Assessing the Effects of Friend-to-Friend Texting on Turnout in the 2018 US Midterm Elections

Aaron Schein
Columbia University
New York, NY, USA

Keyon Vafa
Columbia University
New York, NY, USA

Dhanya Sridhar
Columbia University
New York, NY, USA

Victor Veitch
Columbia University
New York, NY, USA

Jeffrey Quinn
PredictWise
New York, NY, USA

James Moffet
JDM Design
Cambridge, MA, USA

David M. Blei
Columbia University
New York, USA

Donald P. Green
Columbia University
New York, USA

ABSTRACT

Recent mobile app technology lets people systematize the process of messaging 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. However, properly assessing this causal effect is hindered by multiple statistical challenges, including attenuation bias due to mismeasurement of subjects' outcomes and low precision due to two-sided non-compliance with subjects' assignments. We address these challenges, which are likely to impinge upon any study that seeks to randomize authentic friend-to-friend interactions, by tailoring the statistical analysis to make use of additional data about both users and subjects. Using meta-data of users' in-app behavior, we reconstruct subjects' positions in users' queues. We use this information to refine the study population to more compliant subjects who were higher in the queues, and we do so in a systematic way which optimizes a proxy for the study's power. To mitigate attenuation bias, we then use ancillary data of subjects' matches to the voter rolls that lets us refine the study population to one with low rates of outcome mismeasurement. Our analysis reveals statistically significant treatment effects from friend-to-friend mobilization efforts ($\widehat{\text{CACE}} = 8.3$, CI $= (1.2, 15.3)$) that are among the largest reported in the get-out-the-vote (GOTV) literature. While social pressure from friends has long been conjectured to play a role in effective GOTV treatments, the present study is among the first to assess these effects experimentally.

CCS CONCEPTS

• **Applied computing** → **Voting / election technologies**; *Sociology*; • **Human-centered computing** → **Empirical studies in collaborative and social computing**.

WWW '21, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3449800

KEYWORDS

get-out-the-vote, digital field experiments, instrumental variables, non-compliance, outcome mismeasurement

ACM Reference Format:
Aaron Schein, Keyon Vafa, Dhanya Sridhar, Victor Veitch, Jeffrey Quinn, James Moffet, David M. Blei, and Donald P. Green. 2021. Assessing the Effects of Friend-to-Friend Texting on Turnout in the 2018 US Midterm Elections. In *Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3442381.3449800

1 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 scores of campaign workers and volunteers make phone calls, and go door-to-door, encouraging supporters to vote, particularly those in battleground states. Two decades of field experiments have shown that traditional GOTV tactics like phone-banking and door-to-door canvassing substantially increase voter turnout [12].

More recently, campaigns have begun considering scalable alternatives, such as mass texting or emailing, to cut costs and reach more potential voters. Demand for scalable alternatives has further soared in recent months, as the COVID-19 pandemic renders door-to-door canvassing or in-person volunteer training sessions infeasible, since face-to-face contact, especially between strangers, is both limited and unsafe. Unfortunately, meta-analysis shows that many scalable GOTV tactics have only modest effects on voter turnout, typically less than one percentage point [12]. This research suggests that scalable, but impersonal, communication does not substitute for traditional tactics like door-to-door canvassing.

The evolution of mobile communication, however, has created new opportunities for personal yet scalable GOTV tactics 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 message their close contacts to encourage them to vote [28]. The premise of friend-to-friend organizing is that GOTV appeals are especially effective, even via texts or emails, because friends are often welcome and trusted messengers [13, 21].

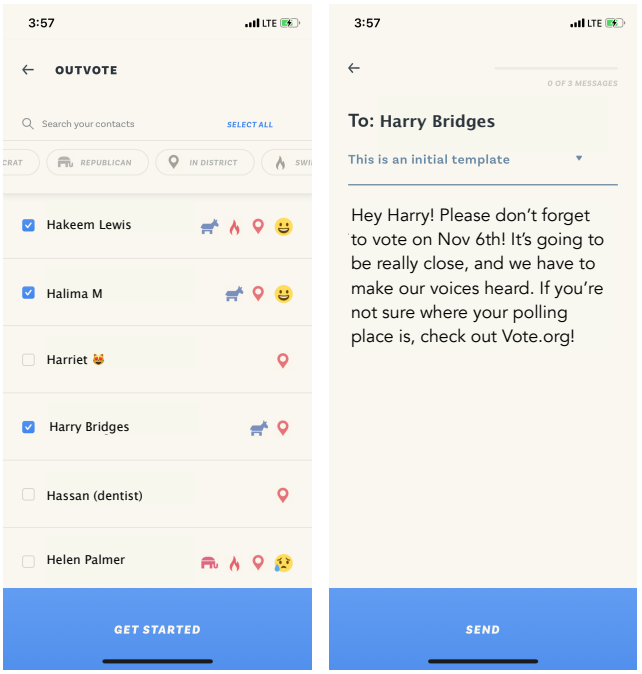
To date, there have been few evaluations of friend-to-friend outreach, and these have either relied on non-experimental designs [7] or experiments involving a relatively small number of participants [11, 14, 27]. This gap in the experimental literature reflects the difficulty of implementing a workable design at scale. 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 produces 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 [12] to authentic friend-to-friend appeals.

This paper uses a large-scale digital field experiment from the 2018 US midterm elections to assess 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 [1], 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 unobtrusive 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, and not take the user to the messaging stage for that person. 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.

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 three interrelated respects. First, using the random skips to define a valid treatment–control split of subjects is challenging. 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. To address this, we define treatment assignment only in terms of a subject's first queue.

Second, due to the way we define the treatment–control split to avoid confounding, a large share of subjects do not comply with their treatment assignment: many subjects assigned to receive treatment do not receive any messages, while others, assigned to not receive treatment, do. Non-compliance generally limits the scope of estimable causal effects, necessitating the adoption of an instrumental variable (IV) framework [2]. Furthermore, the high rate of non-compliance in the present study increases the statistical uncertainty of our IV estimator. To reduce this uncertainty, we reconstruct subjects' positions in users' queues from meta-data of



(a) Queuing stage. Users queue contacts to message. Contacts are sorted alphabetically within tiers.

(b) Messaging stage. For each contact on the queue, users can send a 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 contacts in the queue.

users' in-app behavior. We then use subjects' queue positions as pre-assignment variables to refine the study population to more compliant subjects who were higher in users' queues, and we do so in a principled way that formalizes and optimizes the tradeoff between a higher compliance rate and a larger n.

Third, due to Outvote's limited information about subjects, errors occur when matching them to the database of public information about citizens' voter history, which is necessary to observe their voting outcomes. Unreliable information about voting outcomes biases our estimates of the causal effect towards zero. To mitigate such bias, we obtain ancillary match information on subjects that helps us determine whether a subject's voting outcome was measured reliably, and lets us refine the study population to only such subjects.

Addressing the three statistical challenges presented by Outvote's study, our analysis reveals significant treatment effects from friend-to-friend mobilization efforts ($\widehat{\text{CACE}} = 8.3$, $\text{CI} = (1.2, 15.3)$) that are among the largest reported in the GOTV literature. The statistical challenges we overcome to assess these effects are likely to impinge upon any study that seeks to randomize authentic friend-to-friend interaction, since high compliance and data quality typically come at the cost of obtrusive study designs that discourage users and disrupt natural communication. Rather than forego the advantages of an unobtrusive design, we tailor our statistical analysis to meet the special challenges of evaluating volunteer-driven friend-to-friend outreach efforts.

2 STUDY DETAILS

Setting. The present study was conducted on the mobile app Outvote. Users began by syncing their phone's contacts with the app, which would then match contacts to the voter rolls based on mobile numbers and any available name information in the user's phone. Each user would then be presented with a ranked list of their contacts, each annotated with information about the contact's participation in past elections, party registration, and residence in a battleground state (see Figure 1a). Next, the user could create a queue of friends to whom they intended to send GOTV messages. Once the user finalized this queue, the app would take them to a messaging interface for each contact in the queue, in the order in which the user queued them. In the messaging interface, users could either send a default message or create their own. After the user pressed "Send" (see Figure 1b), they would be 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 would randomly skip contacts in the queue 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 about 132,000 GOTV messages.

Eligible subject pool. During the study period, anyone who an Outvote user added to a queue of length five or greater was subject to randomization. Subjects must also have been matched to the TargetSmart voter rolls database for their voting outcomes to be observed. We further focus only on subjects who were registered to vote prior to the study period. There are 195,118 eligible subjects who meet these criteria.

Messages and treatment. Throughout this paper, we will define receiving the treatment ($D_i = 1$) to mean receiving at least one GOTV text message from an Outvote user during the study period. Users could participate in one or more campaigns, each of which provided their own default messages. Although users were able to craft their own messages from scratch, we estimate from simple text analysis that 98% of all sent messages were the default message or a minor variant (e.g., with an emoticon inserted, or the recipient's name edited). 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 Table 1, we provide the text and campaign of the ten most-used default messages, which accounted for 55% of all sent messages.

IRB approval. This field experiment was conducted by Outvote, which sought to rigorously establish the efficacy of its app. Columbia University's Institutional Review Board approved our access to and storage of the data for research purposes, protocol AAAS8255, provided that no data containing PII are shared publicly.

3 ASSESSING CAUSAL EFFECTS FROM A "LIGHT TOUCH" EXPERIMENT

What were the causal effects of users' messages on getting subjects to turn out to vote in the US 2018 midterm elections? To answer this, we must first address three statistical challenges that arise from idiosyncratic aspects of our data and hamper precise and unbiased estimation of causal effects. These challenges stem directly from Outvote's randomization scheme, which was designed to have a "light touch" as not to interrupt users' natural behavior, but yielded imperfect experimental data as a result. The first challenge is in defining an unconfounded treatment-control split of subjects (Section 3.1) such that the standard causal estimand (defined in Section 3.2) is identified. The second challenge is in mitigating attenuation bias driven by mismeasurement of many subjects' voting outcomes (Section 3.3). The third challenge is in mitigating error driven by many subjects' non-compliance with their treatment assignments (Section 3.4). We overcome these challenges by leveraging additional data and meta-data available to us on users and subjects, which let us refine the study population to one with higher compliance and lower measurement error, and let us do so in a principled and replicable way.

3.1 Defining random treatment assignments

Outvote was programmed to skip contacts in a user's queue, each with probability 0.05, and thus randomly diminish the chances that some subjects received GOTV messages. However, these random skipping events do not necessarily define a valid treatment-control split of subjects. The reason is that users sometimes created multiple queues in succession, adding some of the same contacts to each. In addition, some subjects were queued by multiple users. Overall, some subjects were added to multiple queues and thus subjected to multiple random skipping events.

It is possible and perhaps tempting to define the assigned control group to be subjects who were randomly skipped in every queue they were added to. However, this definition yields a potentially confounded split due to the following hypothetical. Some users may have noticed that a contact in their queue had been skipped, and could have then created any number of new queues repeatedly until the contact was not randomly skipped. The result is that, under this definition of assignment, some subjects may have had no chance of being randomly assigned to the control group, since users were determined to message them despite the app's discouragement. The data support this hypothetical—in Section E, we show that this definition yields an imbalanced treatment-control split.

To sidestep this issue, we consider only the first time a subject is queued by any user, thereby defining the assigned control group ($Z_i = 0$) to be those subjects randomly skipped in their first queue, and defining the assigned treatment group ($Z_i = 1$) to be those not skipped in their first queue. We determine each subject's first queue by examining the timestamps of users' queuing actions. Considering only the first queue ensures a random treatment assignment and prevents possible confounding from subjects being queued multiple times for reasons that may be related to their likelihood of voting. For each subject i, the first queue is considered the moment of entry into the study population, and whether they are skipped in it determines their treatment assignment.

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%

Two-sided non-compliance. This definition of assigned treatment and control groups immediately raises the prospect of non-compliance. 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 was driven by the users, who frequently created long queues (the median length is 22) but stopped before messaging everyone on them. Only 29% of subjects in the assigned treatment group received a message. Some users also re-queued and messaged subjects who had been skipped in their first queue. Of subjects in the assigned control group, 13% received a message. The high rate of non-compliance in this study is an inevitable consequence of how we must define assignments to avoid confounding, and the intentionally unobtrusive nature of Outvote's randomization scheme. We address the challenge to precise estimation presented by non-compliance in Section 3.4.

3.2 Complier average causal effect (CACE)

In this subsection, we introduce our causal estimand, which is routinely targeted in randomized experiments with non-compliance.

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 $\{Z_i, D_i, Y_i\}_{i=1}^{n}$.

We want to estimate causal effects and therefore consider potential (not just recorded) outcomes. 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 then the difference $Y_{i1} - Y_{i0}$.

An ideal randomized experiments with perfect compliance would allow us to estimate the average causal effect, $\mathbb{E}[Y_{i1} - Y_{i0}]$, using the difference of sample means, $\widehat{\text{ACE}} = \mathbb{E}[Y_i | D_i = 1] - \mathbb{E}[Y_i | D_i = 0]$, as an unbiased estimator. However, the presence of non-compliance limits the scope of identifiable average effects, since the recorded receipt D_i may be confounded if users are non-random in choosing which subjects to message.

Instead, experiments with non-compliance are routinely analyzed using an instrumental variables framework, treating the random assignment Z_i as an instrument for treatment receipt D_i [2]. In this framework, we further consider potential (not just recorded) receipts. Let D_{i1} be subject i's potential receipt if assigned to receive treatment and D_{i0} be i's potential receipt if assigned not to. Each subject has two binary potential receipts, implying four basic types of subjects. Some subjects are "compliers" who receive a message if and only if assigned to; the set of all compliers is $C = \{i : D_{i1} = 1 \text{ and } D_{i0} = 0\}$. 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 $D_{i1} = 0$).

The four types of subject have specific interpretations in the present study. Compliers are those who would receive a message if the app does not skip them the first time they are queued, but not otherwise. By contrast, always-takers are people who would always be messaged, even if they are skipped at first. An always-taker might be someone prominently positioned near the top of the ranked list from which users select contacts to queue (see Figure 1a). The user might notice if such a contact were skipped, and then re-queue and message them. Furthermore, if not skipped, such a contact would be high in the queue and among the first to receive a message before the user quits. A never-taker instead may be someone at the bottom of the queue, among those the user fails to message before quitting, who are therefore not messaged even when they are not slated to be skipped by the app. Finally, a defier would be someone that is messaged only if they are skipped in their first queue; but, it is hard to imagine that such subjects exist in large number.

While a study with non-compliance does not generally identify the ACE, it may identify the average causal effect among compliers,

$$\text{CACE} = \mathbb{E}[Y_{i1} - Y_{i0} \mid i \in C]. \tag{1}$$

The CACE is identified under certain assumptions [2] that are reasonable in the Outvote study, as we describe below.

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" [16]. The first condition is "relevance," which stipulates a non-zero association between treatment assignment Z_i and 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" [3], "exchangeability" [16], or "ignorability" which states that the potential outcomes are independent of assignment, i.e., $Y_{i1}, Y_{i0} \perp\!\!\!\perp Z_i$, or informally that the assigned treatment and control groups have the same expected potential outcomes.

A "weak" instrument that meets the three instrumental conditions but is only weakly related to receipt can introduce finite-sample bias. Staiger et al. [26] suggest a rule-of-thumb to diagnose an instrument as "weak" if a regression of D_i on Z_i produces an F-statistic less than 10. There should be no concern of such bias in the present study as this regression on the whole subject pool yields an F-statistic of 348 and similar 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) [22] 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 [25].

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 further augmented with pre-assignment covariates in order to improve the precision with which treatment effects are estimated. We report results in Section 4 using both the unadjusted estimate and the estimate adjusting for 85 pre-assignment covariates (e.g., age, party registration, prior voting history) associated with each subject.

3.3 Using ancillary data to reduce outcome mismeasurement and mitigate bias

As in most GOTV field studies, the recorded outcome Y_i of whether subject i voted is measured by first matching subjects to a voter roll database. Outvote matched subjects to a database using the phone contact information that users shared with the app. However, in keeping with their light-touch approach, Outvote 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 to match her to the voter rolls. Some proportion of subjects in Outvote's study are incorrectly matched and their recorded outcome may be mismeasured.

Mismeasurement of outcomes often introduces attenuation bias when the error is "non-differential" [17, 19, 20]. Let $Y_i^\star$ denote a subject's true outcome. Outcome mismeasurement is non-differential with respect to assignment if the recorded outcome is independent of assignment given the true outcome,

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

This assumption is typically satisfied when assignment Z_i is randomized, as it is in our case. One way this might still fail is if Outvote employed different matching strategies for subjects in the assigned treatment and control groups. However, Outvote matched subjects prior to random assignment, thus allaying this concern. We can safely assume that outcome mismeasurement introduced by matching errors is non-differential in the present study.

It follows, then, that outcome mismeasurement in the present study attenuates the estimated effect. The numerator of the IV estimator in Equation (2), which estimates the ITT, can be written

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

where the bias is the difference of two probabilities,

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

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 assumes that matching is not so inaccurate 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 mismeasurement; we refine the study population to these 30%.

3.4 Reconstructing subjects' queue positions 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 our estimator by systematically filtering subjects by a pre-assignment variable, namely, their positions in users' queues. Doing so improves compliance while still maintaining a large enough n, a tradeoff we formalize and optimize.

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.

Let $Q_i \in \{1, 2, 3, \dots\}$ be the position of subject i in their first queue. Crucially, 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. (We confirm this empirically in Appendix F using balance checks to show 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%, while 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

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

which trades off generalizibility for precision.

What is a principled way to select $q_{\max}$ given that different values imply different estimands? Following Crump et al. [8], who systematically trim their study population to minimize the asymptotic variance of their estimator, we select the $q_{\max}$ which minimizes the expected variance of ours. The variance is affected both by the compliance rate and by n. For instance, although the compliance rate is highest among the subpopulation defined by $q_{\max} = 1$, there are only $n = 2{,}996$ such subjects. Based on a power analysis, detailed in Appendix C, we find that $q_{\max} = 103$ optimizes this tradeoff; it yields a study population of $n = 27{,}464$ with a 25% compliance rate. Our power analysis is summarized in Figure 2, which plots compliance, n, standard error, and power as a function of $q_{\max}$, with power attaining a maximum (or, equivalently, standard error attaining a minimum) at $q_{\max} = 103$.

One wrinkle is that Outvote did not explicitly record subjects' queue position in its database. However, we developed a simple method, based on timestamp meta-data of when users queued subjects, which reconstructs Q_i confidently for 60% of subjects. We detail this method in Appendix D. As further described there, when this method fails to confidently reconstruct Q_i, it is often because the subject had been queued by a user pressing an "Add all" button that simultaneously queued all phone contacts. Unsurprisingly, subjects queued by an "Add all" exhibit very low compliance. Subjects whose Q_i cannot be confidently reconstructed exhibit an 8% compliance rate as compared to 22% for those whose Q_i can be reconstructed. Refining the study population to only those whose Q_i can be reconstructed, as we do, is itself a measure that improves compliance, as well as one that is necessary to further refine based on $q_{\max}$.

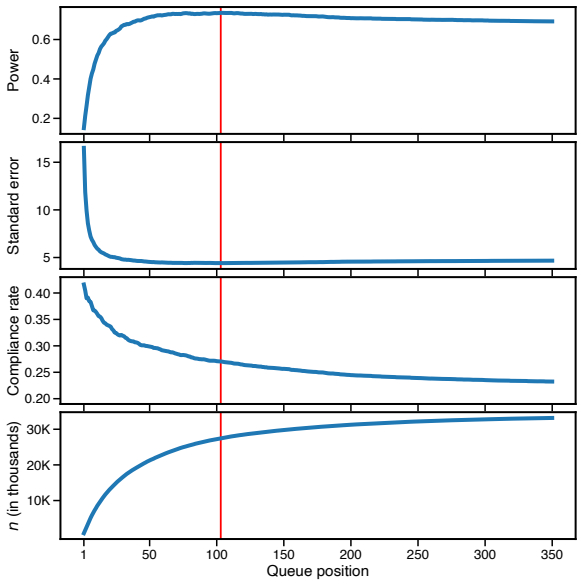


Figure 2: The tradeoff between n and compliance, as the maximum allowable position of subjects in their first queue increases. The study's power is influenced by both. It degrades after $q_{\max} = 103$ (red line), when increasing n fails to compensate for lower compliance.

4 RESULTS

Refined study population. Applying the steps detailed in the previous section to mitigate bias and maximize the precision of our estimator, we obtain the refined study population, which consists of the $n = 27{,}464$ eligible subjects who

(1) were matched by Outvote and Predictwise to the same entry in the voter rolls (see Section 3.3), and
(2) had a (reconstructed) position Q_i in their first queue that was within the top 103 (see Section 3.4).

Of these, 1,454 subjects (5.3%) are part of the assigned control group and the remaining 25,796 are part of the assigned treatment group. Table 2 summarizes the assignments, treatments received, and vote outcomes for the refined study population.

Unadjusted estimates. Using the IV estimator in eq. (2), we estimate the CACE among the refined study population to be:

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

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

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.

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

Using the formulas from Corollary 1 of Aronow and Green [4], 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.

Covariate-adjusted estimates. We can augment the IV estimator using pre-assignment covariates $\mathbf{X}_i$ for each subject. Such information is available from TargetSmart's voter database, to which subjects are matched for their voting outcomes to be measured. We use 85 pre-assignment covariates to adjust CACE estimates. The list includes age, number of previous general and primary election votes, household income, education level, and an assortment of other variables based on multilevel modeling of survey data [18].

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 note that while covariate-adjusted and unadjusted estimates are slightly different, their confidence intervals, which we visualize in Figure 3, closely accord.

Supplemental findings. For the reasons given in Section 3, we surmise that the causal effects among the refined study population are among the most precisely estimable with lowest bias and thus present them as our main findings. However, in Appendix F, we provide a complete breakdown of estimated effects among the segments of the subject pool that we excluded from the refined study population, and we discuss how these ancillary effects support our claims about bias and variance as functions of mismeasurement and non-compliance. 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) which is a large effect, consistent with low attenuation bias, estimated with large error, consistent with that segment's low compliance rate. 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 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 `statsmodels` [24] Python package while the corresponding ITT estimates are obtained using their ordinary least squares (`OLS`) implementation. We have released

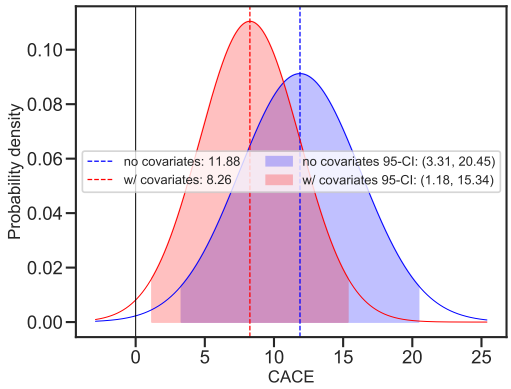


Figure 3: CACE estimates with and without covariates.

data for the entire subject pool, stripped of all personally identifying information (which is sufficient to replicate all ITT/CACE effects without covariates) along with source code for refining the study population and for ITT/CACE estimation.[1]

5 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 [15] estimates their average effect to be 0.44 percentage points, despite the fact that automated distribution systems often deliver 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 [12]. Second, the effects of GOTV appeals may be amplified when the messenger is known to

[1] https://github.com/aschein/outvote

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 [9]. Third, GOTV effects tend to be enhanced when senders exert some degree of social pressure [5], especially when the sender implies that they are counting on the recipient to vote and will be disappointed otherwise [10, 14, 27]. 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 tradeoffs 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 problems, 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 accuracy of 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.

ACKNOWLEDGMENTS

We thank PredictWise for providing voter file match data and thank Roy Adams, Peter Aronow, Josh Kalla, Cyrus Samii, and Otis Reid for helpful discussions.

A ETHICS STATEMENT

As mentioned in the main paper, Columbia University's Institutional Review Board approved our access to and storage of the data for research purposes, protocol AAAS8255, provided that no data containing PII are shared publicly. We take seriously the ethical concerns associated with field experimentation in general and experiments involving elections in particular. We briefly review here the ethical concerns that we considered when reflecting on this research project. The foremost concerns always focus on risk of harm to human subjects. We see no such risks here. We also take precautions to avoid risks associated with a breach of confidential information by not disclosing personally identifying information. Next, we consider threats to subjects' autonomy; here, the subjects are voters who are being contacted by users, and although subjects do not provide active consent, they are being contacted by an authentic friend-to-friend campaign in the same way they ordinarily would in the absence of a research study. Only subjects in the control group are potentially denied contact from a user, and even here the rate at which users follow through with their outreach efforts is low, and users often contacted subjects in the control group anyway. There seems to be no feasible way, prior to the launch of the study, to obtain informed consent from subjects who were not contacted specifically because of the randomization scheme without undermining the light touch character of the study. Finally, we consider harms to society including the possibility of affecting election outcomes. Given the wide geographic range in which the mobilization effort occurred in 2018, these harms seem very unlikely given the low rate at which subjects were assigned to control, assuming they would have been contacted at the same rate by users had they been assigned to treatment.

B PROOF OF ATTENUATION BIAS FROM NON-DIFFERENTIAL MISMEASUREMENT

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

$$\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0]=P(Y_i=1|Z_i=1)-P(Y_i=1|Z_i=0). \quad (7)$$

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

$$=\sum_{y^\star=0}^{1}\Big[P(Y_i=1,Y_i^\star=y^\star|Z_i=1)-P(Y_i=1,Y_i^\star=y^\star|Z_i=0)\Big]. \quad (8)$$

Due to random assignment of Z_i, the mismeasurement is non-differential in the sense that $Y_i \perp\!\!\!\perp Z_i \mid Y_i^\star$. The terms thus factorize $P(Y_i=1,Y_i^\star=y^\star|Z_i=z)=P(Y_i=1\mid Y_i^\star=y^\star)\,P(Y_i^\star=y^\star\mid Z_i=z)$ for both $z\in\{0,1\}$, and the overall expression can be rewritten as

$$=\Big[\underbrace{\mathbb{E}[Y_i|Y_i^\star=1]-\mathbb{E}[Y_i|Y_i^\star=0]}_{\text{bias }\pi_{\text{BIAS}}}\Big]\Big[\underbrace{\mathbb{E}[Y_i^\star|Z_i=1]-\mathbb{E}[Y_i^\star|Z_i=0]}_{\text{true ITT}}\Big]. \quad (9)$$

We assume the mismeasurement error is not extreme, such that

$$\mathbb{E}[Y_i|Y_i^\star=1]>\mathbb{E}[Y_i|Y_i^\star=0], \quad (10)$$

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

C SELECTING $q_{\max}$

As discussed in Section 3.4, setting a maximum allowable queue position $q_{\max}$ defines a subpopulation of subjects whose first queue position is $Q_i \le q_{\max}$. 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 $q_{\max}$,

$$\hat{\beta}_{\text{IV}}(q_{\max})=\frac{\bar{y}_{1q_{\max}}-\bar{y}_{0q_{\max}}}{\bar{d}_{1q_{\max}}-\bar{d}_{0q_{\max}}}, \quad (11)$$

where $\bar{y}_{1q_{\max}} = \hat{\mathbb{E}}[Y_i \mid Z_i = 1, Q_i \leq q_{\max}]$ 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.

By setting $q_{\max}$, we change the estimand to a conditional CACE such that $\hat{\beta}_{\text{IV}}(q_{\max})$ estimates $\mathbb{E}[Y_{i1} - Y_{i0} \mid i \in C \text{ and } Q_i \leq q_{\max}]$. Our approach is analogous to that of Crump et al. [8] 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,

$$\mathbb{V}\left(\hat{\beta}_{\text{IV}}(q_{\max})\right) = f\left(\beta(q_{\max}), \bar{y}_{0q_{\max}}, \bar{d}_{1q_{\max}}, \bar{d}_{0q_{\max}}, n_{q_{\max}}, p\right), \quad (12)$$

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 $\beta(q_{\max})$ and $\bar{y}_{0q_{\max}}$ and empirical values of $\bar{d}_{1q_{\max}}, \bar{d}_{0q_{\max}}$, and $n_{q_{\max}}$. 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 $n_{q_{\max}}$ from the study population, and estimate the average rates of treatment receipt $\bar{d}_{1q_{\max}}$ and $\bar{d}_{0q_{\max}}$ from the pool of subjects excluded from the analysis due to poor match quality[2]. We then select $q_{\max}$ to be,

$$q^*_{\max} \leftarrow \underset{q_{\max}}{\operatorname{argmin}}\ f\left(\beta^\circ, \bar{y}_0^\circ, \bar{d}_{1q_{\max}}, \bar{d}_{0q_{\max}}, n_{q_{\max}}, p\right), \quad (13)$$

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

The form of the expected variance of the IV estimator $f(\ldots)$ can be obtained using the Delta method [6, 29] as

$$\mathbb{V}\left(\hat{\beta}_{\text{IV}}\right) = \frac{\sigma_U^2}{\mu_V^2} - 2\frac{\mu_U}{\mu_V^3}\sigma_{U,V} + \frac{\mu_U^2}{\mu_V^4}\sigma_V^2, \quad (14)$$

where the following quantities are defined,

$$\mu_U = \bar{y}_1 - \bar{y}_0 \quad (15)$$

$$\mu_V = \bar{d}_1 - \bar{d}_0 \quad (16)$$

$$\sigma_U^2 = \tfrac{1}{n_1}\bar{y}_1(1-\bar{y}_1) + \tfrac{1}{n_0}\bar{y}_0(1-\bar{y}_0) \quad (17)$$

$$\sigma_V^2 = \tfrac{1}{n_1}\bar{d}_1(1-\bar{d}_1) + \tfrac{1}{n_0}\bar{d}_0(1-\bar{d}_0) \quad (18)$$

$$\sigma_{U,V} = \beta\,\sigma_V^2. \quad (19)$$

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, 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 $n_1 = p\,n_{q_{\max}}$ and $n_0 = (1-p)\,n_{q_{\max}}$ based on $n_{q_{\max}}$ calculated from the study population. We also plug in $\bar{d}_1 = \bar{d}_{1n_{q_{\max}}}$ and $\bar{d}_0 = \bar{d}_{0n_{q_{\max}}}$ 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:

$$\text{Power}(\hat{\beta}_{\text{IV}}(q_{\max}); \beta^\circ) = P(\text{reject } H_0 : \beta = 0 \mid \beta = \beta^\circ) \quad (20)$$

$$= 1 - \Phi\left(1.96 - \frac{\beta^\circ}{\sqrt{\mathbb{V}\left(\hat{\beta}_{\text{IV}}(q_{\max})\right)}}\right). \quad (21)$$

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 (i.e., 10 percentage points), we see in Figure 2 that power declines significantly after $q_{\max} = 103$ while the corresponding increase in standard error is less evident due to scale of the y-axis.

D RECONSTRUCTING QUEUE POSITIONS

As discussed in Section 3.4, 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 record subjects' queue positions. However, it did record meta-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. Algorithm 2 provides pseudocode for this function.

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

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 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.
  $u^\star \leftarrow u_1$ ▷ 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 ▷ A receiver cannot appear more than once
    Append(SeenReceivers, r_n)
    if $u_n \neq u^\star$ **then**
      return False ▷ Only one user per queue
    if $t^\star$ is not NaN **then**
      if $t_n < t^\star$ **then**
        return False ▷ Queue must be sorted by timestamps
      if $(t_n - t^\star) >$ 1 hour **then**
        return False ▷ Time intervals should not be implausibly long
    $t^\star \leftarrow t_n$
  return True

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 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 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
      $q_n \leftarrow q^\star$ ▷ Assign current queue ID to queuing event
      $t^\star \leftarrow t_n$ ▷ Update current timestamp
  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

Users sometimes pressed an "Add all" button that would instantly add all of their phone contacts to the queue. As evidenced by the timestamp meta-data, many users would hit "Add all", and then immediately exit the queue, and start a new queue to which they would add contacts selectively. This user behavior created many queuing events with nearly simultaneous timestamps that included the same contacts multiple times, thus making it difficult to confidently partition events into 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 confidently partitioned into queues which meet the aforementioned plausibility criteria. This then confidently reconstructs the first queue position of 60% of eligible subjects. Many of the remaining 40% of subjects whose first queue position cannot be reconstructed are subjects who were only introduced to the subject pool by an "Add all". Unsurprisingly, these subjects exhibit a very low compliance rate of 8% (see Figure 4), as would be expected from an accidentally queued subpopulation.

E BALANCE CHECKS

Our data from TargetSmart contains 85 covariates that are either categorical demographic variables (e.g., party registration), binary indicators of voting in past elections, continuous demographic variables (e.g., age), or continuous modeled scores (e.g., Authoritarianism score).

We use these covariates 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,

$$Z_i = w_0 + \boldsymbol{w}^\top \boldsymbol{X}_i + \epsilon_i. \tag{22}$$

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. We say a balance check "fails" when a small p-value indicates that the null hypothesis is unlikely—i.e., that the assigned treatment and control groups do not exhibit balance.

We report the balance check's p-value for all subpopulations in Figure 4. As expected, all checks pass. 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 balance between the two groups.

As mentioned in Section 3.1, 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 [23], employed this confounded definition of treatment–control.

F SUBJECT FLOW DIAGRAM

We depict how subjects were selected for analysis in Figure 4. 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 E).

The first refinement of the eligible subject pool, labeled "Measurement Error" in Figure 4, 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 D). 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: $n = 27{,}464$) 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.

REFERENCES

[1] 2020. Outvote. https://www.outvote.io/. Accessed: 2020-04-21.

[2] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American statistical Association* 91, 434 (1996), 444–455.

[3] Joshua D Angrist and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion.* Princeton University Press.

[4] Peter M Aronow and Donald P Green. 2013. Sharp bounds for complier average potential outcomes in experiments with noncompliance and incomplete reporting. *Statistics & Probability Letters* 83, 3 (2013), 677–679.

[5] Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. 2012. A 61-million-person experiment in social influence and political mobilization. *Nature* 489, 7415 (2012), 295–298.

[6] Paul-Antoine Chevalier. 2017. How can I compute the standard error of the Wald estimator? StackExchange. https://stats.stackexchange.com/questions/219367/how-can-i-compute-the-standard-error-of-the-wald-estimator

[7] Lindsey Cormack. 2019. Leveraging peer-to-peer connections to increase voter participation in local elections. *Politics & Policy* 47, 2 (2019), 248–266.

[8] Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. 2006. *Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand.* Technical Report. National Bureau of Economic Research.

[9] Tiffany C Davenport. 2008. Unsubscribe: Comparing the effects of peer-to-peer and mass email messages on voter turnout. In *Poster presented at the annual meeting of the American Political Science Association, Boston, MA.*

[10] Tiffany C Davenport. 2010. Public accountability and political participation: Effects of a face-to-face feedback intervention on voter turnout of public housing residents. *Political Behavior* 32, 3 (2010), 337–368.

[11] Donald P Green. 2019. *Evaluation of Peer-to-Peer Voter Mobilization Campaign by Alliance for Climate Education during the 2018 Midterm Election.* Technical Report.

[12] Donald P Green and Alan S Gerber. 2019. *Get out the vote: How to increase voter turnout.* Brookings Institution Press.

[13] Donald P Green and Oliver A McClellan. 2020. *Turnout Nation: A pilot experiment evaluating a get-out-the-vote "supertreatment".* Technical Report. https://www.turnoutnation.org/thereport.

[14] Katherine Haenschen. 2016. Social pressure on social media: Using Facebook status updates to increase voter turnout. *Journal of Communication* 66, 4 (2016), 542–563.

[15] Katherine Haenschen and Christopher B. and Mann. 2021. *Short but Mighty: The Effects of SMS Mobilization Message and Timing among 7.1 Million Voters.* Technical Report.

[16] Miguel A Hernan and James M Robins. 2010. Causal inference.

[17] Kosuke Imai and Teppei Yamamoto. 2010. Causal inference with differential measurement error: Nonparametric identification and sensitivity analysis. *American Journal of Political Science* 54, 2 (2010), 543–560.

[18] Tobias Konitzer, Sam Corbett-Davies, and David M. Rothschild. 2017. *Non-representative surveys: modes, dynamics, party, and likely voter space.* Technical Report. https://researchdmr.com/Methods2016.

[19] Peter Kristensen. 1992. Bias from nondifferential but dependent misclassification of exposure and outcome. *Epidemiology* (1992), 210–215.

[20] Arthur Lewbel. 2007. Estimation of average treatment effects with misclassification. *Econometrica* 75, 2 (2007), 537–551.

[21] Todd Rogers, Craig R Fox, Alan S Gerber, and Eldar Shafir. 2013. Rethinking why people vote. *The Behavioral Foundations of Public Policy* 91 (2013).

[22] Donald B Rubin. 1980. Randomization analysis of experimental data: The Fisher randomization test comment. *J. Amer. Statist. Assoc.* 75, 371 (1980), 591–593.

[23] Aaron Schein, Dhanya Sridhar, Keyon Vafa, Victor Veitch, Naseem Makiya, James Moffet, Jeff Quinn, David M Blei, and Donald P Green. 2020. The Effect of Relational Text Messaging on Turnout in the 2018 US Midterm Elections. osf.io/9qwbp

[24] Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference.*

[25] Betsy Sinclair, Margaret McConnell, and Donald P Green. 2012. Detecting spillover effects: design and analysis of multilevel experiments. *American Journal of Political Science* 56, 4 (2012), 1055–1069.

[26] Douglas Staiger, James H Stock, et al. 1997. Instrumental variables regression with weak instruments. *Econometrica* 65, 3 (1997), 557–586.

[27] Holly Teresi and Melissa R Michelson. 2015. Wired to mobilize: The effect of social networking messages on voter turnout. *The Social Science Journal* 52, 2 (2015), 195–204.

[28] Kara Waite. [n.d.]. The Key To Creating Lasting Change? Mobilizing the People You Already Know. *Harper's Bazaar* ([n. d.]). https://www.harpersbazaar.com/culture/politics/a34704263/what-is-relational-organizing/

[29] Larry Wasserman. 2013. *All of statistics: A concise course in statistical inference.* Springer Science & Business Media.

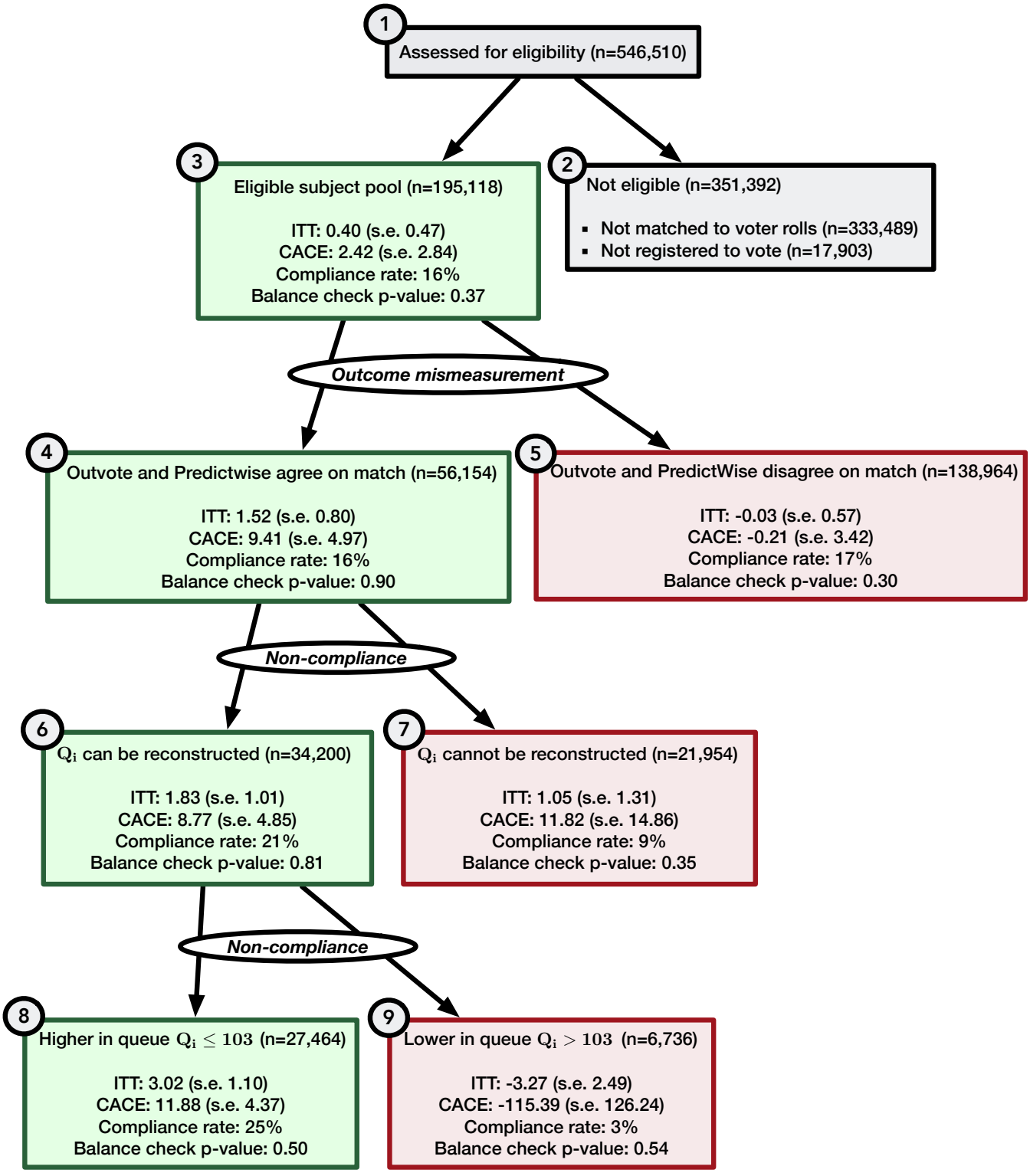


Figure 4: Subject flow diagram.