ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

Case Study: Get out the Vote

Do Phone Calls to Encourage Voting Work?
Why Randomize?

This case study is based on "Comparing Experimental and Matching Methods Using a Large-Scale Field Experiment on Voter Mobilization," by Kevin Arceneaux, Alan S. Gerber, and Donald P. Green, Political Analysis 14: 1-36.

J-PAL thanks the authors for allowing us to use their paper and for sharing their data

Key Vocabulary

Counterfactual: What would have happened to the participants in an intervention had they not received the intervention. The counterfactual cannot be observed from the treatment group; can only be inferred from the comparison group.

Comparison Group: A group that is meant to "represent" the counterfactual. In an experimental design, the comparison group (control group) is a randomly assigned group from the same population that is not intended to receive the intervention.

Impact: the true impact of the intervention is the difference in outcomes between the treatment group and its counterfactual. This is estimated by measuring the difference in outcomes between treatment and comparison groups.

Omitted Variable Bias: statistical bias that occurs when certain variables/characteristics (often unobservable), which are correlated with both the primary outcome and a variable of interest (e.g. participation in an intervention), are omitted from a regression analysis. Because these variables are not included as controls in the regression, one incorrectly attributes the measured impact solely to the program.

Selection Bias: a type of omitted variable bias in which individuals who participate in a program are systematically different from those who don't, and those differences are correlated with the outcome. This can occur when the treatment group is made of deliberately (non-randomly) chosen individuals (either self-selected, or selected by another).

Introduction

In late 2002, a non-partisan civic group, Vote 2002 Campaign, ran a get-out-the-vote initiative to encourage voting in that year's U.S. congressional elections. In the 7 days preceding the election, Vote 2002 placed 60,000 phone calls to potential voters, encouraging them to "come out and vote" on election day.

Did the program work? How can we estimate its impact?

Voter turnout was in decline since the 1960s

While voter turnout (the number of eligible voters that participate in an election) was declining since the 1960s, it was particularly low in the 1998 and 2000 U.S. elections. Only 47 percent of eligible voters voted in the 2000 congressional and presidential elections; the record low was 35 percent in the 1998 mid-term elections.

Vote 2002 get-out-the-vote Campaign

Facing the 2002 midterm election and fearing another low turnout, civic groups in Iowa and Michigan launched the Vote 2002 Campaign to boost voter turnout. In the week preceding the election, Vote 2002 volunteers placed phone calls to 60,000 voters and gave them the following message:

> *"Hello, may I speak with [Mrs. Ida Cook] please? Hi. This is [Carmen Campbell] calling from Vote 2002, a non-partisan effort working to encourage citizens to vote. We just wanted to remind you that elections are being held this Tuesday. The success of our democracy depends on whether we exercise our right to vote or not, so we hope you'll come out and vote this Tuesday. Can I count on you to vote next Tuesday?"*

As telemarketing replaces more traditional face-to-face campaigning, such as door-to-door canvassing, there is considerable debate over its effectiveness. Many believe the decline in voter turnout is a direct result of changing campaign practices. It is therefore worth asking in this context: did the Vote 2002 Campaign work? Did it increase voter turnout at the 2002 congressional elections?

Did the Vote 2002 Campaign work?

What is required in order for us to measure whether a program worked, whether it had impact?

In general, to ask if an intervention works is to ask if it achieves its goal of changing certain outcomes for its participants, and ensure that those changes are not caused by some other factors or events happening at the same time. To show that the intervention causes the observed changes, we need to simultaneously show that if it had not been implemented, the observed changes would not have occurred (or would be different). But how do we know what would have happened? If the intervention happened, it happened. Measuring what would have happened requires entering an imaginary world in which the intervention was never introduced to this group. The outcomes of this group in this imaginary world are referred to as the counterfactual. Since we cannot observe the true counterfactual, the best we can do is to estimate it by constructing ("mimicking") it.

The key challenge of impact evaluation is constructing the counterfactual. We typically do this by selecting a group of people that resemble the participants as much as possible but who did not participate in the intervention. This group is called the comparison group. Because we want to be able to say that it was the intervention and not some other factor that caused the changes in outcomes, it is important that the only difference between the comparison group and the participants is that the comparison group did not participate in the intervention. We then estimate "impact" as the difference in outcomes observed at the end of the intervention between the comparison group and the participants.

The impact estimate is only as accurate as the comparison group is successful at mimicking the counterfactual. If the comparison group poorly represents the counterfactual, the impact is poorly estimated. Therefore the method used to select the comparison group is a key decision in the design of any impact evaluation.

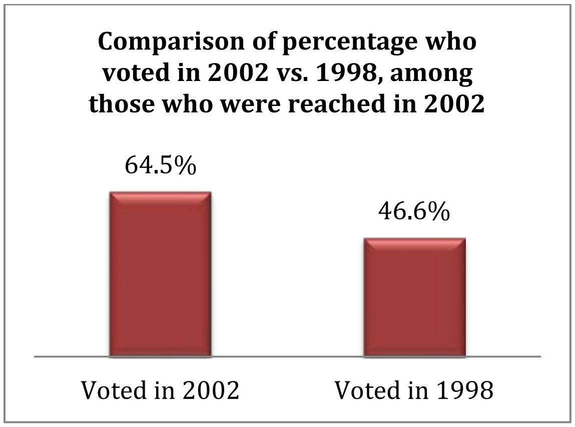
That brings us back to our questions: Did the Vote 2002 Campaign work? What was its impact on voter turnout?

Vote 2002 had access to a list of the telephone numbers of 60,000 people. They called all 60,000, but they were able to speak to only 25,000. For each call, they recorded whether or not the call was completed successfully. They also had census data on the voter's age, gender, household size, whether the voter was newly registered, which state and district the voter was from and data on how competitive the previous election was in that district, and whether the individual had voted in the past. Afterwards, from official voting records, they were able to determine whether, in the end, the voters they had called did actually go out and vote.

What comparison groups can we use? The following newspaper excerpts illustrate different methods of evaluating impact. (Refer to the table on the last page of the case for a list of different evaluation methods).

METHOD 1:

News Release: Vote 2002 Campaign is a huge success



"In 1998, during the last congressional elections, fewer than half of registered voters in Iowa and Michigan showed up on Election Day. This reflects national trends of declining voter turnout. The get-out-the-vote campaign was organized to reverse this trend. And was it ever successful! For the people we called, we saw an 18 percentage point increase in voter turnout."

DISCUSSION TOPIC 1

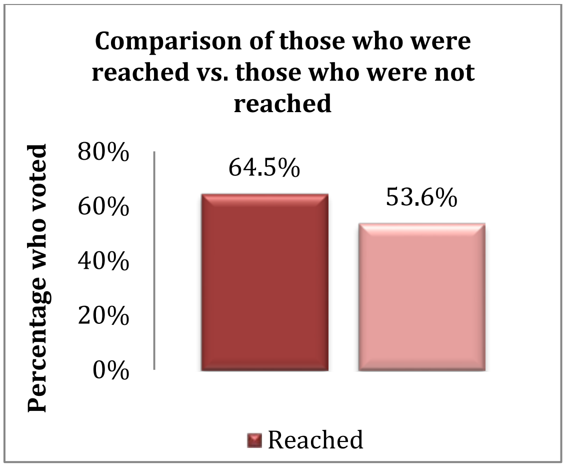
Identifying evaluation

1. What type of evaluation does this new release imply?
2. What represents the counterfactual?
3. What are the problems with this type of evaluation?

METHOD 2:

Opinion: Get-out-the-vote program - good but not great

In a recent news release, the Vote 2002 Campaign claimed to be able to increase voter turnout by nearly 20 percentage points. These estimates are significantly inflated. They are looking at the people they talked to, measuring changes in their rates of voting over time, and attributing the entire difference to their campaign. They are ignoring the possibility that these changes reflect increased political awareness in the country at large, perhaps the result of a declining economy, and escalating concerns over national security. If we compare people who were reached by the campaign's phone calls to those who weren't—both groups that were affected by these national events, and incidentally, both of whom reached the polls in greater numbers this time—we find that the actual impact of the program is 11 percentage points, rather than 18.



DISCUSSION TOPIC 2

Identifying evaluation

1. What type of evaluation does this new release imply?
2. What represents the counterfactual?
3. What are the problems with this type of evaluation?

METHOD 3:

Editorial:

If you haven't been paying close attention, you may have missed the public spat over the effectiveness of the Vote 2002, get-out-the-vote (GOTV) campaign. Campaign organizers claimed to have increased voter turnout by twenty percentage points. An opposing commentator wrote an opinion piece suggesting the impact is closer to half that. However, both analyses managed to get it wrong. The first is wrong in that it doesn't use a comparison group, and simply observes changes in voting patterns. The second uses the wrong metric to measure impact. Voting campaigns are meant to bring *new* voters to the polls, not simply talk to those who vote anyway. The opposing analyst compares the voter turnout among those who were reached with other people who were not reached. Many of those they called were already voting in the prior elections. The analysis should therefore measure *improvement* in voting rates—not the final level. This also helps control for the fact that these two groups had different voting rates in prior elections. When we repeated the analysis using the more-appropriate outcome measure, we find voting rates for those who were reached improved only marginally compared to those not reached (a 10.9 percentage point increase compared to 9 percentage point increase). This 1.9 percentage point difference is still statistically significant, but marginal relative to the other analyses.

Had these evaluators thought to look at the more appropriate outcome, they would recognize that the get-out-the-vote program is not only less successful than it reports, but less successful than even its detractors claim!

DISCUSSION TOPIC 3

Identifying evaluation

1. What type of evaluation does this new release imply?
2. What represents the counterfactual?
3. What are the problems with this type of evaluation?

METHOD 4: REGRESSION

Report: The numbers don't lie, unless your statisticians are asleep

Get-out-the-vote program celebrates victory, estimating a large percentage point improvement in voting rates. Others show almost no impact. A closer look shows that, the truth, as usual, is somewhere in between.

This report uses sophisticated statistical methods to measure the true impact of this campaign. We were concerned about other variables confounding previous results, such as age and household size. For example, it is entirely possible that senior citizens are more likely to vote and more likely to answer the phone. If the group that answered the phone is older on average, then we may expect them to vote at higher rates than those who didn't answer the phone. Indeed, those who answered the phone were on average 56 years old, while those who didn't were 51.

To observe the possible bias caused by omitting key variables, we conducted one analysis without controlling for these differences, and one with controls. This also allowed us to obtain the true impact of the campaign.

Dependent Variable: Voted in 2002

	Reached vs. Not-Reached		Reached vs. Not-Reached	
Reached	**0.1085**	*	**0.0462**	*
	(0.0041)		(0.0035)	
Age			**0.0026**	*
			(0.0001)	
Household Size			**0.0634**	*
			(0.0035)	
Female			**-0.0091**	
			(0.0035)	
Newly registered			**0.0729**	*
			(0.0065)	
From Iowa			**-0.0564**	*
			(0.0037)	
In a competitive district			**0.0334**	*
			(0.0034)	
Voted in 2000			**0.3941**	*
			(0.0041)	
Voted in 1998			**0.2134**	*
			(0.0041)	
Constant	**0.5364**		**-0.0158**	
	(0.0026)		(0.0087)	
Observations	59,972		59,972	

Looking at the above table, we find that the estimate falls by almost 6 percentage points when we control for the appropriate characteristics, showing that most of the change in outcome is being driven by all these other differences between the two groups. This suggests that for every 60 people who were called, and every 25 people who answered the phone, roughly one more person voted. At first glance, that may not appear impressive. But the other way to look at it is: the entire campaign convinced nearly 1,150 more voters to vote. As we saw in the last election, that is more than enough to tip the balance one direction or the other.

DISCUSSION TOPIC 4
Identifying evaluation

1. What type of evaluation does this new release imply?
2. What represents the counterfactual?
3. What are the problems with this type of evaluation?

METHOD 5
Report:

Ronald Coase, a Nobel Prize winning economist, once said: "If you torture the data long enough, it will confess [to anything]." We just witnessed this kind of torture. Analysts of the Vote 2002 Campaign said they were "concerned about other variables confounding previous results, such as age, and household size," and claim that by using a multivariate regression, they are "controlling for" characteristics that make the two groups different, thereby "obtaining the true impact of the campaign". However, there is one critical characteristic that makes the two groups observably different. One group answered the phone, and the other didn't. This is a classic case of selection bias. So no matter how many other variables we control for, as long as we can't fully account for why one group answered and the other didn't (and that unexplained difference is correlated with voting), regression analysis simply cannot remove this selection bias.

Therefore, we suggest another way to estimate the impact of this campaign. We construct a comparison group, not from the set of non-respondents (who didn't answer the phone), but a subset from a larger population who look similar to the people who were called and reached. We have data on two million eligible voters in these states. For each of the 25,000 individuals reached, we find a corresponding individual in the larger population who is identical among all characteristics (i.e., age, gender, location, past voting behavior, etc.). We can then construct a "statistically identical" comparison group with exactly the same observable characteristics.

Using this deliberately constructed comparison group, without any fancy regressions, we find that the group Vote 2002 reached ended up voting at a rate of 65.9%, while the comparison group had a 63.2% voting rate, suggesting an impact of 2.7 percentage points.

DISCUSSION TOPIC 5

Identifying evaluation

1. What type of evaluation does this new release imply?
2. What represents the counterfactual?
3. What are the problems with this type of evaluation?

METHOD 6:

Using randomized experiments

It turns out that from the larger population of about 2 million potential voters, the 60,000 individuals were *randomly* selected. Under the final method, the group that was called (whether reached or not reached) is now called the treatment group and the rest is the comparison group.

Comparing all six methods

Below are the impact estimates of the Vote 2002 Campaign using the six different methods you have discussed in this case study.

Table 1: Comparing all six methods

Method	***Estimated impact***
Pre-Post	17.9 pp*
Simple Difference	10.8 pp*
Difference-in-Differences	1.9 pp*
Multivariate Regression with Panel Data	4.6 pp*
Matching (All Covariates)	2.8 pp*
Randomized Evaluation‡	0.4 pp

NOTES: pp means "percentage points" and * indicates statistically significant at the 5% level

‡ Randomized evaluation estimate is adjusted to reflect that only 25,000 of 60,000 in the treatment were treated (i.e. the Treatment on Treated effect)

As you can see, not all methods give the same result. Hence, the choice of the appropriate method is crucial. The purpose of this case study was not to evaluate one particular voter mobilization campaign, but to evaluate evaluation methods in this particular context.

In the analysis of the Vote 2002 Campaign, we found that people who happened to pick up the phone were more likely to vote in the upcoming (and previous) elections. Even though we statistically accounted for some observable characteristics, including demographics and past voting behavior, there were still some inherent, unobservable differences between the two groups, independent of the get-out-the-vote campaign. Therefore, when our non-randomized methods demonstrated a positive, significant impact, this result was due to "selection bias" (in this case, selection of those who pick up the phone) rather than a successful get-out-the-vote campaign.

	Methodology	Description	Who is in the comparison group?	Required Assumptions	Required Data
Quasi-Experimental Methods	**Pre-Post**	Measure how program participants improved (or changed) over time.	Program participants themselves—before participating in the program.	The program was the only factor influencing any changes in the measured outcome over time.	Before and after data for program participants.
	Simple Difference	Measure difference between program participants and non-participants after the program is completed.	Individuals who didn't participate in the program (for any reason), but for whom data were collected after the program.	Non-participants are identical to participants except for program participation, and were equally likely to enter program before it started.	After data for program participants and non-participants.
	Differences in Differences	Measure improvement (change) over time of program participants *relative to* the improvement (change) of non-participants.	Individuals who didn't participate in the program (for any reason), but for whom data were collected both before and after the program.	If the program didn't exist, the two groups would have had identical trajectories over this period.	Before and after data for both participants and non-participants.
	Multivariate Regression	Individuals who received treatment are compared with those who did not, and other factors that might explain differences in the outcomes are "controlled" for.	Individuals who didn't participate in the program (for any reason), but for whom data were collected both before and after the program. In this case data is not comprised of just indicators of outcomes, but other "explanatory" variables as well.	The factors that were *excluded* (because they are unobservable and/or have been not been measured) do not bias results because they are either uncorrelated with the outcome or do not differ between participants and non-participants.	Outcomes as well as "control variables" for both participants and non-participants.
	Statistical Matching	Individuals in control group are compared to similar individuals in experimental group.	Exact matching: For each participant, at least one non-participant who is identical *on selected characteristics.* Propensity score matching: non-participants who have a mix of characteristics which predict that they would be as likely to participate as participants.	The factors that were *excluded* (because they are unobservable and/or have been not been measured) do not bias results because they are either uncorrelated with the outcome or do not differ between participants and non-participants.	Outcomes as well as "variables for matching" for both participants and non-participants.
	Regression Discontinuity Design	Individuals are ranked based on specific, measureable criteria. There is some cutoff that determines whether an individual is eligible to participate. Participants are then compared to non-participants and the eligibility criterion is controlled for.	Individuals who are close to the cutoff, but fall on the "wrong" side of that cutoff, and therefore do not get the program.	After controlling for the criteria (and other measures of choice), the remaining differences between individuals directly below and directly above the cut-off score are not statistically significant and will not bias the results. A necessary but sufficient requirement for this to hold is that the cut-off criteria are strictly adhered to.	Outcomes as well as measures on criteria (and any other controls).
	Instrumental Variables	Participation can be predicted by an incidental (almost random) factor, or "instrumental" variable, that is uncorrelated with the outcome, other than the fact that it predicts participation (and participation affects the outcome).	Individuals who, because of this close to random factor, are predicted not to participate and (possibly as a result) did not participate.	If it weren't for the instrumental variable's ability to predict participation, this "instrument" would otherwise have no effect on or be uncorrelated with the outcome.	Outcomes, the "instrument," and other control variables.
Experimental Method	**Randomized Evaluation**	Experimental method for measuring a causal relationship between two variables.	Participants are randomly assigned to the control groups.	Randomization "worked." That is, the two groups are statistically identical (on observed and unobserved factors).	Outcome data for control and experimental groups. Control variables can help absorb variance and improve "power".