High traffic on-the-spot vote tripling: relational asks A/B test

Vote Rev Action Fund

Michael Cohn, Aug 2023

Summary

High-traffic on-the-spot vote tripling (HTOTS) has canvassers ask people on the street to text friends or family with a reminder to vote.

In the 2022 Arizona generals, Vote Rev Action Fund added an A/B test to our impact-focused HTOTS program, comparing asks to remind 5 vs. 7 vs. 10 friends.

Asking for 10 friends generated the most relational contacts: **51 contacts per canvasser-hour**, vs. 30-37 per hour when asking for 5 friends. Canvassers talk to fewer mobilizers per hour, but the increase in friends per mobilizer more than makes up for it.

This test was efficient, low marginal effort, and did not reduce the impact of our canvassing program. We encourage other groups to embed A/B tests into their ongoing programs!

Vote Rev provides free support to organizations implementing HTOTS in 2023 and 2024! Please contact Marisa Kanof, Director of Partner Success: marisa@voterev.org.

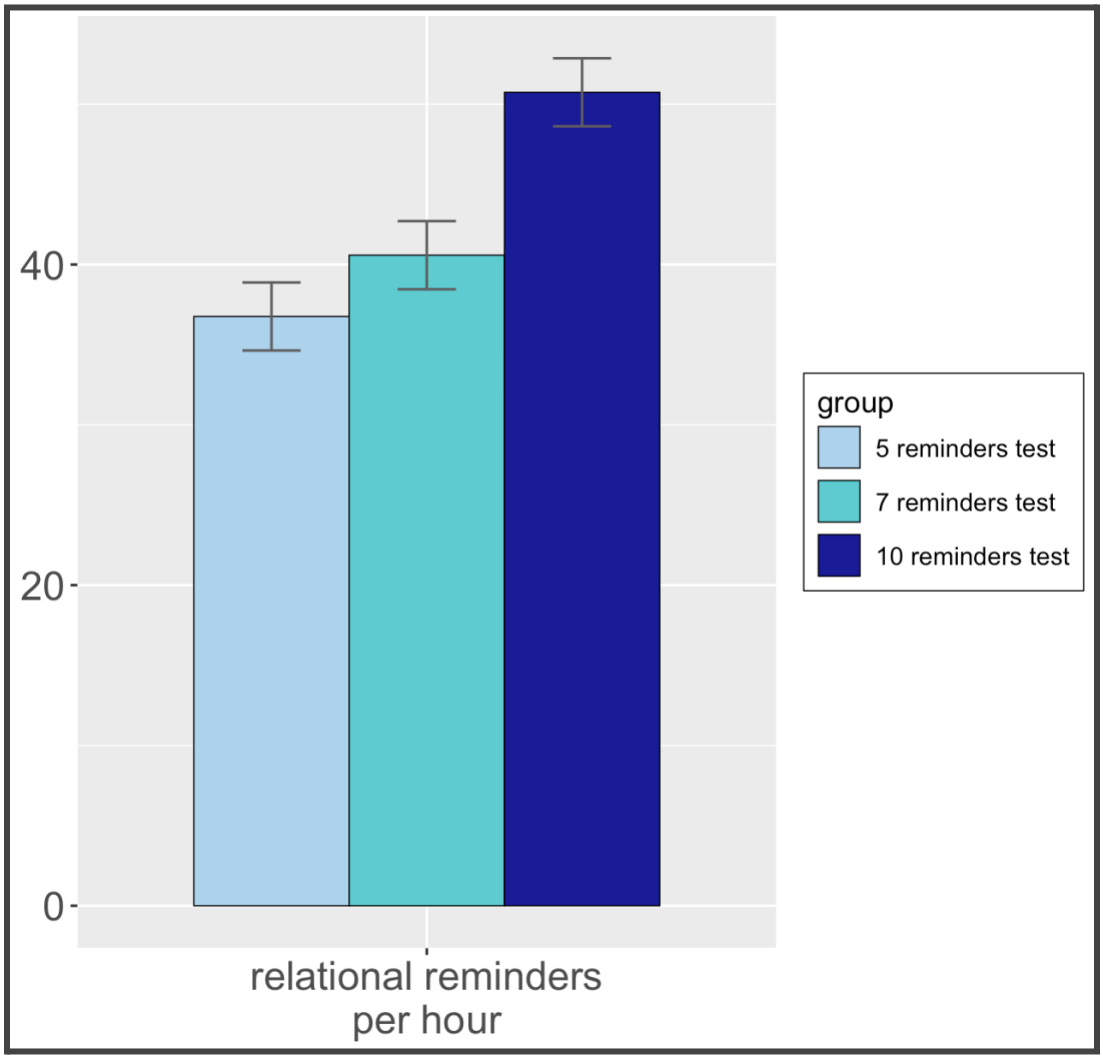

Figure 1: Executive summary: When canvassers changed from asking people to send 5 relational reminders to asking for 10 relational reminders, the number of reminders obtained per hour increased significantly, from 37 to 51. Error bars show standard error.

Background

In high-traffic on-the-spot vote tripling (HTOTS), canvassers station at locations with high foot traffic and ask people ("mobilizers") to send a message to friends or family members ("friends"[1]) to remind them about an upcoming election.

In the 2022 general elections in Arizona, Vote Rev Action Fund conducted an impact-oriented HTOTS program in which each mobilizer was asked to remind 5 friends. We embedded an A/B test examining whether we could increase overall efficiency by asking each mobilizer to name 7 or 10 friends.

Method and data

We collected 5 days of data: Nov 2-4 and 7-8. Canvassers carried out HTOTS using a canvassing data collection app, Grassroots Unwired (GRU), on personal devices. GRU was programmed to give each canvasser one of 3 scripts, which were identical except that one asked the mobilizer to message 5 contacts, one asked for 7, and one for 10. Each canvasser used the same script for an entire day.

Our field director attempted to evenly distribute each canvasser across each script. However, differences in canvasser schedules meant this was not fully possible. Additionally, canvassers were at different locations on different days, and days were inherently different from each other (eg, 11/4 was a Friday, and 11/8 was Election Day). To help control for these effects, we incorporated data from the full Arizona HTOTS program, in which the same canvassers had been doing "standard" HTOTS for the prior few weeks. The standard program also asked pedestrians to remind 5 friends, but because it may have differed from the A/B test in other ways, we treated it as a separate experimental condition.

Note: Many Vote Rev activities during the 2022 elections involved asking mobilizers to match friends to the voter file so we could evaluate turnout effects. This is necessary to test turnout effects but dramatically slows the process and reduces relational contacts per hour. **None of the programs evaluated in this study used voter file matching, so our results are applicable to real-world HTOTS implementations at scale.**

[1] We use "friends" as a shorthand for anyone receiving a reminder; mobilizers in the study were told that they could remind anyone, and suggestions included friends, family members, coworkers, partners, and classmates.

Unobserved variables

Methods for eliciting the list of friends differed between canvassers and between canvass sites. Some canvassers offered the available incentives (usually candy or snacks) up front; some waited until the end; and some offered a small incentive up front and a larger or additional incentive partway through to encourage mobilizers to provide the maximum number of friends. Similarly, some canvassers approached pedestrians asking for the full number of friends up front, while others used a "foot in the door" approach: naming a small number to start with, and then encouraging the mobilizer to keep going. We have no data on these practices, how they may have affected outcomes, or how they might have modified the effects of different scripts.

Analysis

Questions

The main question is: **Does the number of friends requested (5 vs 7 vs 10) affect Relational Contacts** (RCs) **per hour?** We also examined the determinants of RCs / hour:

- mobilizers / hour (regardless of number of RCs they provided)
- Asks / hour
- RCs / mobilizer
- mobilizers / ask (ie, conversion rate)

Due to Unobserved variables (above) we are not able to answer questions about incentive cost per RC or about best practices in carrying out any individual script.

Method

We did not receive reliable data on canvassers' intended schedules, and this data is often unreliable because many canvassers take unrecorded breaks or take time to move between locations. Therefore, we broke our data into half-hourly bins, discarding bins in which the canvasser did not record making any asks but including ones in which they made asks but had zero conversions. For count data (friends, mobilizers, and asks) perform a Poisson regression; for conversion rate treat as linear. Results are interpretable as events per half hour; we convert these to events per hour.

Cleaning and processing

Eliminate any records that are:

- From the first 3 days of the impact program (10/14-10/16), because leads were still working out training and logistics at that time.
- Attributed to canvassers who worked fewer than 10 hours total
- Timestamped before 9:00am or after 9:00pm (no canvassing shifts ran during these times)

Testing approach

- Dependent variables:
    - **Number of RCs** (primary outcome)
        - Group texts to more than 10 people were capped at 10
        - mobilizers who sent more than 20 individual messages were capped at 20
    - Secondary outcomes:
        - Number of mobilizers
        - Number of asks (total count of records, including declines)
        - Conversion rate (yeses / ask)
- Independent variable: Which script version
    - Categorical: 5 vs 7 vs 10 vs impact program
- Covariates
    - Fixed effects:
        - Hour of day
        - Day (categorical)
        - Canvasser

Results

Descriptives

See Appendix 1: descriptives for raw numbers; these numbers are not generalizable results because of imbalances in date and canvasser quality between the conditions.

Significance testing

We ran separate models for each outcome. The table below shows how each script compared to the 5-friend script from the a/b test on each outcome. All differences reported are statistically significant. Because we used Poisson regression for the person-count statistics, coefficients are not directly interpretable and we only indicate

"lower" or "higher." The section Real-world effectiveness estimates below reports on what changes we might expect in real-world contexts.[2]

For most purposes, **friends per hour is the ultimate outcome**, with mobilizers, asks, and conversion as informative process variables. Therefore, in these results **the 10-friend script is strictly superior**, despite the reduction in other metrics (but see Caveats for some potential exceptions).

	friends / hour	mobilizers / hour	Asks / hour	Conversion rate
Standard program	**lower** *than 5-friends a/b test*	**higher** *than 5-friends a/b test*	*no change*	**higher** *than 5-friends a/b test*
7 friends	**higher** *than 5-friends a/b test*	*no change*	**lower** *than 5-friends a/b test*	**higher** *than 5-friends a/b test*
10 friends	**higher** *than 5-friends a/b test*	**lower** *than 5-friends a/b test*	**lower** *than 5-friends a/b test*	*no change*

Table 1: Differences from the 5-friend script

Real-world effectiveness estimates

The **simulated results** below allow us to see **how the conditions would compare if we could remove all imbalances** created by running on different days or with different canvassers.

This was created by using our regression models to calculate the predicted yield for each canvasser-hour as if it had used the specified script.[3] Note that numbers do not multiply out precisely (eg, friends per hour is not equal to mobilizers per hour times friends per mobilizer) because each outcome is being estimated with a separate statistical model.

	reminders (friends) per hour	mobilizers per hour	Asks per hour	friends per mobilizer	Conversion rate

[2] See Appendix: Correlation coefficients for actual coefficients.
[3] For example, we created a copy of our dataset with the outcomes removed and all predictors unchanged except that the condition for every interaction was set to the 5-friend script. We then used our regression model to generate predicted friends, mobilizers, and asks for that dataset. We did the same for the other three conditions.

Standard program(asked for 5)	30.3	**8.4**	21.7	3.6	39%
5 friend script	36.8	7.1	**21.1**	5.2	34%
7 friend script	40.6	6.7	16.4	6.1	**41%**
10 friend script	**50.7**	6.0	17.0	**8.5**	35%

Table 2: *Expected HTOTS outcomes, holding other variables constant*

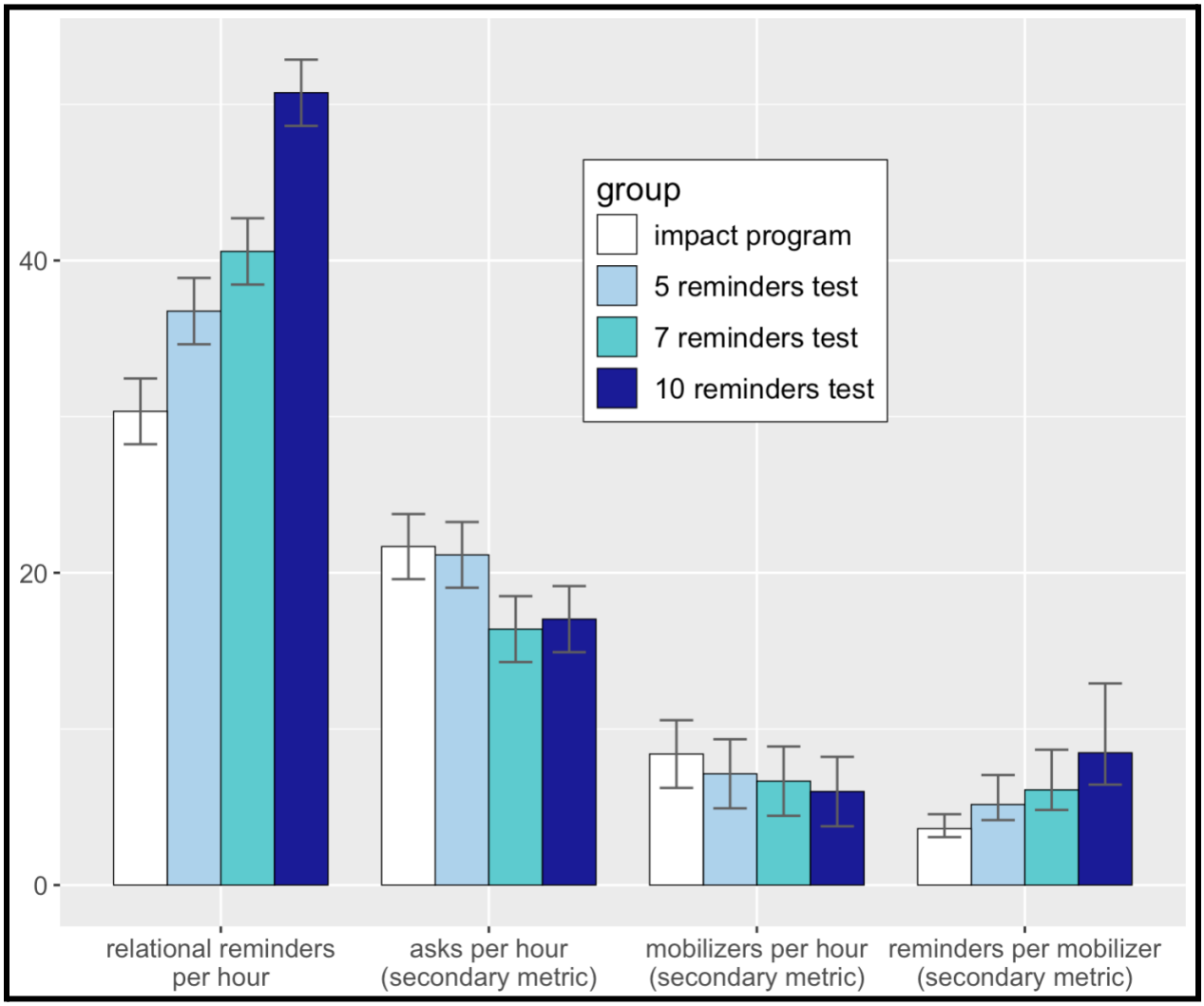

Figure 2: *Total reminders, asks, and people sending reminders in each group. Error bars are standard error. Note that relational reminders (friends) per hour is the most important outcome.*

We note that the 5-friend script substantially outperforms the impact program, even though the impact program also used a 5-friend script. See the Appendix for potential explanations. It's possible that this difference is due to some contextual factor that we were unable to control for, and the 5-friend A/B test script is the best estimate of expected performance in future iterations of this program. Even if that is the case, the 10-friend script still outperforms the 5-friend script by 13.9 friends per hour, a major improvement.

Conclusions

Overall results

- **The 10-friend script produces more relational contacts per hour than the 5-friend script** (high confidence). The 7-friend script may do so as well (low confidence)**.**
    - Because we measure effectiveness in RCs per hour, this bottom-line result outweighs any other differences reported below.
- The 10-friend script produces fewer mobilizers and asks per hour than the 5-friend script (moderate confidence).
- The reduction in mobilizers per hour for the 10-friend script is due to canvassers approaching fewer people and *not* to lower conversion. That is, pedestrians are *not* more likely to refuse when the canvasser asks for more friends (low confidence).
- Conversion rate behaved unexpectedly, and we're not sure what to infer about it. The 5-friend script and the 10-friend script had similar conversion rates, and the impact program and the 7-friend script had substantially higher rates.

The evidence above suggests that the 10-friend script increases *relational contacts per mobilizer,* and does so by more than enough to make up for the lower number of mobilizers.

See the Appendix: Robustness checks for results of alternate model specifications and data cleaning decisions. Regardless of approach, the 10-friend script continues to outperform the 5-friend script.

Practical applications

- **Asking for more friends has a lot of potential!**

- **We plan to investigate whether the first few contacts are more effective than later ones**. We have speculated that the first ~5 relational contacts a person provides may be people they can influence effectively, while contacts 6-10 are more tenuously connected (acquaintances, distant relatives, etc). This might mean that asking for more contacts inflates our contact numbers without equally benefitting turnout.
    - On the other hand, it's also imaginable that contacts 5-10 are *more* valuable, because they're less likely to have already been influenced by the messenger.
    - A relational organizing study by AI and Community Change Action found that relational contact effectiveness dropped around the 30th contact listed; this suggests that our concern may be correct in principle, but not relevant when requesting ≤10 contacts.
- If campaigns ask for more contacts per mobilizer, they should consider **sending ~30% more canvassers to each site**. The reduction in asks/hour may be evidence that canvassers are missing potential conversions due to spending more time with each mobilizer.

Comments on testing

- This A/B test was extremely cheap to run, using canvassers we were already paying and had already trained.
- Instead of using a no-treatment control group, it compared one probably-effective tactic to another. This allowed us to generate important learnings without compromising our impact on AZ voter turnout.
- In many of our trials, we need to match friends to the voter file or otherwise gather information that can be used to check voting outcomes. Doing so adds canvasser and mobilizer burden, slows down the process, and creates a risk that the study results won't be applicable to real-life canvassing. Instead, this study used a proxy outcome: relational contacts per canvasser-hour, based on the knowledge that relational contacts are an effective turnout method. This cost nothing to measure, added no burden, and adds confidence that these results will generalize to other vote tripling canvassing operations.

We are excited about the success of this test and look forward to future opportunities to insert A/B tests into our impact programs!

Caveats

1) The process of recording declined asks in GRU was cumbersome and we think it's likely that at least some canvassers heavily underreported them. We don't have any reason to think that this would have differed between conditions. As a result, we have major concerns about the absolute numbers we generated for asks and conversion rates, but these concerns do not affect our numbers for relational reminders.

2) Our method of estimating canvasser effectiveness only applies to time bins in which canvassers were working consistently. For example, if a canvasser spends a total of 2 hours during a 5-hour shift on breaks, setup, or location-switching, the figure for relational contacts per hour should only be multiplied by 3, not by 5. However, this has no impact on our main question: the difference in per-hour effectiveness between the different scripts.

3) We don't yet know whether reminders sent to group chats differ in effectiveness from those sent to individual friends. We also don't have data on whether asking for larger numbers of friends makes mobilizers more likely to send messages to group chats, or increases the size of the group they choose to send to. These are important topics for further study.

4) This paper focuses on relational contacts per hour. This assumes that the program implementor values all RCs equally. eg, one mobilizer contacting three friends is just as good as three mobilizers contacting one friend each. This is not true in a number of contexts:

    - In a randomized controlled trial (RCT), if we are clustering results by mobilizer, then added mobilizers increase our power much more than more friends per mobilizer.
    - If we believe there's a voting effect on *mobilizers*, that would increase the value of having more mobilizers. However, unless the effect on mobilizers is >2x as large as the effect on friends, the 10-friend script still produces a net gain in expected votes.
    - The mobilizer-effect argument above also applies to any theoretical household spillover effect. We expect that adding more mobilizers will lead to a greater diversity of households being named, with more opportunities for spillover.
    - Finally, a given partner might value having more spread-out impact based on their particular political or community values; they should be aware that

if they ask for fewer friends per mobilizer, they can increase mobilizers in exchange for fewer total reminders.

Appendix: Robustness checks

Statistical models

We re-ran the tests above with several variations:

- With impact program data removed, so that we're only comparing the 3 A/B test conditions
- Using OLS instead of Poisson regression
- Using OLS and treating canvasser as a random instead of a fixed effect

These led to some variation in the magnitude of effects and significance tests. However, the 10-friend script consistently produced more relational contacts per hour than the 5-friend script, and fewer mobilizers and asks per hour. Results for the 7-friend script were unstable, sometimes similar to the 5-script and sometimes more similar to the 10-script.

Work hours

We were uncertain whether canvassers should be counted as working during hours that were within their working period, but had no asks reported. For example, if a canvasser's first ask was recorded at 9:15am and their last ask was at 11:45am, but there were *no* asks recorded between 10 and 11am, our main analysis treated them as not working during the hour of 10-11 (perhaps on break). In the robustness check, we instead treated 10-11 as an hour in which they reported zero asks, zero mobilizers, and zero friends.

The pattern and directionality of results was mostly unchanged, except that: 1) for friends per hour, the 7-script effect was nonsignificant instead of marginal; 2) for mobilizers per hour, the impact program effect was null instead of significant; 3) for asks per hour, the impact program effect was significant and negative, instead of null.

We also analyzed the data using full clock hours, instead of half-hour segments. The results were effectively unchanged.

Robustness check summary

In light of these alternate analyses, we are confident that the 5-friend and 10-friend scripts differ as described above. There is some uncertainty about how the other two scripts (7-friend and impact program) differ from the 5-friend script.

Appendix: Descriptives

These are the actual summary numbers from canvasing. **These are not generalizable outcomes,** because different conditions ran on different dates and with different canvassers.

	hours	friends / hour	mobilizers / hour	asks / hour	friends / mobilizer	conversion rate
Impact program (asked for 5)	3515	30.1	8.3	21.2	3.6	39%
5 friend ask	66.5	40.3	7.7	30.8	5.2	25%
7 friend ask	49.5	58.6	11.2	27.5	5.2	41%
10 friend ask	44.5	55.4	7.4	23.1	7.5	32%

Table 3: Actual summary numbers for each program, not adjusting for covariates

Appendix: Anomalous differences between the impact program and the 5-friend test

The impact program underperformed relative to the 5-friend script, even though the scripts were supposedly the same. As noted in Appendix 1: descriptives, the impact program got an average of 3.6 friends per mobilizer while the 5-friend script got 5.2, meaning that *the average mobilizer exceeded what canvassers were supposed to be asking for.*

A partial explanation is that the A/B test was run in the final days of early voting, meaning that canvassers were at their most experienced, and the poorest canvassers had been let go.

In addition I speculate that canvassers in the A/B test were influenced by having experienced all 3 scripts. On days when they had the 5-friend script, they still knew that we *sometimes* asked mobilizers to contact more than 5 friends, and this may have encouraged them to go "above and beyond" regardless of what the script told them to do. Based on this, I predict that future programs using a 5-friend ask will perform similarly to the impact program (3.6 friends per mobilizer), rather than the 5-friend A/B test. This would also mean that in normal usage, the 7-friend script would outperform the 5-friend script, even though they appeared similar in this study.

Appendix: Correlation coefficients

Below are actual correlation coefficients from the regression models reported in the results. Comparisons are with the 5-friend script in the A/B test. See the Significance testing section for notes on interpretation.

	friends per hour	mobilizers per hour	Asks per hour	Conversion rate
Standard program (asked for 5) relative to 5-friends A/B test	**-.19***	**+.16***	*-0.02*	**+11%***
A/B test: 7 friends relative to 5 friends	**+.10***	*-0.07*	**-.25***	**+8%***
A/B test: 10 friends relative to 5 friends	**+.32***	**-.17***	**-.22***	*-.02*

**=p<.05, ^=p<.1*