J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

GET OUT THE VOTE

Why Randomize

Voting booths with the word "vote" and the American flag. Photo: Shutterstock.

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 as a teaching tool 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.
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. The impact is estimated by measuring the differences in outcomes between the treatment group and the comparison group.
Omitted Variable Bias	Statistical bias that occurs when relevant (and often unobservable) variables/characteristics are left out of the 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 high-income students.
Treatment Group	The group that receives the intervention.
Selection Bias	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? Comparing scores for those who participate and those who don't will produce a biased estimate of the effect of the tutoring program if these groups differ across characteristics that correlate with test scores. For example, those who choose to participate may be more motivated, and may have scored better than non-participants even without the tutoring program. Randomization minimizes 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: - Participants can choose to take up a treatment or refuse it - Participants can choose to leave the study (i.e., attrit/attrition)

	Method	Description	What *assumptions* are required, and how demanding are the assumptions?	Required data
Randomized Evaluation	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.* Examples of 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 comparison 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 difference-in-differences or regressions discontinuity 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 and another group who did not participate in the program after the program took effect.	*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 difference-in-differences or regression discontinuity design is preferred (see below).	Outcome data for program participants as well as another group of non-participants after the program took effect.
	Difference 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	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."
	Instrumental Variables	The design uses an "instrumental variable" that is a predictor of 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."

Learning Objectives

- Introduce various quantitative evaluation methods and demonstrate how each method can provide different estimates
- Identify critical assumptions underpinning different impact evaluation methods
- Provide a deeper understanding of bias and causal inference

Subjects Covered

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

Introduction

What is required in order for us to measure whether a program had an 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 differences in measured outcomes 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 at the outset of 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 a get-out-the-vote campaign in the United States. Using the same data, we will show how different methods may produce different results.

Voter Turnout

Voter turnout–the number of eligible voters that participate in an election–has been declining in many areas around the world since the 1990s (Solijonov 2016). High voter turnout is typically associated with a healthy democracy, whereas low turnout is associated with apathy and mistrust. Some different policies and practices to increase voter turnout include automatic voter registration, encouragement campaigns, and compulsory voting laws. In the United States, voters must register themselves in order to vote, but registration is not compulsory. Voter turnout in the United States was particularly low in 1998 and 2000, hitting a record low of 36 percent of the voting age population in the 1998 congressional elections (U.S. Election Assistance Commission 1998).

Vote 2002's Get Out the Vote Campaign

In late 2002, Vote 2002, a non-partisan civic group, ran a phone-based Get Out the Vote campaign in two states–Iowa and Michigan–to encourage voting in that year's US congressional elections. At the time, as telemarketing campaigns were replacing more traditional face-to-face campaigning, such as door-to-door canvassing, there was considerable debate over their effectiveness. Many believed the decline in voter turnout was a direct result of changing campaign practices. Therefore, in this context it was 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 called 60,000 potential voters in order to deliver the following message:

Phone script

"Hello, may I speak with [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?"

For each call, the potential voter was recorded as having 'responded' if they picked up the phone, listened to the script, and then responded to the question "Can I count on you to vote next Tuesday?" While 60,000 calls were placed, Vote 2002 only received full responses from 25,000 people. They also had data on the voter's age, gender, household size, whether the voter was newly registered, past voting behavior, which state and district the voter was from, and how competitive the previous election was in their district. After the election, the campaign used publicly available official voting records to determine whether the people they called actually voted.

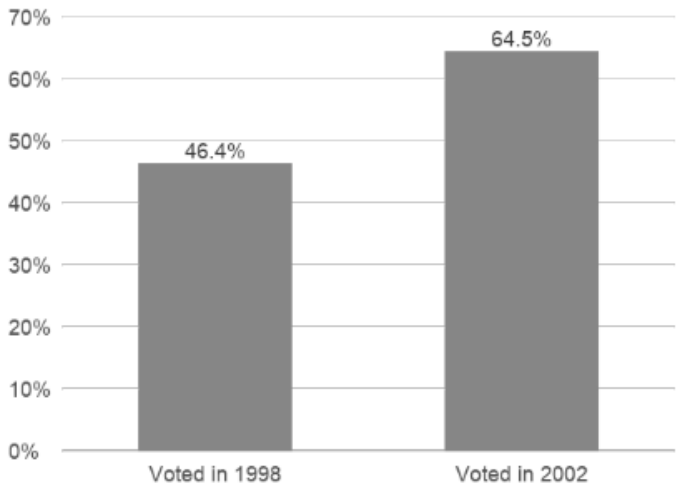
Did the campaign work? The following (fictional) newspaper excerpts illustrate different methods of evaluating impact to answer this question. (Refer to the previous table for a list of different evaluation methods).

Estimating the impact of the Get Out The Vote campaign

Method 1

News Release: Vote 2002 campaign is a huge success

Comparative voting rates among those who responded to the Vote 2002 phone campaign (proportion voting in 1998 vs. 2002)



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 who responded to this phone campaign, we saw an 18 percentage point increase in voter turnout compared to those same voters in 1998.

Discussion Topic 1

1.1 What type of evaluation method does this news release imply?

1.2 How did researchers mimic the counterfactual?

1.3 What assumptions do we have to make to believe this estimate? What might threaten these assumptions?

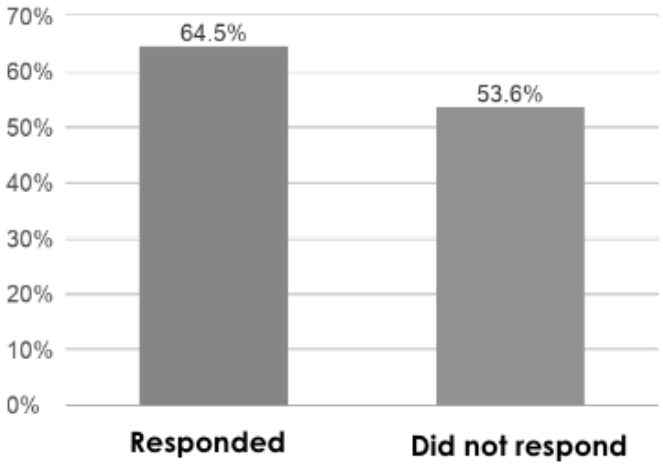
Method 2

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 voter turnout in 2002 of people who responded to this phone campaign to those who did not—both groups that were affected by these national events and, incidentally, 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.

Voting rates in the 2002 election, for people who responded to the Vote 2002 campaign versus people who did not respond



Discussion Topic 2

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

2.2. How did researchers mimic the counterfactual?

2.3. What assumptions do we have to make to believe this estimate? What might threaten these assumptions?

Method 3

Editorial: Analysts of Vote 2002 campaign fail to account for important differences between groups

If you haven't been paying close attention, you may have missed the public spat over the effectiveness of Vote 2002's 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 an external 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 responded with that of people who did not respond. Many of the people who responded to the campaign were already voting in prior elections. The analysis should therefore measure *improvement* in voting rates—not the final voting level. This accounts for the fact that these two groups had different voting rates in prior elections. When we analyze these outcomes to compare differential changes over time, we find voting rates for those who responded to Vote 2002 improved only marginally compared to those who did not respond (a 10.9 percentage point increase versus a 9 percentage point increase). This 1.9 percentage point difference is statistically significant[1] but much smaller than the other estimates. This suggests that the pre-post analysis greatly overestimated the Vote 2002 impact.

[1] For this case study, "statistically significant" means that researchers are confident that the result is correct and meaningful, assuming that the evaluation method is valid.

Had these evaluators taken a more rigorous approach, 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

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

3.2. How did researchers mimic the counterfactual?

3.3. What assumptions do we have to make to believe this estimate? What might threaten these assumptions?

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 people who responded to Vote 2002 are older on average, then we may expect them to vote at higher rates than those who did not respond. Indeed, those who picked up the phone and listened to the Get Out the Vote message were on average 5 years older than those who did not (56 and 51 years old, respectively). To observe the possible bias caused by omitting key variables, we conducted one simple difference analysis without controlling for these differences, and a separate analysis with controls. This helped us to obtain the true impact of the campaign.

Impact of GOTV Campaign on Voter Turnout

	Voter turnout in 2002	Voter turnout in 2002
Responded	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

** Indicates statistically significant at the 5 percent level

Numbers in parentheses are standard errors, a measure of the precision of the estimated size of an effect. Small standard errors suggest that the model describes the dependent variable well.

Discussion Topic 4

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

4.2. How did researchers mimic the counterfactual?

4.3. What assumptions do we have to make to believe this estimate? What might threaten these assumptions?

Method 5

Report: The impact of a phone campaign on voter turnout

The 60,000 individuals who were called by Vote 2002 were actually randomly selected to be called. They were randomly 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, 1,940,000 people).

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 being assigned to the program. After adding controls, 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[2] (0.5), is statistically indistinguishable from 0. In other words, calling voters has no statistically significant effect on voter turnout.

Impact of GOTV Campaign on 2002 Voter Turnout

	Called by GOTV vs. Not called by GOTV
Assigned to be called by GOTV	0.004
	(0.005)
Constant	0.461
	(0.01)

[2] A measure of the precision of the estimated size of an effect.

Observations	1,905,320

** indicates statistically significant at the 5 percent level

Discussion Topic 5

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

5.2. How did researchers mimic the counterfactual?

5.3. What assumptions do we have to make to believe this estimate? What might threaten these assumptions?

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.

Impact estimates of the GOTV campaign on voter turnout under different methods

Method	**Estimated impact**
Pre-Post	18.1 pp**
Simple Difference	10.9 pp**
Differences in Differences	1.9 pp**
Multivariate Regression with Panel Data	4.6 pp**
Randomized Evaluation‡	0.4 pp

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. We always need to think critically about how an impact evaluation method constructs a counterfactual.

The purpose of this case study was not to evaluate one particular voter mobilization campaign, but to compare evaluation methods in the context of this example. A randomized evaluation of the Vote 2002 campaign, found that people who responded to a phone message encouraging voter turnout were *already* more likely to vote in the upcoming (and previous) elections, compared to people who did not pick up the phone or did not respond—and that the phone call intervention did not significantly change people's likelihood of voting. Even when controlling 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. Even though several methods estimated 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.

There are many ways to estimate a program's impact and many reasons why we might choose one method over another. Any method relies on the validity of its underlying assumptions and the possible biases or challenges that these assumptions introduce. Whatever method we use, it is important to think critically about the underlying assumptions.

References

Solijonov, Abdurashid. 2016. "Voter Turnout Trends around the World." International Institute for Democracy and Electoral Assistance (IDEA).

"Turnout in U.S. Has Soared in Recent Elections but by Some Measures Still Trails That of Many Other Countries." 2022. *Pew Research Center* (blog). November 1, 2022. https://www.pewresearch.org/short-reads/2022/11/01/turnout-in-u-s-has-soared-in-recent-elections-but-by-some-measures-still-trails-that-of-many-other-countries/ft_22-10-17_globalturnout_dot/.

"Voter Registration and Turnout - 1998." 1998. U.S. Election Assistance Commission. https://www.eac.gov/sites/default/files/eac_assets/1/6/1998%20Voter%20Registration%20and%20Turnout%20by%20State.pdf.

Reuse and citations

To request permission to reuse this case study or access the accompanying teachers' guide, please email training@povertyactionlab.org. Please do not reuse without permission. To reference this case study, please cite as:

> J-PAL. "Case Study: Get Out the Vote: Why Randomize." Abdul Latif Jameel Poverty Action Lab. 2023. Cambridge, MA.

To reference the original study, 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.