# Assessing the Effects of Friend-to-Friend Texting on Turnout in the 2020 U.S. Presidential Election

Aaron Schein[1], David M. Blei[1], and Donald P. Green[2]

[1]*Data Science Institute, Columbia University*
[2]*Department of Political Science, Columbia University*

## 1 Introduction

Political campaigns in recent elections have started to embrace friend-to-friend organizing, in which volunteers organize and encourage their own close contacts to cast a ballot on Election Day. Unlike traditional "get out the vote" (GOTV) campaigns, which often rely on texts, calls, or visits from strangers, friend-to-friend organizing is premised on the notion that GOTV encouragements are especially effective when delivered by trusted messengers, like friends or family members.

A recent RCT by Schein et al. found large treatment effects of friend-to-friend text-message reminders to vote in the 2018 U.S. midterm elections: $\widehat{\text{CACE}} = 8.3, \text{CI} = (1.2, 15.3)$. This study is the first large-scale experiment to assess the causal effects of friend-to-friend GOTV tactics. A smaller experiment that assessed a friend-to-friend GOTV tactic in the lead-up to the 2019 municipal elections also found large effects [Green and McClellan, 2020]. These results are tantalizing; however, more experimental evidence is necessary to meaningfully compare friend-to-friend tactics to traditional GOTV tactics whose effects have been repeatedly assessed and replicated through two decades of field experiments [Green and Gerber, 2019].

In this paper, we report a follow-up study to that of Schein et al. which we conducted on the same mobile-app platform, Outvote[1], during the 2020 U.S. presidential election. Our study estimates smaller treatment effects—$\widehat{\text{CACE}} = 2.21, \text{CI} = (-2.30, 6.72)$—of friend-to-friend text-message reminders during the 2020 election. We additionally use the data from Schein et al. to estimate the effect that Outvote messages from 2018 had on voter turnout in the 2020 election and find an effect of $\widehat{\text{CACE}} = 5.63, \text{CI} = (-0.98, 12.24)$. Taken together, this new evidence is consistent with moderate treatment effects of friend-to-friend encouragements on turnout in 2020—however, this evidence likely rules out effects in 2020 as large as those found in 2018 and does not rule out null effects (see Fig. 1). Whether these new results should temper excitement about friend-to-friend tactics is debatable. It is generally assumed that treatment effects of GOTV tactics during Presidential elections will be weaker than those during midterms due to ceiling effects from higher baseline turnout, and this may be particularly true for 2020, which had record-high turnout.

The main technical challenge this paper addresses is how to assess causal effects using instrumental variables that are naturally ordinal $Q_i \in \{1, 2, \ldots, L_i\}$ and whose maximum possible level $Q_i$ differs across subjects, possibly endogenously. This challenge arises naturally from our experimental design, which is different from that of Schein et al.. That design relied on randomly "skipping" contacts in users' queues with a small (5%) probability; our new design relies instead on randomizing the order in which users are presented the contacts in their queue. Our estimation strategy relies on automatically selecting a cutoff $K$ and binarizing the ordinal queue positions to obtain binary instruments—$Z_i = \mathbb{1}(Q_i \leq K)$—while also controlling for the confounding effect of queue length $L_i$ using inverse propensity weights.

## 2 Experimental design

**Setting.** The mobile app Outvote provides a streamlined platform for users to message their close contacts with encouragements to vote. The app works in stages. In the first stage, the user creates a queue of contacts they intend to message. After creating a queue, the app then takes the user to a messaging interface for the first contact in the queue, where the user is provided a default message that they can either edit or send as is. After hitting "Send", the app then takes the user to a new messaging interface for the next contact on their queue. The user is also able to "Skip" a contact without messaging them.

[1]The app was renamed to Impactive (www.impactive.io) prior to 2020 but we refer to it as Outvote be consist with existing literature.

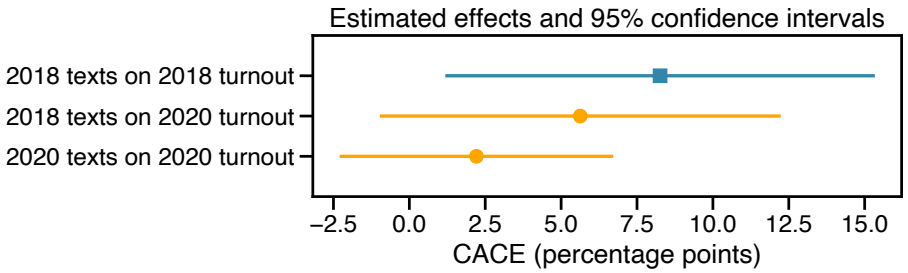


**Figure 1:** We estimate weaker treatment effects of friend-to-friend text messages on turnout in the 2020 presidential election than those previously estimated for turnout in the 2018 midterms. The effects estimated by Schein et al. of texts from 2018 on turnout in 2018 is denoted by the blue square (■). The two new effects this paper reports on turnout in 2020 are denoted by orange circles (●). While these are consistent with moderate effects on turnout in 2020, both confidence intervals include zero and thus do not rule out null effects, while the confidence interval for the effect of texts in 2020 (bottom) does not overlap with the upper half of the confidence interval for 2018 texts on 2018 turnout, likely ruling out the same large effects in 2020.

**Randomization.** Starting on September 21st and lasting through Election Day on November 3rd, 2020, the app randomized the order of contacts in users' queues. This design was premised on the assumption that many users would create long queues and exit the app before messaging everyone on them. Names that were randomly sorted to the end of queues thus had lower probabilities of receiving the treatment.

**Study population.** We consider subjects to be anyone who 1) was queued by a user during the study period, 2) was *confidently*[2] matched to the TARGETSMART voter rolls database, and 3) did not vote early or absentee prior to being queued. A total of $n = 81,204$ subjects meet these conditions. The first two conditions were also applied by Schein et al.. The third is new to this study and necessary since a record number of individuals voted early and absentee in the 2020 election to maintain social distance during the pandemic, among other reasons. Individuals who voted before being queued are essentially "negative controls"—i.e., we know *a priori* the treatment cannot affect their 2020 voting outcome. We note that although it was particularly pertinent in 2020, the same filter could be applied to any GOTV field experiment for which the date on which a ballot was cast is recorded.

## 3 Assessing causal effects

What were the causal effects of OUTVOTE users' messages on their friends' voting outcomes in 2020?

Let $Y_i \in \{0,1\}$ denote subject $i$'s recorded voting *outcome* and $D_i \in \{0,1\}$ denote whether $i$ received the *treatment*—i.e., a message from an OUTVOTE user during the study period. We want to assess causal effects and thus also define subject $i$'s *potential outcomes*, $Y_{i1}$ and $Y_{i0}$, which denote whether $i$ would vote if they did or did not receive the treatment. The causal effect of the treatment on $i$ is then $Y_{i1} - Y_{i0}$. Since $D_i$ was not randomized, the average causal effect $\mathbb{E}[Y_{i1} - Y_{i0}]$ is not identified. However, we can use the randomized queue order to define binary instrumental variables that identify *local* average causal effects.

Let $Q_i \in \{1, 2, \ldots, L_i\}$ be the position of subject $i$ in their first queue and $L_i$ be the length of that queue. Since we randomized the order of queues, $Q_i$ could be considered an ordinal instrument whose levels depend on length: $P(Q_i = q \mid L_i) = L_i^{-1}$. Angrist and Imbens [1995a] show that a weighted average of local average causal effects is identified by data of binary outcomes, binary treatments, and categorical (or ordinal) instruments, and Angrist and Imbens [1995b] show that two-stage least squares (2SLS) consistently estimates it. Tan [2006] and Ogburn et al. [2015] also study local average effects identified by ordinal instruments. However, these

[2]We obtained ancillary data from PREDICTWISE (www.predictwise.com) and consider a contact to be confidently matched if OUTVOTE's matching system matched them to the same voter record in the TARGETSMART database as PREDICTWISE's system. This is the same measure taken by Schein et al. to mitigate attenuation bias stemming from measurement noise. We find that a similar proportion (around 30%) of subjects are confidently matched in this dataset.

approaches are difficult to apply directly in our case due to the presence of positivity violations—i.e., not all subjects have a non-zero probability of taking all observed queue positions since $P(Q_i > q) = 0$ if $q > L_i$.

For simplicity and to guarantee positivity, we instead define binary instrumental variables based on some cutoff $K$—$Z_i = \mathbb{1}(Q_i \leq K)$—and only consider subjects in queues with lengths $L_i > K$. Thus, $P(Z_i = 1 \mid L_i > K) = K/L_i$. With binary instruments, the complier average causal effect is identified,

$$\text{CACE} = \mathbb{E}\left[Y_{i1} - Y_{i0} \mid D_{i0} < D_{i1}\right], \tag{1}$$

where $D_{i1}$ is subject $i$'s *potential receipt* denoting whether they would receive a message if they were before the cutoff ($Z_i = 1$) and $D_{i0}$ denotes whether they would receive a message if they were after ($Z_i = 0$). The subjects for whom $D_{i0} < D_{i1}$ are called "compliers"—they only receive a message if they are before the cutoff. An assumption necessary for identification is "monotonicity," which stipulates that there are no "defiers" for whom $D_{i0} > D_{i1}$. Intuition suggests that no subjects would get messaged *only* if they appeared *after* some cutoff.

Since $P(Z_i = 1 \mid L_i)$ differs across subjects we control for $L_i$ as a confounder using backdoor adjustment in the numerator and denominator of the Wald estimator,

$$\widehat{\text{CACE}} = \frac{\sum_\ell P(L_i = \ell)\Big(\mathbb{E}[Y_i \mid Z_i = 1, L_i = \ell] - \mathbb{E}[Y_i \mid Z_i = 0, L_i = \ell]\Big)}{\sum_\ell P(L_i = \ell)\Big(\mathbb{E}[D_i \mid Z_i = 1, L_i = \ell] - \mathbb{E}[D_i \mid Z_i = 0, L_i = \ell]\Big)}, \tag{2}$$

where $P(L_i = \ell)$ is the proportion of subjects in queues of length $\ell$. Note that the numerator is the inverse probability weighted estimator of the intent-to-treat effect while the denominator is the inverse probability weighted estimator of the first stage.

To select the cutoff $K$, we consider three trade-offs. First, different cutoffs yield different proportions of compliers, as estimated by the denominator in Eq. (2). A low compliance rate increases the variance of estimates and limits their generalizability. Second, since subjects in queues of length $L_i \leq K$ must be excluded, a higher cutoff $K$ means a smaller sample size $n$. Third, cutoffs close to the extremes (e.g., $K = 1$) will yield an imbalanced treatment-control split, which degrades precision. To formalize this trade-off, we follow a similar procedure outlined by Schein et al. to automatically select the cutoff $K$ which minimizes the expected standard error of the estimator—specifically, for every possible value of $K$, we plug in the induced compliance rate, $n$, and $P(Z_i = 1)$ to a proxy for the expected error (which does not depend on $Y_i$) and select the value that minimizes it.

**Results.** The procedure outlined above selected $K = 37$ corresponding to a sub-sample of $n = 43,265$ subjects with $L_i > 37$. We compute the estimator in Eq. (2) on this sub-sample and find an effect of 4.18 (SE=3.13) percentage points. We then recursed by applying the same procedure to the excluded subjects with $L_i \leq 37$. The procedure selected $K = 4$, corresponding to a sub-sample of $n = 33,089$ and an estimated CACE of -2.00 (SE=6.10). Recursing once more, the procedure then selected $K = 1$ on the excluded subjects, corresponding to all remaining subjects and an estimated CACE of 0.74 (SE=4.07). To summarize these three estimates on the three automatically selected disjoint sub-samples, we computed a precision-weighted average which yields an overall estimate of $\widehat{\text{CACE}} = 2.21$ (SE=2.30) percentage points, as shown in Fig. 1.

# 4 Robustness check

As a robustness check, we re-analyze the data using a different way of deriving subjects' binary treatment assignments $Z_i \in \{0, 1\}$ from their ordinal queue positions $Q_i \in \{1, 2, \ldots, L_i\}$. In this analysis, we operate on subjects' normalized queue position $\tilde{Q}_i = Q_i/L_i$, which represents the percentile rank of subject $i$ in their respective queue. Unlike raw queue position, this quantity is bounded on the unit interval, $\tilde{Q}_i \in [\frac{1}{L_i}, 1]$. As before, we then use some cutoff $K \in (0, 1)$ to define binary treatment assignments based on percentile rank, $Z_i = 1$ if $\tilde{Q}_i \leq K$. Thresholding on percentile rank versus thresholding on raw queue position represent two different models of user behavior. Thresholding on raw queue position presupposes that users tend to abandon the app after messaging a certain number of contacts, regardless of how many contacts they queued. Thresholding on percentile rank presupposes that users abandon the app after messaging a similar *percentage* of the contacts in their queue.

| Queue bin $B_i$ | Queue lengths $L_i$ | Description | $n$ subjects | Selected cutoff $K$ | $\widehat{\text{CACE}}$ |
|---|---|---|---|---|---|
| 1 | $\{2, \ldots, 6\}$ | "micro" | 9,043 | 0.84 | 12.83 (SE=9.22) |
| 2 | $\{7, \ldots, 19\}$ | "short" | 14,770 | 0.41 | -8.54 (SE=9.18) |
| 3 | $\{20, \ldots, 53\}$ | "medium" | 22,464 | 0.36 | -4.52 (SE=6.12) |
| 4 | $\{54, \ldots, 147\}$ | "lengthy" | 21,728 | 0.42 | -0.30 (SE=4.06) |
| 5 | $\{148, \ldots, 402\}$ | "long" | 10,176 | 0.23 | 7.27 (SE=4.46) |
| 6 | $\{403, \ldots\}$ | "spammy" | 3,023 | 0.20 | 7.77 (SE=8.66) |

**Table 1:** This analysis partitions subjects into mini-experiments according to similar queue lengths. The mapping we use is the log-transform of queue length, $B_i = \lfloor \log(1 + L_i) \rfloor$.

The previous analysis automatically partitioned the subject pool into mini-experiments, with each having an optimally-selected cutoff $K$ that defines subjects' treatment assignments. In this analysis, we instead partition the subject pool ahead of time into mini-experiments, each involving subjects in queues of similar lengths. The reason we do this is to capture the fact that a percentile cutoff of $K = 0.1$ has a substantially different qualitative interpretation in a queue of length $L_i = 10$ versus a queue of length $L_i = 1,000$. A natural way to bin subjects according to queue length is to use the following log-transform $B_i = \lfloor \log(1 + L_i) \rfloor$. This transformation creates an interpretable mapping of queue lengths to bins that we describe in Table 1.

For each bin, we select an optimal cutoff $K$ using the procedure outlined in the previous section, which trades off the different levels of compliance, different levels of imbalance in the treatment-control split, and different overall sample sizes implied by a selected $K$.

**Results.** For the 1st bin of subjects in "micro" queues, the above procedure selects $K = 0.84$. We provide the selected cutoffs for all other bins in Table 1. The general trend is that earlier cutoffs are selected for longer queues lengths, which makes sense if users rarely message past a certain number of contacts, regardless of queue length. We also report the estimated CACE and standard errors for each bin in Table 1. As before, to summarize these effects, we take a precision-weighted average. This yields an overall estimate of $\widehat{\text{CACE}}$ = 2.17 (SE=2.40), which is similar to the estimate from the previous analysis.

# 5 Discussion

Results from 2020 suggest a smaller CACE than corresponding results from 2018. One interpretation is that GOTV campaigns are less influential in high-salience elections. Using the formulas from Corollary 1 of Aronow and Green [2013], we find that untreated compliers in each of the six bins in Table 1, have implied turnout rates ranging from 74% to 97%, which are all substantially higher than the 67% estimated by Schein et al. in the 2018 study. It remains to be seen whether, in line with this interpretation, results in 2022 resemble the large 2018 estimates.

# References


Joshua Angrist and Guido Imbens. Identification and estimation of local average treatment effects, 1995a.

Joshua D Angrist and Guido W Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American statistical Association*, 90(430):431–442, 1995b.

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

Donald P Green and Alan S Gerber. *Get Out the Vote: How to Increase Voter Turnout*. Brookings Institution Press, 2019.

Donald P Green and Oliver A McClellan. Turnout Nation: A Pilot Experiment Evaluating a Get-Out-The-Vote "Supertreatment". Working paper, 2020.

Elizabeth L Ogburn, Andrea Rotnitzky, and James M Robins. Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):373–396, 2015.

Aaron Schein, Keyon Vafa, Dhanya Sridhar, Victor Veitch, Jeffrey Quinn, James Moffet, David M Blei, and Donald P Green. Assessing the effects of friend-to-friend texting on turnout in the 2018 US midterm elections. In *Proceedings of the Web Conference 2021*.

Zhiqiang Tan. Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association*, 101(476):1607–1618, 2006.