	214	0.002
Political positions from pre-treatment data						
Normalized segregation score	0.191 (0.023)	0.239 (0.017)	0.216 (0.021)	0.194 (0.009)	223	0.098
Party preference score	0.517 (0.018)	0.502 (0.032)	0.521 (0.093)	0.549 (0.049)	223	0.887
Party preference = JxC	0.533 (0.022)	0.551 (0.038)	0.500 (0.119)	0.625 (0.055)	223	0.467

Notes. All variables are measured prior to treatment assignment. “Regular [platform] use = 1” indicates that the participant reports using the specified social media platform more than once per day or continuously throughout the day. Political position variables are calculated as described in Section 2.b. Reported *p*-value comes from a joint hypothesis *F*-test where equality of group means is tested. Standard errors (in parentheses) are clustered at the classroom level.

Table A6. Combined indexes and Δ Cortisol - Detailed regression estimates, full sample

	Combined Index	Combined Index V2	Combined Index V3	Combined Index V4	Δ Cortisol
Twitter user · Treatment: Twitter-Allowed	0.000 (0.081)	0.000 (0.080)	0.000 (0.067)	0.000 (0.065)	0.000 (0.076)
Twitter user · Treatment: Debate-Only	0.156* (0.081)	0.127 (0.083)	0.188** (0.075)	0.147* (0.083)	0.152 (0.109)
Twitter user · Treatment: Twitter-Interact	0.216* (0.113)	0.212* (0.121)	0.301** (0.141)	0.285* (0.146)	0.287*** (0.086)
Twitter user · Treatment: Documentary	0.084 (0.112)	0.061 (0.116)	0.122 (0.129)	0.090 (0.131)	-0.029 (0.090)
Non-Twitter user · Treatment: Debate-Only	-0.275** (0.129)	-0.287** (0.131)	-0.208 (0.142)	-0.227 (0.145)	0.068 (0.072)
Non-Twitter user · Treatment: Documentary	-0.278** (0.133)	-0.331** (0.127)	-0.173 (0.155)	-0.249* (0.144)	0.004 (0.079)
Age	0.292 (0.261)	0.164 (0.259)	0.459 (0.276)	0.288 (0.273)	0.205 (0.274)
Age ²	-0.133 (0.241)	-0.051 (0.243)	-0.276 (0.262)	-0.165 (0.265)	-0.033 (0.254)
Education years	0.006 (0.048)	-0.008 (0.045)	-0.017 (0.050)	-0.031 (0.047)	0.028 (0.042)
Female=1	0.085** (0.039)	0.089** (0.040)	0.072* (0.041)	0.076* (0.041)	0.125** (0.050)
Household head=1	0.058 (0.049)	0.067 (0.045)	0.029 (0.051)	0.042 (0.045)	0.068 (0.040)
Recruitment agency: Online=1	-0.081 (0.061)	-0.075 (0.059)	-0.032 (0.067)	-0.028 (0.064)	0.040 (0.057)
Recruitment agency: Offline=1	-0.035 (0.058)	-0.016 (0.057)	-0.021 (0.059)	0.001 (0.057)	0.052 (0.047)
R ²	0.050	0.047	0.045	0.037	0.087
Adj. R ²	0.032	0.028	0.026	0.018	0.069
Num. obs.	688	688	688	688	659
RMSE	1.073	1.056	1.134	1.104	1.009
Num. clusters	24	24	24	24	24

Notes. This table reports the full regression estimates corresponding to Figure 3. Each column presents results from Eq. 1, estimated on the full sample, for alternative versions of the *Combined Index* and for Δ Cortisol. All outcome variables are standardized so that, within the Twitter-Allowed group, the distribution has a mean of zero and a standard deviation of one. Because the specification omits a constant term and there is no omitted reference category for the treatment group indicators, the corresponding coefficients can be interpreted as group-level means of the standardized outcomes. Standard errors (in parentheses) are clustered at the classroom level. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

Table A7. Differences in the combined indexes and Δ Cortisol between Twitter and non-Twitter participants within the Documentary and Debate-Only conditions, full sample

	Combined Index	Combined Index V2	Combined Index V3	Combined Index V4	Δ Cortisol
Documentary	0.362** (0.155)	0.392** (0.161)	0.295* (0.167)	0.339* (0.176)	-0.032 (0.098)
Debate-Only	0.431*** (0.152)	0.414** (0.158)	0.397** (0.148)	0.374** (0.156)	0.084 (0.126)

Notes. Entries report mean differences between Twitter and non-Twitter participants (Twitter - non-Twitter) within the Documentary and Debate-Only conditions. For the Documentary condition, estimates represent the difference between the coefficients on *Twitter user* $\cdot$ *Treatment: Documentary* and *Non-Twitter user* $\cdot$ *Treatment: Documentary* from the corresponding specification in Table A6. Likewise, Debate-Only estimates represent the difference between the coefficients on *Twitter user* $\cdot$ *Treatment: Debate-Only* and *Non-Twitter user* $\cdot$ *Treatment: Debate-Only*. All differences are obtained using linear hypothesis tests with standard errors (in parentheses) clustered at the classroom level. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

Table A8. Differences in the combined indexes and Δ Cortisol between treatment groups and the Twitter-Allowed benchmark within the Twitter sample

	Combined Index	Combined Index V2	Combined Index V3	Combined Index V4	Δ Cortisol
Debate-Only	0.156 (0.113)	0.127 (0.114)	0.188* (0.099)	0.147 (0.105)	0.152 (0.133)
Twitter-Interact	0.216 (0.140)	0.212 (0.145)	0.301* (0.156)	0.285* (0.160)	0.287** (0.118)
Documentary	0.084 (0.138)	0.061 (0.140)	0.122 (0.145)	0.090 (0.146)	-0.029 (0.121)

Notes. Entries report mean differences for each treatment arm relative to the Twitter-Allowed benchmark within the Twitter sample. Each estimate corresponds to the difference between the coefficient for *Twitter user* $\cdot$ *Treatment: [Group]* and that for *Twitter user* $\cdot$ *Treatment: Twitter-Allowed* from the corresponding specification in Table A6. Differences are computed using linear hypothesis tests with standard errors (in parentheses) clustered at the classroom level. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

Table A9. Individual polarization indexes - Detailed regression estimates, full sample

	Elite	Social Distance	Partisanship	Institutional	Engagement
Twitter user · Treatment: Twitter-Allowed	0.000 (0.029)	0.000 (0.057)	0.000 (0.110)	0.000 (0.022)	0.000 (0.080)
Twitter user · Treatment: Debate-Only	0.337*** (0.080)	0.033 (0.147)	-0.030 (0.110)	0.106 (0.066)	0.004 (0.078)
Twitter user · Treatment: Twitter-Interact	0.336*** (0.080)	0.202* (0.113)	0.072 (0.053)	0.099 (0.122)	-0.087 (0.064)
Twitter user · Treatment: Documentary	0.246** (0.098)	-0.015 (0.140)	0.012 (0.105)	0.032 (0.124)	-0.043 (0.067)
Non-Twitter user · Treatment: Debate-Only	0.071 (0.126)	-0.060 (0.144)	-0.225** (0.100)	-0.269** (0.113)	-0.296*** (0.104)
Non-Twitter user · Treatment: Documentary	0.270** (0.097)	-0.236** (0.086)	-0.299** (0.129)	-0.158 (0.154)	-0.388*** (0.044)
Age	0.568* (0.331)	-0.346 (0.217)	0.579* (0.281)	0.296 (0.262)	-0.238 (0.217)
Age ²	-0.352 (0.311)	0.248 (0.212)	-0.402 (0.246)	-0.161 (0.263)	0.266 (0.203)
Education years	0.048 (0.054)	0.001 (0.036)	0.017 (0.058)	-0.097* (0.054)	0.058 (0.044)
Female=1	0.094** (0.045)	0.112** (0.048)	-0.015 (0.051)	-0.029 (0.032)	0.074* (0.038)
Household head=1	-0.060 (0.060)	0.019 (0.033)	0.032 (0.047)	0.071 (0.046)	0.096* (0.048)
Recruitment agency: Online=1	0.012 (0.080)	0.019 (0.071)	0.003 (0.054)	-0.105** (0.046)	-0.157*** (0.044)
Recruitment agency: Offline=1	-0.070 (0.060)	0.085 (0.068)	0.005 (0.057)	-0.073 (0.056)	-0.048 (0.056)
R ²	0.083	0.036	0.041	0.039	0.091
Adj. R ²	0.065	0.017	0.022	0.020	0.074
Num. obs.	688	687	682	688	688
RMSE	1.204	1.035	1.062	1.016	0.911
Num. clusters	24	24	24	24	24

Notes. This table reports the full regression estimates corresponding to Figure 4. Each column presents results from Eq. 1, estimated on the full sample, for the alternative individual polarization indexes considered. All outcome variables are standardized so that, within the Twitter-Allowed group, the distribution has a mean of zero and a standard deviation of one. Because the specification omits a constant term and there is no omitted reference category for the treatment group indicators, the corresponding coefficients can be interpreted as group-level means of the standardized outcomes. Standard errors (in parentheses) are clustered at the classroom level. *** p < 0.01; ** p < 0.05; * p < 0.1.

Table A10. Differences in the individual polarization indexes between Twitter and non-Twitter participants within the Documentary and Debate-Only conditions, full sample

	Elite	Social Distance	Partisanship	Institutional	Engagement
Documentary	-0.024 (0.116)	0.221 (0.196)	0.311* (0.159)	0.190 (0.176)	0.345*** (†††) (0.083)
Debate-Only	0.265* (0.133)	0.092 (0.155)	0.194 (0.130)	0.375*** (††) (0.113)	0.300** (0.144)

Notes. Entries report mean differences between Twitter and non-Twitter participants (Twitter - non-Twitter) within the Documentary and Debate-Only conditions. For the Documentary condition, estimates represent the difference between the coefficients on *Twitter user · Treatment: Documentary* and *Non-Twitter user · Treatment: Documentary* from the corresponding specification in Table A9. Likewise, Debate-Only estimates represent the difference between the coefficients on *Twitter user · Treatment: Debate-Only* and *Non-Twitter user · Treatment: Debate-Only*. All differences are obtained using linear hypothesis tests with standard errors (in parentheses) clustered at the classroom level. *** p < 0.01, ** p < 0.05, * p < 0.1 (uncorrected). †† p < 0.01, † p < 0.05, † p < 0.1 (Holm-corrected).

Table A11. Differences in the individual polarization indexes between treatment groups and the Twitter-Allowed benchmark within the Twitter sample

	Elite	Social Distance	Partisanship	Institutional	Engagement
Debate-Only	0.337*** ††† (0.083)	0.033 (0.158)	-0.030 (0.156)	0.106 (0.069)	0.004 (0.110)
Twitter-Interact	0.336*** ††† (0.084)	0.202 (0.125)	0.072 (0.125)	0.099 (0.124)	-0.087 (0.103)
Documentary	0.246** (0.102)	-0.015 (0.151)	0.012 (0.153)	0.032 (0.125)	-0.043 (0.103)

Notes. Entries report mean differences for each treatment arm relative to the Twitter-Allowed benchmark within the Twitter sample. Each estimate corresponds to the difference between the coefficient for *Twitter user · Treatment: [Group]* and that for *Twitter user · Treatment: Twitter-Allowed* from the corresponding specification in Table A9. Differences are computed using linear hypothesis tests with standard errors (in parentheses) clustered at the classroom level. *** p < 0.01, ** p < 0.05, * p < 0.1 (uncorrected). †† p < 0.01, † p < 0.05, † p < 0.1 (Holm-corrected).

Table A12. Combined indexes and Δ Cortisol - Detailed regression estimates, considering online segregation

	Combined Index	Combined Index V2	Combined Index V3	Combined Index V4	Δ Cortisol
Twitter user · Treatment: Twitter-Allowed · Non-segregated	-0.223** (0.099)	-0.217** (0.090)	-0.177 (0.103)	-0.173* (0.091)	0.048 (0.129)
Twitter user · Treatment: Twitter-Allowed · Segregated	0.186* (0.105)	0.181 (0.108)	0.147* (0.076)	0.144* (0.081)	-0.037 (0.154)
Twitter user · Treatment: Debate-Only · Non-segregated	-0.245*** (0.085)	-0.280*** (0.073)	-0.179** (0.086)	-0.226*** (0.079)	0.188 (0.130)
Twitter user · Treatment: Debate-Only · Segregated	0.611*** (0.121)	0.592*** (0.120)	0.607*** (0.134)	0.576*** (0.141)	0.097 (0.204)
Non-Twitter user · Treatment: Debate-Only	-0.244* (0.130)	-0.258* (0.132)	-0.179 (0.142)	-0.201 (0.145)	0.078 (0.072)
Twitter user · Treatment: Twitter-Interact · Non-segregated	-0.082 (0.190)	-0.062 (0.195)	0.011 (0.228)	0.025 (0.227)	0.160 (0.102)
Twitter user · Treatment: Twitter-Interact · Segregated	0.557*** (0.041)	0.524*** (0.040)	0.636*** (0.037)	0.583*** (0.027)	0.444*** (0.125)
Twitter user · Treatment: Documentary · Non-segregated	-0.280** (0.133)	-0.264* (0.151)	-0.188 (0.149)	-0.177 (0.170)	-0.187 (0.202)
Twitter user · Treatment: Documentary · Segregated	0.345** (0.123)	0.292** (0.112)	0.343** (0.130)	0.279** (0.118)	0.097 (0.111)
Non-Twitter user · Treatment: Documentary	-0.244* (0.132)	-0.299** (0.126)	-0.142 (0.155)	-0.220 (0.143)	0.012 (0.079)
Age	-0.054 (0.246)	-0.168 (0.245)	0.150 (0.263)	-0.004 (0.261)	0.153 (0.277)
Age ²	0.114 (0.233)	0.185 (0.236)	-0.056 (0.256)	0.043 (0.259)	0.004 (0.256)
Education years	-0.013 (0.050)	-0.026 (0.047)	-0.035 (0.051)	-0.047 (0.047)	0.025 (0.040)
Female=1	0.123*** (0.038)	0.126*** (0.040)	0.106** (0.040)	0.108** (0.042)	0.130** (0.052)
Household head=1	0.067 (0.050)	0.074 (0.046)	0.037 (0.053)	0.048 (0.047)	0.073 (0.043)
Recruitment agency: Online=1	-0.027 (0.063)	-0.022 (0.060)	0.016 (0.068)	0.018 (0.064)	0.041 (0.061)
Recruitment agency: Offline=1	-0.005 (0.056)	0.014 (0.055)	0.004 (0.057)	0.027 (0.054)	0.050 (0.048)
R ²	0.098	0.092	0.080	0.070	0.094
Adj. R ²	0.075	0.069	0.057	0.047	0.070
Num. obs.	688	688	688	688	659
RMSE	1.049	1.034	1.116	1.088	1.009
Num. clusters	24	24	24	24	24

Notes. This table reports the full regression estimates corresponding to Figure 5. Each column presents results from Eq. 2 — which divides the Twitter sample into segregated and non-segregated users — estimated on the full sample, for alternative versions of the *Combined Index* and for Δ Cortisol. All outcome variables are standardized so that, within the Twitter-Allowed group, the distribution has a mean of zero and a standard deviation of one. Because the specification omits a constant term and there is no omitted reference category for the treatment-by-segregation group indicators, the corresponding coefficients can be interpreted as group-level means of the standardized outcomes. Standard errors (in parentheses) are clustered at the classroom level. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

Table A13. Differences in the combined indexes and Δ Cortisol between Twitter and non-Twitter participants within the Documentary and Debate-Only conditions, considering online segregation

	Combined Index	Combined Index V2	Combined Index V3	Combined Index V4	Δ Cortisol
Documentary					
<i>Non-segregated</i>	-0.036 (0.182)	0.034 (0.197)	-0.046 (0.202)	0.044 (0.220)	-0.200 (0.193)
<i>Segregated</i>	0.589*** (0.167)	0.591*** (0.158)	0.486*** (0.164)	0.499*** (0.159)	0.084 (0.134)
Debate-Only					
<i>Non-segregated</i>	-0.002 (0.171)	-0.022 (0.166)	0.001 (0.172)	-0.025 (0.171)	0.110 (0.142)
<i>Segregated</i>	0.855*** (0.193)	0.850*** (0.198)	0.787*** (0.194)	0.776*** (0.199)	0.020 (0.217)

Notes. Entries report mean differences between Twitter and non-Twitter participants (Twitter - non-Twitter) within the Documentary and Debate-Only conditions, conditional on online segregation status. For the Documentary condition, each estimate represents the difference between the coefficients on *Twitter user* · *Treatment: Documentary* · [*Segregation status*] and *Non-Twitter user* · *Treatment: Documentary* from the corresponding specification in Table A12. Likewise, for the Debate-Only condition, estimates correspond to the difference between the coefficients on *Twitter user* · *Treatment: Debate-Only* · [*Segregation status*] and *Non-Twitter user* · *Treatment: Debate-Only*. All differences are obtained using linear hypothesis tests with standard errors (in parentheses) clustered at the classroom level. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

Table A14. Differences in the combined indexes and Δ Cortisol between treatment groups and the Twitter-Allowed benchmark within the Twitter sample, considering online segregation

	Combined Index	Combined Index V2	Combined Index V3	Combined Index V4	Δ Cortisol
<i>Non-segregated</i>					
Debate-Only	-0.022 (0.130)	-0.063 (0.115)	-0.002 (0.134)	-0.053 (0.119)	0.140 (0.193)
Twitter-Interact	0.141 (0.216)	0.155 (0.216)	0.188 (0.255)	0.198 (0.247)	0.112 (0.171)
Documentary	-0.057 (0.162)	-0.047 (0.170)	-0.011 (0.177)	-0.003 (0.187)	-0.236 (0.237)
<i>Segregated</i>					
Debate-Only	0.426** (0.152)	0.412** (0.157)	0.460*** (0.150)	0.431** (0.161)	0.134 (0.256)
Twitter-Interact	0.372*** (0.111)	0.344*** (0.115)	0.489*** (0.081)	0.438*** (0.084)	0.480** (0.197)
Documentary	0.159 (0.155)	0.112 (0.150)	0.196 (0.147)	0.135 (0.139)	0.133 (0.192)

Notes. Entries report mean differences for each treatment arm relative to the Twitter-Allowed benchmark within the Twitter sample. Each estimate corresponds to the difference between the coefficient for *Twitter user* · *Treatment: [Group]* · [*segregation status*] and that for *Twitter user* · *Treatment: Twitter-Allowed* · [*segregation status*] from the corresponding specification in Table A12. Differences are computed using linear hypothesis tests with standard errors (in parentheses) clustered at the classroom level. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

Table A15. Individual polarization indexes - Detailed regression estimates, considering online segregation

	Elite	Social Distance	Partisanship	Institutional	Engagement
Twitter user · Treatment: Twitter-Allowed · Non-segregated	-0.150* (0.078)	-0.058 (0.073)	-0.045 (0.153)	-0.160*** (0.047)	-0.225*** (0.052)
Twitter user · Treatment: Twitter-Allowed · Segregated	0.125** (0.055)	0.048 (0.074)	0.038 (0.110)	0.133*** (0.038)	0.187 (0.131)
Twitter user · Treatment: Debate-Only · Non-segregated	0.191 (0.169)	-0.187 (0.182)	-0.248 (0.161)	-0.174** (0.081)	-0.284** (0.119)
Twitter user · Treatment: Debate-Only · Segregated	0.488*** (0.171)	0.295 (0.187)	0.221** (0.096)	0.422*** (0.130)	0.330** (0.153)
Non-Twitter user · Treatment: Debate-Only	0.094 (0.128)	-0.052 (0.144)	-0.211** (0.099)	-0.246** (0.113)	-0.273** (0.108)
Twitter user · Treatment: Twitter-Interact · Non-segregated	0.075 (0.148)	0.115 (0.117)	-0.064 (0.086)	-0.100 (0.229)	-0.261*** (0.027)
Twitter user · Treatment: Twitter-Interact · Segregated	0.642*** (0.092)	0.300** (0.107)	0.227*** (0.047)	0.325*** (0.097)	0.106 (0.127)
Twitter user · Treatment: Documentary · Non-segregated	-0.053 (0.094)	0.026 (0.166)	-0.155 (0.168)	-0.288 (0.169)	-0.358*** (0.084)
Twitter user · Treatment: Documentary · Segregated	0.464*** (0.149)	-0.053 (0.135)	0.131 (0.097)	0.263* (0.135)	0.185 (0.113)
Non-Twitter user · Treatment: Documentary	0.293*** (0.098)	-0.227** (0.086)	-0.284** (0.130)	-0.133 (0.153)	-0.364*** (0.043)
Age	0.350 (0.319)	-0.443** (0.212)	0.425 (0.287)	0.039 (0.282)	-0.508** (0.217)
Age ²	-0.197 (0.298)	0.317 (0.208)	-0.293 (0.252)	0.023 (0.278)	0.460** (0.206)
Education years	0.037 (0.054)	-0.005 (0.035)	0.008 (0.057)	-0.111* (0.056)	0.044 (0.046)
Female=1	0.119** (0.043)	0.123** (0.049)	0.001 (0.050)	-0.001 (0.035)	0.103** (0.037)
Household head=1	-0.049 (0.059)	0.016 (0.034)	0.035 (0.049)	0.078 (0.046)	0.105** (0.047)
Recruitment agency: Online=1	0.040 (0.083)	0.039 (0.074)	0.028 (0.058)	-0.065 (0.047)	-0.114** (0.046)
Recruitment agency: Offline=1	-0.057 (0.060)	0.097 (0.067)	0.018 (0.058)	-0.052 (0.058)	-0.023 (0.057)
R ²	0.099	0.045	0.052	0.068	0.127
Adj. R ²	0.076	0.021	0.028	0.044	0.105
Num. obs.	688	687	682	688	688
RMSE	1.197	1.033	1.058	1.003	0.895
Num. clusters	24	24	24	24	24

Notes. This table reports the full regression estimates corresponding to Figure 6. Each column presents results from Eq. 2 — which divides the Twitter sample into segregated and non-segregated users — estimated on the full sample, for the alternative individual polarization indexes considered. All outcome variables are standardized so that, within the Twitter-Allowed group, the distribution has a mean of zero and a standard deviation of one. Because the specification omits a constant term and there is no omitted reference category for the treatment-by-segregation group indicators, the corresponding coefficients can be interpreted as group-level means of the standardized outcomes. Standard errors (in parentheses) are clustered at the classroom level. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

Table A16. Differences in the individual polarization indexes between Twitter and non-Twitter participants within the Documentary and Debate-Only conditions, considering online segregation

	Elite	Social Distance	Partisanship	Institutional	Engagement
<i>Documentary</i>					
<i>Non-segregated</i>	-0.347*** ††	0.253	0.129	-0.154	0.006
	(0.105)	(0.222)	(0.222)	(0.238)	(0.084)
<i>Segregated</i>	0.171	0.174	0.414** †	0.397** †	0.549*** †††
	(0.175)	(0.186)	(0.150)	(0.162)	(0.133)
<i>Debate-Only</i>					
<i>Non-segregated</i>	0.097	-0.135	-0.037	0.072	-0.011
	(0.253)	(0.185)	(0.160)	(0.126)	(0.205)
<i>Segregated</i>	0.394*** ††	0.348* †	0.432*** ††	0.667*** †††	0.603*** †††
	(0.136)	(0.201)	(0.150)	(0.166)	(0.171)

Notes. Entries report mean differences between Twitter and non-Twitter participants (Twitter - non-Twitter) within the Documentary and Debate-Only conditions, conditional on online segregation status. For the Documentary condition, each estimate represents the difference between the coefficients on *Twitter user* · *Treatment: Documentary* · [*Segregation status*] and *Non-Twitter user* · *Treatment: Documentary* from the corresponding specification in Table A15. Likewise, for the Debate-Only condition, estimates correspond to the difference between the coefficients on *Twitter user* · *Treatment: Debate-Only* · [*Segregation status*] and *Non-Twitter user* · *Treatment: Debate-Only*. All differences are obtained using linear hypothesis tests with standard errors (in parentheses) clustered at the classroom level. *** p < 0.01, ** p < 0.05, * p < 0.1 (uncorrected). †† p < 0.01, † p < 0.05, † p < 0.1 (Holm-corrected).

Table A17. Differences in the individual polarization indexes between treatment groups and the Twitter-Allowed benchmark within the Twitter sample, considering online segregation

	Elite	Social Distance	Partisanship	Institutional	Engagement
<i>Non-segregated</i>					
<i>Debate-Only</i>	0.342*	-0.129	-0.202	-0.014	-0.059
	(0.174)	(0.186)	(0.230)	(0.085)	(0.126)
<i>Twitter-Interact</i>	0.226	0.173	-0.019	0.060	-0.036
	(0.162)	(0.134)	(0.181)	(0.234)	(0.057)
<i>Documentary</i>	0.097	0.085	-0.109	-0.128	-0.133
	(0.108)	(0.175)	(0.236)	(0.167)	(0.097)
<i>Segregated</i>					
<i>Debate-Only</i>	0.363**	0.247	0.183	0.288**	0.143
	(0.172)	(0.201)	(0.136)	(0.135)	(0.197)
<i>Twitter-Interact</i>	0.517*** †††	0.251*	0.189	0.192*	-0.081
	(0.103)	(0.131)	(0.114)	(0.101)	(0.180)
<i>Documentary</i>	0.339**	-0.101	0.093	0.130	-0.002
	(0.160)	(0.155)	(0.141)	(0.129)	(0.168)

Notes. Entries report mean differences for each treatment arm relative to the Twitter-Allowed benchmark within the Twitter sample. Each estimate corresponds to the difference between the coefficient for *Twitter user* · *Treatment: [Group]* · [*segregation status*] and that for *Twitter user* · *Treatment: Twitter-Allowed* · [*segregation status*] from the corresponding specification in Table A15. Differences are computed using linear hypothesis tests with standard errors (in parentheses) clustered at the classroom level. *** p < 0.01, ** p < 0.05, * p < 0.1 (uncorrected). †† p < 0.01, † p < 0.05, † p < 0.1 (Holm-corrected).

Table A18. Twitter actions - Detailed regression estimates

	Activity Level (IHS)	Contact Segregation	Δ Activity Level (IHS)	Δ Contact Segregation
Twitter user · Treatment: Twitter-Allowed · Non-segregated	-0.212*** (0.052)	-0.212 (0.193)	-0.203*** (0.067)	0.053 (0.049)
Twitter user · Treatment: Twitter-Allowed · Segregated	0.173 (0.095)	0.173 (0.103)	0.166** (0.061)	-0.043 (0.079)
Twitter user · Treatment: Debate-Only · Non-segregated			-0.242** (0.110)	-0.148 (0.097)
Twitter user · Treatment: Debate-Only · Segregated			0.050 (0.125)	0.090 (0.108)
Twitter user · Treatment: Twitter-Interact · Non-segregated	0.981*** (0.102)	0.082 (0.076)	-0.370*** (0.091)	-0.144 (0.161)
Twitter user · Treatment: Twitter-Interact · Segregated	1.320*** (0.100)	0.625*** (0.086)	0.003 (0.143)	0.251** (0.109)
Twitter user · Treatment: Documentary · Non-segregated			-0.105* (0.060)	0.143 (0.092)
Twitter user · Treatment: Documentary · Segregated			-0.123 (0.114)	-0.084 (0.150)
Age	0.230 (0.394)	0.212 (0.620)	-0.443* (0.254)	0.197 (0.414)
Age ²	-0.393 (0.379)	-0.267 (0.604)	0.371 (0.260)	-0.160 (0.375)
Education years	-0.028 (0.083)	0.041 (0.038)	0.068** (0.030)	-0.005 (0.047)
Female=1	0.050* (0.023)	0.093 (0.070)	0.048 (0.050)	0.077* (0.039)
Household head=1	0.112 (0.061)	-0.087 (0.073)	-0.018 (0.061)	-0.026 (0.051)
Recruitment agency: Online=1	0.048 (0.057)	0.042 (0.069)	-0.046 (0.050)	-0.071 (0.073)
Recruitment agency: Offline=1	-0.101 (0.058)	0.056 (0.096)	-0.038 (0.057)	-0.043 (0.074)
R ²	0.567	0.157	0.052	0.031
Adj. R ²	0.544	0.113	0.018	-0.003
Num. obs.	221	221	445	445
RMSE	0.766	0.821	0.945	0.922
Num. clusters	8	8	24	24

Notes. This table reports the full regression estimates corresponding to Figure 7 and Figure 8. Each column presents results from Eq. 2 — which divides the Twitter sample into segregated and non-segregated users — estimated on the Twitter-user sample. Δ Activity Level and Δ Contact Segregation are defined as the value of each measure on the day after the debate minus its value on the day before the debate (following day minus preceding day). All outcome variables are standardized so that, within the Twitter-Allowed group, the distribution has a mean of zero and a standard deviation of one. Because the specification omits a constant term and there is no omitted reference category for the treatment-by-segregation group indicators, the corresponding coefficients can be interpreted as group-level means of the standardized outcomes. Standard errors (in parentheses) are clustered at the classroom level. *** p < 0.01; ** p < 0.05; * p < 0.1.

Table A19. Differences in Twitter actions between treatment groups and the Twitter-Allowed benchmark within the Twitter sample, considering online segregation

	Activity Level (IHS)	Contact Segregation	Δ Activity Level (IHS)	Δ Contact Segregation
<i>Non-segregated</i>				
Debate-Only			-0.039 (0.132)	-0.201** (0.089)
Twitter-Interact	1.194*** (0.105)	0.294 (0.198)	-0.167 (0.104)	-0.198 (0.160)
Documentary			0.099 (0.081)	0.090 (0.086)
<i>Segregated</i>				
Debate-Only			-0.116 (0.142)	0.133 (0.131)
Twitter-Interact	1.146*** (0.127)	0.452** (0.129)	-0.163 (0.159)	0.294** (0.133)
Documentary			-0.289** (0.129)	-0.041 (0.161)

Notes. Entries report mean differences for each treatment arm relative to the Twitter-Allowed benchmark within the Twitter sample. Each estimate corresponds to the difference between the coefficient for $Twitter\ user \cdot Treatment: [Group] \cdot [segregation\ status]$ and that for $Twitter\ user \cdot Treatment: Twitter-Allowed \cdot [segregation\ status]$ from the corresponding specification in Table A18. Δ Activity Level and Δ Contact Segregation are defined as the value of each measure on the day after the debate minus its value on the day before the debate (following day minus preceding day). Differences are computed using linear hypothesis tests with standard errors (in parentheses) clustered at the classroom level. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

C. Appendix figures

Figure A1. Experiment’s recruitment Twitter account with pinned tweet



Notes. In English, the pinned tweet reads as follows: “Are you on Twitter and want to take part in an on-site academic study about political ideas and voter preferences on Sunday 10/13? (As a thank-you and to make up for your time, we'll give away a purchasing card to each participant) For more info, visit: <url_link>”

Figure A2. Twitter-Interact screen display

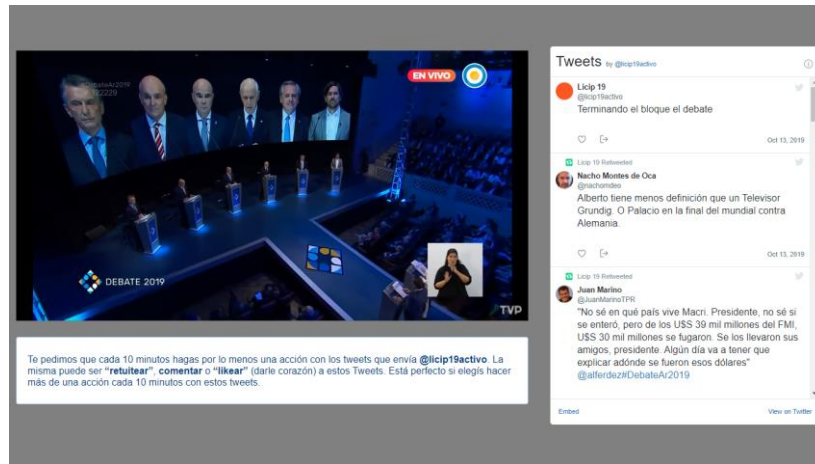
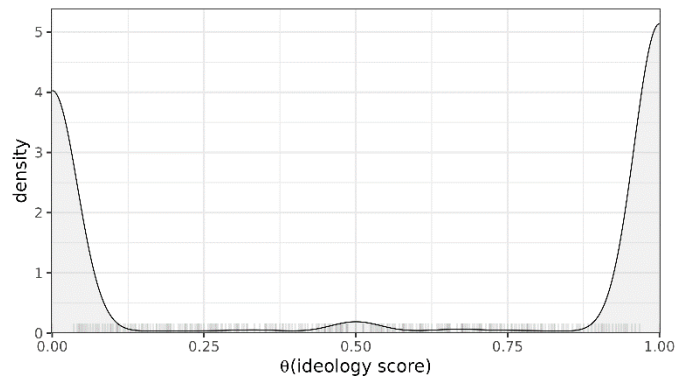
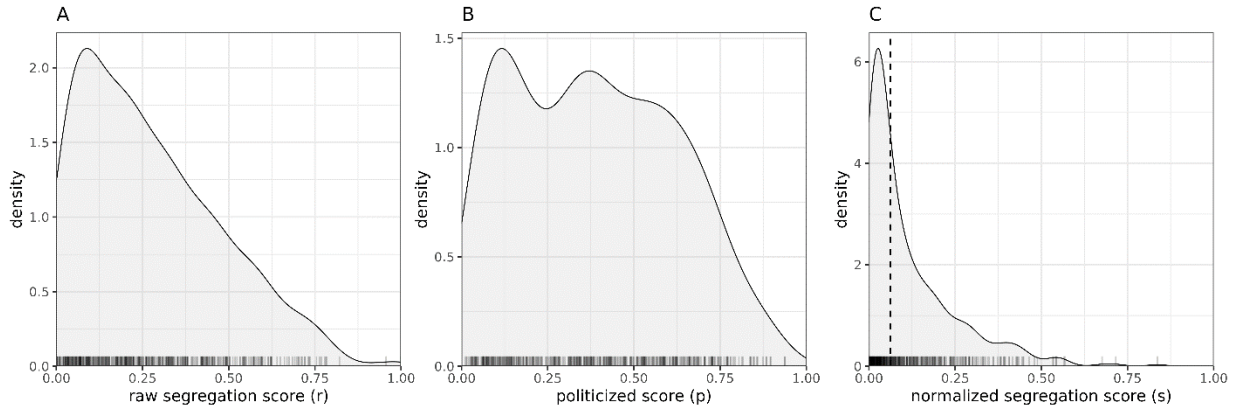


Figure A3. Ideology score (θ) distribution across the Argentine political landscape network nodes



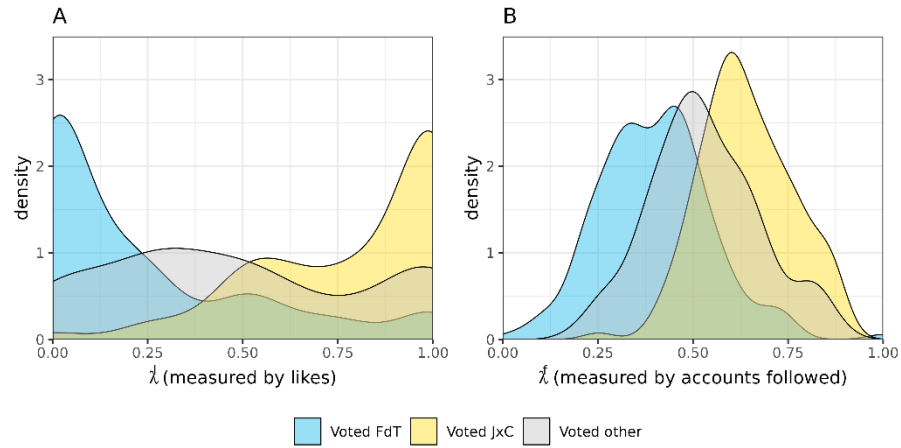
Notes. The density plot (gray shaded area) represents the smooth distribution of the ideology score (θ) across the nodes in the network (for both seed and non-seed accounts). A rug plot at the bottom highlights the individual data points, with each tick mark corresponding to a specific node's ideology score.

Figure A4. Distribution of normalized segregation scores and its components



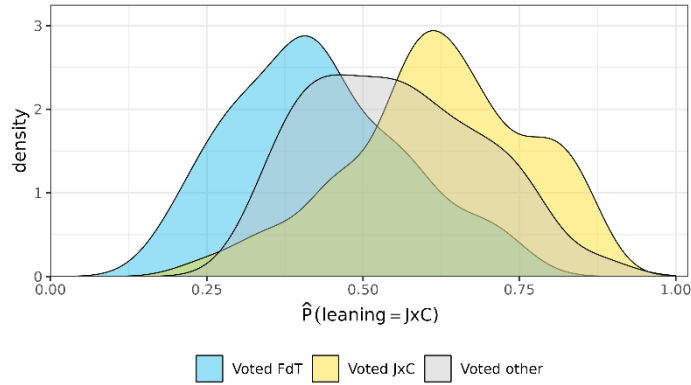
Notes. Dashed line in panel C indicates the position of the median score. The distribution of λ_i^f is shown in Figure A5. A rug plot at the bottom of each panel highlights the individual data points, with each tick mark representing the score of a single participant.

Figure A5. Distribution of Twitter users' party preference scores (likes and following)



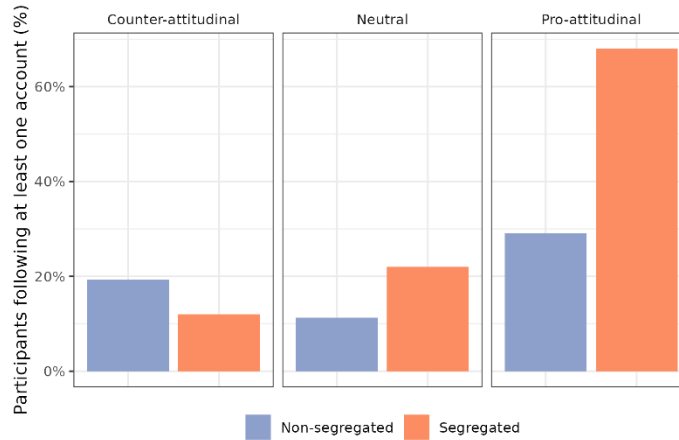
Notes. Panel A shows participants' party preference scores estimated from their "likes" (λ^l), while Panel B shows scores estimated from the accounts they follow (λ^f). The two measures are strongly correlated, with a Pearson correlation coefficient of 0.719. Twitter-based indicators of political orientation closely align with self-reported primary vote choices, with particularly strong correspondence for the measure based on likes.

Figure A6. Predicted JxC-leaning among non-Twitter participants by primary vote choice



Notes. The figure shows the distribution of the predicted probability of leaning toward JxC for non-Twitter participants, separately by their self-reported primary vote choice (PASO). The measure is estimated using a classifier trained to predict party leaning in the Twitter sample. The observed patterns indicate that the model exhibits strong discriminative performance even for individuals without Twitter accounts.

Figure A7. Percentage of participants who already followed accounts retweeted by the experiment's active account, classified as counter-attitudinal, neutral, or pro-attitudinal



Notes. For each participant, we identify whether the accounts retweeted by the experiment's active account can be classified as counter-attitudinal (e.g., a *FdT* retweeted account for a JxC-leaning participant), neutral (e.g., a retweet originating from a fact-checker account) or pro-attitudinal (e.g., a *FdT* retweeted account for a *FdT*-leaning participant). Then, we calculate for both the segregated and non-segregated groups the percentage of participants who followed at least one counter-attitudinal, neutral, or pro-attitudinal account retweeted by the experiment's active account. Most of the non-segregated participants followed neither counter-attitudinal nor neutral accounts. However, most segregated participants followed at least one pro-attitudinal account.

D. Robustness checks for the online segregation measure

This appendix presents a set of robustness checks for our measure of online segregation and for the main results that rely on it. We first examine whether the heterogeneity patterns obtained when splitting Twitter users into segregated and non-segregated groups are driven by raw segregation or politicization considered separately. We then decompose the Twitter-activity outcomes from Section 3.c to determine which component of *Contact Segregation* accounts for the reported effects. Finally, we study whether the backfire effects associated with online segregation remain once treatment effects are allowed to vary with the main pre-treatment characteristics correlated with segregation.

D.1. Raw segregation and politicization considered separately when splitting the Twitter-user sample

Our *normalized segregation score* used to classify participants as segregated or non-segregated is constructed as the product of two components: a *raw segregation score* and a *politicization score* (see Section 2.b.2). Hence, the backfire effects observed among segregated Twitter users might stem from having just one of these characteristics, rather than from the joint presence of both. To assess this possibility, we re-estimated the analyses from Figure 5, considering each component of the index separately.

In a first step, we defined as segregated those users with a *raw segregation score* above the median value of the *raw segregation score* (calculated over our Twitter-user sample), and as non-segregated otherwise. Using this definition, we find a small backfire effect in our *Combined Index* for the group now classified as non-segregated in the Debate-Only condition (statistically significant at the 10% level) and another mild effect for the group now considered segregated in the Twitter-Interact condition (also significant at the 10% level). The 5% statistically significant effect for Δ *Cortisol* remains, but notably, it now appears for the group considered non-segregated Twitter-Interact under this new definition. In other words, we no longer observe effects solely within the segregated group, suggesting that this split fails to capture the same phenomenon as our main characterization. See the first two columns of Table A20 for detailed results.

Table A20. Differences in outcomes when classifying segregation by raw segregation or politicization scores

	By raw segregation score		By politicization score	
	Combined Index	Δ Cortisol	Combined Index	Δ Cortisol
<i>Non-segregated</i>				
Debate-Only	0.149* (0.077)	0.274 (0.203)	0.114 (0.119)	0.096 (0.165)
Twitter-Interact	0.171 (0.129)	0.233** (0.104)	0.225 (0.210)	0.170 (0.184)
Documentary	-0.011 (0.145)	0.038 (0.172)	0.149 (0.204)	-0.194 (0.225)
<i>Segregated</i>				
Debate-Only	0.145 (0.223)	0.043 (0.257)	0.253 (0.177)	0.195 (0.268)
Twitter-Interact	0.347* (0.189)	0.394 (0.278)	0.198 (0.126)	0.404** (0.191)
Documentary	0.173 (0.186)	-0.086 (0.235)	-0.041 (0.154)	0.118 (0.198)

Notes. Entries report estimated mean differences for each treatment group relative to the Twitter-Allowed benchmark within the Twitter sample, under two alternative definitions of segregation. In the first panel, segregation status is determined by the median of the *raw segregation score*. In the second, by the median of the *politicization score*. Each estimate corresponds to the difference between the coefficient for *Twitter user* · *Treatment*: [*Group*] · [*segregation status*] and that for *Twitter user* · *Treatment*: *Twitter-Allowed* · [*segregation status*] from the underlying specifications based on Eq. 2. Detailed regressions are available upon request. Standard errors (in parentheses) are clustered at the classroom level. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

In a second step, we adopted an alternative definition of segregation based only on participants' politicization scores. Those with a *politicization score* above the sample median were classified as segregated, and those at or below the median as non-segregated. Under this alternative criterion, we find no significant effects for the *Combined Index*. However, the backfire effect on Δ *Cortisol* remains statistically significant at the 5% level among segregated participants in the Twitter-

Interact group. Again, these results suggest that politicization alone does not capture the same underlying polarization dynamics as our composite segregation index. See the second two columns of Table A20 for detailed results.

Taken together, these analyses indicate that neither the *raw segregation score* nor the *politicization score* alone fully accounts for the patterns observed in our main results. Rather, the joint presence of high raw segregation and high politicization appears essential for explaining the observed backfire dynamics.

D.2. Raw activity segregation and activity politicization considered separately in the analysis of Twitter users' online activity

In Section 3.c, we reported results for *Contact Segregation* and Δ *Contact Segregation*. As defined in Section 2.c.2, *Contact Segregation* is the product of two components: users' *raw activity segregation* and their *activity politicization*. Here, we examine whether the patterns reported in Figure 7, Figure 8, and Table A19 are driven primarily by one of these two components. To do so, we estimate treatment effects separately for *raw activity segregation* and *activity politicization*, both during the debate and for post-treatment changes measured as the day-after minus day-before difference. For ease of comparison, we also reproduce the corresponding estimates for *Contact Segregation* and Δ *Contact Segregation*.

Table A21. Differences in the components of contact segregation between treatment groups and the Twitter-Allowed benchmark within the Twitter sample

Treatment condition	Raw Activity Segregation	Activity Politicization	Contact Segregation	Δ Raw Activity Segregation	Δ Activity Politicization	Δ Contact Segregation
<i>Non-segregated</i>						
Debate-Only				-0.474*** (0.095)	-0.051 (0.124)	-0.201** (0.089)
Twitter-Interact	0.342 (0.221)	0.710*** (0.161)	0.294 (0.198)	-0.146 (0.201)	-0.088 (0.161)	-0.198 (0.160)
Documentary				0.113 (0.117)	0.029 (0.108)	0.090 (0.086)
<i>Segregated</i>						
Debate-Only				0.137 (0.139)	0.091 (0.136)	0.133 (0.131)
Twitter-Interact	0.511*** (0.110)	0.701*** (0.086)	0.452** (0.129)	0.225*** (0.079)	0.113 (0.171)	0.294** (0.133)
Documentary				-0.049 (0.139)	-0.194 (0.141)	-0.041 (0.161)

Notes. The table reports estimated differences for each treatment arm relative to the Twitter-Allowed benchmark within the Twitter sample, disaggregated by segregation status. The first three columns report treatment effects on *raw activity segregation*, *activity politicization*, and *Contact Segregation* during the debate. These outcomes are only defined for groups allowed to use Twitter during the experimental session, so the Debate-Only and Documentary entries are not applicable for the first three columns. The last three columns report treatment effects on the corresponding post-debate changes, defined as the value of each measure on the day after the debate minus its value on the day before the debate. Estimates are computed using the same specification as in Table A19. Detailed regressions are available upon request. Standard errors (in parentheses) are clustered at the classroom level. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

Table A21 reports the resulting estimates. Focusing first on Twitter activity during the debate, the table shows that the Twitter-Interact treatment increases *activity politicization* for both segregated and non-segregated users. This pattern is consistent with compliance with the experimental design, since participants in that treatment were instructed to interact with an experimenter-run account whose content was overwhelmingly political. However, the pattern differs for *Raw activity segregation*. Among non-segregated users, the coefficient is positive but imprecisely estimated, whereas among segregated users it is positive, large, and statistically significant. Thus, although both groups increased the share of their activity directed at political accounts, only segregated users also concentrated that activity more strongly on accounts from one side of the ideological spectrum. Consistent with this decomposition, the increase in *Contact Segregation* reported in Section 3.c is concentrated among segregated users.

Turning to post-debate Twitter behavior, Table A21 shows no significant treatment effects on Δ *Activity Politicization*. By contrast, treatment effects do appear for Δ *Raw Activity Segregation*. In particular, the Debate-Only treatment generates a negative and statistically significant effect among non-segregated users, while the Twitter-Interact treatment generates a positive and statistically significant effect among segregated users. These movements map directly into the

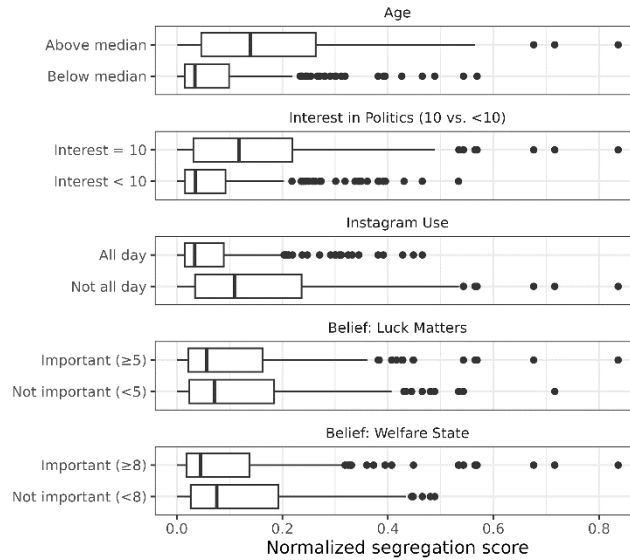
corresponding results for Δ Contact Segregation: the decline among non-segregated Debate-Only participants and the increase among segregated Twitter-Interact participants are driven primarily by changes in the ideological concentration of interactions rather than by changes in the overall share of activity directed toward political accounts. Overall, the decomposition suggests that the post-debate dynamics documented in Section 3.c reflect changes in ideological concentration within political activity, not changes in politicization per se.

D.3. Backfire effects accounting for correlated characteristics

An additional potential concern with our findings on segregation is that the backfire effects observed among segregated participants might reflect other characteristics correlated with segregation rather than segregation itself. To assess this possibility, we implement a data-driven robustness exercise in two steps. First, following a procedure analogous to that described in Section 2.b.3, we train a random forest classifier to identify the pre-treatment variables that best predict online segregation status (rather than party leaning, as described earlier). Among all predictors, the model ranks the following five as the most important (listed in descending order of their mean decrease in Gini importance): i) age, ii) interest in politics, iii) Instagram use, iv) beliefs about the role of luck in poverty, and v) the importance attributed to the welfare state.

To illustrate these relationships, Figure A8 displays the distribution of the *normalized segregation score* when the sample is divided along each of the five predictive variables. Concretely, each panel splits the sample into two groups such that roughly half of participants fall on each side of the cutoff — namely, age above the sample median, interest in politics equal to 10, Instagram use reported as “all day,” belief that luck matters for poverty rated 5 or higher, and importance assigned to the welfare state rated 8 or higher. The figure shows that older individuals and those reporting maximum interest in politics tend to be more segregated. In contrast, frequent Instagram users (a behavior negatively correlated with age) exhibit lower levels of online segregation. Finally, participants who believe that luck plays little role in poverty or who assign low importance to the welfare state also tend to display higher segregation levels.

Figure A8. Normalized segregation score distributions by top predictive pre-treatment characteristics



Notes. The figure displays the distribution of the *normalized segregation score* across subgroups defined by the five pre-treatment variables most predictive of segregation status, as identified by a random forest classifier. Each variable is dichotomized so that approximately half of the sample falls into each group. The panels are ordered by the predictors' relative importance, ranked in decreasing order according to the model's mean decrease in Gini.

Having identified these five predictors, we next examine whether the backfire effects observed among segregated users remain once treatment effects are allowed to vary with these correlated characteristics. We do so by estimating a series of regressions based on Eq. 2, using the same baseline specification as in Figure 5 for the *Combined Index*. Concretely, for each predictor m we augment Eq. 2 as follows:

$$Y_i = \sum_{k,l} \beta_{k,l} \cdot (S_{i,l} \cdot T_{i,k}) + \sum_{k,j} \delta_{k,j}^m \cdot (Z_i^m \cdot \text{Twitter}_{i,j} \cdot T_{i,k}) + \gamma \cdot X_i + \epsilon_i.$$

Here, Z_i^m denotes one dichotomized predictor of online segregation, and $\text{Twitter}_{i,j}$ is a set of dummy variables indexing Twitter-user status, where $j \in \{\text{Twitter}, \text{non-Twitter}\}$, as in Eq. 1. This distinction matters because the Debate-Only and Documentary treatments include both Twitter users and non-Twitter participants. The augmented specification preserves the original segregation-by-treatment structure while allowing the association between the added predictor and the outcome to vary across treatment arms, separately for Twitter users and non-Twitter participants. The predictors are dichotomized using the same thresholds as in Figure A8, for example, age above versus below the median, interest in politics equal to 10 versus below 10, Instagram use all day versus not all day, and so on. All other controls and estimation settings remain unchanged relative to the baseline specification.

Table A22 reports the estimated differences in the *Combined Index* between each treatment group and its corresponding Twitter-Allowed benchmark within each segregation group. Each column presents the relevant treatment contrasts from a separate regression in which the binary predictor listed in the header is interacted with the treatment-by-user-type cells. Across all specifications, the positive contrasts for the Debate-Only and Twitter-Interact groups among segregated participants remain sizable and statistically significant, whereas the corresponding contrasts among non-segregated participants are generally smaller and statistically insignificant. The only additional finding arises in the specification that interacts treatment with the importance attributed to the welfare state. In that column, the contrast for the segregated Documentary group is also positive and statistically significant at the 5% level, while the contrast for the non-segregated Twitter-Interact group becomes positive and marginally significant at the 10% level.

Table A22. Treatment effects on Combined Index by segregation status accounting for interactions with predictors of online segregation

Treatment condition	Additional variable interacted				
	Age (Above / below median)	Interest in politics (10 / <10)	Instagram use (All / not all day)	Belief: luck matters (≥5 / <5)	Belief: welfare state (≥8 / <8)
<i>Non-segregated</i>					
Debate-Only	-0.048 (0.153)	0.105 (0.149)	-0.127 (0.121)	0.115 (0.184)	0.178 (0.157)
Twitter-Interact	0.079 (0.236)	0.266 (0.226)	-0.027 (0.274)	0.070 (0.226)	0.416* (0.211)
Documentary	-0.072 (0.185)	0.003 (0.179)	-0.200 (0.193)	-0.066 (0.224)	0.135 (0.199)
<i>Segregated</i>					
Debate-Only	0.402* (0.216)	0.645** (0.269)	0.374** (0.148)	0.535*** (0.155)	0.561*** (0.120)
Twitter-Interact	0.247* (0.138)	0.566** (0.247)	0.286** (0.122)	0.304*** (0.107)	0.599*** (0.115)
Documentary	0.135 (0.252)	0.233 (0.210)	0.048 (0.168)	0.137 (0.173)	0.332** (0.151)

Notes. The table reports estimated differences in the *Combined Index* between each treatment group and its corresponding Twitter-Allowed benchmark, disaggregated by segregation status. Estimates are obtained from Eq. 2 augmented with interactions between one predictive feature at a time and the treatment-by-user-type cells, using the same dichotomizations as in Figure A8. Detailed regressions are available upon request. Each column reports the resulting treatment contrasts from a separate regression. Standard errors (in parentheses) are clustered at the classroom level. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

This analysis suggests that the backfire effects observed among segregated participants are unlikely to be driven by demographic or attitudinal factors correlated with segregation. The persistence of these effects across specifications

that allow treatment effects to vary with the main pre-treatment predictors of segregation strengthens the interpretation that online segregation, and the informational environments it entails, plays a central role in shaping polarization responses to our interventions.