J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

CASE STUDY 2: GET OUT THE VOTE

Why Randomize?

This case study is based on:
Arceneaux, Kevin, Alan S. Gerber, and Donald P. Green. 2006. "Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment." *Political Analysis* 14 (1): 37-62. https://doi.org/10.1093/pan/mpj001.

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

KEY VOCABULARY	
Comparison Group:	A group that is as similar as possible to the treatment group in order to be able to learn about the counterfactual. In an experimental design, the comparison group (also called the control group) is a group from the same population as the treatment group that, by random assignment, is not intended to receive the intervention.
Counterfactual:	What would have happened to the participants of an intervention had they not received the intervention. The counterfactual can never be observed; it can only be inferred from a comparison group different from the treatment group.
Estimate:	In statistics, a "best guess" about an unknown value in a population (such as the effect of a program on an outcome) according to a rule (known as the "estimator") and the values observed in a sample drawn from that population.
Impact:	The impact of the intervention is the effect of the treatment on the whole population. The impact is estimated by measuring the differences in outcomes between the treatment group and its counterfactual, i.e., by measuring the difference in outcomes between treatment and comparison groups.
Omitted Variable Bias:	Statistical bias that occurs when relevant (and often unobservable) variables/characteristics are left out of the regression analysis. When these variables are correlated with both the primary outcome and a variable of interest (e.g., participation in an intervention), their omission can lead to incorrectly attributing the measured impact solely to the program. For example, omitting socioeconomic status, which is correlated with test scores, could lead to overestimating the impact of a tutoring intervention on a group of wealthy students.
Treatment Group:	The group that receives the intervention.
Selection Bias:	Selection bias is bias that occurs when the individuals who receive the program are systematically different from those who do not. For example, consider an elective after-school tutoring program. Is it effective at raising children's exam scores? If we compare those who take up the tutoring program to those who don't, we will get a biased estimate of the effect of the tutoring program, because those who chose to participate are likely different from those who don't (for example, those who took it up may be more motivated, or they may be weaker students). Randomization removes selection bias because it breaks the link between characteristics of the individual and their treatment status. Selection bias can occur in other ways in a randomized evaluation. For example, consider a situation where an intervention is making a phone call to a landline: - Callers may be unable to reach certain participants (for example, participants in rural areas may have poor cell phone service and may be more likely to have landlines than those in urban areas). - Some participants may be less likely to pick up the phone depending on the time of day they are called (for example, calling a home phone during standard business hours).

LEARNING OBJECTIVES

To identify evaluation methods and how they estimate impact differently. To better understand issues of bias and causal inference and think through how to use evaluation methods to measure impact.

SUBJECTS COVERED

Causality, counterfactual, impact, comparison groups, selection bias, omitted variables, randomization, and balance.

	Method	Description	What assumptions are required, and how demanding are the assumptions?	Required data
Randomization	Randomized Evaluation/ Randomized Control Trial	Measure the differences in outcomes between randomly assigned program participants and non-participants after the program took effect.	*The outcome variable is only affected by program participation itself, not by assignment to participate in the program or by participation in the randomized evaluation itself.* Examples for such confounding effects could be information effects, spillovers, or experimenter effects. As with other methods, the sample size needs to be large enough so that the two groups are statistically comparable; the difference being that the sample size is chosen as part of the research design.	Outcome data for randomly assigned participants and non-participants (the treatment and control groups).
Basic Non-Experimental Comparison Methods	Pre-Post	Measure the differences in outcomes for program participants before the program and after the program took effect.	*There are no other factors (including outside events, a drive to change by the participants themselves, altered economic conditions, etc.) that changed the measured outcome for participants over time besides the program*. In stable, static environments and over short time horizons, the assumption might hold, but it is not possible to verify that. Generally, a diff-in-diff or RDD design is preferred (see below).	Data on outcomes of interest for program participants before program start and after the program took effect.
	Simple Difference	Measure the differences in outcomes between program participants after the program took effect and another group who did not participate in the program.	*There are no differences in the outcomes of participants and non-participants except for program participation*, and both groups were equally likely to enter the program before it started. This is a demanding assumption. Non-participants may not fulfill the eligibility criteria, live in a different location, or simply see less value in the program (self-selection). Any such factors may be associated with differences in outcomes independent of program participation. Generally, a diff-in-diff or RDD design is preferred (see below).	Outcome data for program participants as well as another group of non-participants after the program took effect.
	Differences in Differences	Measure the differences in outcomes for program participants before and after the program *relative* to non-participants.	*Any other factors that may have affected the measured outcome over time are the same for participants and non-participants, so they would have had the same time trajectory absent the program*. Over short time horizons and with reasonably similar groups, this assumption may be plausible. A "placebo test" can also compare the time trends in the two groups before the program took place. However, as with "simple difference," many factors that are associated with program participation may also be associated with outcome changes over time. For example, a person who expects a large improvement in the near future may not join the program (self-selection).	Data on outcomes of interest for program participants as well as another group of non-participants before program start and after the program took effect.

	Method	Description	What assumptions are required, and how demanding are the assumptions?	Required data
More advanced statistical non-experimental methods	Multivariate Regression/OLS	The "simple difference" approach can be—and in practice almost always is—carried out using multivariate regression. Doing so allows accounting for other observable factors that might also affect the outcome, often called "control variables" or "covariates." The regression filters out the effects of these covariates and measures differences in outcomes between participants and non-participants while holding the effect of the covariates constant.	Besides the effects of the control variables, *there are no other differences between participants and non-participants that affect the measured outcome*. This means that any unobservable or unmeasured factors that do affect the outcome must be the same for participants and non-participants. In addition, the control variables cannot in any way themselves be affected by the program. While the addition of covariates can alleviate some concerns with taking simple differences, limited available data in practice and unobservable factors mean that the method has similar issues as simple difference (e.g., self-selection).	Outcome data for program participants as well as another group of non-participants, as well as "control variables" for both groups.
	Statistical Matching	Exact matching: participants are matched to non-participants who are identical based on "matching variables" to measure differences in outcomes. Propensity score matching uses the control variables to predict a person's likelihood to participate and uses this predicted likelihood as the matching variable.	Similar to multivariable regression: *there are no differences between participants and non-participants with the same matching variables that affect the measured outcome*. Unobservable differences are the main concern in exact matching. In propensity score matching, two individuals with the same score may be very different even along observable dimensions. Thus, the assumptions that need to hold in order to draw valid conclusions are quite demanding.	Outcome data for program participants as well as another group of non-participants, as well as "matching variables" for both groups.
	Regression Discontinuity Design (RDD)	In an RDD design, eligibility to participate is determined by a cutoff value in some order or ranking, such as income level. Participants on one side of the cutoff are compared to non-participants on the other side, and the eligibility criterion is included as a control variable (see above).	*Any difference between individuals below and above the cutoff (participants and non-participants) vanishes closer and closer to the cutoff point.* A carefully considered regression discontinuity design can be effective. The design uses the "random" element that is introduced when two individuals who are similar to each other according to their ordering end up on different sides of the cutoff point. The design accounts for the continual differences between them using control variables. The assumption that these individuals are similar to each other can be tested with observables in the data. However, the design limits the comparability of participants further away from the cutoff.	Outcome data for program participants and non-participants, as well as the "ordering variable" (also called "forcing variable").
	Instrumental Variables	The design uses an "instrumental variable" that is a predictor for program participation. The method then compares individuals according to their predicted participation, rather than actual participation.	*The instrumental variable has no direct effect on the outcome variable. Its only effect is through an individual's participation in the program*. A valid instrumental variable design requires an instrument that has no relationship with the outcome variable. The challenge is that most factors that affect participation in a program for otherwise similar individuals are also in some way directly related to the outcome variable. With more than one instrument, the assumption can be tested.	Outcome data for program participants and non-participants, as well as an "instrumental variable".

INTRODUCTION

What is required in order for us to measure whether a program had impact and, if so, how much of an impact?

This is the same as asking whether changes in certain outcomes can be attributed directly to the intervention, which in turn requires ensuring that these measured outcome changes are not caused by other factors or events happening at the same time. Ideally, evaluators would do this by following the progress of a group of people as they participate in a program, measure any changes that occur, and then go back in time and measure the same group's progress without the program in place. This second set of outcomes is called the **counterfactual**. Since we cannot observe the true counterfactual, the best we can do is to approximate it by constructing (or mimicking) it.

The key challenge of an impact evaluation is constructing the counterfactual. We typically do this by selecting a group of people who 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 comparison group and the participant group are, on average, as similar as possible so that we can attribute any differences in outcomes to the intervention. We then estimate the **impact** as the difference in outcomes observed at the end of the intervention between the comparison group and the **treatment group**.

An accurate impact estimate can only be attained if the comparison group is a good representation of the counterfactual, or what the treatment group would have looked like had the intervention not happened. If the comparison group poorly represents the counterfactual, then the estimated impact will be **biased**. Therefore, the method used to select, construct, or estimate the comparison group is a key decision in the design of any impact evaluation.

This case study will explore different methods for measuring impact by looking at the Vote 2002 campaign. Using the same data, we will show how different methods may produce different results.

VOTE 2002 CAMPAIGN

While voter turnout (the number of eligible voters that participate in an election) has been in decline since the 1960s, it was particularly low in the 1998 and 2000 elections in the United States. 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.

In late 2002, a non-partisan civic group, the Vote 2002 campaign, ran a get-out-the-vote initiative in Iowa and Michigan to encourage voting in that year's U.S. congressional elections. As telemarketing campaigns such as Vote 2002 replace more traditional face-to-face campaigning, such as door-to-door canvassing,

there is considerable debate over their effectiveness. Many believe the decline in voter turnout is a direct result of changing campaign practices. Therefore, in this context it is worth asking: Did the Vote 2002 campaign work? In other words, did calling potential voters increase voter turnout in the 2002 congressional elections?

IMPACT OF VOTE 2002 CAMPAIGN

In the week preceding the election, Vote 2002 placed phone calls to 60,000 potential voters and gave them the following message:

> *"Hello, may I speak with [Mrs. Ida Cook], please? 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?"*

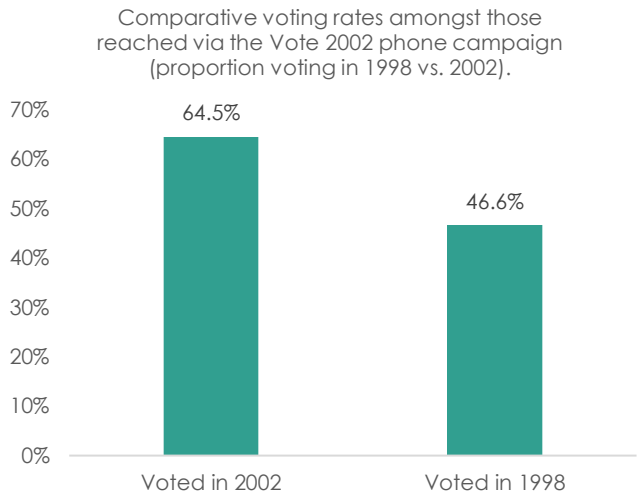
While Vote 2002 called all 60,000 people, they were able to speak to only 25,000. For each call, they recorded whether or not the call was completed successfully. They also had census data on the voter's age, gender, household size, whether the voter was newly registered, which state and district the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

Did the campaign work? The following newspaper excerpts illustrate different methods of evaluating impact to answer this question. (Refer to the previous table for a list of different evaluation methods.)

ESTIMATING THE IMPACT OF THE GET OUT THE VOTE PROJECT

METHOD 1

News Release: Vote 2002 campaign is a huge success



In 1998, during the last congressional elections, fewer than half of Iowa and Michigan's registered voters 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

1. What type of evaluation method does this news release imply?

2. What represents the counterfactual?

3. What are some potential problems with this type of evaluation method? List some confounding factors.

METHOD 2

Opinion: Get Out the Vote program—good but not great

In a recent news release, the Vote 2002 campaign claimed to increase voter turnout by nearly 20 percentage points. These estimates are significantly inflated. Why is this? They are looking at the people they talked to, measuring changes in their rates of voting over time, and then attributing the entire difference to their campaign. They are ignoring the possibility that these observed changes might be driven by factors outside the intervention. For example, they may 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 called and picked up the phone to those who were called but did *not* pick up the phone—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 percentage points.

DISCUSSION TOPIC 2

1. What type of evaluation method is this opinion piece employing?

2. What represents the counterfactual?

3. What are some potential problems with this type of evaluation method? List some confounding factors.

METHOD 3

Editorial:

If you haven't been paying close attention, you may have missed the public spat over the effectiveness of Vote 2002, the Get Out the Vote campaign. Campaign organizers claim to have increased voter turnout by 20 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 voter turnout among those who were reached with that of people who were not reached. Many of those they called were already voting in prior elections. The analysis should therefore measure *improvement* in voting rates—not the final voting level. This also helps control for the fact that these two groups (voters that did and did not answer the phone when called by the campaign) had different voting rates in prior elections. When we repeat the analysis using the more appropriate outcome measure, we find voting rates for those who picked up improved only marginally compared to those who did not pick up (10.9 percentage point increase for those who picked up versus 9 percentage point increase for those who did not pick up). This 1.9 percentage point difference is still statistically significant but marginally relative to the other analyses. This suggests that the pre-post analysis greatly overestimated the Vote 2002 impact.

Had these evaluators taken a more rigorous approach, including measuring a more appropriate outcome, they would recognize that the Get Out the Vote program is not only less successful than reported, but less successful than even its detractors claim!

DISCUSSION TOPIC 3

1. What type of evaluation method is this letter using?

2. What represents the counterfactual?

3. What are some potential problems with this type of evaluation method? List some confounding factors.

METHOD 4

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 statistical methods to measure the true impact of this campaign. We were concerned about other variables, such as age and household size, confounding previous results. 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 5 years older than those who didn't (56 and 51 years old, respectively). To observe the possible bias caused by omitting key variables, we conducted one "naïve" analysis without controlling for these differences, and a separate analysis with controls. This helped us to obtain the true impact of the campaign.

	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

DISCUSSION TOPIC 4

1. What type of evaluation method is utilized in this report?

2. What represents the counterfactual?

3. What are some potential problems with this type of evaluation method? List some confounding factors.

METHOD 5:

Report: Matching Individuals

The 60,000 individuals who were called by Vote 2002 were actually randomly selected to be called. They were chosen from a larger population of about 2 million potential voters. We can split this population of 2 million potential voters into the treatment group (60,000 individuals selected to be called) and the control group (the rest of the population).

Of the 60,000 people called by the campaign, only 25,000 people picked up the phone and listened to the full message. As such, we estimate the intention-to-treat (ITT) effect, which measures the impact of the program. Controlling for state and competitiveness, we find that the treatment group was 0.4 percentage points more likely to vote, a figure that when taking into account the standard error (0.5) is statistically indistinguishable from 0. The addition of other controls increases this effect to 0.5 percentage points with a standard error of 0.4. In other words, randomly calling voters has no statistically significant effect on inducing them to vote.

	Reached vs. Not-
Reached	0.004
	(0.005)
Constant	0.461
	(0.01)
Observations	1,905,320

DISCUSSION TOPIC 5

1. What type of evaluation method is used in this report?

2. What represents the counterfactual?

3. What are some potential problems with this type of evaluation method? List some confounding factors.

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**
Pre-Post	17.9 pp**
Simple Difference	10.8 pp**
Difference-in-Differences	1.9 pp**
Multivariate Regression with Panel Data	4.6 pp**
Randomized Evaluation‡	0.4 pp

NOTE: pp denotes "percentage points" and ** indicates statistically significant at the 5 percent level

As you can see, not all methods yield the same result. Hence, the choice of method is crucial.

The purpose of this case study was not to evaluate one particular voter mobilization campaign, but to compare evaluation methods in this particular context. In the analysis of the Vote 2002 campaign, we found that people who picked up the phone were more likely to vote in the upcoming (and previous) elections. Even when we controlled for some observable characteristics, including demographics and past voting behavior, there were still some unobservable and systematic differences between the two groups, independent of the Get Out the Vote campaign. Therefore, while our non-randomized methods demonstrated a positive, significant impact, this result was due to "selection bias" (in this case, inherent differences between those who pick up the phone versus those who do not) rather than a successful Get Out the Vote campaign.

REUSE AND CITATIONS

To request permission to access the accompanying teachers' guide, please email training@povertyactionlab.org. To reference this case study, please cite as:

> J-PAL. "Case Study: Get Out the Vote: Why Randomize?" Abdul Latif Jameel Poverty Action Lab. 2019. Cambridge, MA.

To reference the original study by Heller and coauthors, please cite as:

> Arceneaux, Kevin, Alan S. Gerber, and Donald P. Green. 2006. "Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment." Political Analysis 14 (1): 37-62. https://doi.org/10.1093/pan/mpj001.