A Digital Field Experiment Reveals Large Effects of Friend-to-Friend Texting on Voter Turnout

Aaron Schein,[1*] Keyon Vafa,[2] Dhanya Sridhar,[1] Victor Veitch,[3]
Jeffrey Quinn,[5] James Moffet,[6] David M. Blei,[1,2,3] Donald P. Green[4]

[1]Data Science Institute, Columbia University, New York, NY 10027
[2]Department of Computer Science, Columbia University, New York, NY 10027
[3]Department of Statistics, Columbia University, New York, NY 10027
[4]Department of Political Science, Columbia University, New York, NY 10027
[5]PredictWise, https://www.predictwise.com, New York, NY 10036
[6]JDM Design, Cambridge, MA 02139

[*]Address correspondence to aaron.schein@columbia.edu.

Two decades of field experiments on get-out-the-vote tactics suggest that impersonal tactics, like mass emails, have only a modest or negligible effect on voter turnout, while more personal tactics, like door-to-door canvassing, are more effective. However, the COVID-19 pandemic threatens to upend the vast face-to-face voter mobilization efforts that have figured prominently in recent presidential election campaigns. If campaigns can no longer send canvassers to voters' doors, what tactics can they turn to in order to mobilize their supporters? This paper evaluates a promising alternative to face-to-face get-out-the-vote tactics: mobile app technology that enables millions of people to message their friends to urge them to vote. Prior to the most recent US midterm elections in 2018, the mobile app OUTVOTE randomized an aspect

of their system, hoping to unobtrusively assess the causal effect of their users' messages on voter turnout. We develop a statistical methodology to address the challenges of such data, and then analyze the OUTVOTE study. Our analysis reveals evidence of very large and statistically significant treatment effects from friend-to-friend mobilization efforts ($\widehat{\text{CACE}}$= 8.3, CI = (1.2, 15.3)). Further, the statistical methodology can be used to study other friend-to-friend messaging efforts. These results suggest that friend-to-friend texting, which is a personal voter mobilization effort that does not require face-to-face contact, is an effective alternative to conventional voter mobilization tactics.

Introduction

Political campaigns in the United States spend enormous resources on "get out the vote" (GOTV) interventions to nudge potential voters to the polls. Such efforts are especially large in the weeks leading up to a presidential election, when millions of campaign workers and volunteers descend on battleground states to encourage supporters to vote. Randomized control trials (RCTs) show that face-to-face encouragements such as these substantially increase voter turnout (*1*). However, the COVID-19 epidemic makes door-to-door canvassing infeasible because face-to-face contact, especially between strangers, is both limited and unsafe. If campaigns cannot use one of their most reliable tactics for mobilizing supporters, what scalable alternatives can they turn to? One easy alternative is to use mass texting or email. But meta-analysis shows that many scalable GOTV tactics such as these have only modest effects on voter turnout, typically less than one percentage point (*1*). Scalable, but impersonal, communication is not a substitute for door-to-door canvassing.

The evolution of mobile communication, however, has created new opportunities for effective GOTV strategies that do not require face-to-face contact. In particular, political campaigns

have begun to embrace friend-to-friend organizing, in which volunteers are encouraged to send messages to their close contacts to encourage them to vote. The premise of friend-to-friend organizing is that GOTV appeals are especially effective, particularly over mass texts or emails, because friends are often welcome and trusted messengers.

But there have been few evaluations of friend-to-friend outreach, and none conducted on a large scale (*2–4*). The reason is that friend-to-friend organizing is a fundamentally challenging object of experimental study. At first glance, designing a suitable randomized evaluation may seem simple: a campaign directs volunteers to send a scripted GOTV message to a randomly assigned subset of their phone contacts and not to any others. This approach does produce a random treatment–control split of each volunteer's contacts, but it risks studying something other than the effects of organic communication between friends. The issue is that it directs volunteers to send a message to contacts they may not have otherwise chosen, like bosses or ex-boyfriends. Such GOTV messages may feel like impersonal spam, not personal appeals, and so the study may mistakenly attribute the typically small effect that spam has on turnout (*1*) to authentic friend-to-friend appeals.

This paper uses a large-scale field study from the 2018 US midterm elections to measure the causal effect of friend-to-friend messages on voter turnout. (This study was performed before the COVID-19 pandemic but, for the reasons described above, its results are particularly pertinent now.) The study was conducted by OUTVOTE (*5*), a mobile phone app designed to systematize the process of encouraging one's friends to vote. When opened, the app first prompts its user to create a queue of all the phone contacts they intend to message. Then, for each queued contact, the app takes the user to a message interface where the user can either select a default message—e.g., "Don't forget to vote this Tuesday!"—or craft an individualized message, and then send it. Figure 1 shows screenshots of an OUTVOTE user's typical workflow.

During the months prior to the 2018 midterm elections, OUTVOTE deployed an unobtru-

sive randomization scheme between the queuing of friends and sending of messages. For any queues of at least five people, the app would randomly skip some people in the queue, each with probability 0.05, not taking the user to the messaging stage for them. This skipping procedure was designed to minimally degrade the user's experience while still injecting enough randomness into natural friend-to-friend interaction to facilitate evaluation of friend-to-friend contact. In this paper, we study the data produced by OUTVOTE to assess the causal effect of receiving a GOTV message. We find evidence of large and significant treatment effects from friend-to-friend mobilization efforts.

OUTVOTE's study was designed to have a "light touch," but the constraints of conducting an experiment that is invisible to users complicates the estimation of causal effects in several respects. First, if a user noticed someone had been skipped, the user could add that person to a new queue in which they might not be skipped, thereby overriding the random assignment process and inviting confounding. We therefore define a treatment–control split of subjects in a manner that is unconfounded even when subjects may be added to many queues but only randomly skipped in some.[1] Second, a large share of subjects do not comply with their treatment assignment, as we define it: some subjects assigned to receive messages do not, while others assigned not to receive messages do. Non-compliance limits the scope of estimable causal effects and increases the statistical uncertainty of the estimators. We address non-compliance by working within a standard instrumental variable (IV) framework (*6*) and defining the causal quantity of interest to be the average effect of the treatment on a subset of the overall subject pool, namely, those who would comply with their assigned treatment. We further refine this subset using subjects' positions in the queue in ways that improve the precision of the IV esti-

[1]It is possible and perhaps tempting to define treatment–control splits that are confounded. For instance, defining the control group to be subjects who were skipped in every queue they were added to leads to a confounded split, because there may be some subjects whom users will re-queue any number of times to message for reasons systematically related to the outcome. In the appendix, we provide empirical evidence that this definition is confounded as it fails to maintain symmetry between the treatment and control groups.

mator. Third, due to Outvote's limited information about subjects, errors occur when matching them to a database of public information about citizens' voter history. Unreliable information about voting biases estimates of the causal effect towards zero. We address such measurement error using ancillary information, refining the study population to subjects whose voter turnout is measured reliably. We show that when these estimation issues are addressed, the estimated effects of friend-to-friend encouragements to vote are large and statistically robust.

The subject pool, treatment assignment, and non-compliance. The subject pool consists of people who OUTVOTE users queued to message, provided that the queue contained at least five subjects. The treatment of interest is whether a subject received a GOTV text, and the outcome of interest is whether or not they voted in the 2018 US midterm election.

To estimate the causal effect, we must address several statistical challenges. Some arise when users repeatedly create queues with the same contacts, or when the same subject appears on more than one user's queue. To sidestep these, we consider only the first time a subject is queued by any user, thereby defining the assigned control group to be those subjects randomly skipped in their first queue, and defining the assigned treatment group to be those not skipped in their first queue. Considering only the first queue ensures a random treatment assignment—in particular, it prevents possible confounding from subjects being queued multiple times for reasons related to their likelihood of voting. For each subject, the first queue is considered the moment of entry into the study population, and whether they are skipped in it determines their treatment assignment.

This definition of the assigned treatment and control groups immediately raises the prospect of "non-compliance." Generally speaking, non-compliance occurs when subjects in the assigned treatment group do not receive treatment, or when subjects in the assigned control group do. In this study, subjects' non-compliance is driven by the users, who typically created long

queues (the median length is 22) and often stopped before messaging everyone. Only 29% of all subjects in the assigned treatment group received a message. Users sometimes also re-queued and messaged subjects who had been skipped in their first queue; 13% of subjects in the assigned control group received a message.

Formally, let Z_i denote the assignment to receive ($Z_i = 1$) or not receive ($Z_i = 0$) the treatment. Let D_i denote the treatment receipt, whether subject i received a text ($D_i = 1$) or did not ($D_i = 0$). Let Y_i denote the recorded outcome, whether subject i voted ($Y_i = 1$) or did not ($Y_i = 0$). The data are $\{Z_i, D_i, Y_i\}_{i=1}^n$.

Defining and estimating the complier average causal effect. We want to estimate the causal effect of receiving a GOTV text on voting. Let Y_{i1} be subject i's potential outcome of voting if they receive a GOTV text ($D_i = 1$) and Y_{i0} be their potential outcome if they do not ($D_i = 0$). The causal effect of subject i receiving a text is the difference Y_{i1}–Y_{i0}. In an ideal randomized experiment, the average causal effect (ACE) across all subjects—i.e., $\mathbb{E}\left[Y_{i1} - Y_{i0}\right]$—would be identified; however, the presence of non-compliance limits the scope of identified average effects.

Experiments with non-compliance are routinely analyzed by treating the random assignment Z_i as an instrumental variable for the treatment receipt D_i (*6*). In the face of non-compliance, the recorded receipt D_i may be confounded—users may be systematic in choosing which subjects they message. Following (*6*), let D_{i1} be subject i's potential receipt if they are assigned to receive treatment and D_{i0} be their potential receipt if they are assigned not to.

There are four types of subjects, defined by the four combinations of the potential receipts. Some subjects are "compliers" who receive a message if and only if assigned to the treatment group; the set of all compliers is $\mathcal{C} = \{i : D_{i1} = 1 \text{ and } D_{i0} = 0\}$. The three remaining types of subject are "always-takers" ($D_{i0} = D_{i1} = 1$), "never-takers" ($D_{i0} = D_{i1} = 0$), and "defiers" ($D_{i0} = 1$ and $D_{i1} = 0$).

In the OUTVOTE study, a complier is someone the user would message only if they were not skipped in the first queue. By contrast, an always-taker is someone the user would always message, even if that person were skipped. An always-taker could be someone prominently positioned near the top of the ranking of phone contacts from which users select whom to queue (see fig. 1a). The user notices if they are skipped and then re-queues and messages them. If not skipped, they are among those high in the queue whom the user messages before quitting. A never-taker instead may be someone near the bottom of the queue, whom the user would not message, even if they were not slated to be skipped. Finally, a defier is someone who would be messaged only if they are skipped in the first queue—it is hard to imagine that such subjects exist in large number.

While a study with non-compliance cannot generally identify the average causal effect among all subjects, it may identify the average causal effect among compliers (CACE),

$$\text{CACE} = \mathbb{E}\left[Y_{i1} - Y_{i0} \,|\, i \in \mathcal{C}\right]. \tag{1}$$

Identifying the CACE requires several assumptions (*6*). One assumption is "monotonicity," i.e., that there are no defiers. In the OUTVOTE study, we assume there are no subjects who would only receive a message if skipped in a user's queue.

Another assumption is that the assignment Z_i satisfies the three "instrumental conditions" (*7*). The first is "relevance," which stipulates a non-zero association between assignment Z_i and treatment receipt D_i. This condition is met in the OUTVOTE data; recall that 29% of the assigned treatment group received messages, as compared to 13% of the assigned control group. The second condition is the "exclusion restriction," which stipulates that Z_i must only affect the outcome Y_i through the mediating effect of the treatment D_i. Intuition suggests that this condition holds: whether a subject is skipped in a user's queue (Z_i) should not affect their voting behavior (Y_i) except by influencing the user to send them a reminder to vote (D_i). The

third condition stipulates that the assignment Z_i and outcome Y_i do not share causes. This condition holds because the assignment Z_i is randomized. Note that the second and third instrumental conditions guarantee a stronger condition known as "independence" (*8*), "exchangeability" (*7*), or "ignorability." It states that the potential outcomes are independent of assignment, i.e., $Y_{i1}, Y_{i0} \perp\!\!\!\perp Z_i$, and informally means the assigned treatment and control groups have the same expected potential outcomes.

Note that an instrument that meets all of the above conditions but is only weakly related to receipt can introduce finite-sample bias. The rule-of-thumb of Staiger and Stock (1997) (*9*) diagnoses an instrument as "weak" if a regression of D_i on Z_i produces an F-statistic less than 10. No such bias is present in the OUTVOTE data—this regression on the whole subject pool yields an F-statistic of 348 and comparable values for all subsets we analyze.

The final assumption is that messages sent to subject i have no effect on any other subject j. This is a formulation of the "stable unit treatment value assumption" (SUTVA) (*10*) that is commonly made in the experimental literature. This assumption is violated by "spillover effects," which are typically of concern when subjects form a densely-connected social network. Here, however, 97% of subjects were queued by a single user, suggesting that OUTVOTE's user base and, by extension, the overall subject pool are not densely connected. Previous work assessing spillover effects of get-out-the-vote appeals within and across households suggests that such bias tends to be negligible (*11*).

Under these assumptions, the CACE is consistently estimated by the IV estimator,

$$\widehat{\text{CACE}} = \frac{\mathbb{E}[\,Y_i \mid Z_i = 1] - \mathbb{E}[\,Y_i \mid Z_i = 0]}{\mathbb{E}[D_i \mid Z_i = 1] - \mathbb{E}[D_i \mid Z_i = 0]}. \tag{2}$$

The numerator estimates the effect of assignment, or the intent-to-treat (ITT) effect. Under the monotonicity assumption (no defiers), the denominator estimates the proportion of compliers in the subject pool. The IV estimator can be augmented with covariates in order to improve the

precision with which treatment effects are estimated. Below we present both the unadjusted estimate and the estimate adjusting for 85 pre-assignment covariates (e.g., age, party registration, prior voting history).

Using ancillary information to reduce measurement error when measuring voter turnout. As in most GOTV field studies, the recorded outcome of whether subject i voted (Y_i) is measured by matching subjects to a voter roll database. OUTVOTE matched subjects to a database using the information in the phone contacts that users shared with the app. However, in keeping with their light-touch approach, they did not ask users to enter any missing details. If a user's phone listed a contact as "Alice," with no last name, then OUTVOTE's matching algorithm only used Alice's first name and mobile number. Thus, some proportion of subjects in OUTVOTE's study are incorrectly matched and their recorded outcome may be misclassified.

Measurement error in outcomes often introduces attenuation bias when the error is "non-differential" (*12–14*). Denote a subject's true outcome as $Y_i^\star$. Measurement error on outcomes is non-differential with respect to assignment if the recorded outcome is independent of the assignment given the true outcome,

$$Y_i \perp\!\!\!\perp Z_i \,|\, Y_i^\star. \tag{3}$$

We can safely assume the measurement error introduced by matching errors is non-differential since Z_i was randomized and because matching occurred prior to assignment.

We surmise, then, that in the present study measurement error attenuates the estimated effect. The numerator of the IV estimator in eq. (2), which estimates the ITT, can be written

$$\mathbb{E}[\,Y_i \,|\, Z_i = 1] - \mathbb{E}[\,Y_i \,|\, Z_i = 0] = \pi_{\text{BIAS}} \times \Big(\underbrace{\mathbb{E}[\,Y_i^\star \,|\, Z_i = 1] - \mathbb{E}[\,Y_i^\star \,|\, Z_i = 0]}_{\text{true ITT}} \Big), \tag{4}$$

where the bias is the difference of two probabilities,

$$\pi_{\text{BIAS}} = P(Y_i = 1 \,|\, Y_i^\star = 1) - P(Y_i = 1 \,|\, Y_i^\star = 0). \tag{5}$$

We include a proof in the appendix. If the conditional probability of correct measurement is greater than the probability of mismeasurement, i.e., $P(Y_i = 1 \mid Y_i^\star = 1) > P(Y_i = 1 \mid Y_i^\star = 0)$, then the bias attenuates the true effect, $\pi_{\text{BIAS}} \in (0, 1]$. Informally, this assumption implies that the measurement is not so poor that a non-voter has a greater chance than a voter of being classified as having cast a ballot.

In the OUTVOTE study, the attenuation bias of the estimate may be severe given the inherent difficulty of accurately matching contacts based on incomplete information. To mitigate this issue, we obtained data from the data vendor PREDICTWISE that links millions of mobile phone numbers to voter roll entries. PREDICTWISE relies on commercially-available marketing data to associate complete name and demographic information with mobile phone numbers; such information is helpful in matching phone numbers to the voter rolls. OUTVOTE did not rely on such information when it performed matching—thus, their two approaches often yield different results. We found that 30% of subjects were matched to the same entry by both OUTVOTE and PREDICTWISE. We assume (and provide evidence in the appendix) that this subset of subjects exhibits substantially less measurement error and refine the study population to these 30%.

Using queue position to improve compliance and reduce error. Low compliance rates reduce the precision of the CACE estimator. The denominator of eq. (2) implies that only 16% of subjects are compliers. We can improve the precision of the estimator by filtering the subjects by a pre-assignment variable that improves compliance while still maintaining a large enough n.

Non-compliance in the assigned treatment group is driven by users abandoning long queues before messaging everyone on them. Whether subjects "complied" with their treatment assignment thus depends on their position in the queue. OUTVOTE did not save subjects' queue position in its database. However, a simple method based on timestamp information of when users queued subjects, detailed in the appendix, confidently reconstructs the queue position for

60% of subjects. As described in the appendix, when this method fails to reconstruct queue position, it is often because users pressed an "Add all" button that queued all of their phone contacts simultaneously. Those added to the subject pool by the "Add all" function exhibit very low compliance: subjects whose position cannot be reconstructed have an 8% compliance rate as compared to 22% among those whose position can be. Refining the study population to those whose queue position can be reconstructed, as we do, both improves compliance rates and facilitates further refinements based on queue position.

Let $Q_i \in \{1, 2, 3, \dots\}$ be the position of subject i in their first queue. Note that Q_i is a pre-assignment variable since assignment is determined by subject i's first queue only. Moreover, Q_i is unchanged by subjects' random assignments: if the first two people in a queue are randomly skipped, the third person will still have $Q_i = 3$. Thus, any subset of subjects defined by levels of Q_i maintains the symmetry between the assigned treatment and control groups. (In the appendix, we confirm empirically that Q_i is unrelated to assignment Z_i.)

Subjects with lower values of Q_i were more likely to receive messages. For example, the contact rate among subjects who were added first ($Q_i = 1$) is 47%, and the rate among those added within the top ten ($Q_i \leq 10$) is 34%. Refining the study population based on a maximum allowable queue position q_{max} yields a higher share of compliers, for whom causal effects are more precisely estimable. Refining the study population in this way changes the estimand to

$$\text{CACE}_{q_{\text{max}}} = \mathbb{E}[Y_{i1} - Y_{01} \mid i \in \mathcal{C} \text{ and } Q_i \leq q_{\text{max}}], \tag{6}$$

which trades off generalizibility for precision.

We select q_{max} to maximize the expected precision of the estimator. The precision is affected both by the compliance rate and by n. For example, although the compliance rate is highest among the subpopulation defined by $q_{\text{max}} = 1$, there are only $n = 2,996$ such subjects. Based on a power analysis (in the appendix), we find that $q_{\text{max}} = 103$ minimizes a proxy for the expected

sampling variability of the CACE estimator; it yields a study population of $n = 27,464$ with a 25% compliance rate. Figure 2 shows compliance, n, and power as a function of $q_{\max}$. This procedure is directly analogous to that of (*15*), which also trims the sample systematically to minimize the asymptotic variance of the estimator.

Study Details

Setting An OUTVOTE user first syncs his or her phone's contacts with the app, which then matches them to the voter rolls. The user is then presented with a ranked list of contacts, each annotated with information about the contact's participation in past elections, party registration, and residence in a battleground state (see fig. 1a). The user then creates a queue of friends they intend to message. Once a queue is set, the app takes the user to a messaging interface for each contact in the queue, in the order in which the user queued each contact. In the messaging interface, users can either send a default message or create their own. After the user presses "Send" (see fig. 1b), they are taken immediately to the messaging interface for the next contact.

Randomization and study period During the study period from August 3, 2018 until the day of the US midterm elections on November 6, 2018, the app randomly skipped queued contacts when taking users to the messaging stage. Contacts were only skipped in queues of length five or greater, and each was slated to be skipped independently with probability 0.05. During this period, approximately 5,000 unique OUTVOTE users added 500,000 unique phone contacts to queues of length five or greater and ultimately sent approximately 132,000 GOTV messages.

Treatment We define the treatment ($D_i = 1$) to be receiving at least one GOTV text message from an OUTVOTE user during the study period. Users were able send a default message or craft their own. We estimate from simple text analysis that 98% of all sent messages were

the default or minor variants (e.g., with an emoticon inserted, or the recipient's name edited). Users could participate in one or more campaigns, each of which provided their own defaults. Approximately 88% of all messages sent were associated with non-partisan campaigns like OUTVOTE's "Text Every Voter" campaign or a campaign hosted by Vote.org. The remaining 12% of messages were associated with partisan campaigns, either those on behalf of a candidate (e.g., Alexandria Ocasio-Cortez for Congress) or broadly partisan campaigns (e.g., Swing Left). In the appendix, we provide the text and campaign of the ten most-used default messages, which account for 55% of all messages; in table 1, we show the top three.

Eligible subject pool During the study period, anyone whom an OUTVOTE user added to a queue of length five or greater was subject to randomization. Subjects must have also been successfully matched to the TARGETSMART voter rolls database for their outcomes to be observed. We further restrict our attention to subjects who were registered to vote prior to the study period. There are 195,118 subjects who meet these criteria.

Table 1: Examples of default messages.

Message template	Campaign	# of messages
Hey {FIRST_NAME}, I'm reminding all my friends to vote on Tue, Nov 6th! You can find your polling place at polls.vote.org. I'm using the Vote.org app. It takes 2 mins! https://votedotorg.outvote.io/vote	Vote.org	45,802
Hey, I'm using this app called Outvote to make sure my friends are registered to vote and get a reminder on an election day with their polling place. There are only a few days left, tell your friends! https://campaigns.outvote.io/outvote	OUTVOTE	5,867
OK, I'm voting this year. You? Election Day's Nov. 6 but MoveOn lets you text VOTE to 668366 if you want info on voting early or absentee. What do you think?	MoveOn.org	3,668

Assigned treatment and control groups A subject is classified as assigned to the control group if they were skipped in the first queue of length five or greater to which they were added ($Z_i = 0$); otherwise, they were classified as part of the treatment group ($Z_i = 1$). We determine each subject's first queue by examining the timestamps of users' queuing actions.

Refined study population To mitigate the bias introduced by mismatching subjects to voter rolls, we obtained ancillary match data from PREDICTWISE. Both OUTVOTE and PREDICTWISE matched 56,154 subjects to the same entry in the voter rolls.

To reduce the precision losses due to low compliance with assignment, we reconstructed subjects' queue positions from timestamp information indicating when users added subjects to queues. The position in the first queue can be confidently inferred for 60% of eligible subjects. A power analysis indicates that the subset of subjects who were within the top 103 positions in their first queue minimizes a proxy for the expected error in the CACE estimator (see fig. 2), but the results are largely unchanged when the threshold is raised or lowered.

There are $n = 27,464$ subjects in the refined study population who were (1) matched by OUTVOTE and PREDICTWISE to the same entry in the voter rolls, and (2) in the top 103 positions in their first queue. Of these, $1,454$ subjects (5.3%) are part of the assigned control group and the remaining $25,796$ part of the assigned treatment group. Table 2 summarizes the assignments, treatments received, and vote outcomes for the refined study population.

Covariates Pre-assignment covariate information on subjects is available from TARGETSMART's voter database, to which subjects were matched. We use 85 pre-assignment covariates to adjust CACE estimates. A full description of these covariates is provided in the appendix. The list includes age, number of previous general and primary election votes, household income, and an assortment of other variables based on multilevel modeling of survey data (*16*).

Table 2: The refined study population of $n = 27,464$ subjects.

	Assigned Treatment ($Z_i = 1$)			Assigned Control ($Z_i = 0$)		
	Overall	Messaged ($D_i = 1$)	Not messaged ($D_i = 0$)	Overall	Messaged ($D_i = 1$)	Not messaged ($D_i = 0$)
n	26,010	11,105	14,905	1,454	251	1,203
n voted	20,517	8,804	11,713	1,103	202	901
% voted	78.88	79.28	78.58	75.86	80.48	74.90

Results

Focusing solely on the 27,464 subjects who (1) were likely to be accurately matched to vote outcomes and (2) had reconstructed queuing positions of 103 or less, we estimate the CACE using the IV estimator in eq. (2):

$$\widehat{\text{ITT}} = 78.88 - 75.86 = 3.02 \text{ (s.e. } 1.10) \text{ percentage points.}$$

$$\widehat{\text{CACE}} = \frac{\widehat{\text{ITT}}}{\frac{11,105}{26,010} - \frac{251}{1,454}} = 11.88 \text{ (s.e. } 4.37) \text{ percentage points.}$$

The effect of assignment (ITT) is estimated to be 3.02 percentage points with a standard error of 1.10. The average causal effect among compliers is then estimated by dividing the $\widehat{\text{ITT}}$ by the estimated share of compliers, 25.43%, which yields 11.88 percentage points.

Using the formulas from corollary 1 of (*17*), we find that untreated compliers have an implied turnout rate of 66.88%, whereas treated compliers have an implied turnout rate of 78.48%. Given the high base rate of voting among compliers in this study, it is interesting that friend-to-friend appeals elevated turnout so profoundly.

Adding covariates to the model produces somewhat smaller estimates and standard errors. The estimated ITT is 2.09 percentage points with a standard error of 0.91 percentage points. The covariate-adjusted estimate of the CACE is 8.26 percentage points (SE = 3.61), which is still quite substantial by the standards of other large-scale GOTV experiments. We depict the confidence intervals of the covariate-adjusted and unadjusted estimates of the ITT and CACE

in figs. 3a and 3b.

The appendix presents results from other segments of the subject pool that we excluded from the main analysis. For example, the estimated CACE among those whose voting records were reliably accessed but whose queue position cannot be reconstructed is 11.90 (SE = 14.96). Across almost all segments of the data, well-matched records yield strong but often noisily estimated CACEs, while subjects who were poorly matched to voter turnout records produce estimated CACEs close to zero, consistent with attenuation bias.

Materials and Methods

CACE estimates and standard errors with and without covariates are obtained using the two-stage least squares implementation (`IV2SLS`) in the Python library `statsmodels` (*18*) while ITT estimates and standard errors are obtained using the ordinary least squares (`OLS`) implementation in the same library. We will release source code for ITT/CACE estimation, along with code for reconstructing queue position, and code for refining the study population from the general subject pool. We will also release data for the entire subject pool, stripped of all personally identifying information, which is sufficient to replicate all ITT/CACE effects without covariates.

Discussion

GOTV campaigns have become increasingly reliant on text messaging. But large-scale texting efforts typically originate from organizations rather than friends. These tactics often simulate a person-to-person conversation—if the recipient replies to the text, a campaign worker will engage in conversation—but actual exchanges are rare, and the effects on turnout tend to be modest. A recent meta-analysis (*1*) estimates their average effect to be 0.29 percentage points, despite the fact that the automated distribution system successfully delivered texts to more than 90% of targeted voters.

The decentralized friend-to-friend texting effort evaluated here produced much larger turnout effects. Although fewer than one-third of the intended targets actually received messages, the estimated ITT effect (2.09 percentage points) is nevertheless many times larger than the apparent effect of automated texting. The estimated effect among compliers is 8.3 percentage points, one of the strongest effects to emerge from a large randomized GOTV trial.

What aspects of friend-to-friend texting might account for this unusually strong effect? Three hypotheses suggest themselves. First, this finding is consistent with a substantial body of experimental evidence suggesting that personal appeals to vote (e.g., authentic conversations in person or by phone) tend to be more effective than recorded messages or mass emails (*1*). Second, the effects of GOTV appeals may be amplified when the messenger is known to the receiver. For example, although email GOTV messages tend to be ineffective, some small RCTs suggest that email encouragements from friends can increase turnout substantially (*19*). Third, GOTV effects tend to be enhanced when senders exert some degree of social pressure, either by suggesting that they are counting on the recipient to vote and will be disappointed otherwise (*20*). Testing these causal mechanisms by systematically adding or subtracting aspects of personalization, close personal ties between users and receivers, and exertion of social pressure is a fruitful line of future inquiry.

Researchers who endeavor to investigate causal mechanisms or the effectiveness of friend-to-friend texting with different target populations are likely to face a number of trade-offs akin to those that OUTVOTE faced. OUTVOTE elegantly designed its randomized experiment so as to minimally degrade the user experience. The upside of this unobtrusive approach is the naturalistic way in which users communicated with the contacts whom they queued. But the downside is a host of statistical impediments to assessing causal effects: low compliance rates decrease precision, and mismatched outcome data introduce bias.

While one can imagine more obtrusive approaches that might have mitigated these prob-

lems, they come with changes to the user experience that might preclude any useful data. The app could have dropped contacts at a higher rate (e.g., 10%), prevented users from re-queuing contacts, or nagged users to finish messaging everyone they queued. These steps might have increased compliance, but also might have led users to less natural behavior or even to abandon the service altogether. The app could have also asked users to furnish information about their contacts, a step that might have improved the rate of accurate matches to vote outcomes. But requesting this information also changes the user experience and might drive users away.

Answering causal questions about authentic friend-to-friend behavior constitutes an important but challenging class of applications. Experiments suited to answer such questions require light-touch encouragements; these help ensure a large n to offset the debilitating statistical consequences of low compliance rates and unreliable outcome measurement. In the face of such statistical challenges, the methods used here are essential to accurately estimating causal effects among informative subsets of the subjects.

References

1. D. P. Green, A. S. Gerber, *Get out the vote: How to increase voter turnout* (Brookings Institution Press, 2019).

2. H. Teresi, M. R. Michelson, *The Social Science Journal* **52**, 195 (2015).

3. K. Haenschen, *Journal of Communication* **66**, 542 (2016).

4. D. P. Green, Evaluation of peer-to-peer voter mobilization campaign by alliance for climate education during the 2018 midterm election, *Tech. rep.*, Working Paper (2019).

5. Outvote, https://www.outvote.io/ (2020). Accessed: 2020-04-21.

6. J. D. Angrist, G. W. Imbens, D. B. Rubin, *Journal of the American statistical Association* **91**, 444 (1996).

7. M. A. Hernan, J. M. Robins, Causal inference (2010).

8. J. D. Angrist, J.-S. Pischke, *Mostly harmless econometrics: An empiricist's companion* (Princeton university press, 2008).

9. D. Staiger, J. H. Stock, *et al.*, *Econometrica* **65**, 557 (1997).

10. D. B. Rubin, *Journal of the American statistical association* **75**, 591 (1980).

11. B. Sinclair, M. McConnell, D. P. Green, *American Journal of Political Science* **56**, 1055 (2012).

12. P. Kristensen, *Epidemiology* pp. 210–215 (1992).

13. A. Lewbel, *Econometrica* **75**, 537 (2007).

14. K. Imai, T. Yamamoto, *American Journal of Political Science* **54**, 543 (2010).

15. R. K. Crump, V. J. Hotz, G. W. Imbens, O. A. Mitnik, Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand, *Tech. rep.*, National Bureau of Economic Research (2006).

16. T. Konitzer, S. Corbett-Davies, D. M. Rothschild, Non-representative surveys: Modes, dynamics, party, and likely voter space, *Tech. rep.*, Working Paper (2017).

17. P. M. Aronow, D. P. Green, *Statistics & Probability Letters* **83**, 677 (2013).

18. S. Seabold, J. Perktold, *9th Python in Science Conference* (2010).

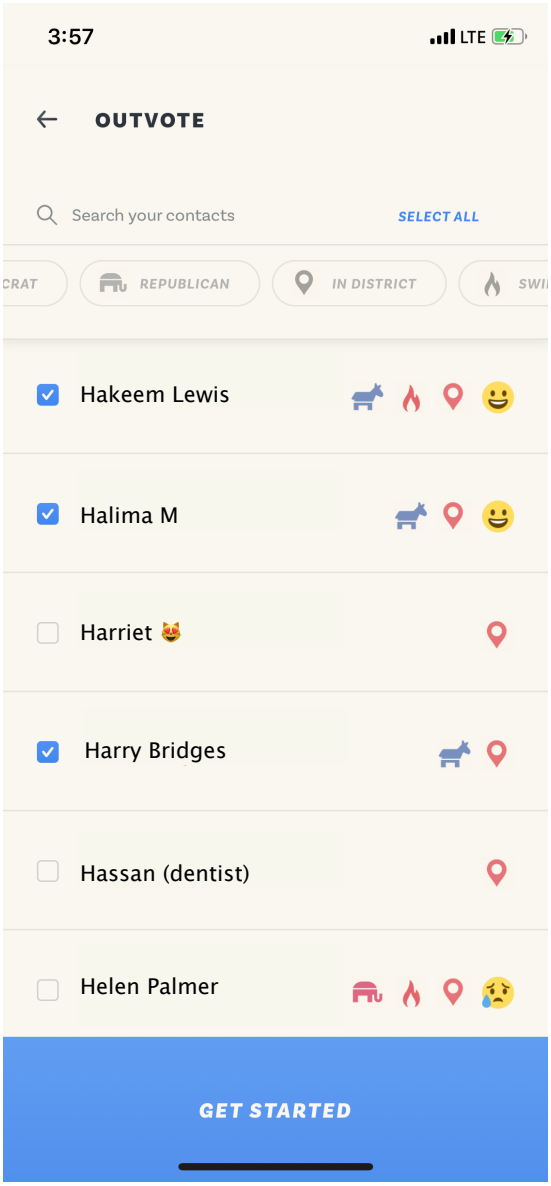
19. T. C. Davenport, *Poster presented at the annual meeting of the American Political Science Association, Boston, MA* (2008).

20. T. C. Davenport, *Political Behavior* **32**, 337 (2010).

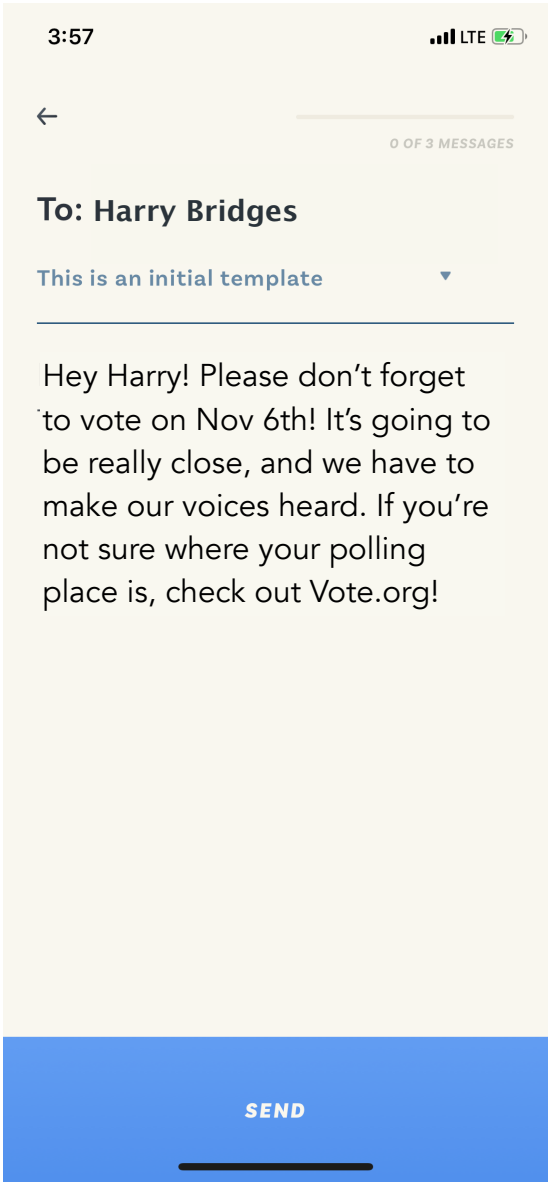
Acknowledgments

We thank PredictWise for providing voter file match data and thank Roy Adams and Otis Reid for helpful discussions.

Supplementary materials



(a) Queuing stage. Users select contacts to message from a ranked list. Contacts are ranked by tiers within which they are sorted alphabetically. The top tier are contacts registered in a district with an upcoming election.



(b) Messaging stage. Users are taken to a messaging interface for each contact on their queue. The user can use the default message or craft their own.

Figure 1: The two stages of an OUTVOTE user's workflow. The app injected randomness between the queuing stage (a) and the messaging stage (b) by skipping selected contacts in the user's queue.

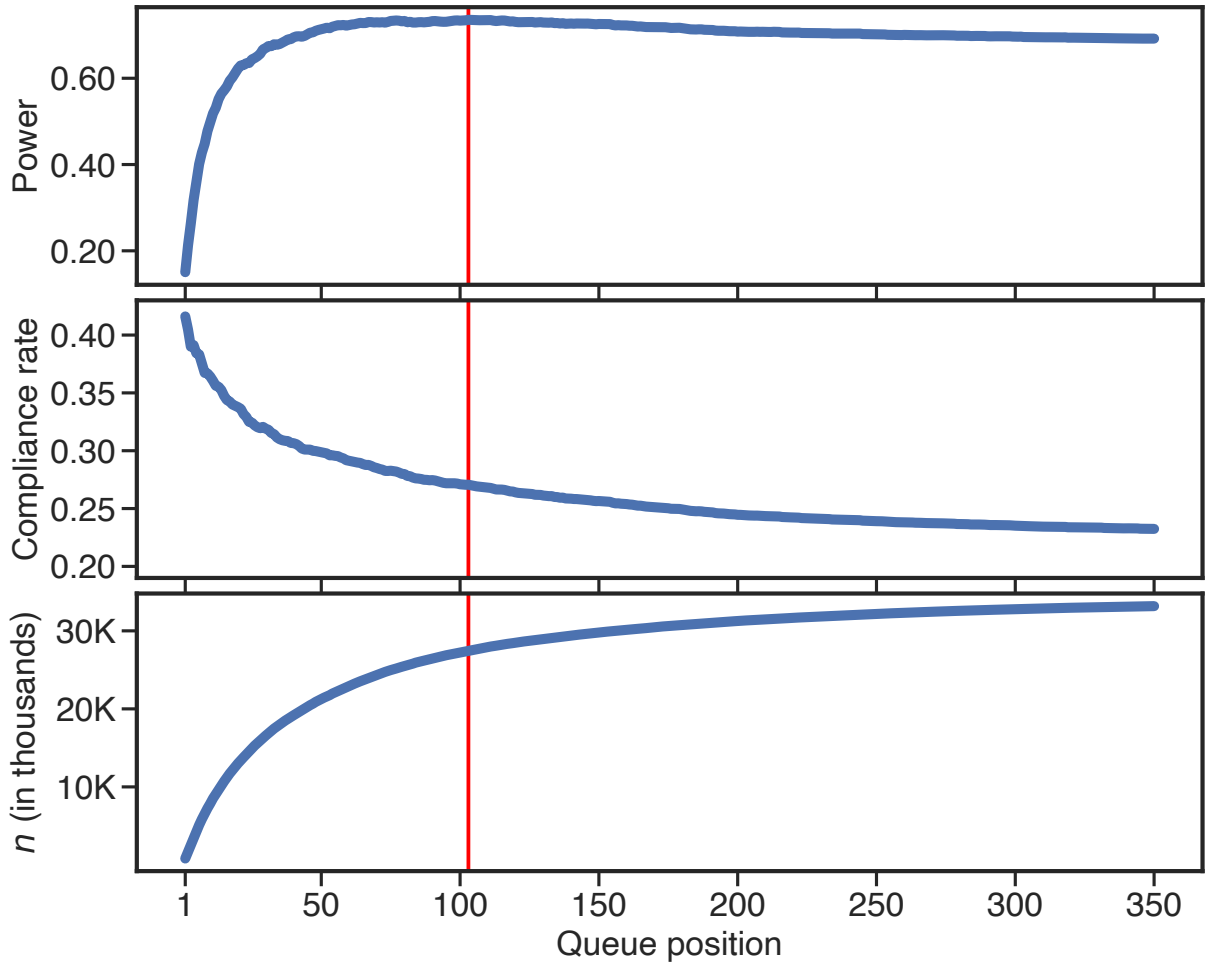
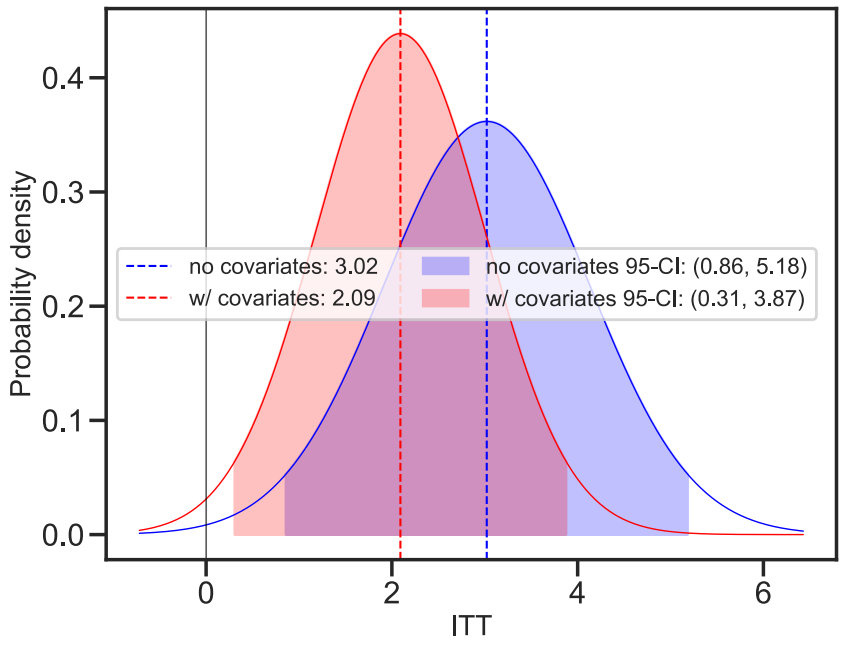
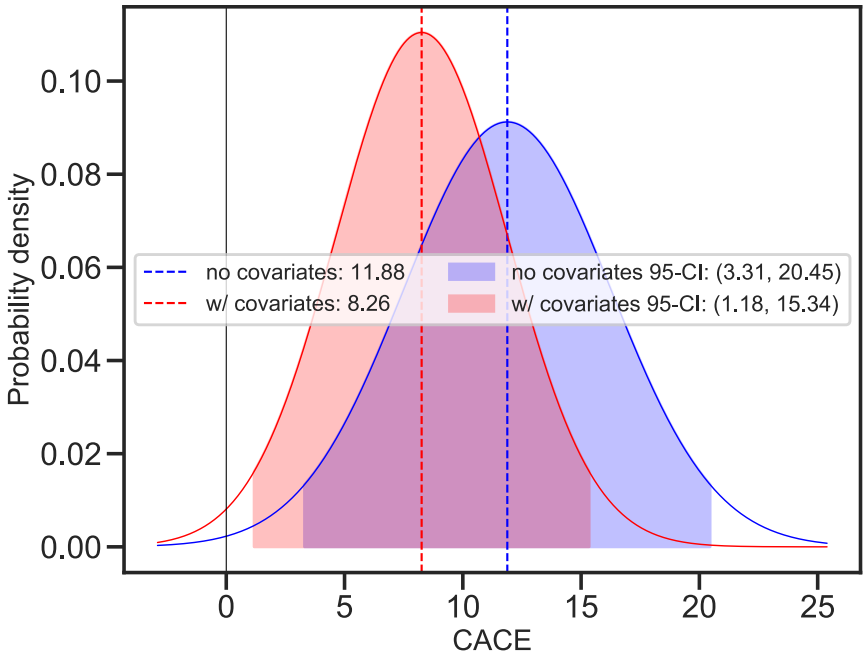


Figure 2: The trade-off between n and compliance, as the maximum allowable position of subjects in their first queue increases. The study's power is influenced by both. It degrades after $q_{\max} = 103$ (red line), when increasing n fails to compensate for lower compliance rates.



(a) ITT estimates and 95% confidence intervals with and without covariates.



(b) CACE estimates and 95% confidence intervals with and without covariates.

Figure 3: ITT and CACE estimates with and without covariates.

Supplementary Materials for "A Digital Field Experiment Reveals Large Effects of Friend-to-Friend Texting on Voter Turnout"

Aaron Schein,[1*] Keyon Vafa,[2] Dhanya Sridhar,[1] Victor Veitch,[3]
Jeffrey Quinn,[5] James Moffet,[6] David M. Blei,[1,2,3] Donald P. Green[4]

[1]Data Science Institute, Columbia University, New York, NY 10027
[2]Department of Computer Science, Columbia University, New York, NY 10027
[3]Department of Statistics, Columbia University, New York, NY 10027
[4]Department of Political Science, Columbia University, New York, NY 10027
[5]PredictWise, https://www.predictwise.com, New York, NY 10036
[6]JDM Design, Cambridge, MA 02139

[*]To whom correspondence should be addressed; E-mail: aaron.schein@columbia.edu.

Supplementary Text

A Proof of attenuation bias under non-differential outcome mismeasurement

When outcomes are binary, the numerator of the IV estimator, which estimates the ITT, equals

$$\mathbb{E}[\,Y_i \,|\, Z_i{=}1] - \mathbb{E}[\,Y_i \,|\, Z_i{=}0] = P(Y_i{=}1 \,|\, Z_i{=}1) - P(Y_i{=}1 \,|\, Z_i{=}0). \tag{1}$$

When measurement error is present, we introduce the true outcome $Y_i^\star$ as a latent variable that is marginalized out:

$$= \sum_{y^\star=0}^{1} \Big(P(Y_i{=}1, Y_i^\star = y^\star \,|\, Z_i{=}1) - P(Y_i{=}1, Y_i^\star = y^\star \,|\, Z_i{=}0) \Big). \tag{2}$$

Due to random assignment of Z_i, the mismeasurement is non-differential in the sense that $Y_i \perp Z_i \mid Y_i^\star$ and therefore:

$$= \sum_{y^\star=0}^{1} \Big(P(Y_i{=}1 \mid Y_i^\star = y^\star)\, P(Y_i^\star = y^\star \mid Z_i{=}1) - P(Y_i{=}1 \mid Y_i^\star = y^\star)\, P(Y_i^\star = y^\star \mid Z_i{=}0) \Big) \tag{3}$$

$$= \Big(\underbrace{P(Y_i{=}1 \mid Y_i^\star = 1) - P(Y_i{=}1 \mid Y_i^\star = 0)}_{\text{bias term } \pi_{\text{BIAS}}} \Big) \Big(\underbrace{P(Y_i^\star = 1 \mid Z_i{=}1) - P(Y_i^\star = 1 \mid Z_i{=}0)}_{\text{true ITT}} \Big). \tag{4}$$

If we assume that the magnitude of the measurement error is not too large, so that

$$P(Y_i{=}1 \mid Y_i^\star{=}1) > P(Y_i{=}1 \mid Y_i^\star{=}0), \tag{5}$$

then the bias term $\pi_{\text{BIAS}} \in (0, 1]$ only attenuates the true ITT.

B Reconstructing queue positions

As discussed in the main paper, a subject's position Q_i in their first queue is a pre-assignment variable that lets us refine the study population to improve its compliance rate.

OUTVOTE did not explicitly save queue positions. However, it did save data that allows us to reconstruct them. Specifically, OUTVOTE's database stored "queuing events", each of which is a 3-tuple (t, u, r) consisting of the timestamp t at which user u queued phone contact r. OUTVOTE only recorded a queuing event if it was subject to randomization—i.e., part of a queue of length five or greater and during the study period. Queue positions can be reconstructed from the timestamps since the order of subjects in the queue was the order in which the user queued them.

Our approach to reconstructing queue position involves two basic functions. The first function inputs a set of queuing events and outputs whether this set meets a minimal plausibility criteria of being a queue. These criteria are: 1) there are 5 or more queuing events, 2) all events involve the same user, 3) no phone contact is queued more than once, 4) events are

sorted chronologically, and 5) the time between any two adjacent events is not implausibly long (greater than 1 hour). Algorithm 1 provides pseudocode for this function.

The second function partitions all queuing events involving the same user into separate queues and then calls the first function to check that each proposed queue meets the minimal plausibility criteria. If any one of the proposed queues does not meet the plausibility criteria, then all proposed queues for that user are nullified. Otherwise, the positions of each subject in the queues are returned. The pseudocode for this function is given in algorithm 2.

One wrinkle is that users could press an "Add all" button that would instantly add all of their phone contacts to the queue. Many users would hit this, then immediately exit the queue, and start a new queue to which they would add contacts selectively. This user behavior would create many queuing events with very similar timestamps that included the same contacts twice, thus making it difficult to partition the queues. To account for this behavior, our criteria for partitioning queuing events into queues is strict: we consider two events more than 3 seconds apart to be part of different queues. This criteria often successfully separates the accidental "Add all" events from the subsequent selective queues. It was rare for users to take longer than 1 second between queuing events within the same queue, as evidenced by the average time between events for users who only ever queued 9 or fewer contacts (which must all be part of the same queue since each queue had five or more contacts); thus this strict criteria should rarely partition a single queue into multiple ones.

With the strict 3-second criteria, the queuing events for 85% of users can be partitioned into queues that meet the plausibility criteria and the queue position for 60% of eligible subjects can be reconstructed. The remaining 40% of subjects whose first queue position cannot be reconstructed consist mainly of subjects who were added by an accidental "Add all"—these subjects exhibit a very low compliance rate of 8% (see fig. 2) as would be expected from an accidentally queued subpopulation.

Algorithm 1 Check if a sequence of queuing events meets the minimal criteria to constitute a valid queue

Input: QUEUINGEVENTS $= [(t_1, u_1, r_1), (t_2, u_2, r_2), \dots]$, a list of queuing events where each event is a 3-tuple of the timestamp t_n when user u_n queued the receiver r_n.

Output: True or False, whether the list of queuing events constitutes a valid queue.

procedure ISVALIDQUEUE(QUEUINGEVENTS)
  if LENGTH(QUEUINGEVENTS) < 5 **then**
    return FALSE ▷ Queues must be length 5 or greater.
  end if
  $u^\star \leftarrow u_1$ ▷ Get user of first event
  $t^\star \leftarrow$ NaN ▷ Initialize previous timestamp
  SEENRECEIVERS $\leftarrow [\,]$ ▷ Initialize set of receivers in this queue
  for (t_n, u_n, r_n) in QUEUINGEVENTS **do**
    if r_n in SEENRECEIVERS **then**
      return FALSE ▷ The same person cannot appear on a queue more than once
    end if
    APPEND(SEENRECEIVERS, r_n)
    if $u_n \neq u^\star$ **then**
      return FALSE ▷ Only one user per queue
    end if
    if $t^\star$ is not NaN **then**
      if $t_n < t^\star$ **then**
        return FALSE ▷ Queue must be sorted by timestamps
      end if
      if $(t_n - t^\star)$ ¿ 1 hour **then**
        return FALSE ▷ Time between any two queuing events shouldn't be implausibly long
      end if
    end if
    $t^\star \leftarrow t_n$
  end for
  return TRUE
end procedure

Algorithm 2 Get queue IDs for queuing events

Input: QUEUINGEVENTS $= [(t_1, u_1, r_1), (t_2, u_2, r_2), \dots]$, a list of queuing events where each event is a 3-tuple of the timestamp t_n when user u_n queued the receiver r_n.

Output: QUEUINGIDS $= [q_1, q_2, \dots]$, a list of queue IDs for every input queuing event.

procedure GETQUEUEIDS(QUEUINGEVENTS)
    $q^\star \leftarrow 0$ ▷ Initialize current queue ID
    for user $u^\star$ in UNIQUEVALUES($[u_1, u_2, \dots]$) **do**
        $t^\star \leftarrow$ NaN ▷ Initialize current timestamp
        for (t_n, u_n, r_n) in QUEUINGEVENTS such that $u_n = u^\star$ **do**
            if $t^\star$ is NaN or $(t_n - t^\star) > 3$ seconds **then**
                $q^\star \leftarrow q^\star + 1$ ▷ Update current queue ID to a new queue
            end if
            $q_n \leftarrow q^\star$ ▷ Assign current queue ID to queuing event
            $t^\star \leftarrow t_n$ ▷ Update current timestamp
        end for
    end for
    for queue $q^\star$ in UNIQUEVALUES($[q_1, q_2, \dots]$) **do**
        if not ISVALIDQUEUE($[(t_n, u_n, r_n)$ such that $q_n = q^\star]$) **then**
            for q_n such that $u_n = u^\star$ **do**
                $q_n \leftarrow$ NaN ▷ If clustering yields an invalid queue, undo queue IDs for that user
            end for
        end if
    end for
end procedure

C Selecting $q_{\max}$

As discussed in the main paper, setting a maximum allowable queue position $q_{\max}$ defines a subpopulation of subjects whose first queue position is $Q_i \leq q_{\max}$. A smaller value for $q_{\max}$ defines a subpopulation that is more compliant but also smaller. The variance of the IV estimator is a function of both the compliance rate and n—thus, there is a tradeoff in selecting $q_{\max}$ to minimize the expected variance. We can write the IV estimator as an explicit function of $q_{\max}$,

$$\hat{\beta}_{\text{IV}}(q_{\max}) = \frac{\bar{y}_{1q_{\max}} - \bar{y}_{0q_{\max}}}{\bar{d}_{1q_{\max}} - \bar{d}_{0q_{\max}}}, \tag{6}$$

where $\bar{y}_{1q_{\max}} = \hat{\mathbb{E}}\left[Y_i \mid Z_i = 1, Q_i \leq q_{\max}\right]$ is the mean voting outcome among treatment subjects in the subpopulation defined by $q_{\max}$, $\bar{y}_{0q_{\max}}$ is the corresponding mean among control subjects, and $\bar{d}_{1q_{\max}}$ and $\bar{d}_{0q_{\max}}$ are the analogous means of receipt.

As mentioned in the main paper, we emphasize that by setting $q_{\max}$ we change the estimand to a conditional CACE—i.e., $\hat{\beta}_{\text{IV}}(q_{\max})$ estimates $\mathbb{E}\left[Y_{i1} - Y_{i0} \mid i \in \mathcal{C} \text{ and } Q_i \leq q_{\max}\right]$. Our approach is analogous to that of Crump et al. (2006) (*1*) who also propose a systematic way to trim their sample in order to minimize the expected variance of the estimator. These approaches seek to estimate the estimand most precisely estimable albeit with a loss of generalizability. In our case, we do not have reason to believe that the treatment effect among compliers high in the queue should be different than those low in the queue.

We select $q_{\max}$ to minimize a proxy for the expected variance of the IV estimator, or equivalently, to maximize its expected power. The expected variance can be written as a function (whose form is given later) of the following quantities,

$$\mathbb{V}\left(\hat{\beta}_{\text{IV}}(q_{\max})\right) = f\big(\beta(q_{\max}), \bar{y}_{0q_{\max}}, \bar{d}_{1q_{\max}}, \bar{d}_{0q_{\max}}, n_{q_{\max}}, p\big), \tag{7}$$

where $\beta(q_{\max})$ is the true effect, $n_{q_{\max}}$ is the size of the subpopulation, and $p = P(Z_i = 1)$ is the probability of being assigned to receive treatment, which is $p = 0.95$ in our case.

We plug-in assumed values for $\beta(q_{\text{max}})$ and $\bar{y}_{0q_{\text{max}}}$ and empirical values of $\bar{d}_{1q_{\text{max}}}, \bar{d}_{0q_{\text{max}}},$ and $n_{q_{\text{max}}}$. Using the circle notation $(\circ)$ to denote assumed quantities, we assume a large true effect size $\beta^{\circ} = 0.1$ (10 percentage points), and a control voting rate of $\bar{y}_0^{\circ} = 0.7$, both of which are constant across q_{max}. For each value of q_{max} we calculate $n_{q_{\text{max}}}$ from the study population, and estimate the average rates of treatment receipt $\bar{d}_{1q_{\text{max}}}$ and $\bar{d}_{0q_{\text{max}}}$ from the pool of subjects excluded from the analysis due to poor match quality[1]. We then select q_{max} to be,

$$q^*_{\text{max}} \leftarrow \underset{q_{\text{max}}}{\text{argmin}}\ f\big(\beta^{\circ}(=0.1), \bar{y}_0^{\circ}(=0.7), \bar{d}_{1q_{\text{max}}}, \bar{d}_{0q_{\text{max}}}, n_{q_{\text{max}}}, p(=0.95)\big), \tag{8}$$

and get a value of $q^*_{\text{max}} = 103$. This analysis finds the q_{max} that minimizes the expected variance under the assumption that the only thing which varies by q_{max} is n_{max} and the compliance rate.

The form of the expected variance of the IV estimator $f(\dots)$ can be obtained using the Delta method (*2, 3*) as

$$\mathbb{V}\left(\hat{\beta}_{\text{IV}}\right) = \frac{\sigma_U^2}{\mu_V^2} - 2\frac{\mu_U}{\mu_V^3}\sigma_{U,V} + \frac{\mu_U^2}{\mu_V^4}\sigma_V^2, \tag{9}$$

where the following quantities are defined,

$$\mu_U = \bar{y}_1 - \bar{y}_0 \tag{10}$$

$$\mu_V = \bar{d}_1 - \bar{d}_0 \tag{11}$$

$$\sigma_U^2 = \tfrac{1}{n_1}\bar{y}_1(1-\bar{y}_1) + \tfrac{1}{n_0}\bar{y}_0(1-\bar{y}_0) \tag{12}$$

$$\sigma_V^2 = \tfrac{1}{n_1}\bar{d}_1(1-\bar{d}_1) + \tfrac{1}{n_0}\bar{d}_0(1-\bar{d}_0) \tag{13}$$

$$\sigma_{U,V} = \beta\,\sigma_V^2. \tag{14}$$

Here n_1 and n_0 are the number of subjects in treatment and control groups, $\bar{y}_1$ and $\bar{y}_0$ are the outcome means for subjects in the treatment and control groups, $\bar{d}_1$ and $\bar{d}_0$ are the receipt means,

[1] We estimate these quantities using the subjects excluded from the analysis to avoid the poor optics "double-dipping" the same data to both select q_{max} and estimate effects. However, we do not believe this is necessary—i.e., using estimates of $\bar{d}_{1q_{\text{max}}}$ and $\bar{d}_{0q_{\text{max}}}$ to select q_{max} should not introduce confounding since Q_i is a pre-assignment variable.

and β is the true CACE. For our analysis, we plug in assumed values for $\beta = \beta^\circ = 0.1$ and $\bar{y}_0 = \bar{y}_0^\circ = 0.7$, which are constant across proposed values of q_{max}. We plug in $n_1 = p\, n_{q_{\text{max}}}$ and $n_0 = (1-p)\, n_{q_{\text{max}}}$ based on $n_{q_{\text{max}}}$ calculated from the study population. We also plug in $\bar{d}_1 = \bar{d}_{1 n_{q_{\text{max}}}}$ and $\bar{d}_0 = \bar{d}_{0 n_{q_{\text{max}}}}$ based on sample means from the subjects excluded from analysis due to poor match quality. The value of $\bar{y}_1$ is determined by the other plug-in values—i.e., $\bar{y}_1 = \beta(\bar{d}_1 - \bar{d}_0) + \bar{y}_0$.

To better interpret the results, we report the expected power. Our power analysis asks: what is the probability of rejecting the null hypothesis of a zero effect, given that the true effect is large? Power is a function of the expected variance of the estimator and the assumed true effect:

$$\text{Power}(\hat{\beta}_{\text{IV}}(q_{\text{max}}); \beta^\circ) = P(\text{reject } H_0 : \beta = 0 \,|\, \beta = \beta^\circ) = 1 - \Phi\Big(1.96 - \frac{\beta^\circ}{\sqrt{\mathbb{V}\big(\hat{\beta}_{\text{IV}}(q_{\text{max}})\big)}}\Big). \quad (15)$$

Power incorporates the scale of the standard error in relation to the scale of the effect size. A standard error of 1 is small if the true effect size is 15 but large if the true effect size is 0.1. Power synthesizes the expected error and the assumed effect size into a single number between 0 and 1. Assuming a large CACE of 0.1, we see in fig. 1 that power declines significantly after $q_{\text{max}} = 103$ while the corresponding increase in standard error is less evident due to scale of the y-axis.

D Balance checks

We use the pre-assignment covariates from the TARGETSMART database to perform balance checks on subpopulations of the subject pool which test the null hypothesis that the covariates are no better than random at predicting subjects' assignments. For any subset of subjects, we fit the following ordinary least squares (OLS) regression,

$$Z_i = w_0 + \boldsymbol{w}^\top \boldsymbol{X}_i + \epsilon_i. \quad (16)$$

An F-test of this regression is a test of the null hypothesis that the covariates $\boldsymbol{X}_i$ are not predictive of the dependent variable Z_i; we report the p-value of this F-test. A balance check fails

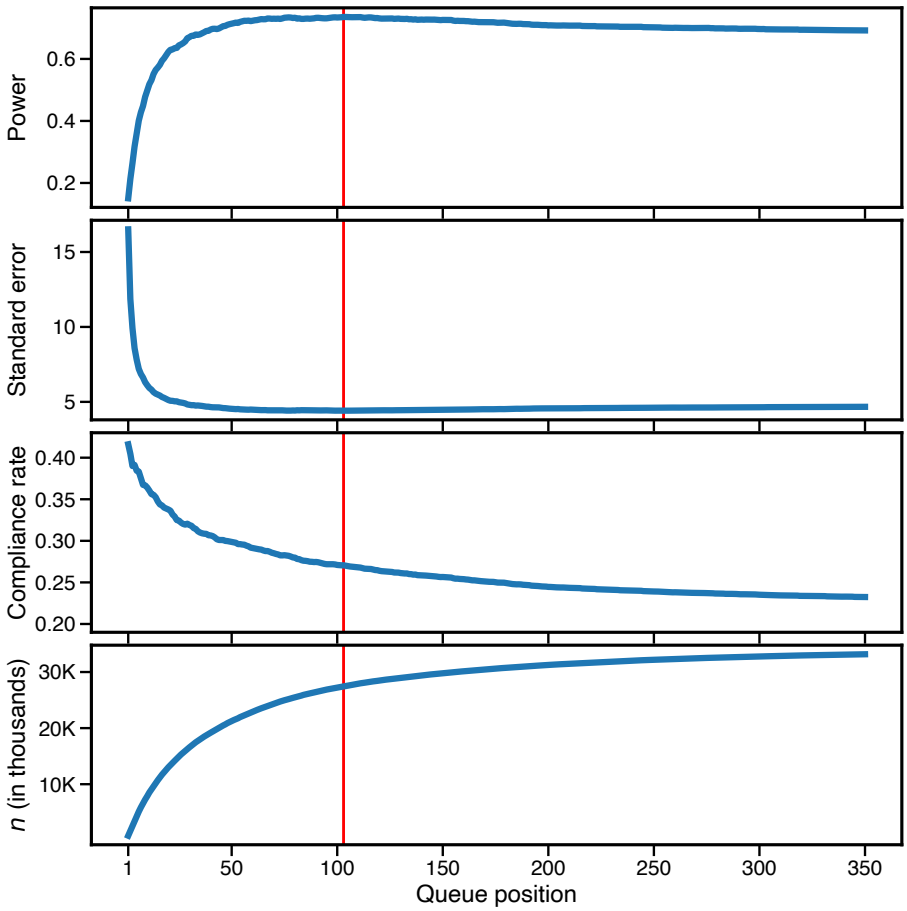


Figure 1: Tradeoff.

when a small p-value indicates that that the null hypothesis is unlikely and thus that the assigned treatment and control groups do not exhibit symmetry.

We report the balance check's p-value for all subpopulations in fig. 2—all pass the check. Since Z_i is randomized, we do not expect any subpopulation to fail the balance check; this exercise serves simply to provide empirical support that randomization was correctly implemented and our definition of the treatment-control split maintains symmetry between the two groups.

As mentioned in footnote 1 of the main paper, a possibly tempting alternative definition of the treatment–control split defines the control group to consist of subjects who were skipped in *all* queues in which they appeared (as opposed to only their first queue, the definition we adopt

in this study). In this case, the probability of being assigned to receive treatment depends on how many queues a subject appeared in k—i.e., $P(Z_i = 1) = 1 - 0.05^k$. However, this definition introduces a confounder, since every queue after the first is potentially affected by the user noticing the subject being skipped (or not skipped) the first time. Indeed, the study population defined in this way fails the balance check, producing a p-value of 0.01, suggesting that the covariates $\boldsymbol{X}_i$ are predictive of assignment Z_i and thus that Z_i is not random. This is likely due to the fact that users were more likely to notice if a subject higher in the queue was skipped and thus were more likely to re-queue higher-ranked subjects. Since OUTVOTE's ranking of subjects in the queuing phase used covariate information from the voter rolls, systematically re-queuing higher-ranked subjects induces confounding that is detected by a balance check. A previous version of this study, described in a pre-analysis plan registered with EGAP[2] (this link may be down due to EGAP migrating to Open Science Framework)., employed this confounded definition of treatment–control.

E Subject flow diagram

We depict how subjects were selected for analysis in fig. 2. Box 1 represents the total number of unique mobile phone numbers that users added to queues ($n = 546,510$). Of these, only $n = 195,118$ met the minimal eligibility criteria of having been successfully matched to public voter rolls and being registered to vote prior to the study (Box 3). Every node in the flow diagram below and including Box 3, represents a specific subset of eligible subjects. Each node is annotated with: 1) the number n of subjects in that subset, 2) the ITT and CACE estimates for that subset, and 3) the p-value from a balance check that tests the null hypothesis of symmetry between the assigned treatment and control groups (see appendix D).

The first refinement of the eligible subject pool, labeled "Measurement Error" in fig. 2,

[2]https://egap.org/content/effect-relational-text-messaging-turnout-2018-us-midterm-elections

involves subjects who were well-matched (Box 4: $n = 56,154$) versus poorly matched (Box 5: $n = 138,964$) to the public voter rolls. As described in the main text, we consider a subject to be well-matched if both OUTVOTE and PREDICTWISE matched them to the same entry—we assume that interannotator agreement in a subject's match is an indicator of the match's accuracy. The data support this assumption: the ITT estimate in the poorly matched population (Box 5) is nearly zero, -0.03 (s.e. 0.57), which is consistent with attenuation bias expected under measurement error.

The next two refinements, labeled "Non-compliance", seek to improve the compliance rates of the study population. The first refinement considers subjects whose position in their first queue can (Box 6: $n = 34,200$) versus cannot (Box 7: $n = 21,954$) be reconstructed from timestamp information of when users added subjects to their queues. The major contributing factor to why timestamps do not always reconstruct subjects' queue position is due to users accidentally pressing an "Add all" button (see appendix B). The compliance rate among the subjects whose queue position cannot be reconstructed (Box 7) is low (9%), as would be expected from a subpopulation of subjects who were accidentally added to queues.

The final refinement considers subjects who were within the top $q_{\max} = 103$ positions in their first queue (Box 8: $m = 27,464$) versus those who were not (Box 9: $n = 6,736$). The compliance rate among subjects whose position was greater than 103 (Box 9) is very low (3%)—as a result, the standard error of the CACE estimate for that subpopulation is very high (126.24 percentage points). Box 8 represents the refined study population which we use to estimate the results presented in the main text.

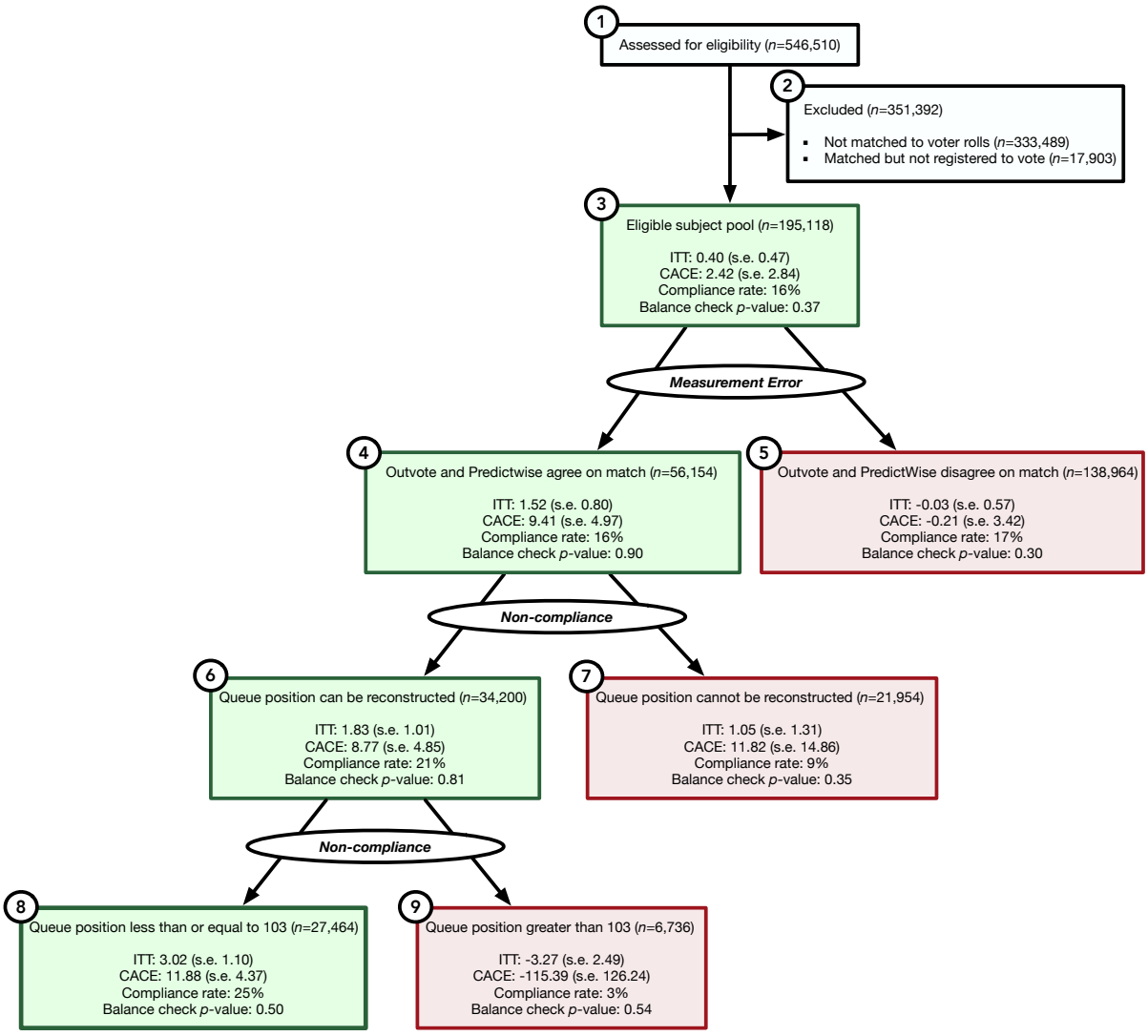


Figure 2: Subject flow diagram.

F Further details on matching subjects to voter rolls

OUTVOTE's matching algorithm was based on the subject's mobile phone number and the first and last name fields stored in the user's phone contacts. OUTVOTE's algorithm first checked whether there was any unique match of a person with the exact first and last name registered in the precinct associated with the mobile phone number's area code. If this failed, the algorithm entered the mobile phone number into an API that sometimes returned the first and last name of the person who paid the mobile phone number's bill, or who was otherwise associated with the phone number. The algorithm would try again to find a unique match using this alternative name. The algorithm also relied on a gazetteer of common name shortenings (e.g., Richard: Dick, Rich, Ricky,...) to find non-exact matches, when the previous attempts failed.

PREDICTWISE has on a number of commercially-available marketing databases that associates mobile phone numbers with an extensive set of demographic information that is useful for matching phone numbers to the voter rolls. The data we obtained from PREDICTWISE contains millions of pairs of mobile phone numbers and voter roll IDs. A phone number may be associated with multiple voter roll IDs in this data.

We assume that instances where OUTVOTE's matches appear in PREDICTWISE's data, are much less likely to be misclassified than matches that do not. One reasonable question is why not simply use PREDICTWISE's matches, ignoring OUTVOTE's. First, a give subject (i.e., mobile phone number) is associated with multiple voter roll entries by PREDICTWISE—it's not clear how to decide among them, if we are not using OUTVOTE's match to disambiguate. Second, OUTVOTE's match was made at the time of the study, while PREDICTWISE may have matched the same mobile phone number to a voter ID years earlier, when the phone number belonged to someone else. Third, it's important to condition on which identity OUTVOTE matched a subject to, even if that identity was incorrect: since OUTVOTE presented that information to

users and used it to rank subjects, it still likely helps reduce error of the CACE, when conditioning on covariates.

G Messages

Users could either send default messages or craft their own. The vast majority of messages were either the exact default message or a minor variant (e.g., with an emoticon inserted, or the recipient's name edited). Based on simple string matching to default templates, we estimate that about 98% of the 132,318 messages sent were the default message or minor variants.

Users participated in different campaigns, each of which provided their own default message. The vast majority of messages are part of a campaign by Vote.org or OUTVOTE's "Text Every Voter" campaign. The most-used default message accounts for 45,802 messages (about 35% of all sent messages); we show the 10 most-used default messages in table 1, which account for over 55% of all sent messages.

Users participated in 72 unique campaigns during the study period. While 67 of these campaigns were partisan, the vast majority of user activity was accounted for by the non-partisan campaigns. About 88% of sent messages were part of a non-partisan campaign, like Vote.org or OUTVOTE's "Text Every Voter".

Table 1: Top ten most-used default messages in descending order. These account for 55% of all sent messages.

Message template	Campaign	# of messages	% of all messages
Hey {FIRST_NAME}, I'm reminding all my friends to vote on Tue, Nov 6th! You can find your polling place at polls.vote.org. I'm using the Vote.org app. It takes 2 mins! https://votedotorg.outvote.io/vote	Vote.org	45,802	35.6%
Hey, I'm using this app called Outvote to make sure my friends are registered to vote and get a reminder on an election day with their polling place. There are only a few days left, tell your friends! https://campaigns.outvote.io/outvote	OUTVOTE	5,867	4.4%
Hey, I'm using this app called Outvote to make sure my friends are registered to vote and get a reminder on an election day with their polling place. If you want one just text "vote" to (202) 868-8683.	OUTVOTE	5,761	4.4%
OK, I'm voting this year. You? Election Day's Nov. 6 but MoveOn lets you text VOTE to 668366 if you want info on voting early or absentee. What do you think?	MoveOn.org	3,668	2.8%
Hey {FIRST_NAME}, Hey are you sure you're registered to vote at the right address? I'm reminding all my friends to double check: https://www.vote.org/am-i-registered-to-vote/	Vote.org	3,328	2.5%
Hey, I'm using this app called Outvote to remind my friends to vote! If you want to help, it tells you who you know in swing districts too.	OUTVOTE	2,610	2.0%
Hey! Please don't forget to vote on Nov 6th! It's going to be really close, and we have to make our voices heard. If you're not sure where your polling place is, check out vote.org.	Vote.org	2,501	1.9%
I know you're gonna vote on November 6th DUH, but make sure to remind your friends, too! Download Outvote and find out who you know in swing districts :D https://campaigns.outvote.io/outvote	OUTVOTE	2,248	1.7%
Hey, I'm reminding my friends to check that they're registered at the right address this year just in case they forget. You can check in a few minutes right here: https://www.vote.org/am-i-registered-to-vote/	Vote.org	1,132	0.9%
Hey, are you registered to vote? It only takes 2 minutes: https://www.vote.org/register-to-vote/	Vote.org	1,073	0.8%

H Covariates from TARGETSMART voter roll database

A rich set of covariates on subjects is available from the TARGETSMART voter roll database. OUTVOTE matched subjects to the voter roll database prior to users adding subjects to queues. We use a set of 85 covariates that were recorded prior to random assignment of subjects; these covariates are listed and described in tables 2 to 11.

Table 2: Pre-assignment covariates. Descriptions are from TARGETSMART documentation.

			Mean by group	
Variable name	Variable support	Description	Z_i=1	Z_i=0
vb.education	$\{1, \ldots, 6\}$	Indicates the highest level of education completed by the individual	2.38	2.32
vb.discretionary_income_decile	$\{0, \ldots, 10\}$	Household-level ranking within a geographic area of the annual amount of discretionary income of the household. Ranking of the amount of cash a household has after paying for essentials such as food, electricity, taxes, education and mortgage. 1 - Highest 10 - lowest	3.71	3.70
vb.recession_sensitivity_decile	$\{0, \ldots, 10\}$	Household-level ranking within a geographic area of how sensitive the household is to the recession	3.68	3.74
vb.vehicle_value_decile	$\{0, \ldots, 10\}$	Decile ranking of household's propensity to invest highly in automobiles (1 = most likely, 10 = least likely)	2.81	2.81
voter_score_index	$\{0, \ldots, 4\}$	Roll-up grouping of voters based on prior election turnout	2.84	2.79
vb.voterbase_general_votes	$\{0, 1, \ldots\}$	Indicates the number of elections a voter has voted in within all available general elections	5.59	5.45
vb.voterbase_primary_votes	$\{0, 1, \ldots\}$	Indicates the number of elections a voter has voted in within all available primary elections	2.12	1.98
vb.voterbase_age	$\{18, 19, \ldots\}$	Indicates the individual's age, if available, and calculated from the VoterBase Date of Birth	46.64	46.76
vb.number_of_adults_in_household	$\{0, 1, \ldots\}$	Indicates the number of adults 18 years or older within the household	1.70	1.74
vb.number_of_children_in_household	$\{0, 1, \ldots\}$	Indicates the number of children age 0-17 living within the household	0.37	0.40
vb.household_income_amount	$[0, \infty)$	Indicates the estimated income amount of a household in $1000 increments	98.88	99.61
vb.mortgage_amount	$[0, \infty)$	Estimated mortgage loan amount in $1000 increments	22.85	23.80
vb.home_value_amount	$[0, \infty)$	Estimated home value in $1000 increments	55.29	56.62
vb.home_equity_amount	$[0, \infty)$	Estimated amount of equity held in home in $1000 increments	36.54	38.71
vb.length_of_residence_in_years	$\{0, 1, \ldots\}$	Indicates the length of residence at current address in years	5.37	5.35

Table 3: Pre-assignment covariates (continued). Descriptions are from TARGETSMART documentation.

			Mean by group	
Variable name	Variable support	Description	Z_i=1	Z_i=0
voted_g2008	$\{0, 1\}$	Voted in 2008 General election (1=Yes, 0=No)	0.66	0.66
voted_g2010	$\{0, 1\}$	Voted in 2010 General election (1=Yes, 0=No)	0.49	0.49
voted_g2012	$\{0, 1\}$	Voted in 2012 General election (1=Yes, 0=No)	0.75	0.73
voted_g2014	$\{0, 1\}$	Voted in 2014 General election (1=Yes, 0=No)	0.52	0.51
voted_g2016	$\{0, 1\}$	Voted in 2016 General election (1=Yes, 0=No)	0.85	0.85
voted_p2008	$\{0, 1\}$	Voted in 2008 Primary election (1=Yes, 0=No)	0.20	0.19
voted_p2010	$\{0, 1\}$	Voted in 2010 Primary election (1=Yes, 0=No)	0.20	0.18
voted_p2012	$\{0, 1\}$	Voted in 2012 Primary election (1=Yes, 0=No)	0.19	0.19
voted_p2014	$\{0, 1\}$	Voted in 2014 Primary election (1=Yes, 0=No)	0.19	0.19
voted_p2016	$\{0, 1\}$	Voted in 2016 Primary election (1=Yes, 0=No)	0.38	0.36
ts.tsmart_presidential_general_turnout_score	$[0, 100]$	Presidential General Turnout Score: These models were created to predict the likelihood that an individual will vote in each specific type of Election year. They are meant to replace the yearly turnout models we have created in the past and will now update automatically as we receive new demographic and vote history data from similar past elections.	65.32	63.89
ts.tsmart_presidential_primary_turnout_score	$[0, 100]$	Presidential Primary Turnout Score: These models were created to predict the likelihood that an individual will vote in each specific type of Election year. They are meant to replace the yearly turnout models we have created in the past and will now update automatically as we receive new demographic and vote history data from similar past elections.	38.48	37.04
ts.tsmart_midterm_general_turnout_score	$[0, 100]$	Midterm General Turnout Score: These models were created to predict the likelihood that an individual will vote in each specific type of Election year. They are meant to replace the yearly turnout models we have created in the past and will now update automatically as we receive new demographic and vote history data from similar past elections.	52.97	52.21
ts.tsmart_offyear_general_turnout_score	$[0, 100]$	Off-Year General Turnout Score: These models were created to predict the likelihood that an individual will vote in each specific type of Election year. They are meant to replace the yearly turnout models we have created in the past and will now update automatically as we receive new demographic and vote history data from similar past elections.	31.74	31.36
ts.tsmart_non_presidential_primary_turnout_score	$[0, 100]$	Non-Presidential Year Primary Turnout Score: These models were created to predict the likelihood that an individual will vote in each specific type of Election year. They are meant to replace the yearly turnout models we have created in the past and will now update automatically as we receive new demographic and vote history data from similar past elections.	39.55	38.38

Table 4: Pre-assignment covariates (continued). Descriptions are from TARGETSMART documentation.

			Mean by group	
Variable name	Variable support	Description	Z_i=1	Z_i=0
ts.tsmart_local_voter_score	[0, 100]	An ensemble method classifier model was created to predict the likelihood that an individual will vote in local, county, and municipal elections. The model was constructed using live telephone interviews conducted in the first quarter of 2016. A representative sample of the nation was collected by calling both landline and wireless numbers in every state. Interviewees were asked if they vote in local, county, and municipal elections where very few people vote. The model scores are expressed from 0-100, with the score representing the probability that person will vote in a local election. The model was used to score over 240 million voting age persons nationwide.	35.08	34.56
ts.tsmart_activist_score	[0, 100]	Predicts the likelihood that an individual is an activist	51.15	50.51
ts.tsmart_campaign_finance_score	[0, 100]	Model score (0-100) constructed from live and IVR survey responses, collected nationwide in early 2016, indicating support for the following proposal: "Limiting the amount of money that any individual or group can donate to a political campaign." Higher scores are more likely to favor this position.	59.38	59.70
ts.tsmart_catholic_raw_score	[0, 100]	For the Catholic score, voters are assigned a likelihood rating from 1 (very unlikely) to 4 (very likely). These ratings are based on models of survey-reported religious affiliation, which predict responses as a function of demographic characteristics, local patterns of religious affiliation and attendance, and other individual- and aggregate-level data. Predicted probabilities greater than 50% are coded as 4 ("very likely"), those between 25% and 50% as 3 ("somewhat likely"), those between 10% and 25% as 2 ("somewhat unlikely"), and those below 10% as 1 ("very unlikely"). The probabilities across all categories sum to 100%, so with these thresholds, a voter can have at most one "very likely" religion.	18.03	17.64
ts.tsmart_children_present_score	[0, 100]	The Presence of Children score represents the probability that an individual voter lives in a household with children under age 18. The score is based on models predicting responses from more than 49,000 live interviews with voters nationwide from January 2013 to present. The scores are on a scale from 0-100, with higher scores representing a higher probability that a person lives in a household with children.	28.81	29.33
ts.tsmart_climate_change_score	[0, 100]	Model score (0-100) constructed from live and IVR survey responses, collected nationwide in early 2016, indicating support for the following proposal: "Imposing tougher environmental regulations in order to combat climate change." Higher scores are more likely to favor this position.	76.53	75.82
ts.tsmart_college_funding_score	[0, 100]	Model score (0-100) constructed from live and IVR survey responses, collected nationwide in early 2016, indicating support for the following proposal: "Raising government spending on student aid to make college more affordable." Higher scores are more likely to favor this position.	66.97	67.56
ts.tsmart_college_graduate_score	[0, 100]	Model score (0-100) constructed from live and IVR survey responses, collected nationwide in early 2016, indicating support for the following proposal: "Raising government spending on student aid to make college more affordable." Higher scores are more likely to favor this position.	60.93	59.68

Table 5: Pre-assignment covariates (continued). Descriptions are from TARGETSMART documentation.

			Mean by group	
Variable name	Variable support	Description	Z_i=1	Z_i=0
ts.tsmart_evangelical_raw_score	[0, 100]	For the Evangelical score, voters are assigned a likelihood rating from 1 (very unlikely) to 4 (very likely). These ratings are based on models of survey-reported religious affiliation, which predict responses as a function of demographic characteristics, local patterns of religious affiliation and attendance, and other individual- and aggregate-level data. Predicted probabilities greater than 50% are coded as 4 ("very likely"), those between 25% and 50% as 3 ("somewhat likely"), those between 10% and 25% as 2 ("somewhat unlikely"), and those below 10% as 1 ("very unlikely"). The probabilities across all categories sum to 100%, so with these thresholds, a voter can have at most one "very likely" religion.	18.24	19.31
ts.tsmart_govt_privacy_score	[0, 100]	Model score (0-100) constructed from live and IVR survey responses, collected nationwide in early 2016, indicating opposition to the following proposal: "Allowing increased surveillance of US citizens phones and emails in order to prevent terror attacks." Higher scores are more likely to oppose this position.	52.51	52.81
ts.tsmart_gun_control_score	[0, 100]	Model score (0-100) constructed from live and IVR survey responses, collected nationwide in early 2016, indicating support for the following proposal: "Limiting access to guns through legislation which tightens background checks and restricts the purchase of military-style weapons." Higher scores are more likely to favor this position.	74.63	74.01
ts.tsmart_gunowner_score	[0, 100]	Predicts the likelihood that an individual is a gun owner	39.85	40.74
ts.tsmart_high_school_only_score	[0, 100]	The High School Education Only score represents the probability that an individual voter has a high school education or less. The score is based on models predicting responses from more than 49,000 live interviews with voters nationwide from January 2013 to present. The scores are on a scale from 0-100, with higher scores representing a higher probability that a person has only a high school education.	32.10	33.58
ts.tsmart_ideology_score	[0, 100]	An ensemble method classifier model was created to predict the likelihood that an individual supports liberal ideology. The model was constructed using live telephone interviews conducted in the first quarter of 2016. A representative sample of the nation was collected by calling both landline and wireless numbers in every state. Interviewees were asked if they are liberal, moderate, or conservative. The model scores are expressed from 0-100, with the score representing the probability that person supports liberal ideology. The model was used to score over 190 million voters nationwide.	67.26	66.86
ts.tsmart_income_rank_score	[0, 100]	The Income Rank score represents the probability that an individual voter has a high income. The scores are presented on a scale from 0-100, with higher scores representing a greater probability that a person has a high income.	52.34	52.89
ts.tsmart_marriage_score	[0, 100]	The Marital Status score represents the probability that an individual voter is married. The score is based on models predicting responses from more than 49,000 live interviews with voters nationwide from January 2013 to present. The scores are on a scale from 0-100, with higher scores representing a higher probability that a person is married.	41.49	42.40

Table 6: Pre-assignment covariates (continued). Descriptions are from TARGETSMART documentation.

			Mean by group	
Variable name	Variable support	Description	Z_i=1	Z_i=0
ts.tsmart_midterm_general_enthusiasm_score	[0, 100]	Description left blank in TARGETSMART documentation	0.06	0.05
ts.tsmart_minimum_wage_score	[0, 100]	Model score (0-100) constructed from live and IVR survey responses, collected nationwide in early 2016, indicating support for the following proposal: "Raising the minimum wage paid by all employers nationwide." Higher scores are more likely to favor this position.	63.37	63.77
ts.tsmart_moral_authority_score	[0, 100]	An ensemble method classifier model was created to predict the likelihood (0-100, ascending) that an individual aligns with the Authority pillar, defined as follows: "shaped by our long primate history of hierarchical social interactions. It underlies virtues of leadership and followership, including deference to legitimate authority and respect for traditions." The models were constructed using results from a telephone survey conducted between February 24 - March 2, 2016.	34.24	34.93
ts.tsmart_moral_care_score	[0, 100]	An ensemble method classifier model was created to predict the likelihood (0-100, ascending) that an individual aligns with the Care pillar, defined as follows: "an ability to feel (and dislike) the pain of others. It underlies virtues of kindness, gentleness, and nurturance." The models were constructed using results from a telephone survey conducted between February 24 - March 2, 2016.	49.34	49.47
ts.tsmart_moral_equality_score	[0, 100]	An ensemble method classifier model was created to predict the likelihood (0-100, ascending) that an individual aligns with the Equality pillar, defined as follows: "feelings of reactance and resentment people feel toward those who dominate them and restrict their liberty. Its intuitions are often in tension with those of the authority foundation." The models were constructed using results from a telephone survey conducted between February 24 - March 2, 2016.	52.74	52.64
ts.tsmart_moral_equity_score	[0, 100]	An ensemble method classifier model was created to predict the likelihood (0-100, ascending) that an individual aligns with the Equity pillar, defined as follows: "related to the evolutionary process of reciprocal altruism. It generates ideas of justice, rights, and autonomy." The models were constructed using results from a telephone survey conducted between February 24 - March 2, 2016.	43.50	44.12
ts.tsmart_moral_loyalty_score	[0, 100]	An ensemble method classifier model was created to predict the likelihood (0-100, ascending) that an individual aligns with the Loyalty pillar, defined as follows: "related to our long history as tribal creatures able to form shifting coalitions. It underlies virtues of patriotism and self-sacrifice for the group. It is active anytime people feel that it's 'one for all, and all for one.'" The models were constructed using results from a telephone survey conducted between February 24 - March 2, 2016.	47.32	47.61

Table 7: Pre-assignment covariates (continued). Descriptions are from TARGETSMART documentation.

			Mean by group	
Variable name	Variable support	Description	Z_i=1	Z_i=0
ts.tsmart_moral_purity_score	$[0, 100]$	An ensemble method classifier model was created to predict the likelihood (0-100, ascending) that an individual aligns with the Purity pillar, defined as follows: "shaped by the psychology of disgust and contamination. It underlies religious notions of striving to live in an elevated, less carnal, more noble way. It underlies the widespread idea that the body is a temple which can be desecrated by immoral activities and contaminants (an idea not unique to religious traditions)." The models were constructed using results from a telephone survey conducted between February 24 - March 2, 2016.	37.80	38.01
ts.tsmart_nonchristian_raw_score	$[0, 100]$	For the Non-Christian score, voters are assigned a likelihood rating from 1 (very unlikely) to 4 (very likely). These ratings are based on models of survey-reported religious affiliation, which predict responses as a function of demographic characteristics, local patterns of religious affiliation and attendance, and other individual- and aggregate-level data. Predicted probabilities greater than 50% are coded as 4 ("very likely"), those between 25% and 50% as 3 ("somewhat likely"), those between 10% and 25% as 2 ("somewhat unlikely"), and those below 10% as 1 ("very unlikely"). The probabilities across all categories sum to 100%, so with these thresholds, a voter can have at most one "very likely" religion.	26.37	26.38
ts.tsmart_otherchristian_raw_score	$[0, 100]$	For the Christian (other) score, voters are assigned a likelihood rating from 1 (very unlikely) to 4 (very likely). These ratings are based on models of survey-reported religious affiliation, which predict responses as a function of demographic characteristics, local patterns of religious affiliation and attendance, and other individual- and aggregate-level data. Predicted probabilities greater than 50% are coded as 4 ("very likely"), those between 25% and 50% as 3 ("somewhat likely"), those between 10% and 25% as 2 ("somewhat unlikely"), and those below 10% as 1 ("very unlikely"). The probabilities across all categories sum to 100%, so with these thresholds, a voter can have at most one "very likely" religion.	24.61	24.96
ts.tsmart_paid_leave_score	$[0, 100]$	Model score (0-100) constructed from live and IVR survey responses, collected nationwide in early 2016, indicating support for the following proposal: "Requiring all employers, including small businesses, to provide paid time off when their workers are ill, have a new child, or need to care for sick family members." Higher scores are more likely to favor this position.	63.95	64.49
ts.tsmart_partisan_score	$[0, 100]$	An ensemble method classifier model was created to predict the likelihood that an individual supports the Democratic Party. The model was constructed using live telephone interviews conducted in the first quarter of 2016. A representative sample of the nation was collected by calling both landline and wireless numbers in every state. Interviewees were asked a generic two-way congressional vote choice question and a generic two-way presidential vote choice question. The model scores are expressed from 0-100, with the score representing the probability that person supports the Democratic Party. The model was used to score over 240 million voting age persons nationwide.	79.84	79.37

Table 8: Pre-assignment covariates (continued). Descriptions are from TARGETSMART documentation.

			Mean by group	
Variable name	Variable support	Description	Z_i=1	Z_i=0
ts.tsmart_path_to_citizen_score	$[0, 100]$	Model score (0-100) constructed from live and IVR survey responses, collected nationwide in early 2016, indicating support for the following proposal: "Creating a path to citizenship for undocumented immigrants." Higher scores are more likely to favor this position.	72.81	72.25
ts.tsmart_prochoice_score	$[0, 100]$	Model score (0-100) constructed from live and IVR survey responses, collected nationwide in early 2016, indicating opposition to the following proposal: "Placing more restrictions on abortion, such as waiting periods and mandatory ultrasounds." Higher scores are more likely to oppose this position.	77.14	76.45
ts.tsmart_tax_on_wealthy_score	$[0, 100]$	Model score (0-100) constructed from live and IVR survey responses, collected nationwide in early 2016, indicating support for the following proposal: "Increasing taxes on the wealthy in order to improve infrastructure and government services and reduce the budget deficit." Higher scores are more likely to favor this position.	67.14	67.39
ts.tsmart_teaparty_score	$[0, 100]$	An ensemble method classifier model was created to predict the likelihood that an individual supports the Tea Party movement. The model was constructed using live telephone interviews conducted in the first quarter of 2016. A representative sample of the nation was collected by calling both landline and wireless numbers in every state. Interviewees were asked if they agree, disagree, or have no opinion concerning the Tea Party movement. The model scores are expressed from 0-100, with the score representing the probability that person supports the Tea Party movement. The model was used to score over 240 million voting age persons nationwide.	30.81	30.74
ts.tsmart_trump_resistance_score	$[0, 100]$	Predicts the likelihood that an individual is part of the Trump resistance	55.50	55.08
ts.tsmart_trump_support_score	$[0, 100]$	Predicts the likelihood that an individual supports Donald Trump	26.52	27.81
ts.tsmart_veteran_score	$[0, 100]$	Predicts the likelihood that an individual is a Military Veteran or Active Military Service Member	27.99	28.10
ts.tsmart_working_class_score	$[0, 100]$	Description left blank in TARGETSMART documentation	44.87	45.82
predictwise.authoritarianism_score	$[0, 100]$	The Predictwise authoritarianism score measures the probability that an individual prefers central authority. The model was created using questions about free speech, obedience, and respect for authority. The score ranges from 0-100, with 0 being not at all authoritarian, and 100 being extremely authoritarian.	39.05	39.07
predictwise.compassion_score	$[0, 100]$	The Predictwise compassion score aims to define an individual's compassion for those who are less fortunate than them. The model was built off questions about disability, homelessness, and reformed criminals. The score ranges from 0-100, with 0 being not at all compassionate, and 100 being extremely compassionate.	70.34	70.13

Table 9: Pre-assignment covariates (continued). Descriptions are from TARGETSMART documentation.

			Mean by group	
Variable name	Variable support	Description	Z_i=1	Z_i=0
predictwise.economic_populism_score	[0, 100]	The Predictwise economic populism score measures the likelihood of an individual being an economic populist, meaning that they believe that the government should provide assistance to the struggling working class. The model was build off questions about unions, big business, and social safety nets. The score ranges from 0-100, with 0 being not at all populist, and 100 being extremely populist.	65.67	65.53
predictwise.free_trade_score	[0, 100]	The Predictwise free trade score measures the probability that an individual supports free trade. The model was build using questions about the trade-offs between more/cheaper goods and quality control, stress on the border, and jobs moving from established industries to newer industry. The score ranges from 0-100, with 0 being entirely unsupportive of free trade, and 100 being fully supportive of free trade.	58.00	58.08
predictwise.globalism_score	[0, 100]	The Predictwise Globalism score measures the probability that an individual has globalist beliefs, meaning they believe that the US should be open as a society. The model was built off questions about trade, automation, and fears about a global economy. The score ranges from 0-100, with 0 meaning that an individual is not at all globalist, and 100 meaning that an individual is extremely globalist.	44.81	44.67
predictwise.guns_score	[0, 100]	The Predictwise guns score measures the probability of an individual supporting the freedom to own and/or buy guns. The model was build using questions about individuals' support for citizens owning assault weapons, carrying concealed weapons, and for the government to register gun owners. The score ranges from 0-100, with 0 meaning that an individual is extremely nonsupporting of the freedom to own and buy guns, and 100 meaning that an individual fully supports the freedom to buy and own guns.	45.02	45.07
predictwise.healthcare_score	[0, 100]	The Predictwise score measures the likelihood that an individual supports government provided healthcare. The model was built off individuals' attitudes toward government subsidized healthcare for the poor, people with pre-existing conditions, the elderly, and mothers and newborns. The score ranges from 0-100, with 0 meaning an individual does not support government provided healthcare and 100 meaning an induvial is supportive of government provided healthcare.	55.81	55.97
predictwise.healthcare_women_score	[0, 100]	The Predictwise women's healthcare score measures the probability of an individual supporting women's healthcare and reproductive rights. The model was built off individuals' attitudes towards womens' and girls' healthcare, access to birth control without parental consent, comprehensive sexual education, and access to abortion in different circumstances. The score ranges from 0-100, with 0 meaning an individual does not support women's healthcare and reproductive rights, and 100 meaning an individual is fully supportive of women's healthcare and reproductive rights.	78.38	78.27

Table 10: Pre-assignment covariates (continued). Descriptions from TARGETSMART documentation.

			Mean by group	
Variable name	Variable support	Description	Z_i=1	Z_i=0
predictwise.immigrants_score	$[0, 100]$	The Predictwise immigrant score measures the probability that someone is supportive of immigrants. The model was built off questions that gauged individuals' opinions on legal versus illegal immigration, refugees, the trade-offs with national security and job security and open immigration. The score ranges from 0-100, with 0 meaning an individual does not support immigration, and 100 meaning an individual is fully supportive of immigration.	57.17	57.05
predictwise.military_score	$[0, 100]$	The Predictwise military score measures the probability that someone supports an expansive military role. The model was built based off questions about securing US territories, protecting trade routes, spreading democracy, and countering terrorism. The model ranges from 0-100, with 0 meaning an individual does not support an expansive military role, and 100 meaning an individual is fully supportive of an expansive military role.	43.02	43.33
predictwise.poor_score	$[0, 100]$	The Predictwise safety net for poor score measures the probability that an individual is supportive of a safety net for the poor. The model was built based on attitudes about education, healthcare, shelter, food, and employment for the poor. The model ranges from 0-100 with 0 meaning an individual is not supportive of a safety net for the poor, and 100 meaning an individual is fully supportive of a safety net for the poor.	58.39	58.43
predictwise.populism_score	$[0, 100]$	The Predictwise anti-elitist populism score measures the probability that an individual is an anti-elitist populist, meaning that they believe there is a struggle between "the people" and "the elite". The model was built using questions about trust in elites, the distribution of power, and thoughts on the "system". The score ranges from 0-100, with 0 meaning that an individual is not at all populist, and 100 meaning an individual is extremely populist.	51.15	51.28
predictwise.presidential_score	$[0, 100]$	The Predictwise presidential score measures the probability that an individual views Trump as presidential. The model was built based off questions explicitly about President Trump, including his appropriateness, honestly, morality, competence, and work ethic. The score ranges from 0-100, with 0 meaning an individual does not view Trump as presidential at all, and 100 meaning an individual views trump as extremely presidential.	41.66	41.97
predictwise.racial_resentment_score	$[0, 100]$	The Predictwise racial resentment score measures the probability that an individual is racially resentful. The model was built off questions pertaining to black work ethic, historical discrimination, crime, and black protest. The score ranges from 0-100, with 0 meaning an individual is not at all racially resentful, and 100 meaning an individual is extremely racially resentful.	62.80	62.61
predictwise.regulation_score	$[0, 100]$	The Predictwise regulation model measures the probability that an individual is supportive of government regulation. The model was built off opinions on environmental, workplace, financial, food, and drug safety regulations. The score ranges from 0-100, with 0 meaning an individual is not supportive of government regulation, and 100 meaning an individual is fully supportive of government regulation.	52.96	53.25

Table 11: Pre-assignment covariates (continued). Descriptions from TARGETSMART documentation.

			Mean by group	
Variable name	Variable support	Description	Z_i=1	Z_i=0
predictwise.religious_freedom_score	$[0, 100]$	The Predictwise religious freedom score measures the probability that an individual supports religious freedom to deny equal rights. This model was based on attitudes about conflict between discrimination and religious freedom—for example, adoption rights for LGBT couples, and denying procedures at religious hospitals. The score ranges from 0-100, with 0 meaning and individual is not supportive of religious freedom denying equal rights, and 100 meaning an individual is extremely supportive of religious freedom denying equal rights.	38.97	39.06
predictwise.taxes_score	$[0, 100]$	The Predictwise taxes score measures the probability that an individual supports tax raises. The model was based off attitudes regarding taxation on the wealthy and middle class, inheritance taxes, corporate taxes and capital gains taxes. The score ranges from 0-100, with 0 meaning an individual does not support tax raises, and 100 meaning an individual fully supports tax raises.	71.94	71.72
predictwise.traditionalism_score	$[0, 100]$	The Predictwise traditionalism score measures the probability that in individual is a traditionalist. The model was built based off questions about corporal punishment in school and at home, generational divides, morality, and religious depth. The score ranges from 0-100, with 0 meaning an individual is not at all traditionalist, and 100 meaning an individual is extremely traditionalist.	44.57	44.75
predictwise.trust_in_institutions_score	$[0, 100]$	The Predictwise trust in institutions score measures the probability that an individual trusts institutions. The model was built off questions about media, political government, and the intelligence community. The model ranges from 0-100, with 0 meaning an individual does not trust institutions at all, and 100 meaning an individual is extremely trusting in institutions.	52.38	52.36

References

1. R. K. Crump, V. J. Hotz, G. W. Imbens, O. A. Mitnik, Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand, *Tech. rep.*, National Bureau of Economic Research (2006).

2. L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference* (Springer Science & Business Media, 2013).

3. C. Validated, How can I compute the standard error of the Wald estimator?, https://stats.stackexchange.com/questions/219367/how-can-i-compute-the-standard-error-of-the-wald-estimator (2017).