The impact of digital advertising on turnout during the 2020 US presidential election: evidence from a massive campaign-level field experiment

Minali Aggarwal,[1†] Jennifer Allen,[2†] Alexander Coppock,[1†]
Dan Frankowski,[3†] Solomon Messing,[4*†] Kelly Zhang,[6†]
James Barnes,[5] Andrew Beasley,[9] Harry Hantman,[7] Sylvan Zheng[8]

Affiliations:

[1]Yale University, Department of Political Science, New Haven, CT, US
[2]MIT Sloan School of Management, Boston, MA, US
[3]Independent researcher at Boxydog, Minneapolis, MN, US
[4]Georgetown University, McCourt School of Public Policy, Washington, DC, US
[5]Condorsay, New York, NY, US
[6]Busara Center for Behavioral Economics, Nairobi, Kenya (Research conducted prior to joining)
[7]Greenlight, New York, NY, US
[8]New York University, Department of Politics, New York, NY, US
[9]Macmillan Publishers, New York, NY, US

[†] Contributed equally
[*]To whom correspondence should be addressed; E-mail: sm3563@georgetown.edu.

November 15, 2022

Abstract

We present the results of a massive, $8.9 million campaign-wide field experiment, conducted among 2 million moderate and low-information "persuadable" voters in five battleground states during the 2020 US Presidential election. Treatment group subjects were exposed to an eight-month-long advertising program delivered via social media, designed to persuade people to vote against Donald Trump and for Joe Biden. On average, the program neither increased or decreased turnout. We find evidence of differential turnout effects by modeled level of Trump support: the campaign increased voting among Biden leaners by 0.4 percentage points (SE: 0.2pp) and decreased voting among Trump leaners by 0.3 percentage points (SE: 0.3pp), for a difference-in-CATES of 0.7 points ($t(1035571) = -2.09$, $p = 0.036$, $\widehat{DIC} = 0.7$ points, 95% CI = $[-0.014, -0.00]$). An important but exploratory finding is that the strongest differential effects appear in *early voting* data, which may inform future work on early campaigning in a post-COVID electoral environment. Our results indicate that differential mobilization effects of even large digital advertising campaigns in presidential elections are likely to be modest.

Introduction

Isolating the causal effects of the billions spent on political advertising in the United States each election cycle has proven to be one of the most significant research design challenges in the social sciences. Observational studies are vulnerable to confounding: for example, comparisons of vote returns in localities exposed to different levels of advertising suffer from selection bias because ads tend to be bought in competitive districts [1, 2]. Survey experiments [3, 4, 5] address selection bias but typically measure the immediate effects of a single ad or "dose" of advertising. Survey experiments may overstate advertising effects due to exaggerated compliance (attention to an ad), experimenter demand effects, or unmeasured decay [6, 7, 8].

Field experiments often address those measurement concerns by separating the delivery of advertising treatments from the collection of outcomes, either via survey or inspection of electoral returns. While classic work in this area found large but short-lived effects of ads in a gubernatorial election [9], a recent meta-analysis of the field experimental evidence concludes that campaign contact has very small effects on vote choice that mostly cannot be distinguished from zero [10]. A persistent worry, however, is that field experiments understate advertising effects because they measure the consequences of only small advertising doses delivered in competitive information environments. A prominent pollster put the critique crisply: "One group ate a single potato chip, the other had none. Each person was then retested. Would you expect to find that eating a single potato chip affected the health of your subjects [11]?"

Similar challenges arise when studying the questions of whether negative campaigns and political coverage demobilize the electorate. Negative advertisements are hypothesized by some to lower turnout, either because subjects are persuaded to abstain rather than vote for a criticized candidate or because subjects are turned off to politics in general. Using a regression discontinuity design, [12] find that exposure to partisan media increases turnout for in-partisans but decreases turnout for out-partisans. Some lab experiments have found that exposure to negative advertisements lowers reported intention to vote, and aggregate analysis of Senate races has found a negative association between tone and turnout [13], though other work has disputed the generalizability of those findings [14] (see also [15] for the authors' reply). Others claim to find the opposite: an observational analysis of Senate campaigns in 1990 finds that negative campaigning is associated with *higher* turnout [16]. Comparing survey respondents who reported television habits would have exposed them to more or less negative advertising, [17, 18] conclude that negative advertisements stimulate turnout. Still others claim to find no relationship at all between negative advertising volumes and turnout measured in surveys [19].

More recent observational work attempts to reconcile these conflicting results by suggesting that demobilization only occurs after a voter decides on a candidate to support, and only when negativity centers on the candidate [20]. Of course, much of this literature describes television advertising rather than digital ads, and the collection of studies cited here hardly covers the extensive literature on the effects of negative campaigning on turnout; for a review see [21].

Our design contributes to this literature by evaluating the differential mobilization hypothesis at scale with a massive campaign-level randomized field experiment. In particular, we measure the cumulative impact of an entire \$8.9 million digital advertising campaign aired in battleground states, with an average of 754 ad impressions per treated subject. Our experiment is notable for its sample size and sustained application/dosage of digital advertising.

Our results provide limited support for the differential mobilization hypothesis. According to our pre-registered regression specification to estimate conditional average treatment effects (CATEs), we find that the campaign increased voting among Democrats by 0.5 percentage points (SE: 0.5pp) and decreased voting among Republicans by 0.9 percentage points (SE: 0.5pp), for a difference in CATEs of 1.7 points ($t(254813) = -2.53$, $p = 0.012$, $\widehat{DIC} = 1.7$ points, 95% CI = $[-0.031, -0.00]$). However, party registration has low coverage in our sample, so we also esitmate this CATE using an imputed "Trump Support Score" (details below). Among Biden leaners (those with modeled Trump support scores between 30 and 40) by 0.4 percentage points and decreased voting among Trump leaners (Trump support score between 60 and 70) by 0.3 percentage points, resulting in a difference-in-CATEs of 0.7 percentage points ($t(1035571) = -2.09$, $p = 0.036$, $\widehat{DIC} = 0.7$ points, 95% CI = $[-0.014, -0.00]$).

Results

We present the main results of the campaign-level experiment in Figure 1. Focusing on the preregistered specification (middle column of facets), we find that the overall effect on turnout (ATE) is -0.06 percentage points, with a robust standard error of 0.12 points ($t(1999277) = -0.52$, $p = 0.60$, $\hat{ATE} = -0.0006$, 95% CI: [-0.0030, 0.0017]). Using a very narrow equivalence range (plus or minus one-third of a percentage point), we can affirm our overall estimate is effectively equivalent to zero using the Two One-Sided Tests (TOST) procedure ($p = 0.013$) [22, 23]. We also observe small conditional average effect estimates by age, gender, race, and vote margin in 2016 that are not statistically significant.

Turning next to the effects by partisanship and Trump support, we find evidence in favor of the differential mobilization hypothesis. Among those with Trump support scores between 60 and 70, the average effect is a 0.3 point decrease and among those with scores between 30 and 40, the effect is a 0.4 point increase. As shown in Figure 2, top row, the difference-in-CATEs by Trump support score group is 0.7 percentage points ($t(1035571) = -2.09$, $p = 0.036$, $\widehat{DIC} = 0.7$ points, 95% CI = $[-0.014, -0.00]$) according to the PAP adjustment set (see methods section for details). We find similar effects by party: the campaign increased voting among Democrats by 0.5 percentage points (SE: 0.5pp) and decreased voting among Republicans by 0.9 percentage points (SE: 0.5pp), for a difference of 1.7 points that is distinguishable from zero ($t(254813) = -2.53$, $p = 0.012$, $\widehat{DIC} = 1.7$ points, 95% CI = $[-0.031, -0.00]$).

Next we examine early and in-person voting in 2020. The next two rows in Figure 2 report estimates of the difference-in-CATEs. Here we see that differential effects of the program are stronger in our early voting data (1.0 percentage points favoring Biden) than in the in-person voting data (0.3 percentage points favoring Trump, linear hypothesis test of equality of the differences-in-CATEs, accounting for the covariance of the estimates: $F(1, 1035571) = 20.99$, $p = 0.000$).

An alternative test using a regression model across the entire range of the Trump Support Score variable reveals qualitatively similar if slightly larger results compared with the difference-in-CATEs analysis above. The bottom three panels of Figure 2 report the interaction term from OLS regressions that linearly interact Trump support with the treatment indicator.

While differential turnout with respect to Trump support is mild, it is strongest for early voting. The scale of the effects can be best appreciated by inspection of Figure 3. The horizontal axis arrays subjects by level of Trump support and the vertical axis shows average rates of turnout for early voting, in-person voting, and all voting. The plots show the CATE estimates within 1-point bin. The estimates themselves are somewhat imprecise, owing to the relatively small size of the control group (as reflected in the wider confidence intervals on the untreated means). Overlaid on the CATEs is the predicted average marginal effects from the PAP specification of the interaction term (see the bottom middle facet of Figure 2). We emphasize that this finding was not pre-registered and should be considered exploratory.

We discuss treatment assignment, balance, and example ads in the methods section below. A description of the treatment assignment process can be found in Figure 4 below. Figure 5 shows the differences in pre-treatment covariate means by treatment condition. Figure 6 shows example ads used in the manuscript.

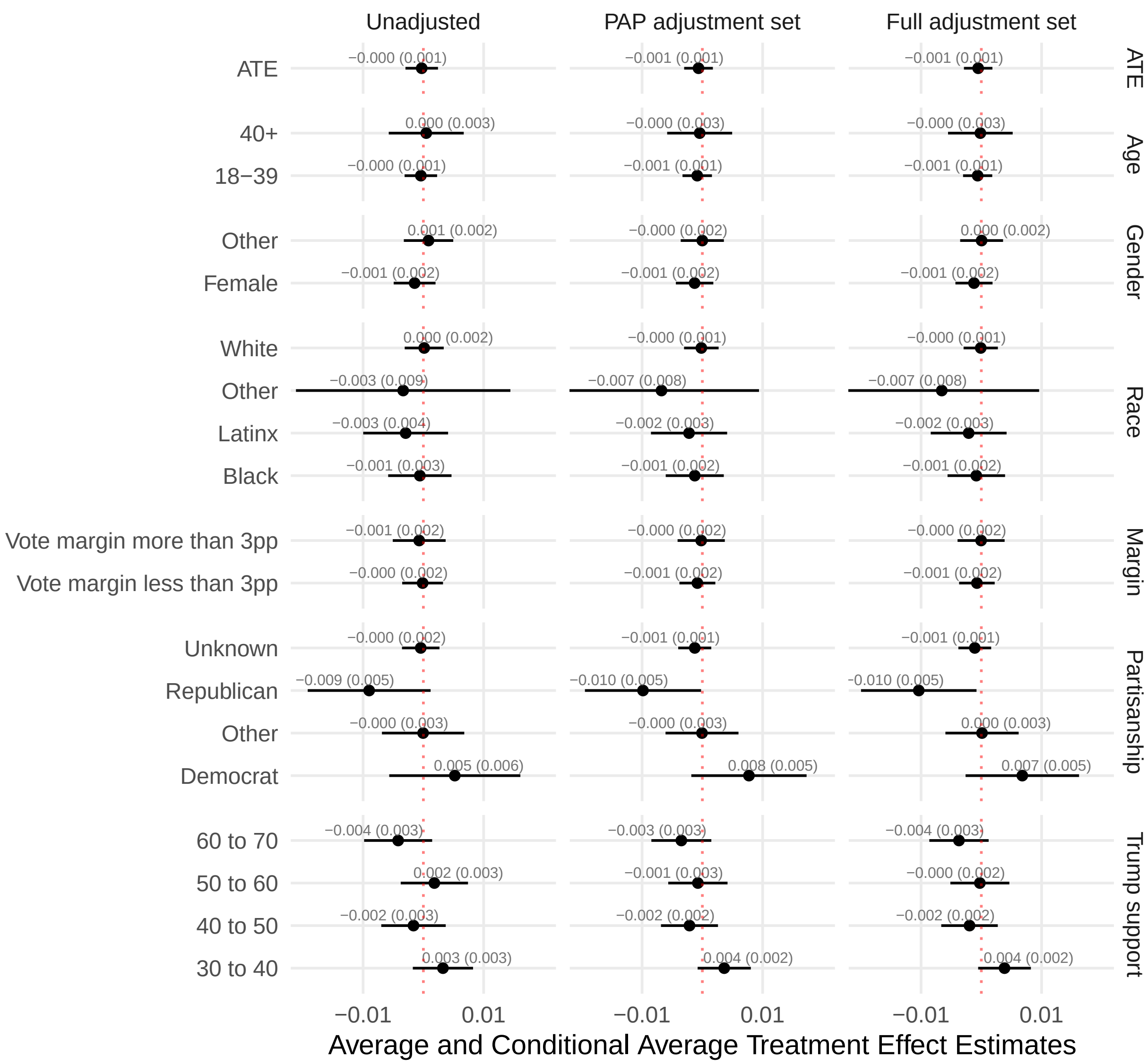


Figure 1: Average and conditional average treatment effects of treatment on 2020 turnout under three inverse probability weighted regression specifications: unadjusted, the PAP adjustment set, and the full adjustment set. Point estimates and standard errors are reported above each estimate. Error bars represent 95% confidence intervals. Inferential statistics for each estimate are reported in supplementary table 1. Ns: Total: 1,999,282. Age: 18-39: 1,379,017; 40+: 620,265. Gender: Female: 978,041; Other: 1,021,241. Race: Black: 233,546; Latinx: 179,036; White: 1,531,129; Other: 55,571. Margin: less than 3pp: 1,337,057; more than 3pp: 662,225. Partisanship: Democrat: 182,945; Republican: 71,875; Unknown: 1,442,071; Other: 302,391. Trump support: 30 to 40: 522,918; 40 to 50: 485,371; 50 to 60: 478,333; 60 to 70: 512,660.

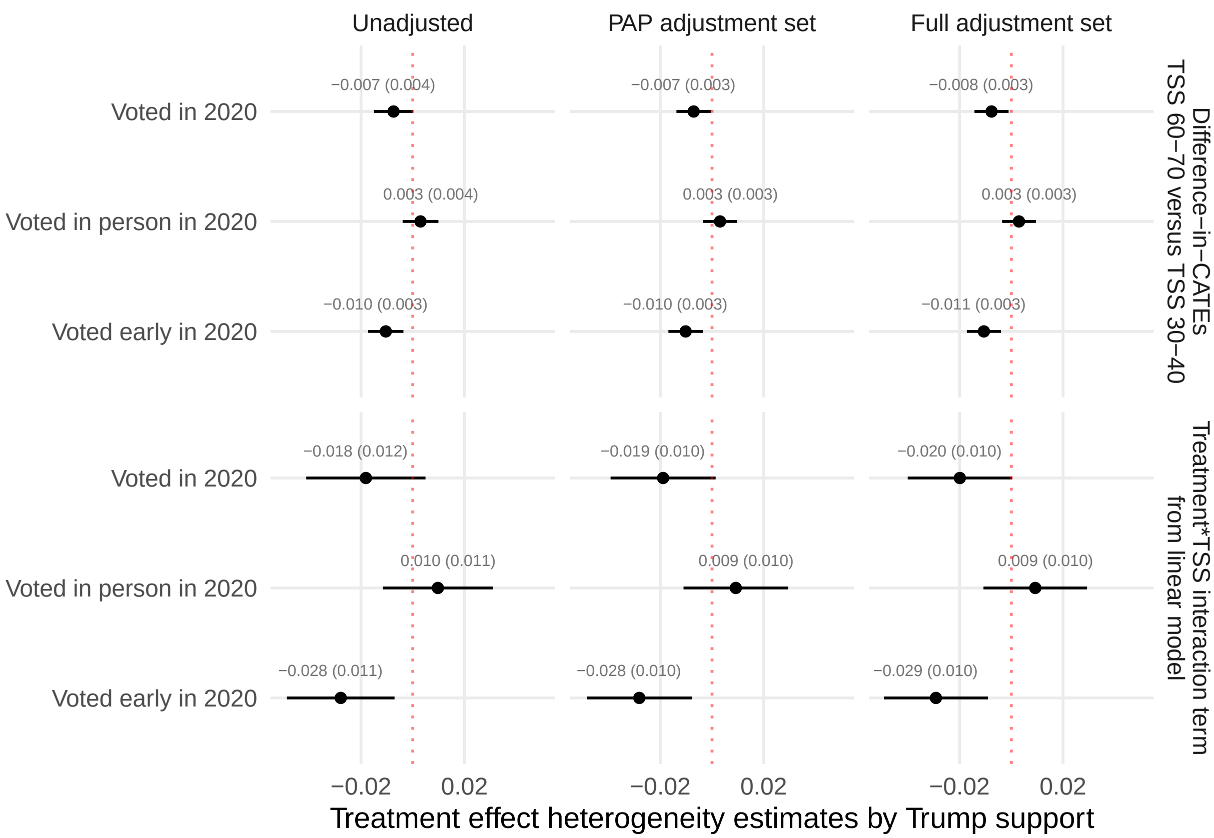


Figure 2: The heterogeneous effects of treatment on voting, in-person voting, and early voting by Trump support under three inverse probability weighted regression specifications: unadjusted, the PAP adjustment set, and the full adjustment set. Point estimates and standard errors are reported above each estimate. Error bars represent 95% confidence intervals. Inferential statistics for each estimate are reported in supplementary table 2. N for difference-in-cates estimates = 1,035,578. N for interaction estimates = 1,999,282.

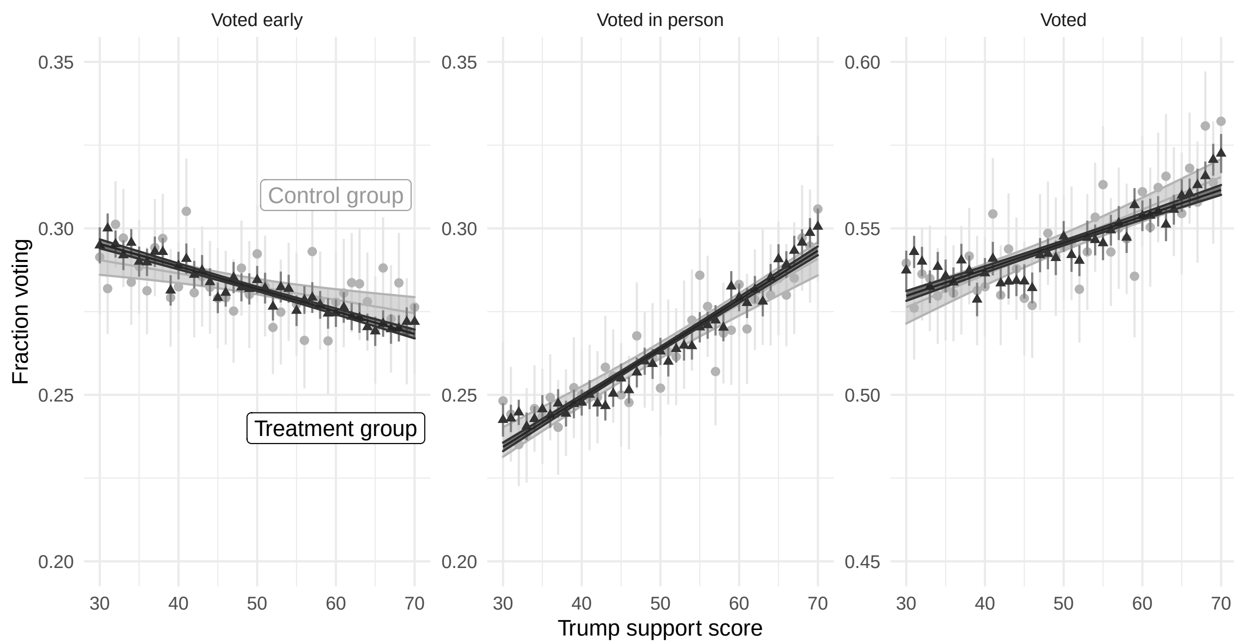


Figure 3: 2020 turnout rates by 1-point bins of Trump support score and condition in early voting (left), election day voting (middle), and voting regardless of mode (right). Error bars represent 95% confidence intervals. Linear predictions from the unadjusted models reported in the bottom left facet of Figure 2 overlaid on the binned means, with shaded 95% confidence regions. The vertical scales of all three facets cover a 15 percentage point range, but the ranges differ across facets to emphasize the relevant variation. N = 1,999,282.

Discussion

The experimental literature on the effects of advertising to date has relied on survey experiments that force exposure and measure outcomes immediately, and field experiments that deploy doses of treatment that some scholars and practitioners consider to be too small. The challenge in the study of political advertising is to amplify the intensity of treatment to politically meaningfully levels in the field, while maintaining experimental control. In our view, our design does not suffer from the "potato chip" critique described in the introduction. In this project, we randomized exposure to the full weight of an 8-month digital advertising campaign, deployed in real-time during a contentious election season in battleground states.

What do we learn from this design? First, we provide evidence that persuasion campaigns can indeed cause small differential turnout effects—much smaller than pundits and media commentators often assume, but our field experimental study is large enough to show those effects are distinct from zero. We find both small mobilizing effects among Biden leaners and small demobilizing effects among Trump leaners. These results shed light on the long-hypothesized causal connection between messages critical of a preferred candidate and decreased turnout.

Second, the strongest differential effects appear in our *early voting* data, suggesting that it may be increasingly important to advertise early in a post-COVID electoral environment. Because Acronym began its advertising campaign far earlier than other political organizations, our data are particularly well-suited to help shed light on this question. We know from prior research that ads airing closer to election day tend to have weaker effects, possibly due to saturation and people having already decided how to vote [10]. At the same time, decay effects [9] mean that early ads should be expected to have weaker effects over time. Thus, the program's early ads may have had a stronger effect on early voting because they reached voters before they made up their minds, before the media environment became completely saturated, and/or before the effects decayed.

Third, this notable dose of advertising had no overall effect on turnout on average. While we test the full weight of an entire advertising campaign, the fact that this campaign took place in battleground states during a presidential campaign means that exposure to background political advertising and media was relatively high. Acronym's campaign increased that exposure, but not as much as might be expected in a less-visible election.

We can interpret these small differential turnout effects in two ways. Under one interpretation, the difference reflects persuasion: the advertisements may have lowered evaluations of Trump so much that voters who initially leaned towards voting for him abstained in the end. Under an alternative interpretation, the differential turnout effects might reflect decreased levels of enthusiasm among Trump leaners but increased levels among Biden leaners, without shifting evaluations of Trump. We cannot assess the effects of the treatment on vote choice in this study so our design cannot distinguish between these two channels, though we think either or a combination is plausible.

Zooming out to the broader implications of our study, our results suggest that influencing voter turnout in presidential elections via digital advertisements is expensive. An apples-to-apples comparison to persuasion campaigns in other media is not possible as we lack the requisite information on vote choice. Given our findings, however, the popular narrative that Russia's $150,000 Facebook ad expenditure in 2016 could have caused enough differential turnout to affect the outcome of the election is implausible [24]. This campaign-level field experiment, conducted over eight months in five battleground states, shows that digital advertisements yielded small returns for presidential campaigns in the 2020 general election. We extrapolate from our study is that future digital campaigns of similar scope and size will yield similarly small returns. The quality of this generalization may depend on a number of factors, including how engaging the treatments are (to ensure compliance) and proximity to the election (to mitigate decay).

It is worth noting that this campaign had an emphasis on promoted news stories, alongside conventional political video ads. Past survey and observational research finds that emotional appeals may play an important role in political persuasion [25], in part because emotional appeals can be more memorable [26]; however, more recent large-scale experimental evidence shows the critical importance of information gain in persuasion—in other words, providing a low-information audience with information-rich persuasive content is the most powerful way to influence political behavior in the aggregate [27].

One reasonable question for our study is how well our findings would generalize to voters who were not eligible for our experimental program or to other electoral contexts. Conventional wisdom holds that individuals with higher levels of support for one candidate or another are harder to persuade and mobilize than those in the middle, suggesting that any turnout effects should expected to be smaller among those not eligible for the program. By contrast, the differential mobilization hypothesis holds we should expect larger demobilization effects among those with stronger attachments to the criticized candidate. Our speculation is that effects would be similarly small at the extremes along the full range of Trump support, though of course we cannot know for sure. Turning to questions of context, it could be that the 2020 election was exceptional because of COVID and the idiosyncrasies of the candidates, so perhaps digital advertising would have larger effects in more typical settings. We can see the logic of this speculation, though our beliefs about "larger effects" are on the scale of single percentage points, not three or five. Ultimately, learning the answers to these generalizability questions will require further experimentation.

Materials and Methods

Research Origins, Processes, and Ethics The experimental design described herein was originally conceived in order to allow Acronym—a prominent left-leaning 501(c)(4) non-profit organization—to gauge the overall impact of its "Soften the ground (STG) persuasion" advertising program for business reporting purposes. To accomplish this, the organization created a "holdout group" – a randomly assigned set of people who were not exposed to any of Acronym's ads. The research team (including individuals authoring this manuscript) designed and implemented the holdout, and helped administer the holdout group, collect and curate the data, and conduct the analysis below.

The research team (with the exception of Coppock) contributed to the design and/or implementation of the experiment while employed by Acronym from January 2020 to January 2021. After the termination of their employment in January 2021, but prior to working on this manuscript, the other primary authors signed explicit third party data sharing agreements in February, 2021 to allow data access while they conducted the scientific analysis and reporting presented here, with other institutions.

Yale University's Institutional Review Board reviewed this research and issued a waiver because the data were collected by a third-party, non-academic organization. In the absence of this research, all people in this study would have received the treatment. The "intervention" in this case was to remove a random subset of voters from the treatment program. The main ethical consequence of the research activity was that some subjects were not delivered Acronym's advertising but instead were delivered whichever ads that Facebook and the other ad platforms might have chosen to show them instead (See Supplementary Section B for a description of the political advertising environment experienced by the control group).

Preregistration and PAP deviations The research team pre-registered the design with OSF `https://osf.io/3evfp` in November 2020. We report all analyses in that pre-registration. However, because our primary interest is differential turnout, and because we lack party registration information for much of the sample, we submitted an update in December to examine turnout by Trump Support Score, which has full coverage (`https://osf.io/jkush/`). This occurred *after* seeing early vote data but *before* seeing final turnout data.

Furthermore, in the preregistration, we describe one regression specification that includes controls for Trump support score, Presidential turnout score, and a count of vote history. We report that specification in the main text and two others: an unadjusted specification and a fuller specification that includes Trump support score, Presidential turnout score, strata fixed effects, indicators for voting in any even-year election between 2000 and 2018, party membership indicators (Republican, Democrat, or Unknown, relative to Other).

Subgroups In the PAP, we specified that we would consider treatment effect heterogeneity by Age, Race, Gender, 2016 Vote Margin, and Party registration. We do report all those analyses but because we lack party registration information for much of the sample, we also included heterogeneity analyses by Trump support score, which is available for all subjects. We submitted an update to the registration *after* seeing early voting data (but before seeing final turnout data) from the voter file to use Trump Support Score (TSS) instead of party registration (`https://osf.io/jkush/`)

Analysis In the PAP, we describe one regression specification that includes controls for Trump support score, Presidential turnout score, and a count of vote history. We report that specification in the main text and two others: an unadjusted specification and a fuller specification that includes Trump support score, Presidential turnout score, strata fixed effects, indicators for voting in any even-year election between 2000 and 2018, party membership indicators (Republican, Democrat, or Unknown, relative to Other).

Field experimental design In the eight months leading up to the 2020 Presidential election, Acronym, a prominent left-leaning non-profit organization, conducted an $8.9 million digital messaging persuasion campaign with the intention of reducing support for Donald Trump and increasing support for Joe Biden in five battleground states: Arizona, Wisconsin, Michigan, North Carolina, and Pennsylvania. The campaign ran paid advertising on Facebook, Instagram, and Outbrain ad networks.

The full experimental design is visualized in the diagram shown in Figure 4. In February of 2020, eligible subjects were randomly assigned to a treatment group that received the messaging program or to a "hold out" control group of subjects who were never shown any Acronym advertising for the whole of the 2020 presidential campaign. The random assignment process was unusual and oversampled specific subgroups into the control group. First, a sample from the total population of registered voters was drawn, then successive samples from important subgroups (young people, Black and Latinx voters, and women) were drawn with replacement. A voter was assigned to the control group if sampled at one or more of these steps. The assignment process results in 18 demographic strata, with probability of assignment to treatment shown in Table 1.

This assignment process differs subtly from standard block random assignment in which fixed numbers of units in each block are assigned to treatment or control. Here, because of the overlapping sample draws, the number of units assigned to treatment in each stratum could differ across realizations of the randomization. This procedure still generates unbiased treatment effect estimates but is likely higher variance than standard block random assignment. Moreover, we are missing race data for 4% of the sample where race was "uncoded." At the time of random assignment, Acronym used information from a third-party data provider to infer race categories for these individuals, but these data are no longer available due to the provider's dynamic prediction models. As a result, we cannot calculate the denominators in steps 4 and 5 of the assignment procedure shown in Figure 4. We therefore estimate the probabilities of assignment from fraction treated in each group. While these assignment probability estimates are not exact, they are unbiased and quite precise.

Acronym employed specific treatment targeting criteria to focus on a subset of centrist voters thought to be persuadable. In particular, they directed advertising to voters modeled to have mid-range Trump support scores (TSS) and turnout scores. In July of 2020, they narrowed the criteria to exclude voters with above median political knowledge. We restrict both the treatment and control groups on the basis of all these criteria.

Due to an oversight during the implementation of the design we further restrict our analysis to subjects between 18 and 55 years old, those with a college education score below 50, and a presidential turnout score above 20. After random assignment, the IDs of voters assigned to the control group were saved, but the full set of treatment IDs were not. Fortunately, an Acronym employee happened to save an exact list of treatment group subjects comprising the subset above in February of 2020. By starting from the voter file and filtering on all relevant variables, we were able to recover all treatment and control IDs in this subset, resulting in an analysis sample of 1,999,282 subjects. The reason for this oversight was operational: Acronym planned to deliver advertisements to everyone who met their criteria except those held out to form the control group, so they had no special need to keep a separate list of treatment IDs. Over the campaign, the organization made significant changes to its data systems. Important data describing the treatment group and treatment strata were lost, most notably, some imputations of race and Hispanic ethnicity. Despite extensive effort, our attempts to exactly reconstruct these strata and thus the full dataset were unsuccessful, as indicated by significant imbalance on pre-treatment covariates in the full dataset. Thus, we analyze the "archival subset" below.

As shown in Figure 5, the random assignment procedure generated treatment and control groups that are balanced on pre-treatment characteristics in this subset. Out of 21 covariates, only 2 exhibit statistically significant imbalance (Trump support score, $t(1999280) = -2.260$, $p = 0.024$, $\hat{ATE} = -0.001$, 95% CI = [-0.001, -0.000]; TSS between 60 and 70, $t(1999280) = -2.609$, $p = 0.009$, $\hat{ATE} = -0.003$, 95% CI = [-0.006, -0.001]). Neither of these estimates remains significant after a Benjamini-Hochberg correction [28] to control the false discovery rate (p = 0.262 and 0.200,

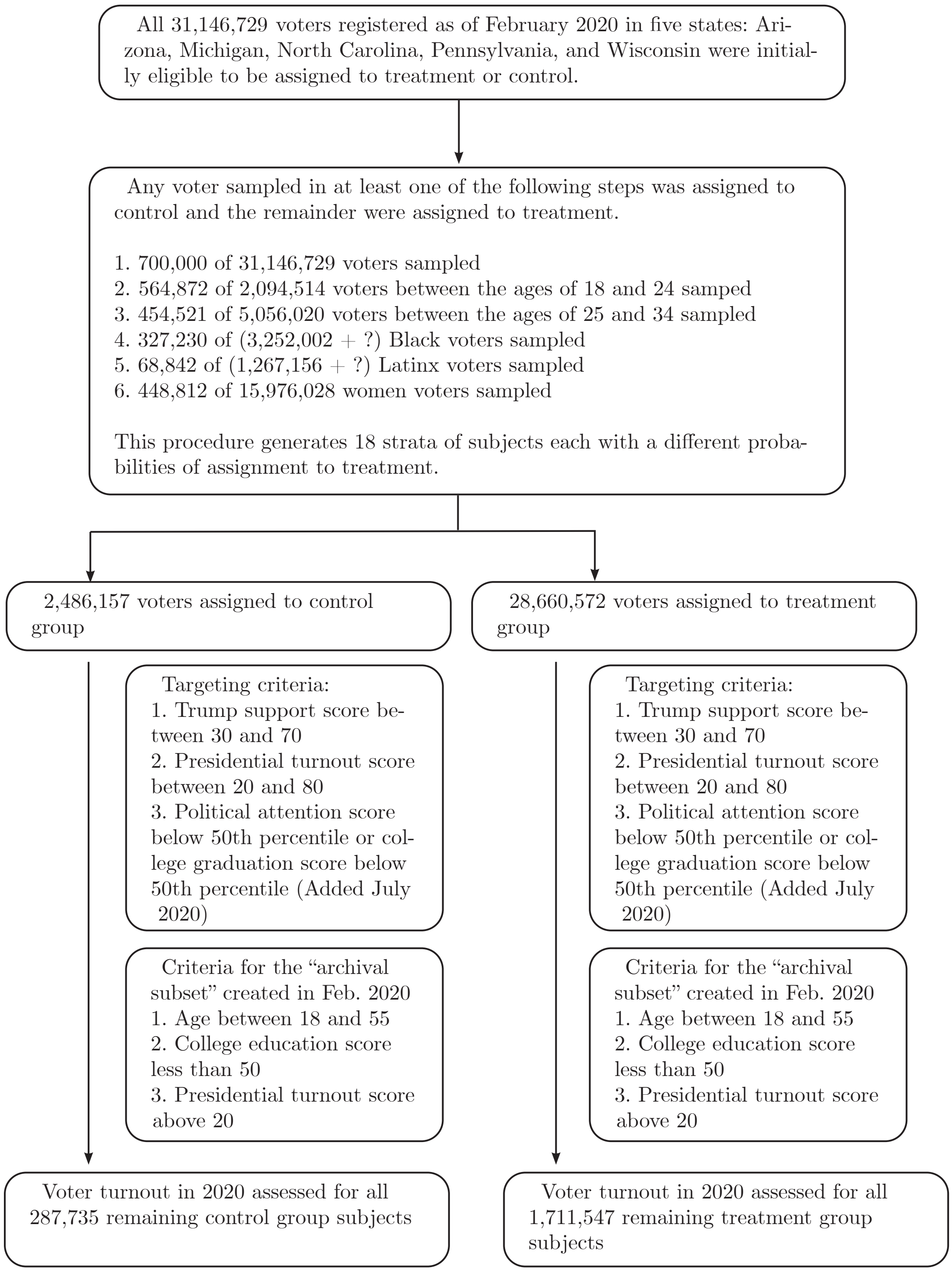


Figure 4: Diagram describing the experimental sampling, assignment, and measurement procedures.

Table 1: Experimental Strata

			Group size			2020 Voting Rate	
Gender	Race	Age	Control	Treatment	Pr(treat)	Control	Treatment
Female	Black	18-24	4,508	7,605	0.628	48.0%	46.4%
Female	Black	25-34	3,473	11,952	0.775	42.2%	41.3%
Female	Black	Other	4,564	26,783	0.854	51.4%	51.2%
Female	Latinx	18-24	4,831	9,741	0.668	52.2%	51.7%
Female	Latinx	25-34	3,074	14,743	0.827	46.0%	45.0%
Female	Latinx	Other	2,727	26,699	0.907	50.7%	50.8%
Female	Other	18-24	60,999	139,081	0.695	66.0%	66.2%
Female	Other	25-34	34,406	219,742	0.865	57.0%	57.0%
Female	Other	Other	19,998	383,115	0.950	62.8%	62.5%
Other	Black	18-24	14,004	25,313	0.644	36.9%	36.9%
Other	Black	25-34	11,122	44,850	0.801	27.8%	28.1%
Other	Black	Other	9,501	69,871	0.880	36.7%	36.9%
Other	Latinx	18-24	11,145	23,709	0.680	39.3%	39.2%
Other	Latinx	25-34	5,974	34,290	0.852	33.1%	33.1%
Other	Latinx	Other	2,835	39,268	0.933	42.6%	42.0%
Other	Other	18-24	60,726	151,473	0.714	57.9%	57.8%
Other	Other	25-34	28,067	227,248	0.890	48.3%	48.1%
Other	Other	Other	5,781	256,064	0.978	57.9%	58.6%
		Total	287,735	1,711,547	0.856	54.6%	54.6%

respectively). Further, when we assess whether we can predict treatment assignment from the covariates, we find that we cannot. An F-test comparing a regression of treatment status on covariates with strata fixed effects versus a restricted model predicting treatment status from strata fixed effects is non-significant ($F(1999264, -18) = 1.0266$, $p = 0.425$). This omnibus test gives us confidence that within this subset, the experimental design works as expected. The main way in which the subset differs from the full sample is that it unfortunately excludes voters over 55.

Advertising campaign and content The messaging program consisted of 536 unique paid ad spots. These ads were largely comprised of promoted news—social media posts with links to news articles that render with branding and formatting from the originating news source—and more traditional video and info-graphic ads included prior to August. Examples of a typical promoted news ad and a typical traditional video ad can be found in Figure 6. Acronym produced all ad-spots, conducted audience targeting, and purchased all ad-inventory for the program. Further examples of treatment stimuli can be found on our OSF site here: `https://osf.io/ex3kq/`

Using the Facebook Ad Library API, we are also able to calculate the lower bound of spending by Acronym on ads containing the words "Biden" or "Trump". Over experimental period, ACRONYM spend at minimum $368,800 on ads containing the word "Biden", $3,254,600 on ads containing the word "Trump", and $244,500 on ads containing both words. Turning to ad formats, Acronym spend $1,275,000 on promoted news ads, $2,288,900 on video ads, and $304,000 on other ad formats (e.g. images). While these numbers only represent a minimum spend due to limitations of Facebook's API and further do not include spending on Instagram and Outbrain, they are representative of Acronym's advertising focus. A more detailed description of the content in the program can be found in Supplementary Figure 1.

The program content changed over time. Early in the campaign, the treatment ads were a mix of promoted news and traditional video ads, but after August 2020, the treatment persuasion program switched to almost entirely promoted news content. Early in the campaign, the ads were mostly Anti-Trump but later ads were a mix of Anti-Trump and Pro-Biden content (Supplementary Figures 1 and 2).

Using data provided by the Wesleyan Media Project [29], we show that since Acronym's spending started earlier than many other political actors, its share of the total volume of political advertising on Facebook was higher earlier in the

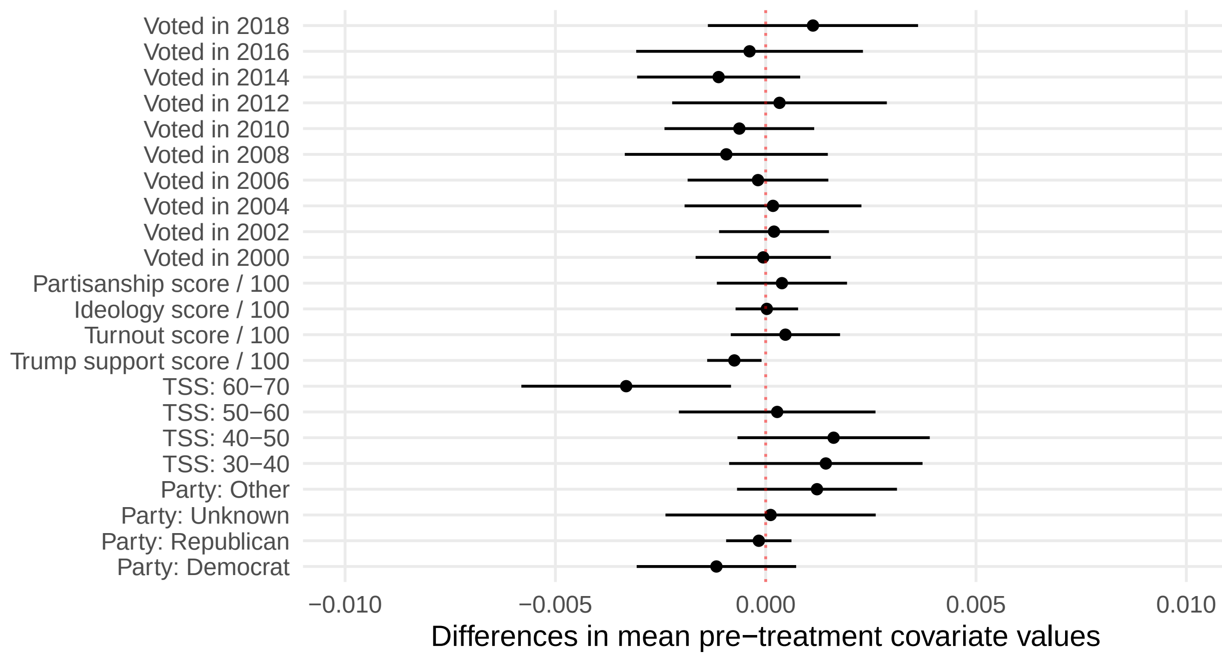


Figure 5: Balance on pre-treatment covariates. Point estimates from IPW regressions of the covariate on the treatment indicator. Error bars represent 95% confidence intervals. Inferential statistics reported in supplementary table 3. N = 1,999,282.

campaign than near Election Day.

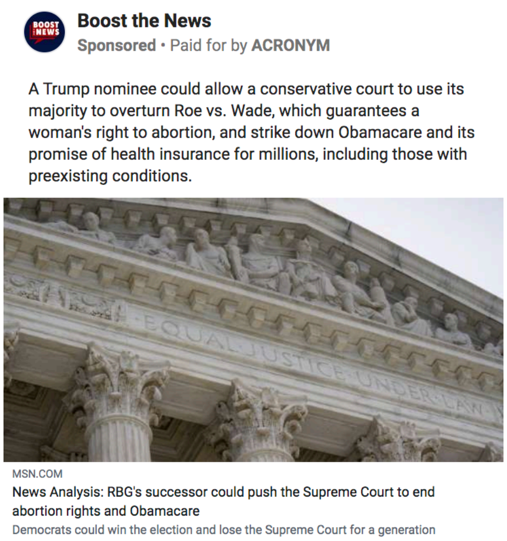
Political activists have long leveraged news to achieve political ends. Recent work has shown just how powerfully the set of news we consume can affect attitudes and beliefs—when a large audience of Fox News viewers were paid to watch CNN for a month instead, they responded with less conservative answers to questions about an incoming Democratic administration and public health issues [30]. Promoted news in digital political campaigns has been a tactic since at least 2014, when the House GOP created a network of "local news" domains and promoted them using Google ads [31]. In 2018, "Well News" promoted stories about prominent Blue Dog democrats [32]. The use of promoted news ads in political campaigns was common enough in 2020 that Facebook went out of its way to clarify that its political ad ban that year (1-week before and in the weeks after the election) applied to its promoted news ad product [33].

While there are similarities between promoted news and conventional social media ads, one key distinction advertisers often point to is that the "messenger" in promoted news is a trusted news source rather than a political campaign. For example, Working America ran a boosted news study during the 2020 election, and found that promoted news was as effective as traditional ad copy—and in fact *more effective* among Working America union members, a difference they suggest is due to source cues [34].

Dosage and treatment delivery We gain some appreciation for the dosage of the treatment by considering how subjects interacted with the treatment materials. For the promoted news ads appearing on Facebook from May 5 through election day, we find a click-through-rate of 1%. The video view rate (which is defined as the fraction of videos that play for at least two seconds while at least halfway on the screen) was 53%. Even if subjects did not click the ad or watch the video, however, they were nevertheless exposed to headlines, photographs, and accompanying text (see Figure 6).

As is usually the case with targeted digital campaigns, not all treatment group subjects could be successfully identified and served ads. Facebook reported that 60% of our treatment group was successfully matched, but does not reveal which units were and were not matched for privacy reasons. In the supplementary information, we describe an exploratory analysis that suggests that the 60% matched likely includes some false positives. Formally, this implies that our experiment encountered two-sided noncompliance: a large fraction of the assigned treatment group was untreated and small fraction of the assigned control group was probably treated. We conduct all our analyses according to the intention-to-treat principle.

One potential concern is that even though the treatment group was exposed to more political advertising than the control group, the control group was nevertheless exposed to some. We think of this issue as a further manifestation of treatment noncompliance. Since political advertising is a small fraction of overall advertising (an estimated 3 percent of Facebook's Q3 revenue in 2020) [35], exposure in the control group was likely to be small, at until the final weeks before Election Day.



(a) A promoted news ad

(b) A traditional video ad

Figure 6: Examples of typical ad content run in Acronym's persuasion program

The average matched subject received 754 ad impressions over the eight full months between March 2020 and Election Day. In comparison with most field experimental investigations of the effects of political advertisements, this intervention represents a large dose of pro-Biden, anti-Trump information. The full cost of the advertising campaign was $8.9 million dollars spread out over a treatment audience of 1,993,216 million (3,322,027 program-eligible voters $\times$ our 60% match rate), amounting to $4.46 of advertising expenditure per voter.

Statistical software We implement all data processing and analysis in `R` (4.1.1) [36]. For data cleaning, processing and data visualization, we use `tidyverse` (1.3.1) [37]. For all models, we estimate HC2 robust standard errors, implemented using `estimatr` (0.30.6) [38]. To compare early voting effects to election day effects, we use the `linearHypothesis` function in the `car` (3.1.0) [39].

Data availability

An anonymized replication dataset is available via Dataverse at the following URL `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/YMKVA1`. TargetSmart generously agreed to make replication data available for this paper. By downloading replication data, researchers agree to use the data only for academic research, agree not to share the data with outside parties, and agree not to attempt to re-identify individuals in the data set in order to download the data.

Code availability

Replication scripts are available via Dataverse at the following URL `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/YMKVA1`.

Acknowledgments

We thank Don Green, Fredrik Sävje, Melissa Michelson, Betsy Sinclair, Ethan Porter, Lynn Vavreck, Yamil Velez, Josh Kalla, and David Broockman for generous early feedback. We also thank Erika Franklin Fowler and the Wesleyan Media Project for generously sharing data. The authors received no specific external funding for this work. However, Acronym played a role in conceptualizing and designing the study presented here—all of the authors except Coppock contributed to the design and/or implementation of the experiment while employed by Acronym from January 2020 to January 2021. After the termination of their employment in January 2021, but prior to working on this manuscript, the other primary authors signed explicit third party data sharing agreements on February 8, 2021 to allow data access while they conducted the scientific analysis and reporting presented here, with other institutions. Acronym and authors agreed in writing to the publication of this manuscript in advance of manuscript analysis and preparation. Authors agreed to provide Acronym with a draft of the manuscript prior to publication.

Author Contributions Statement

All of the authors except Coppock contributed to the design and/or implementation of the experiment while employed by Acronym from January 2020 to January 2021. After the termination of their employment in January 2021, but prior to working on this manuscript, the other primary authors signed explicit third party data sharing agreements on February 8, 2021 to allow data access while they conducted the scientific analysis and reporting presented here, with other institutions.

SM and JB conceived of the initial research. SM designed and implemented the original experiment. AB and HH administered the experiment. DF, MA, KZ, SZ, JA, and SM contributed to data collection and curation. MA, JA, and AC analyzed the results of the experiment. AC and SM drafted the initial manuscript and figures. All authors contributed to the revision and editing of the manuscript.

Competing Interests Statement

All researchers were employed by Acronym or contractors thereof during the 2020 election cycle. JA, SZ, and HH have a significant financial interest in Facebook.

Acronym played a role in conceptualizing and designing the study presented here—all of the authors except Coppock contributed to the design and/or implementation of the experiment while employed by Acronym from January 2020 to January 2021. After the termination of their employment in January 2021, but prior to working on this manuscript, the other primary authors signed explicit third party data sharing agreements on February 8, 2021 to allow data access while they conducted the scientific analysis and reporting presented here, with other institutions. Acronym and authors agreed in writing to the publication of this manuscript in advance of manuscript analysis and preparation. Authors agreed to provide Acronym with a draft of the manuscript prior to publication.

References

[1] Romer, D., Kenski, K., Winneg, K., Adasiewicz, C. & Jamieson, K. H. *Capturing campaign dynamics, 2000 and 2004: The national Annenberg election survey* (University of Pennsylvania Press, 2006).

[2] Sides, J., Vavreck, L. & Warshaw, C. The effect of television advertising in united states elections. *American Political Science Review* **116**, 702–718 (2022).

[3] Druckman, J. N. The Implications of Framing Effects for Citizen Competence. *Political Behavior* **23**, 225–256 (2001).

[4] Druckman, J. N. & Holmes, J. W. Does Presidential Rhetoric Matter? Priming and Presidential Approval. *Presidential Studies Quarterly* **34**, 755–778 (2004).

[5] Coppock, A., Hill, S. J. & Vavreck, L. The small effects of political advertising are small regardless of context, message, sender, or receiver: Evidence from 59 real-time randomized experiments. *Science advances* **6** (2020).

[6] Chong, D. & Druckman, J. N. A theory of framing and opinion formation in competitive elite environments. *Journal of communication* **57**, 99–118 (2007).

[7] Zaller, J. R. *The Nature and Origins of Mass Opinion* (Cambridge University Press, 1992).

[8] Vavreck, L. *The Message Matters: The Economy and Presidential Campaigns* (Princeton University Press, Princeton, NJ, 2009).

[9] Gerber, A. S., Gimpel, J. G., Green, D. P. & Shaw, D. R. How large and long-lasting are the persuasive effects of televised campaign ads? results from a randomized field experiment. *American Political Science Review* **105**, 135–150 (2011).

[10] Kalla, J. L. & Broockman, D. E. The minimal persuasive effects of campaign contact in general elections: Evidence from 49 field experiments. *American Political Science Review* **112**, 148–166 (2018).

[11] Mellman, M. Are we doing it all wrong? *The Hill* (2017). URL `https://thehill.com/opinion/campaign/353720-mark-mellman-are-we-doing-it-all-wrong`.

[12] Spenkuch, J. L. & Toniatti, D. Political advertising and election results. *The Quarterly Journal of Economics* **133**, 1981–2036 (2018).

[13] Ansolabehere, S., Iyengar, S., Simon, A. & Valentino, N. Does attack advertising demobilize the electorate? *American political science review* **88**, 829–838 (1994).

[14] Wattenberg, M. P. & Brians, C. L. Negative campaign advertising: Demobilizer or mobilizer? *American political science review* 891–899 (1999).

[15] Ansolabehere, S. D., Iyengar, S. & Simon, A. Replicating experiments using aggregate and survey data: The case of negative advertising and turnout. *American political science Review* **93**, 901–909 (1999).

[16] Kahn, K. F. & Kenney, P. J. Do negative campaigns mobilize or suppress turnout? clarifying the relationship between negativity and participation. *The American Political Science Review* **93**, 877–889 (1999). URL `http://www.jstor.org/stable/2586118`.

[17] Freedman, P. & Goldstein, K. Measuring media exposure and the effects of negative campaign ads. *American Journal of Political Science* **43**, 1189–1208 (1999). URL `http://www.jstor.org/stable/2991823`.

[18] Goldstein, K. & Freedman, P. Campaign advertising and voter turnout: New evidence for a stimulation effect. *Journal of Politics* **64**, 721–740 (2002). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/0022-3816.00143. https://onlinelibrary.wiley.com/doi/pdf/10.1111/0022-3816.00143.

[19] Finkel, S. E. & Geer, J. G. A spot check: Casting doubt on the demobilizing effect of attack advertising. *American Journal of Political Science* **42**, 573–595 (1998). URL http://www.jstor.org/stable/2991771.

[20] Krupnikov, Y. When does negativity demobilize? tracing the conditional effect of negative campaigning on voter turnout. *American Journal of Political Science* **55**, 797–813 (2011).

[21] Lau, R. R. & Rovner, I. B. Negative campaigning. *Annual review of political science* **12**, 285–306 (2009).

[22] Schuirmann, D. J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics* **15**, 657–680 (1987).

[23] Lakens, D. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science* **8**, 355–362 (2017).

[24] Politico. Facebook: Russian-linked accounts bought $150,000 in ads during 2016 race. URL https://www.politico.com/story/2017/09/06/facebook-ads-russia-linked-accounts-242401.

[25] Redlawsk, D. *Feeling politics: Emotion in political information processing* (Springer, 2006).

[26] Albertson, B., Dun, L. & Kushner Gadarian, D. The emotional aspects of political persuasion. *The Oxford handbook of electoral persuasion* 169 (2020).

[27] Broockman, D. & Kalla, J. When and why are campaigns' persuasive effects small? evidence from the 2020 us presidential election. *OSF* (2020).

[28] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289–300 (1995).

[29] Neumann, M. *et al.* Identifying and Classifying Federal Election Activity in Online Ads. *Working paper* .

[30] Broockman, D. & Kalla, J. The manifold effects of partisan media on viewers' beliefs and attitudes: A field experiment with fox news viewers (2022). URL osf.io/jrw26.

[31] House gop campaign arm created misleading fake news sites. *Talking Points Memo* (2014). URL https://talkingpointsmemo.com/livewire/nrcc-fake-news-sites.

[32] Blue dog-affiliated website blurs line between newsroom and political advertiser. *Sludge* URL https://readsludge.com/2021/09/17/blue-dog-affiliated-website-blurs-line-between-newsroom-and-political-advertiser/.

[33] Fischer, S. Facebook says new pre-election political ad rules apply to boosted posts. *Axios* (2020). URL https://www.axios.com/facebook-says-new-pre-election-political-ad-rules-apply-to-boosted-posts-ecbab5e5-48c8
html.

[34] 2020: Working america. *Fight for a Better America* (2020). URL https://www.fightforbetter.org/workingamerica.

[35] Levy, A. & Rodriguez, S. Why political campaigns are flooding facebook with ad dollars (2020). URL https://www.cnbc.com/2020/10/08/trump-biden-pacs-spend-big-on-facebook-as-election-nears.html.

[36] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2022). URL https://www.R-project.org/.

[37] Wickham, H. *et al.* Welcome to the tidyverse. *Journal of Open Source Software* **4**, 1686 (2019).

[38] Blair, G., Cooper, J., Coppock, A., Humphreys, M. & Sonnet, L. *estimatr: Fast Estimators for Design-Based Inference* (2022). URL `https://CRAN.R-project.org/package=estimatr`. R package version 0.30.6.

[39] Fox, J. & Weisberg, S. *An R Companion to Applied Regression* (Sage, Thousand Oaks CA, 2019), third edn. URL `https://socialsciences.mcmaster.ca/jfox/Books/Companion/`.

[40] What To Know About Facebook Advertising Around the Election. *Meta for Business* URL `https://www.facebook.com/business/news/facebook-ads-restriction-2020-us-election`.

[41] Ridout, T. N., Fowler, E. F. & Franz, M. M. Spending Fast and Furious: Political Advertising in 2020. *The Forum* **18**, 465–492 (2021). URL `https://www.degruyter.com/document/doi/10.1515/for-2020-2109/html`. Publisher: De Gruyter.

Supplementary Information

Contents

A Matching treatment voters to Facebook users

In order to deliver advertising content to subjects in the treatment group, Acronym uploaded personally identifiable information (PII) from the voter file to Facebook's Custom Audience targeting system. Facebook conducts a "waterfall match," an algorithm that attempts to find Facebook users using the data in the order *provided by the advertiser*, stopping once a match is made. For example, if email is provided first and is sufficient to indicate a match, Facebook does not use secondary, tertiary, or other subsequent fields in the matching process.

To better understand match quality, we leverage the fact that the order in which advertisers provide data fields to Facebook can dramatically change the mechanics of the matching process. While audience match rates were always around 60%, regardless of PII scheme utilized, we hypothesized that prioritization of some types of PII – namely wireless phones and personal emails – would lead to higher match precision. We evaluated match precision by stratifying survey recruitment, targeting by VF age groups, and comparing stratified age group labels to self-reported age from surveys. As we predicted, we found that voter file age and self-reported age were more likely to match when wireless phones and personal email addresses were evaluated earlier in the waterfall matching process. This pattern was especially strong for younger voters. We saw an 82% overall match rate when wireless phones and personal emails were prioritized versus an 81% match rate when name and address were prioritized; for younger voters, the difference was 64% versus 59%. The lack of age data consistency, which ranged from 59 - 82%, suggests at least some degree of false-positive matches. In principle, control units could have be inadvertently received our ads who we did not intend to target.

Formally, this matching process induces two-sided noncompliance: some treatment group units were not matched, so did not receive treatment and some control units could have been matched, so did receive treatment. Our best guess is that being assigned to the treatment group causes a 60 percentage point increase in the probability of being treated, which is important because even with some noncompliance, our assignment to treatment still causes big shifts in who is actually treated. We conduct all analyses according to the intention to treat principle.

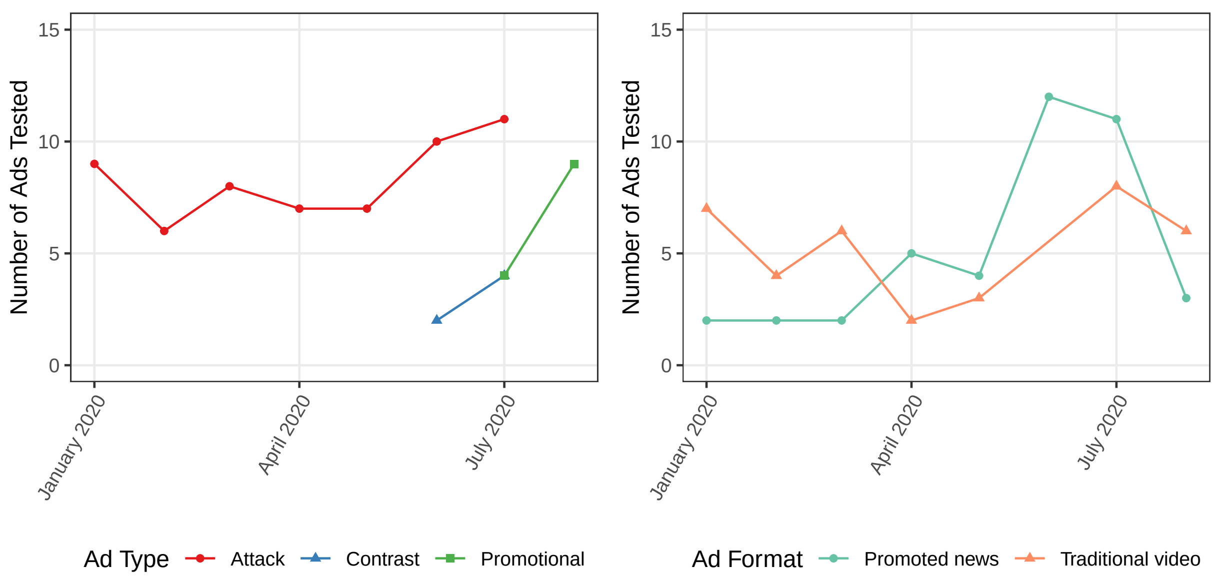
B Description of Ad Content

We describe two separate analyses meant to shed light on the content included in Acronym's advertising program. The analyses show that the program leaned primarily on anti-Trump attack ads, especially early in the year before Joe Biden was the presumptive Democratic nominee, and that Acronym employed a mix of promoted news and traditional video formats, shifting to more news ads closer to the election.

B.1 Acronym Ad Testing Program

Acronym ran in-house message testing of its ads on Facebook and Amazon Mechanical Turk in order to determine which ads were most effective at lowering Trump approval and increasing Biden approval. The ad treatments tested in this program can be found on our OSF site: `https://osf.io/ex3kq/`. These tests informed the allocation of spending in the broader Acronym persuasion program, and, except for a few ads which appeared to inspire backlash in testing, were representative of the overall content of the program. We only include data until August 2020, after which the message testing program focused exclusively on turnout content (which was not shown to the audience included in this paper's analysis).

Figure 1 (left panel) shows the breakdown of ads tested by whether they were anti-Trump attack ads, pro-Biden promotional ads, or contrast ads. As can be seen, most ads tested early in the campaign were anti-Trump attack ads, with more contrast ads and promotional ads tested later in the campaign after Biden became the Democratic nominee. The right panel shows the formats of the ads tested. It shows that Acronym tested more promoted news later in the campaign season.



Supplementary Figure 1: Ads tested in Acronym's in-house message testing program, broken down by ad type and format

B.2 Facebook Ad Library

In addition to the in house testing program, we also use Facebook's Ad Archive API, which allows researchers to search political and issue ads run on the platform, to quantify the amount spent by Acronym running ads of different types (see: `https://www.facebook.com/ads/library/api/`. We did this by querying the API for ads that ran with Acronym's paid-for disclaimer during the time period of the campaign for ads containing the keywords "Biden" or "Trump", and excluding explicit turnout content. While this method allows us to get a rough estimate of the spend of the campaign, there are some drawbacks. First, the API only returns spend information in buckets: <100, 100-499, 500-999, 1K-5K, 5K-10K, 10K- 50K, 50K-100K, 100K-200K, 200K-500K, >1M. Thus, precise estimates of spend are impossible to calculate, and we only focus on the lower bound of estimates. Second, we only look at ads that contained either "Biden" or "Trump" in the ad, which unfortunately excludes persuasion content that might not contain those keywords (e.g. ads highlighting Kamala Harris). Third, this spending only includes that on Facebook and not on other platforms like Outbrain and Instagram. Despite these limitations, we believe that this data is broadly representative of the spending done by Acronym on its persuasion program. Figure 2 shows the month-by-month spend broken down by keyword and ad format. Early in the campaign, Acronym's ads focused on Trump before including some pro-Biden content later in the campaign. Earlier ads were also mainly traditional videos, while later in the campaign, Acronym shifted its emphasis to promoted news.

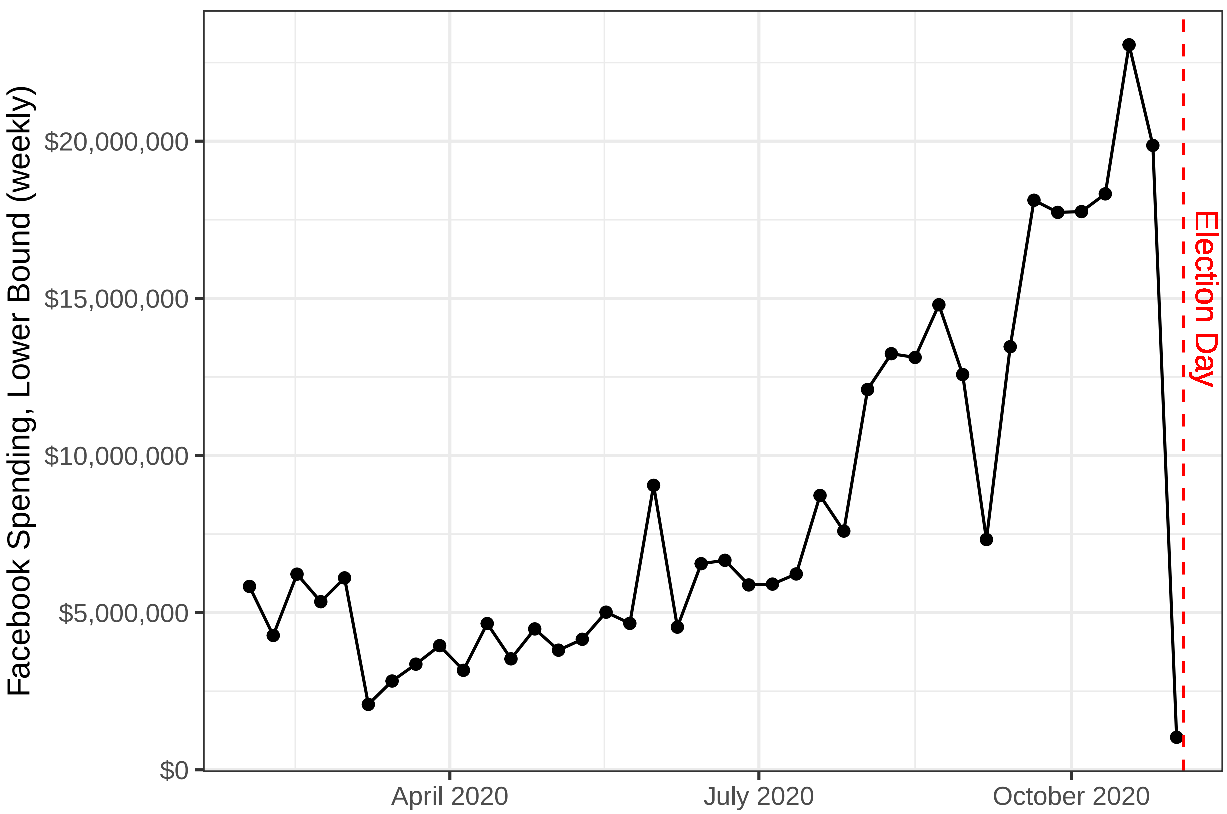
Supplementary Figure 2: Lower bound of spending by Acronym on ads containing the words "Biden" or "Trump" on Facebook, over time, by format and keyword

C Analysis of Overall Facebook Ad Environment

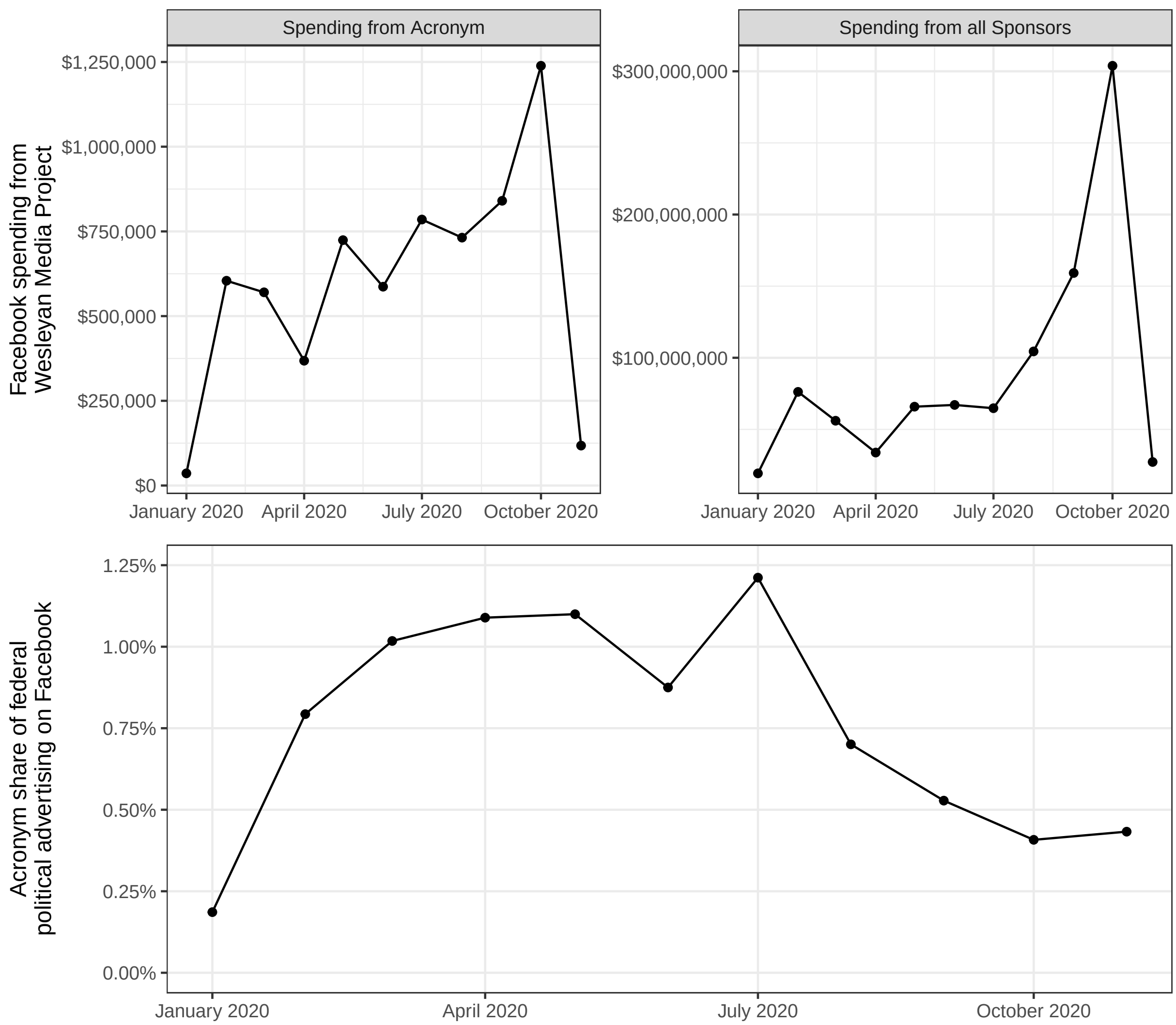
We again leverage the Facebook Ad API to get a lower bound of an estimate of all political ad spending in Acronym's targeted states in order to better understand the ad environment on Facebook. To do this, we queried the Facebook Ad API for all ads containing "Biden" or "Trump" that ran from 2/1/2020 - 11/3/2020 on Facebook, where the ads were targeted to at least one of Acronym's program's states: AZ MI, NC, PA, or WI. The lower bound of total spend by all advertisers that meet this criteria is $349,006,000; unfortunately, we cannot calculate an upper bound since the largest bucket (>$1 million) is boundless. The results of this analysis can be found in Figure 3.

As can be seen, spending on political ads increases sharply in the leadup to Election Day, with the highest spend coming in October. Note that Facebook prevented new ads from being uploaded after 10/27/2020 (although already existing ad campaigns could continue to run), and froze all political ads after election day, 11/3/2020 [40]. Acronym spent approximately $3,867,900 to $5,921,963 on persuasion ads containing the words "Biden" or "Trump" during this period. Comparing both the lower bound of Acronym spend and overall spend as specified above ($349,006,000), we find that Acronym constituted approximately 1% of spending on Facebook on the presidential campaign in our target states.

An analysis of data shared with us by the Wesleyan Media Project also confirms the case that most spending on Facebook in the 2020 election was heavily concentrated in the last weeks of the campaign [29]. They shared a dataset of pages on Facebook who advertised at least once in a federal election from 9/1/2020 to 11/3/2020 (election day) and their weekly spending on Facebook. The spending totals are much more precise than that given by the Facebook Ad API and cover a larger population of advertisements than the above analysis. According to this data, there was a total of approximately $977,761,865 spent on federal races on Facebook from 2/1/2020 to 11/3/2020, $6,603,488 (0.67%) of which was from Acronym. Figure 4 shows the total spending over time, with most of the spending concentrated later in the year (particularly in October).



Supplementary Figure 3: Lower bound of spending by all advertisers on Facebook on ads with the words "Biden" or "Trump" and targeted to at least one of AZ, GA, MI, NC, PA, or WI. Source: Facebook Ad Library



Supplementary Figure 4: Spending on Facebook by pages that advertised in a 2020 federal election. Data was collected and shared by [29]

Even though spending by Acronym constituted a relatively low percentage of total spending on political ads on Facebook, we still have reason to believe that Acronym delivered a much higher dose of pro-Biden / anti-Trump persuasion ads to the treatment group than what was seen by the control group. Political advertising is a small fraction of overall advertising (an estimated 3 percent of Facebook's Q3 US ad revenue in 2020) [35], which leads us to infer that the holdout group likely saw a smaller dose of political advertisement than our treatment group, which is the condition we need to hold in order for our experiment to speak to the political science theories under investigation.

A number of other factors make it highly unlikely that that Acronym's holdout audience simply saw a comparable dose of left-leaning political ads, which have to do with the nature of digital advertising. What they saw depends on how ad systems work and what ads other advertising clients were airing. End users of Facebook, Instagram, and Outbrain see a fixed number of ads. The ad served depends on whether the user matches advertiser targeting criteria and if so, the outcome of a second-price auction for the ad slot in question. Advertisers with different targeting criteria will not be in direct competition for ad slots for the same audience. Finally, the modal ad served on Facebook, Instagram, and Outbrain is a commercial ad, for which a clear ROI can be quickly and easily computed to optimize ad bidding, not a political ad, for which the ROI is impossible to compute prior to an election.

These facts about the Facebook ad environment means that other left-leaning advertisers would need to be targeting the *same audience at the same time with a higher bid than commercial advertisers* for the holdout to contain similar content to the treatment group.

There are a number of reasons this is unlikely. First, traditional campaigns generally target high-education, high-turnout centrist voters for persuasion campaigns, but Acronym targeted an *unconventional audience*, with lower vote-propensity and lower levels of political interest than the conventional "persuasion" audience.

Furthermore, both our analysis and other research shows that campaigns in 2020 spent the majority of funds raised in the two weeks before the election [41], perhaps based on a widespread belief that persuasive effects decay, which has some support in the literature [9].

However, Acronym spent a great deal more ad dollars *earlier* in the campaign than other left-leaning political advertisers in a gamble that 2020 would see unprecedented early voting. In fact, we observe significant early spending (Acronym spent more on persuasion in July than in any other month, see main Figure **??**), and we find much stronger effects on early voting than day of voting (see Figure 3 in the main text). This pattern is consistent with a world in which messaging in the final weeks of the campaign is completely saturated, but ad dollars spent earlier in the campaign were more effectual and did not decay among early voters.

D Pre-analysis Plan

Here we present the original pre-analysis plan registered by Acronym on November 22, 2020, prior to accessing any early vote data at `https://osf.io/3evfp/`.

This PAP covered the analysis plans for two of Acronym's messaging programs (Persuasion and Turnout), but we only study the Persuasion program in this paper. At the moment and we do not plan to write up the turnout experiment due to a number of complications. The decision to allocate resources towards persuasion rather than turnout mean that both the dose and audience were significantly smaller than for the persuasion program (roughly half the budget, half the number of participants, and lower ad-delivery frequency). Additionally, the complete set of IDs included in the treatment group was lost after the employees terminated their employment with Acronym. Combined with the relatively small size of the holdout (control) group, this means that the experiment was not sufficiently large and well-powered to provide good evidence about the impact of campaigns on voter turnout.

In some places, we deviate from the PAP and we note those deviations below.

STG Global Holdout Pre-Analysis Plan

Minali Aggarwal

November 11, 2020

Background

This pre-analysis plan (PAP) builds on our standard operating procedure (SOP) for analysis of "Barometer" experiments running from January to November 2020.

This analysis focuses on the global holdout that we set up in March 2020. Voters in the holdout did not receive ads from our STG Persuasion or Turnout programs. The goal of the final global holdout analysis is to measure the cumulative effects of STG Persuasion and STG Turnout Programs. In particular, we aim to understand whether either program had an effect on verified vote.

STG Persuasion

The STG Persuasion Program ran ads that focused on reducing support for Donald Trump and increasing support for Joe Biden from March to November 2020. We tested many of the ads (and message tracks) in Barometer field experiments and found positive effects on reversed Trump approval and the horse race variable. We also ran 3 global holdout check in surveys throughout the year to measure the program's cumulative effects on our audience, in which we also saw positive effects. We believe that these positive cumulative effects would have an impact on verified vote.

STG Turnout

The STG Turnout Program ran ads that focused on mobilizing left-leaning voters. This program ran from August 2020 to November 3, 2020.

Experimental Design

This experiment falls under the Barometer SOP/PAP for experimental design. Since March, we defined an audience of 5M swing-state voters to target in the STG Persuasion program. We defined a holdout audience that would not receive any ads throughout the rest of the program (450K voters). The rest of the 5M voters were targeted with Acronym ads regularly through November 2020.

We also defined an audience of 1.8M swing-state voters for the STG Turnout program. We defined the holdout audience that would not receive any turnout ads throughout the program (370K voters), while the rest of the audience received ads through November 2020.

To analyze the effect of our program on verified vote, we will compare the holdout audience to the audience that was delivered ads.

Treatment

Treatments for the treatment audience ran from March 2020 to November 2020.

- Cell 0: Control
- Cell 1: Treatment, received STG Persuasion ads

The treatment topics covered the following topic categories:

1. Conservative Messenger
2. Climate
3. Economy
4. Healthcare
5. COVID
6. Racial Justice

Some ads were anti-Trump focused, while others were pro-Biden or Trump/Biden contrast. They also varied by medium: video versus boosted news from sources like Fox News, CNN, Reuters, and more.

Data and Sample

Global Holdout Sampling

- We randomly selected voters from Target Smart Voter File (VF) for the holdout audience, oversampling women, Black, Hispanic, and young people. The remaining voters in AZ, MI, WI, NC, PA, GA were eligible to receive our ads.
- For the STG Persuasion program evaluation, we will filter the audience on the following criteria:
  - Presidential Turnout Score between 20-100
  - TSS between 30-70
  - College Education Score below 50th percentile
  - Political Attention Score below 50th percentile (from Civis)
  - States: AZ, MI, WI, NC, PA
- For the STG Turnout program evaluation, we will filter the audience on the following criteria:
  - Presidential Turnout Score between 0-50
  - TSS between 0-30
  - College Education Score below 50th percentile
  - Political Attention Score below 50th percentile (from Civis)
  - States: AZ, MI, WI, NC, PA, GA

Voter Returns

Because we sampled directly from the Voter File, we can join the Voter Returns by voterbase_id to identify voters and non-voters.

Models

We use the same models as outlined in the Barometer SOP/PAP.

Variables

Outcomes

- Verified Vote, where *verified_vote_any* is:
  - binary indicator for voting in the 2020 presidential election, either by mail or in person
- Verified Mail Vote, where *verified_vote_mail* is:
  - binary indicator for voting by mail in the 2020 presidential election
- Verified In-Person Vote, where *verified_vote_in_person* is:
  - binary indicator for voting in-person in the 2020 presidential election

Subgroups

- Demographic strata
  - Age
    - 18-39
    - 40+
  - Race
    - Black
    - Latinx *Combine Black and Latinx into "Non-white" if Black voters $<$ 10% of sample
    - White
  - Gender, where *is_female*:
    - Binary indicator for whether gender is female
- 2016 Vote Margins, where *vote_margin_over_three_pts*:
  - Binary indicator for whether Trump's vote margin was greater than 3 points in the 2016 election
    - Vote margins $>$ 3% in GA, NC, AZ
    - Vote margins $<$ 3% in PA, WI, MI
- Party registration, where *vb_vf_party* is:
  - 3 categories
    - Republican
    - Democrat
    - Unaffiliated

Controls

- Trump Support Score, where *trump_support_score* is:
  - ts_tsmart_trump_support_score from the VF
- Turnout Score, where *turnout_score* is:
  - ts_tsmart_presidential_general_turnout_score from the VF
- Vote History, where *num_times_voted* is:
  - Number of times voted out of 3 for the 2012, 2016, and 2018 elections

```
data <- data %>% mutate(
  voted_in_2012 = ifelse(!is.na(vb_vf_g2012), 1, 0),
  voted_in_2016 = ifelse(!is.na(vb_vf_g2016), 1, 0),
  voted_in_2018 = ifelse(!is.na(vb_vf_g2018), 1, 0),
  num_times_voted = voted_in_2012 + voted_in_2016 + voted_in_2018
)
```

For anything not described here, we default to the practices and guidelines outlined in our Barometer SOP/PAP.

D.1 PAP Deviations

Here we detail the deviations from the pre-registration document.

Subgroups In the PAP, we specified that we would consider treatment effect heterogeneity by Age, Race, Gender, 2016 Vote Margin, and Party registration. We do report all those analyses but because we lack party registration information for much of the sample, we also included heterogeneity analyses by Trump support score, which is available for all subjects. We submitted an update to the registration *after* seeing early voting data (but before seeing final turnout data) from the voter file to use Trump Support Score (TSS) instead of party registration (`https://osf.io/jkush/`)

Analysis In the PAP, we describe one regression specification that includes controls for Trump support score, Presidential turnout score, and a count of vote history. We report that specification in the main text and two others: an unadjusted specification and a fuller specification that includes Trump support score, Presidential turnout score, strata fixed effects, indicators for voting in any even-year election between 2000 and 2018, party membership indicators (Republican, Democrat, or Unknown, relative to Other).

Regression Discontinuity Design None of the regression discontinuity analyses were pre-specified.

Archival subset As described in the main text, our final analysis sample differs from the intended sample in the PAP. The reason for this is that when we analyse the full sample, we find clear evidence of experimental imbalance, leading us to believe that something is incorrect in the construction of the full sample. We are able to fully reconstruct the process by which the "archival subset" was created, so we base our inferences on the subset in which we can have confidence. As shown in the main text, balance in this subset is within the normal levels expected in a randomized experiment of this design.

E Randomization code

Here we report the SQL code used to randomly sample units from the voter file in to the holdout control group.

```
drop table if exists "Acronym"."stg_global_holdout_stage1";
drop table if exists "Acronym"."stg_global_holdout_stage2";
create table "Acronym"."stg_global_holdout_stage1" as (
select vb_voterbase_id
    , case when vb_voterbase_race = 'African-American' or (vb_voterbase_race = 'Uncoded' ar
    , case when vb_voterbase_race = 'Hispanic' or (vb_voterbase_race = 'Uncoded' and civis_
    , ntl.vb_voterbase_age
    , case when ntl.vb_voterbase_age <18 then 'Under 18'
        when ntl.vb_voterbase_age between 18 and 24 then '18 - 24'
        when ntl.vb_voterbase_age between 25 and 34 then '25 - 34'
        when ntl.vb_voterbase_age between 35 and 44 then '35 - 44'
        when ntl.vb_voterbase_age between 45 and 54 then '45 - 54'
        when ntl.vb_voterbase_age between 55 and 64 then '55 - 64'
        when ntl.vb_voterbase_age > 64 then '65+'
    END as age_category
    , vb_tsmart_city
    , vb_tsmart_state
    , vb_voterbase_gender
    , vb_vf_yob
    , vb_voterbase_dob
    , ntl.vb_tsmart_first_name
    , ntl.vb_tsmart_last_name
    , ntl.vb_tsmart_zip
    , vb_voterbase_phone
    , vb_voterbase_phone_wireless
    , vb_vf_phone
    , email.voterbase_email
    , ts_tsmart_trump_support_score
    , ts_tsmart_presidential_general_turnout_score
    , vb_voterbase_registration_status
from ts.ntl_current ntl
LEFT JOIN tmc.email_current email ON ntl.vb_voterbase_id = email.voterbase_id
where vb_vf_voter_status is not null
    and vb_voterbase_registration_status = 'Registered'
    and vb_voterbase_deceased_flag is null
    and vb_vf_source_state in ('MI','PA','WI','AZ','NC')
    and (vb_voterbase_age >= 18 or vb_voterbase_age is null)
group by 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20
);
grant all on "Acronym"."stg_global_holdout_stage1" to group Acronym;
select * from "Acronym"."stg_global_holdout_stage1";

drop table if exists "Acronym"."stg_global_holdout_stage2";
create table "Acronym"."global_holdout_stage2" as (
    (select vb_voterbase_id
        , vb_tsmart_first_name
        , vb_tsmart_last_name
        , vb_tsmart_zip
```

```
    , vb_tsmart_city
    , vb_tsmart_state
    , vb_voterbase_age
    , black
    , hispanic
    , vb_voterbase_gender
    , age_category
    , vb_vf_yob
    , vb_voterbase_dob
    , vb_voterbase_phone
    , vb_voterbase_phone_wireless
    , vb_vf_phone
    , voterbase_email
    , ts_tsmart_trump_support_score
    , ts_tsmart_presidential_general_turnout_score
    , vb_voterbase_registration_status
from "Acronym"."stg_global_holdout_stage1"
ORDER BY RANDOM() LIMIT 700000)
--Oversample '18 - 24'
UNION ALL
(select vb_voterbase_id
    , vb_tsmart_first_name
    , vb_tsmart_last_name
    , vb_tsmart_zip
    , vb_tsmart_city
    , vb_tsmart_state
    , vb_voterbase_age
    , black
    , hispanic
    , vb_voterbase_gender
    , age_category
    , vb_vf_yob
    , vb_voterbase_dob
    , vb_voterbase_phone
    , vb_voterbase_phone_wireless
    , vb_vf_phone
    , voterbase_email
    , ts_tsmart_trump_support_score
    , ts_tsmart_presidential_general_turnout_score
    , vb_voterbase_registration_status
from "Acronym"."stg_global_holdout_stage1"
where age_category = '18 - 24'
ORDER BY RANDOM() LIMIT 564872)
UNION ALL
  --Oversample '25 - 34'
  (select vb_voterbase_id
    , vb_tsmart_first_name
    , vb_tsmart_last_name
    , vb_tsmart_zip
    , vb_tsmart_city
    , vb_tsmart_state
    , vb_voterbase_age
    , black
    , hispanic
```

```
    , vb_voterbase_gender
    , age_category
    , vb_vf_yob
    , vb_voterbase_dob
    , vb_voterbase_phone
    , vb_voterbase_phone_wireless
    , vb_vf_phone
    , voterbase_email
    , ts_tsmart_trump_support_score
    , ts_tsmart_presidential_general_turnout_score
    , vb_voterbase_registration_status
  from "Acronym"."stg_global_holdout_stage1"
  where age_category = '25 - 34'
  ORDER BY RANDOM() LIMIT 454521)
UNION ALL
  --Oversample AfAm
  (select vb_voterbase_id
    , vb_tsmart_first_name
    , vb_tsmart_last_name
    , vb_tsmart_zip
    , vb_tsmart_city
    , vb_tsmart_state
    , vb_voterbase_age
    , black
    , hispanic
    , vb_voterbase_gender
    , age_category
    , vb_vf_yob
    , vb_voterbase_dob
    , vb_voterbase_phone
    , vb_voterbase_phone_wireless
    , vb_vf_phone
    , voterbase_email
    , ts_tsmart_trump_support_score
    , ts_tsmart_presidential_general_turnout_score
    , vb_voterbase_registration_status
  from "Acronym"."stg_global_holdout_stage1"
  where black = 'black'
  ORDER BY RANDOM() LIMIT 327230)
UNION ALL
  --Oversample Hispanic American
  (select vb_voterbase_id
    , vb_tsmart_first_name
    , vb_tsmart_last_name
    , vb_tsmart_zip
    , vb_tsmart_city
    , vb_tsmart_state
    , vb_voterbase_age
    , black
    , hispanic
    , vb_voterbase_gender
    , age_category
    , vb_vf_yob
    , vb_voterbase_dob
```

```
        , vb_voterbase_phone
        , vb_voterbase_phone_wireless
        , vb_vf_phone
        , voterbase_email
        , ts_tsmart_trump_support_score
        , ts_tsmart_presidential_general_turnout_score
        , vb_voterbase_registration_status
      from "Acronym"."stg_global_holdout_stage1"
      where hispanic = 'hispanic'
      ORDER BY RANDOM() LIMIT 68842)
    UNION ALL
      --Oversample Women
      (select vb_voterbase_id
        , vb_tsmart_first_name
        , vb_tsmart_last_name
        , vb_tsmart_zip
        , vb_tsmart_city
        , vb_tsmart_state
        , vb_voterbase_age
        , black
        , hispanic
        , vb_voterbase_gender
        , age_category
        , vb_vf_yob
        , vb_voterbase_dob
        , vb_voterbase_phone
        , vb_voterbase_phone_wireless
        , vb_vf_phone
        , voterbase_email
        , ts_tsmart_trump_support_score
        , ts_tsmart_presidential_general_turnout_score
        , vb_voterbase_registration_status
      from "Acronym"."stg_global_holdout_stage1"
      where vb_voterbase_gender = 'Female'
      ORDER BY RANDOM() LIMIT 448812)
);
grant all on "Acronym"."global_holdout_stage2" to group Acronym;
select * from "Acronym"."global_holdout_stage2";
```

F Regression tables

References

[1] Romer, D., Kenski, K., Winneg, K., Adasiewicz, C. & Jamieson, K. H. *Capturing campaign dynamics, 2000 and 2004: The national Annenberg election survey* (University of Pennsylvania Press, 2006).

[2] Sides, J., Vavreck, L. & Warshaw, C. The effect of television advertising in united states elections. *American Political Science Review* **116**, 702–718 (2022).

[3] Druckman, J. N. The Implications of Framing Effects for Citizen Competence. *Political Behavior* **23**, 225–256 (2001).

[4] Druckman, J. N. & Holmes, J. W. Does Presidential Rhetoric Matter? Priming and Presidential Approval. *Presidential Studies Quarterly* **34**, 755–778 (2004).

[5] Coppock, A., Hill, S. J. & Vavreck, L. The small effects of political advertising are small regardless of context, message, sender, or receiver: Evidence from 59 real-time randomized experiments. *Science advances* **6** (2020).

[6] Chong, D. & Druckman, J. N. A theory of framing and opinion formation in competitive elite environments. *Journal of communication* **57**, 99–118 (2007).

[7] Zaller, J. R. *The Nature and Origins of Mass Opinion* (Cambridge University Press, 1992).

[8] Vavreck, L. *The Message Matters: The Economy and Presidential Campaigns* (Princeton University Press, Princeton, NJ, 2009).

[9] Gerber, A. S., Gimpel, J. G., Green, D. P. & Shaw, D. R. How large and long-lasting are the persuasive effects of televised campaign ads? results from a randomized field experiment. *American Political Science Review* **105**, 135–150 (2011).

[10] Kalla, J. L. & Broockman, D. E. The minimal persuasive effects of campaign contact in general elections: Evidence from 49 field experiments. *American Political Science Review* **112**, 148–166 (2018).

[11] Mellman, M. Are we doing it all wrong? *The Hill* (2017). URL `https://thehill.com/opinion/campaign/353720-mark-mellman-are-we-doing-it-all-wrong`.

[12] Spenkuch, J. L. & Toniatti, D. Political advertising and election results. *The Quarterly Journal of Economics* **133**, 1981–2036 (2018).

[13] Ansolabehere, S., Iyengar, S., Simon, A. & Valentino, N. Does attack advertising demobilize the electorate? *American political science review* **88**, 829–838 (1994).

[14] Wattenberg, M. P. & Brians, C. L. Negative campaign advertising: Demobilizer or mobilizer? *American political science review* 891–899 (1999).

[15] Ansolabehere, S. D., Iyengar, S. & Simon, A. Replicating experiments using aggregate and survey data: The case of negative advertising and turnout. *American political science Review* **93**, 901–909 (1999).

[16] Kahn, K. F. & Kenney, P. J. Do negative campaigns mobilize or suppress turnout? clarifying the relationship between negativity and participation. *The American Political Science Review* **93**, 877–889 (1999). URL `http://www.jstor.org/stable/2586118`.

[17] Freedman, P. & Goldstein, K. Measuring media exposure and the effects of negative campaign ads. *American Journal of Political Science* **43**, 1189–1208 (1999). URL `http://www.jstor.org/stable/2991823`.

[18] Goldstein, K. & Freedman, P. Campaign advertising and voter turnout: New evidence for a stimulation effect. *Journal of Politics* **64**, 721–740 (2002). URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/0022-3816.00143`. `https://onlinelibrary.wiley.com/doi/pdf/10.1111/0022-3816.00143`.

[19] Finkel, S. E. & Geer, J. G. A spot check: Casting doubt on the demobilizing effect of attack advertising. *American Journal of Political Science* **42**, 573–595 (1998). URL `http://www.jstor.org/stable/2991771`.

[20] Krupnikov, Y. When does negativity demobilize? tracing the conditional effect of negative campaigning on voter turnout. *American Journal of Political Science* **55**, 797–813 (2011).

[21] Lau, R. R. & Rovner, I. B. Negative campaigning. *Annual review of political science* **12**, 285–306 (2009).

[22] Schuirmann, D. J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics* **15**, 657–680 (1987).

[23] Lakens, D. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science* **8**, 355–362 (2017).

[24] Politico. Facebook: Russian-linked accounts bought $150,000 in ads during 2016 race. URL `https://www.politico.com/story/2017/09/06/facebook-ads-russia-linked-accounts-242401`.

[25] Redlawsk, D. *Feeling politics: Emotion in political information processing* (Springer, 2006).

[26] Albertson, B., Dun, L. & Kushner Gadarian, D. The emotional aspects of political persuasion. *The Oxford handbook of electoral persuasion* 169 (2020).

[27] Broockman, D. & Kalla, J. When and why are campaigns' persuasive effects small? evidence from the 2020 us presidential election. *OSF* (2020).

[28] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289–300 (1995).

[29] Neumann, M. *et al.* Identifying and Classifying Federal Election Activity in Online Ads. *Working paper* .

[30] Broockman, D. & Kalla, J. The manifold effects of partisan media on viewers' beliefs and attitudes: A field experiment with fox news viewers (2022). URL `osf.io/jrw26`.

[31] House gop campaign arm created misleading fake news sites. *Talking Points Memo* (2014). URL `https://talkingpointsmemo.com/livewire/nrcc-fake-news-sites`.

[32] Blue dog-affiliated website blurs line between newsroom and political advertiser. *Sludge* URL `https://readsludge.com/2021/09/17/blue-dog-affiliated-website-blurs-line-between-newsroom-and-political-advertiser/`.

[33] Fischer, S. Facebook says new pre-election political ad rules apply to boosted posts. *Axios* (2020). URL `https://www.axios.com/facebook-says-new-pre-election-political-ad-rules-apply-to-boosted-posts-ecbab5e5-48c8` `html`.

[34] 2020: Working america. *Fight for a Better America* (2020). URL `https://www.fightforbetter.org/workingamerica`.

[35] Levy, A. & Rodriguez, S. Why political campaigns are flooding facebook with ad dollars (2020). URL `https://www.cnbc.com/2020/10/08/trump-biden-pacs-spend-big-on-facebook-as-election-nears.html`.

[36] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2022). URL `https://www.R-project.org/`.

[37] Wickham, H. *et al.* Welcome to the tidyverse. *Journal of Open Source Software* **4**, 1686 (2019).

[38] Blair, G., Cooper, J., Coppock, A., Humphreys, M. & Sonnet, L. *estimatr: Fast Estimators for Design-Based Inference* (2022). URL `https://CRAN.R-project.org/package=estimatr`. R package version 0.30.6.

[39] Fox, J. & Weisberg, S. *An R Companion to Applied Regression* (Sage, Thousand Oaks CA, 2019), third edn. URL `https://socialsciences.mcmaster.ca/jfox/Books/Companion/`.

[40] What To Know About Facebook Advertising Around the Election. *Meta for Business* URL `https://www.facebook.com/business/news/facebook-ads-restriction-2020-us-election`.

[41] Ridout, T. N., Fowler, E. F. & Franz, M. M. Spending Fast and Furious: Political Advertising in 2020. *The Forum* **18**, 465–492 (2021). URL `https://www.degruyter.com/document/doi/10.1515/for-2020-2109/html`. Publisher: De Gruyter.

Supplementary Table 1: Average and conditional average treatment effects. These estimates correspond to Figure 1 in the main text.

Covariate	Level	Adjustment	Estimate	SE	df	t	p-value	95%CI lower	95%CI upper
ATE	ATE	Unadjusted	-0.000	0.001	1999280	-0.200	0.842	-0.003	0.002
ATE	ATE	PAP adjustment set	-0.001	0.001	1999277	-0.525	0.600	-0.003	0.002
ATE	ATE	Full adjustment set	-0.001	0.001	1999248	-0.436	0.663	-0.003	0.002
Age	18-39	Unadjusted	-0.000	0.001	1379015	-0.304	0.761	-0.003	0.002
Age	18-39	PAP adjustment set	-0.001	0.001	1379012	-0.682	0.495	-0.003	0.002
Age	18-39	Full adjustment set	-0.001	0.001	1378983	-0.493	0.622	-0.003	0.002
Age	40+	Unadjusted	0.000	0.003	620263	0.150	0.881	-0.006	0.007
Age	40+	PAP adjustment set	-0.000	0.003	620260	-0.162	0.871	-0.006	0.005
Age	40+	Full adjustment set	-0.000	0.003	620243	-0.059	0.953	-0.006	0.005
Gender	Female	Unadjusted	-0.001	0.002	978039	-0.821	0.411	-0.005	0.002
Gender	Female	PAP adjustment set	-0.001	0.002	978036	-0.812	0.417	-0.004	0.002
Gender	Female	Full adjustment set	-0.001	0.002	978016	-0.775	0.438	-0.004	0.002
Gender	Other	Unadjusted	0.001	0.002	1021239	0.408	0.683	-0.003	0.005
Gender	Other	PAP adjustment set	-0.000	0.002	1021236	-0.005	0.996	-0.004	0.004
Gender	Other	Full adjustment set	0.000	0.002	1021216	0.025	0.980	-0.004	0.004
Race	Black	Unadjusted	-0.001	0.003	233544	-0.218	0.827	-0.006	0.005
Race	Black	PAP adjustment set	-0.001	0.002	233541	-0.513	0.608	-0.006	0.004
Race	Black	Full adjustment set	-0.001	0.002	233524	-0.342	0.732	-0.006	0.004
Race	Latinx	Unadjusted	-0.003	0.004	179034	-0.818	0.413	-0.010	0.004
Race	Latinx	PAP adjustment set	-0.002	0.003	179031	-0.685	0.493	-0.009	0.004
Race	Latinx	Full adjustment set	-0.002	0.003	179014	-0.655	0.513	-0.008	0.004
Race	Other	Unadjusted	-0.003	0.009	55569	-0.369	0.712	-0.021	0.014
Race	Other	PAP adjustment set	-0.007	0.008	55566	-0.821	0.411	-0.023	0.009
Race	Other	Full adjustment set	-0.007	0.008	55549	-0.797	0.426	-0.023	0.010
Race	White	Unadjusted	0.000	0.002	1531127	0.085	0.932	-0.003	0.003
Race	White	PAP adjustment set	-0.000	0.001	1531124	-0.115	0.908	-0.003	0.003
Race	White	Full adjustment set	-0.000	0.001	1531107	-0.069	0.945	-0.003	0.003
Margin	Vote margin less than 3pp	Unadjusted	-0.000	0.002	1337055	-0.073	0.942	-0.004	0.003
Margin	Vote margin less than 3pp	PAP adjustment set	-0.001	0.002	1337052	-0.543	0.587	-0.004	0.002
Margin	Vote margin less than 3pp	Full adjustment set	-0.001	0.002	1337023	-0.488	0.625	-0.004	0.002
Margin	Vote margin more than 3pp	Unadjusted	-0.001	0.002	662223	-0.310	0.757	-0.005	0.004
Margin	Vote margin more than 3pp	PAP adjustment set	-0.000	0.002	662220	-0.096	0.924	-0.004	0.004
Margin	Vote margin more than 3pp	Full adjustment set	-0.000	0.002	662191	-0.019	0.985	-0.004	0.004
Partisanship	Democrat	Unadjusted	0.005	0.006	182943	0.939	0.348	-0.006	0.016
Partisanship	Democrat	PAP adjustment set	0.008	0.005	182940	1.585	0.113	-0.002	0.017
Partisanship	Democrat	Full adjustment set	0.007	0.005	182914	1.416	0.157	-0.003	0.016
Partisanship	Other	Unadjusted	-0.000	0.003	302389	-0.013	0.990	-0.007	0.007
Partisanship	Other	PAP adjustment set	-0.000	0.003	302386	-0.019	0.985	-0.006	0.006
Partisanship	Other	Full adjustment set	0.000	0.003	302360	0.037	0.970	-0.006	0.006
Partisanship	Republican	Unadjusted	-0.009	0.005	71873	-1.729	0.084	-0.019	0.001
Partisanship	Republican	PAP adjustment set	-0.010	0.005	71870	-2.003	0.045	-0.019	-0.000
Partisanship	Republican	Full adjustment set	-0.010	0.005	71844	-2.123	0.034	-0.020	-0.001
Partisanship	Unknown	Unadjusted	-0.000	0.002	1442069	-0.276	0.783	-0.004	0.003
Partisanship	Unknown	PAP adjustment set	-0.001	0.001	1442066	-0.901	0.367	-0.004	0.001
Partisanship	Unknown	Full adjustment set	-0.001	0.001	1442040	-0.783	0.434	-0.004	0.002
Trump support	30 to 40	Unadjusted	0.003	0.003	522916	1.274	0.203	-0.002	0.008
Trump support	30 to 40	PAP adjustment set	0.004	0.002	522913	1.613	0.107	-0.001	0.008
Trump support	30 to 40	Full adjustment set	0.004	0.002	522884	1.724	0.085	-0.001	0.008
Trump support	40 to 50	Unadjusted	-0.002	0.003	485369	-0.601	0.548	-0.007	0.004
Trump support	40 to 50	PAP adjustment set	-0.002	0.002	485366	-0.885	0.376	-0.007	0.003
Trump support	40 to 50	Full adjustment set	-0.002	0.002	485337	-0.818	0.413	-0.007	0.003
Trump support	50 to 60	Unadjusted	0.002	0.003	478331	0.640	0.522	-0.004	0.007
Trump support	50 to 60	PAP adjustment set	-0.001	0.003	478328	-0.296	0.767	-0.006	0.004
Trump support	50 to 60	Full adjustment set	-0.000	0.002	478299	-0.098	0.922	-0.005	0.005
Trump support	60 to 70	Unadjusted	-0.004	0.003	512658	-1.454	0.146	-0.010	0.001
Trump support	60 to 70	PAP adjustment set	-0.003	0.003	512655	-1.373	0.170	-0.008	0.001
Trump support	60 to 70	Full adjustment set	-0.004	0.003	512626	-1.477	0.140	-0.009	0.001

Supplementary Table 2: The heterogeneous effects of treatment by Trump support. These estimates correspond to Figure 2 in the main text.

Target	Outcome	Adjustment	Estimate	SE	df	t	p-value	95%CI lower	95%CI upper
Difference-in-CATEs TSS 60-70 versus TSS 30-40	Voted early in 2020	Unadjusted	-0.010	0.003	1035574	-3.002	0.003	-0.017	-0.004
Difference-in-CATEs TSS 60-70 versus TSS 30-40	Voted early in 2020	PAP adjustment set	-0.010	0.003	1035571	-3.005	0.003	-0.017	-0.004
Difference-in-CATEs TSS 60-70 versus TSS 30-40	Voted early in 2020	Full adjustment set	-0.011	0.003	1035542	-3.151	0.002	-0.017	-0.004
Difference-in-CATEs TSS 60-70 versus TSS 30-40	Voted in person in 2020	Unadjusted	0.003	0.004	1035574	0.847	0.397	-0.004	0.010
Difference-in-CATEs TSS 60-70 versus TSS 30-40	Voted in person in 2020	PAP adjustment set	0.003	0.003	1035571	0.922	0.356	-0.003	0.010
Difference-in-CATEs TSS 60-70 versus TSS 30-40	Voted in person in 2020	Full adjustment set	0.003	0.003	1035542	0.890	0.374	-0.004	0.009
Difference-in-CATEs TSS 60-70 versus TSS 30-40	Voted in 2020	Unadjusted	-0.007	0.004	1035574	-1.933	0.053	-0.015	0.000
Difference-in-CATEs TSS 60-70 versus TSS 30-40	Voted in 2020	PAP adjustment set	-0.007	0.003	1035571	-2.093	0.036	-0.014	-0.000
Difference-in-CATEs TSS 60-70 versus TSS 30-40	Voted in 2020	Full adjustment set	-0.008	0.003	1035542	-2.268	0.023	-0.014	-0.001
Treatment*TSS interaction term from linear model	Voted early in 2020	Unadjusted	-0.028	0.011	1999278	-2.624	0.009	-0.049	-0.007
Treatment*TSS interaction term from linear model	Voted early in 2020	PAP adjustment set	-0.028	0.010	1999276	-2.712	0.007	-0.048	-0.008
Treatment*TSS interaction term from linear model	Voted early in 2020	Full adjustment set	-0.029	0.010	1999247	-2.840	0.005	-0.049	-0.009
Treatment*TSS interaction term from linear model	Voted in person in 2020	Unadjusted	0.010	0.011	1999278	0.897	0.370	-0.012	0.031
Treatment*TSS interaction term from linear model	Voted in person in 2020	PAP adjustment set	0.009	0.010	1999276	0.892	0.373	-0.011	0.029
Treatment*TSS interaction term from linear model	Voted in person in 2020	Full adjustment set	0.009	0.010	1999247	0.907	0.365	-0.011	0.029
Treatment*TSS interaction term from linear model	Voted in 2020	Unadjusted	-0.018	0.012	1999278	-1.542	0.123	-0.041	0.005
Treatment*TSS interaction term from linear model	Voted in 2020	PAP adjustment set	-0.019	0.010	1999276	-1.826	0.068	-0.039	0.001
Treatment*TSS interaction term from linear model	Voted in 2020	Full adjustment set	-0.020	0.010	1999247	-1.935	0.053	-0.040	0.000

Supplementary Table 3: Balance estimates. These estimates correspond to Figure 5 in the main text.

Covariate	Estimate	SE	df	t	p-value	p-value (BH correction)	95%CI lower	95%CI upper
Party: Democrat	-0.001	0.001	1999280	-1.211	0.226	0.810	-0.003	0.001
Party: Republican	-0.000	0.000	1999280	-0.414	0.679	0.944	-0.001	0.001
Party: Unknown	0.000	0.001	1999280	0.091	0.927	0.944	-0.002	0.003
Party: Other	0.001	0.001	1999280	1.257	0.209	0.810	-0.001	0.003
TSS: 30-40	0.001	0.001	1999280	1.219	0.223	0.810	-0.001	0.004
TSS: 40-50	0.002	0.001	1999280	1.385	0.166	0.810	-0.001	0.004
TSS: 50-60	0.000	0.001	1999280	0.229	0.819	0.944	-0.002	0.003
TSS: 60-70	-0.003	0.001	1999280	-2.609	0.009	0.200	-0.006	-0.001
Trump support score / 100	-0.001	0.000	1999280	-2.260	0.024	0.262	-0.001	-0.000
Turnout score / 100	0.000	0.001	1999280	0.707	0.480	0.944	-0.001	0.002
Ideology score / 100	0.000	0.000	1999280	0.070	0.944	0.944	-0.001	0.001
Partisanship score / 100	0.000	0.001	1999280	0.489	0.625	0.944	-0.001	0.002
Voted in 2000	-0.000	0.001	1999280	-0.072	0.943	0.944	-0.002	0.002
Voted in 2002	0.000	0.001	1999280	0.296	0.767	0.944	-0.001	0.002
Voted in 2004	0.000	0.001	1999280	0.161	0.872	0.944	-0.002	0.002
Voted in 2006	-0.000	0.001	1999280	-0.215	0.829	0.944	-0.002	0.001
Voted in 2008	-0.001	0.001	1999280	-0.761	0.447	0.944	-0.003	0.001
Voted in 2010	-0.001	0.001	1999280	-0.690	0.490	0.944	-0.002	0.001
Voted in 2012	0.000	0.001	1999280	0.252	0.801	0.944	-0.002	0.003
Voted in 2014	-0.001	0.001	1999280	-1.132	0.258	0.810	-0.003	0.001
Voted in 2016	-0.000	0.001	1999280	-0.279	0.780	0.944	-0.003	0.002
Voted in 2018	0.001	0.001	1999280	0.881	0.378	0.944	-0.001	0.004