ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

Massachusetts Institute of Technology

Courtesy of Flickr user theocean

Case 1: 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

> The 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 has been in decline since the 1960s

While voter turnout (the number of eligible voters that participate in an election) has been 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. The campaign employed telemarketing techniques commonly used in modern elections. In the week preceding the election, Vote 2002 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 telephone campaigns replace many of the more traditional face-to-face interventions, there is considerable debate over their effectiveness. Many believe the decline in voter turnout is directly related to the reduction in more personal methods of campaigning. 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 a program works is to ask if the program achieves its goal of *changing certain outcomes* for its participants. To say, validly, that a program changes outcomes, we need to establish three things: (1) that outcomes have changed; (2) that the observed changes occurred among participants of the program and did not occur among non-participants; and (3) that it is not something else, some other event happening at the same time as the program, that drove the observed changes. In other words, we need to show that the program *causes* the observed changes.

To show that the program causes the changes, we need to simultaneously show that if the program had not been implemented, the observed changes would not have happened. What is called the "counterfactual" is the imaginary state of the world that program participants would have experienced if they had not participated in the program. It does not represent the state in which would-be participants receive absolutely no services, but rather the state of the world in which life goes on as before, the participants receive whatever services they would have received had they not participated in the program; it represents life without participating in the program.

The impact of the program, then, is the difference between the observed outcomes and what those outcomes would have been in the absence of the program, under the counterfactual. Thus we need to know the counterfactual to determine impact. But the fact is the program was implemented; we can never observe the counterfactual. Because we cannot directly observe the true counterfactual, we cannot actually determine impact. The best we can do is to estimate it, and we do so by *mimicking* the counterfactual.

The key challenge of program impact evaluation is constructing or mimicking 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 program. This group is called the comparison group. Because we want to be able to say that it was the program and not some other factor that caused the changes in outcomes—condition (3) above—we want to be able to say that the only difference between the comparison group and the participants is that the comparison group did not participate in the program. We then estimate "impact" as the difference observed at the end of the program between the outcomes of the comparison group and the outcomes of the program 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 (in most circumstances) 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?

In this case, the targeted behavior is to "get out and vote," and the outcome measure is voter turnout. So, when we ask if the Vote 2002 Campaign worked, we are asking if it increased voter turnout in the 2002 congressional elections. The impact is the difference between voter turnout on that Tuesday in 2002 and what voter turnout would have been if Vote 2002 had never existed.

What comparison groups can we use?

Estimating the impact of the Vote 2002 Campaign

Your team is doing pro-bono consulting for Vote 2002. Your task is to estimate the impact of the Vote 2002 Campaign. 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.

There are a number of methods available to your team to estimate the impact. In this case, we will compare their validity and identify the circumstances under which a given method can be used or not.

Method 1: Using a simple difference

Discussion Topic 1: Using simple differences: comparing voter turnout between the "reached" and "not reached"

Method 1: Comparing voter turnout between reached and not reached. Assume the 25,000 households who received the full message constitute the participant group and the 35,000 households who were called but not reached represent the comparison group. If you want to see what the impact of receiving a call has on voter turnout, you could check whether those who were reached were more likely to vote than those who were not reached. Estimate impact by comparing the proportion of people who voted in the treatment group and that of the comparison group, as shown in the following table:

	Voter turnout by group		***Impact Estimate***
	Reached	Not reached	
Method1: Simple difference	64.5%	53.6%	10.8 pp*

Discuss whether this method gives you an accurate estimate of the effect of the program. What might be the possible sources of biases? In other words, what is likely to make the comparison group a poor approximation of the true counterfactual?

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

Method 2: Using multivariate regression to control for inherent differences

Discussion Topic 2: Using multivariate regression

You were concerned that people reached might have different inherent characteristics from those who were not reached. Indeed, when you compare the two groups, you observe significant differences:

Characteristics of Reached and Not-Reached Groups

	Reached	***Not Reached***	***Difference***
Household Size	1.56	1.50	0.06
Average age	55.8	51.0	4.8
Percent female	56.2%	53.8%	2.4 pp*
Percent newly registered	7.3%	9.6%	-2.3 pp*
Percent from a competitive district	50.3%	49.8%	0.5 pp
Percent from Iowa	54.7%	46.7%	8.0 pp*
Sample Size	25,043	34,929	

1. Can you overcome the problems of method 1 by taking a random sample from the participant group and a random sample from the comparison group?

Method 2: Using multivariate regression to control for differences between reached and not-reached.
You could control for these differences by using a multivariate regression as follows: The participant and comparison groups are defined in the same way as in method 1. To estimate the impact of the program, you run a regression where the "dependent variable" is a zero/one variable indicating whether the person voted or not (i.e., 0 = did not vote, 1 = voted). The "key explanatory variable" is a zero/one variable indicating whether the person received the call or not (i.e., 0 = did not receive the call, 1 = received a call). Potential differences in characteristics can be controlled for using other "explanatory variables" such as age, gender, newly registered voter, etc. The coefficient on the key explanatory variable (i.e., received the call) represents the "controlled" estimated impact of the program.

Using multivariate regression to control for the characteristics shown in the table below, you estimate the impact to be 6.1 pp (percentage points), significant at the 5% level.

2. Why do you think the estimated impact using method 2 is lower than the 10.8 pp impact you estimated using method 1?

3. For method 2, discuss whether it is reasonable to expect that the estimated impact represents the true causal effect of Vote 2002 on voter participation. What remaining biases could there be?

4. Using the data described above, can you think of more convincing methods to estimate the impact of the Vote 2002 Campaign?

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

Method 3: Using panel data—tracking the same people over time

You are still concerned about differences in characteristics between the reached and non-reached. You decide to use panel data, that is, track the same person over time.

Discussion Topic 3: Using panel data

Method 3: Using panel data to track the same people over time. It turns out that staff members of Vote 2002 also had data on whether the person voted in the previous elections (1998 and 2000). Past voting behavior is thought to be a strong predictor of future voting behavior. The table below indicates past voting behavior for the group of people who were reached by the Vote 2002 Campaign and the group of people who were called but not reached.

Voter turnout in 1998 and 2000 elections between the reached and not-reached

	2002 Reached	***2002 Not Reached***	***Difference***
Voted in 2000	71.7%	63.3%	8.3 pp*
Voted in 1998	46.6%	37.6%	9.0 pp*

1. How can these data on past voting behavior be used to improve your analysis?

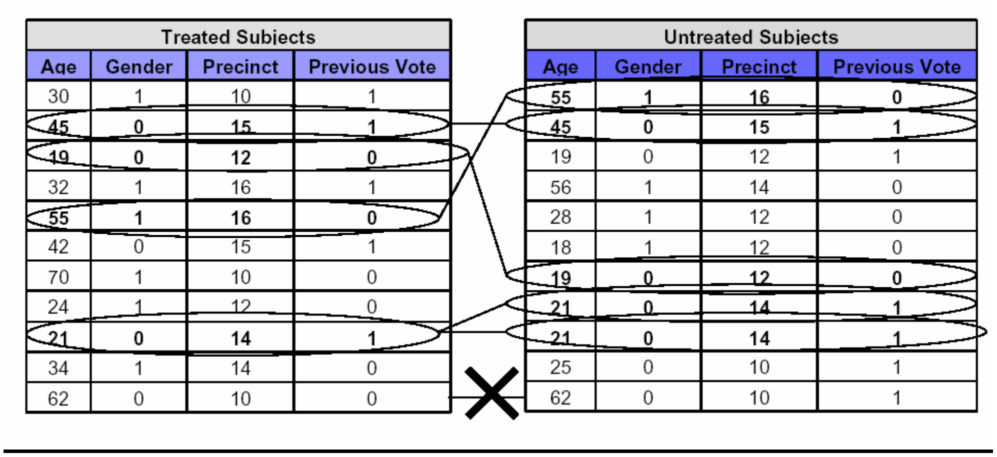
2. Given the information in the above table, would you expect that controlling for past voting behavior in method 2 would result in a higher or lower estimate of the impact of the Vote 2002 Campaign on voter turnout than the 6.1 pp found without controlling for it?

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

Method 4: Using matching

One way to estimate the impact of the Vote 2002 Campaign is to select as a comparison group a subset of non-participants who look similar to the participant group (the 25,000 called and reached). To select this subset, researchers often employ a statistical procedure called *matching*. While there are many ways to do matching, it turns out that in this context it is possible to do *exact matching* for almost all the individuals in the sample. The lists from which these 60,000 individuals were selected and tracked include data on another 2 million eligible voters. Therefore, for each of the 25,000 individuals reached, we can select another individual who has the exact same characteristics (i.e., age, gender, etc.). In this way, the participant and comparison groups will have exactly the same observable characteristics. Figure 1 shows exact matching.

Figure 1: Exact Matching



Treated Subjects

Age	Gender	Precinct	Previous Vote
30	1	10	1
45	**0**	**15**	**1**
19	**0**	**12**	**0**
32	1	16	1
55	**1**	**16**	**0**
42	0	15	1
70	1	10	0
24	1	12	0
21	**0**	**14**	**1**
34	1	14	0
62	0	10	0

Untreated Subjects

Age	Gender	Precinct	Previous Vote
55	**1**	**16**	**0**
45	**0**	**15**	**1**
19	0	12	1
56	1	14	0
28	1	12	0
18	1	12	0
19	**0**	**12**	**0**
21	**0**	**14**	**1**
21	**0**	**14**	**1**
25	0	10	1
62	0	10	1

Source: Arceneaux, Gerber, and Green (2004)

Discussion Topic 4: Exact Matching

Method 4: Matching. Matching was performed and then the impact of the Vote 2002 Campaign was estimated by taking the difference between the voter turnout rate in the participant group and the voter turnout rate in the comparison group created through matching (the "matched" group). The results are shown in the table.

Matching Analysis

Number of Covariates matched on:	Subset of Matched Reached	Subset of Matched Not-Reached Individuals	*Impact*
4 (HH size, age, newly registered, state)	64.5%	60.8%	3.7 pp*
6 (HH size, age, newly registered, state in a competitive district, voted in 2000)	64.5%	61.5%	3.0 pp*
All	65.9%	63.2%	2.8 pp*

1. Assess whether it is reasonable to expect that the impact estimated using this method represents the true causal effect of Vote 2002 on voter participation.

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

a. All: household size, age, newly registered, county, state senate district, state house district, from a competitive district, voted in 2000, voted in 1998. Using all covariates, only 90% of the reached-individuals had exact matches in the comparison group.

Method 5: 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.

Discussion Topic 5: Randomized Experiment

Method 5: Randomized Experiment. You can exploit this randomization to estimate the impact of the Vote 2002 Campaign. The idea is that the 60,000 individuals Vote 2002 called (now called the treatment group) should be statistically identical to the 2,000,000 individuals (called the control group) in everything (observable and unobservable) except for the fact that the first group was called by the Vote 2002 Campaign.

Compares the treatment and control groups on observable characteristics

	Treatment	Control	*Difference*
Voted in 2000	56.7%	56.4%	0.4 pp
Voted in 1998	22.7%	23.1%	-0.5 pp
Household Size	1.50	1.50	0.0
Average age	52.0	52.2	-0.2
% Female	54.6%	55.2%	-0.6 pp
% Newly registered	11.6%	11.7%	0.0 pp
Total people in group	14,972	1,153,072	

1. Notice that the two groups look very similar. Is this what you would expect?

Comparing voter turnout in the experimental treatment and the control groups

	Treatment (60,000 called)	**Control (2M not called)**	***Impact***
Simple Difference	58.2%	58.0%	0.2 pp
Difference after controlling for observable characteristics (multivariate regression)			0.2 pp

2. Notice that the impact estimates are not statistically significant. This result is different than those obtained with the previous methods. How do you explain this difference in results?

3. In the above analysis, we compare the 60,000 *who were called* to the 2,000,000 not called by the Vote 2002 Campaign. Why don't we compare just the 25,000 who were *reached* to the same control group?

Adjusting estimate to remove "dilution" of impact from those not reached

	Impact
Difference after adjusting for the fact that only 25,000 of 60,000 in the treatment group were reached ("Treatment Effect on the Treated")*	0.4 pp

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

* This corresponds to an instrumental variable regression that estimates the effect of the treatment "on the treated."

Comparing all five methods

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

Table 1: Comparing all five methods

Method	*Estimated impact*
Simple Difference	10.8 pp*
Multivariate Regression	6.1 pp*
Multivariate Regression with Panel Data	4.5 pp*
Matching (All Covariates)	2.8 pp*
Randomized experiment with adjustment to reflect that only 25,000 of 60,000 in the treatment were treated	0.4 pp

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

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.

Discussion Topic 6: Selection bias

Selection bias is a problem that arises in many program evaluations. Think about some of the non-randomized development programs you have, or have seen, evaluated. Discuss how the participant group was selected, and how "selection" may have affected the ability to estimate the true impact of the program.

References:

Gerber, Alan and Donald Green, 2000. "The Effects of Canvassing, Telephone calls, and Direct mail on Voter Turnout: A Field Experiment" *American Political Science Review* 94 (3): 653-663

Arceneaux, Kevin, Alan Gerber, and Donald Green 2004. "Comparing Experimental and Matching Methods using a Large-Scale Field Experiment on Voter Mobilization" *Preliminary Draft*