summit

What *Would Have* Happened: Experimental and Quasi-Experimental Methods for Program Evaluation

Edward Seiler, George Cave, China Layne and Shane Thompson

Summit Consulting, LLC

June 2017

Table of Contents

Summary

Evaluating a treatment, intervention or policy change ideally would entail comparing what *actually* happened to a "counterfactual"–what *would have* happened in the absence of the treatment. Since we cannot observe both situations, the objective of causal inference in evaluation thus is to estimate a counterfactual.

In this white paper, we summarize some of the counterfactual estimation designs used widely by evaluators.

We start the paper by discussing randomized controlled trials (RCTs)–the gold standard experimental approach. We discuss what can go wrong with RCTs and why they are rarely used now in complex domestic policy evaluation settings. We then summarize some of the more widely used observational designs, and provide examples of how to implement these methods based on Summit evaluators' experience. We focus on three methods:

- Regression discontinuity designs;
- Difference-in-differences and synthetic controls; and
- Propensity score matching.

We conclude the paper with a summary table that provides a "cheat-sheet" for these methods.

Randomized Controlled Trials

Randomized Controlled Trials (RCTs) often are "considered the gold standard approach for estimating the effects of treatments, interventions, and exposures." [1] At "baseline", an evaluation sample is split into two groups: A "treatment group" is assigned to receive the treatment being evaluated, and a "comparison group" is assigned to serve as the counterfactual. When assignment is random and evaluation is prospective rather than retrospective, the comparison group is known as a "control group", and the evaluation is known as a randomized controlled trial. Both groups are followed up for a period long enough so the treatment can be substantially completed. Average differences in outcomes between the two groups at follow-up are known as "impacts" or "average treatment effects" (ATEs).

For example, consider a pharmaceutical RCT to evaluate a new blood-pressure drug. Suppose a sample of 200 clinic patients at baseline is split evenly into a treatment group and a control group using randomization. The treatment group is given access to a twice-daily regimen of blood-pressure pills, while the pills are withheld from the control group. Three months later, blood pressure measurements are taken for every baseline patient who can be found, and ATEs are calculated.

What are the advantages of such an RCT?

Eliminate selection bias: There are two main reasons average outcomes in the two groups might differ at follow-up. First, the treatment may have made a difference. Second, the two groups may have differed systematically in baseline characteristics related to blood pressure, before any treatment took place. One group may have had more severe blood pressure elevation, or might have had higher prevalence of protective factors ("improvers") or hypertension "worseners" such as smoking. Randomization tends to balance these potential baseline "confounders" between

[1] Austin, Peter (2011): "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies," Multivariate Behavioral Research, 46: 399-424.

treatment and control groups, so that treatment becomes "exogenous" and differences at follow-up can be attributed with some validity to the treatment rather than to pre-existing differences in confounders.

What might go wrong in an RCT?

Incomplete treatment: Some members of the treatment group might not fully "take up" the treatment, even though they are eligible for it. Researchers cannot force human beings to do everything researchers want them to do, especially if they are outpatients. Some treatment group members will miss dosages, stop taking pills entirely, or never even take a single pill.

Control group contamination: Some members of the control group may receive the same or similar treatment from other sources, like other clinics. Some RCTs try to prevent such weakening of the control contrast with "placebos", indistinguishable from the pills given to treatment group members. In general, placebo controls are permitted only in situations where there is no standard treatment for a disease. In other situations, controls generally must be given an effective alternative to a drug being evaluated.

Small sample sizes: Many RCTs provide inconclusive results because recruitment of patients can be slow and expensive, and eventuate in relatively small sample sizes with inadequate statistical power when deadlines and/or funds are exhausted.

Differential attrition: Even though selection bias may have been eliminated by randomization at baseline, it can creep back in at follow-up through differential attrition. Typically, a higher percentage of treatment group members than control group members provide follow-up data. If the unknown outcomes for missing controls differ from the known outcomes for controls available at follow-up, impact estimates may be biased beyond what might be counteracted by missing outcomes among treatment group members.

Internal / external validity trade-offs: RCTs may trade better "internal validity" (high sample take-up rates, low rates of similar treatments for sample controls. Low differential attrition) for worse "external validity" (generalizability of RCT sample results to easily observed and passive populations).

Why are alternatives to RCTs commonly used?

While RCTs have been used in specific cases in economic settings, observational design alternatives are more commonly used to evaluate programs outside medicine. Why? [2]

Practical concerns: "Social Experiments" outside medical research face all the statistical and ethical problems of medical RCTs plus some additional practical problems. Services such as job training may be available readily from multiple sources, so that controls turned away from a program being evaluated may end up receiving the same or very similar services from other sources. There are unlikely to be "placebo" programs available to lessen this risk, and, even if there were, it would be unethical to use them.

Ethical risks: For example, the need to recruit control groups and attempts to increase "take up" rates by screening for high motivation to accept program opportunities might result in some scarce program slots going unused. Random assignment generally is considered ethically acceptable only when it results in no diminution of service delivery to a community, ideally acting as a fair way to distribute scarce and heavily oversubscribed program resources.

Implementation issues: If relatively small fractions of the treatment group receive the treatment, and/or relatively large fractions of the control group receive the treatment or something like the treatment, then the usual "intent-to-treat"

[2] RCTs are being used in growing number of instances in economics. For the interested reader, we suggest Abhijit Banerjee's and Esther Duflo's 2011 book "Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty" that uses an evidence-based RCT approach to report on the effectiveness of solutions to alleviate poverty (see http://www.pooreconomics.com).

ATEs may not suffice. Policymakers then might be more interested in the effects of treatment on the treated ("TOT" impacts).[3] Strong assumptions may be necessary to calculate such TOT impacts, and the simplicity of RCTs may be compromised.

When RCTs Don't Work: Allowing Subjects To Decide Whether To Participate

Selection bias may be particularly problematic in observational studies when eligibility criteria limit entry into the intervention or because of characteristics of subjects or environments that influence the choice of who will receive the treatment.

In an RCT study to estimate the benefit of financial counseling on mortgage loan defaults, borrowers who were candidates to receive a loan modification were randomly assigned into treatment and control groups. However, legal requirements meant that treatment group borrowers could not be forced to participate in counseling post assignment. Borrowers were given the choice to opt-out of receiving counseling.

The direction of the selection bias introduced by this legal constraint depends on how the decision to accept the counseling is correlated with factors that matter for future loan performance. On one hand, if borrowers with more difficult financial situations are more likely to seek help and thus enroll in the counseling sessions, we would bias down the effect of financial counseling. On the other hand, borrowers experiencing more permanent financial hardships may be more pessimistic and thus opt-out of the program, knowing they will likely default soon. This would have the opposite effect and magnify the bias.

In the rest of this paper, we summarize some of the more widely used observational designs (used when RCTs are not feasible), and provide examples of how to implement these methods based on Summit evaluators' experience. We focus on three methods:

- Regression discontinuity designs;
- Difference-in-differences and synthetic controls; and
- Propensity score matching.

We conclude the paper with a summary table that provides a "cheat-sheet" for these methods, and discusses succinctly technical processes and when each method is preferable for use.

Regression Discontinuity Design

Sometimes, when prioritizing potential clients for program services, an agency uses a strict quantitative standard with a sharp cutoff threshold. For example, a scholarship program for disadvantaged C+/B- students (unlikely to be able to afford community college on their own or get another scholarship from another agency) may use an entrance test score. Those just above and just below the cutoff score arguably have similar pre-scholarship characteristics, and minor differences in their entrance test scores may be considered almost random. This situation is quite like an RCT, where randomization determines which applicants with common baseline characteristics are eligible for program services.

Summit authored a report for a Department of Labor (DOL) Occupational Safety and Health Administration (OSHA) evaluation using such a regression discontinuity design (RDD) scheme.[4]

[3] To further illustrate ITT and TOT, an example can be seen at: http://people.bu.edu/lang/itt-tot.pdf.
[4] This report is available at https://www.dol.gov/asp/evaluation/completed-studies/SST_Evaluation_Final_Report.pdf.

OSHA's Site Specific Targeting Program (SST) is a planned inspection program that prioritizes worksites for inspection based on their rates of injury and illness in a prior year. Worksites just above and just below the threshold for inspection in each year arguably are similar in pre-inspection characteristics, and small differences in prior injury and illness rates that lead some worksites to be inspected but not others arguably are distributed nearly randomly. "High rate letters" informing managers that their worksites ranked among the highest in their industry in illnesses and injuries were a second treatment in addition to (or in place of) follow-up SST inspection.

To determine whether the SST program has an impact on improving regulatory compliance and workplace health and safety, the DOL's Chief Evaluation Office (CEO) and OSHA contracted Summit to evaluate the program. This evaluation, started by IMPAQ International in 2010 and taken over by Summit in late 2013, assessed the impacts of SST on two main outcomes of interest:

- Regulatory Compliance, measured by the probability of OSHA citing a worksite for a violation during a follow-up inspection; and
- Health and Safety, measured by the follow-up injury/illness rate.

Summit applied both a RCT design and a RDD to assess impacts on these outcomes. In the RCT experimental design, worksites were assigned randomly to one of two treatments (a high-rate letter or both a high-rate letter and an SST inspection) or to a control group. 2,520 worksites were included in this experimental study. The treatments occurred in 2011 and outcomes were observed in 2012-2015. The RDD relied mainly on an annual OSHA Data Initiative survey of 7,045 worksites.

The original research design and analysis plan included only the RCT. Summit supplemented the experimental study with a quasi-experimental design to offset implementation and data limitations of the RCT and to capitalize on data OSHA already had collected. Overall, neither the RCT nor the RDD study found statistically significant impacts. However, the lack of statistically significant impacts from our RCT and RDD studies does not mean that OSHA's enforcement actions had no beneficial impact on workplace health and safety. The results of this study only imply that if the enforcement actions had any impact on worksites' regulatory compliance and health and safety, that impact is unlikely to be larger than what the estimated confidence intervals indicate. Even if this study can only rule out the existence of some large impacts, the knowledge and techniques developed through our data preparation may help enhance OSHA's research infrastructure and could be used to facilitate and support future evaluation and analytic work with OSHA data.

Regression Kink Design (RKD)

As alluded to in its name, RDD is based on a *discontinuity* in the likelihood of being treated at some threshold point. RKD is based on a change in slope at the threshold leading to a "*kinked*" function. For example, state weekly unemployment benefits often are an increasing linear function, up to some maximum amount, of highest quarterly earnings over some past period. At the maximum, there is a kink in the benefit-earnings schedule, so that the slope changes from positive to zero.

This situation is like RDD, where there is a jump in the probability of receiving services, and evaluators look for a corresponding jump in outcomes. In an RKD design, there is a jump in the first derivative of the probability of receiving services, and evaluators look for a corresponding jump in the first derivative of outcomes. (Technical details can be seen in: Card, David, David S. Lee, Zhuan Pei and Andrea Weber (2015): "Inference on Causal Effects in a Generalized Regression Kink Design," Econometrica, 83(6): 2453–2483).

Difference-in-Differences and Synthetic Controls

Difference-in-differences (DiD) and synthetic control methods use longitudinal data to determine treatment effects. Each method establishes baseline similarity between treatment and control groups in the pre-treatment period(s), and then tracks post-treatment outcomes. If the baseline similarity in the pre-treatment period diverges in the post-treatment period, we attribute the divergence to the treatment.

Researchers choose DiD designs or synthetic control methods depending on the available data and the underlying program. If there are many treatment and control units, DiD is best. If there is only one treatment unit and relatively few control units, the synthetic control method is appropriate.

Difference-in-differences designs:

1. Identify control units that have a statistically equal pre-program outcome trend to the treatment group (determined via regression).
2. Test whether that statistical equivalence still holds after the program.
3. If the equivalence still holds, conclude that the program had no statistically significant effect on the outcome.
4. If it does not, attribute the difference to the program.

Synthetic control methods:

1. Identify the treatment group and the donor pool of potential control groups.
2. Construct a synthetic control group, which is a weighted combination of several control groups from the donor pool, which is approximately equal to the treatment group in pre-program characteristics and outcomes.
3. If a difference in outcomes exists between the treatment group and its synthetic control group after the program, attribute the difference to the program.

Summit is currently estimating state and county level outcomes to analyze the effect of a statewide shock to public education funding. At the county level, we are using a difference-in-differences design since we have many treatment units (counties in the state with the funding shock) and many control units (counties from other states). At the state level, we are using synthetic control methods since we have only one treatment unit (the state with the funding shock) and a small pool of 49 potential control units (the other states).

The figures below show how DiD designs and the synthetic control method differ in identifying the effect of the shock. This hypothetical illustration assumes student outcomes would improve after a funding shock.

Pre-treatment:

- DiD designs require a parallel trend in outcomes between treated and control counties.
- Synthetic control methods require approximate equality in outcomes between the treated state and the synthetic control state.

Post-treatment:

- DiD designs measure the treatment effect as the difference between the treatment counties' outcomes and the control counties' outcomes, after removing the baseline difference from the pre-treatment period.
- Synthetic control methods measure the treatment effect as the difference between the state outcome and the synthetic state outcome.

Figure 1: Difference-in-Difference

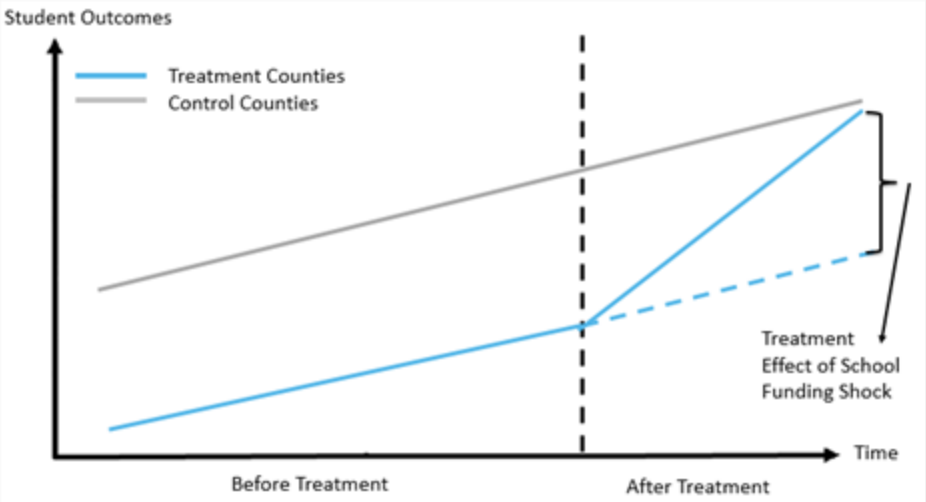
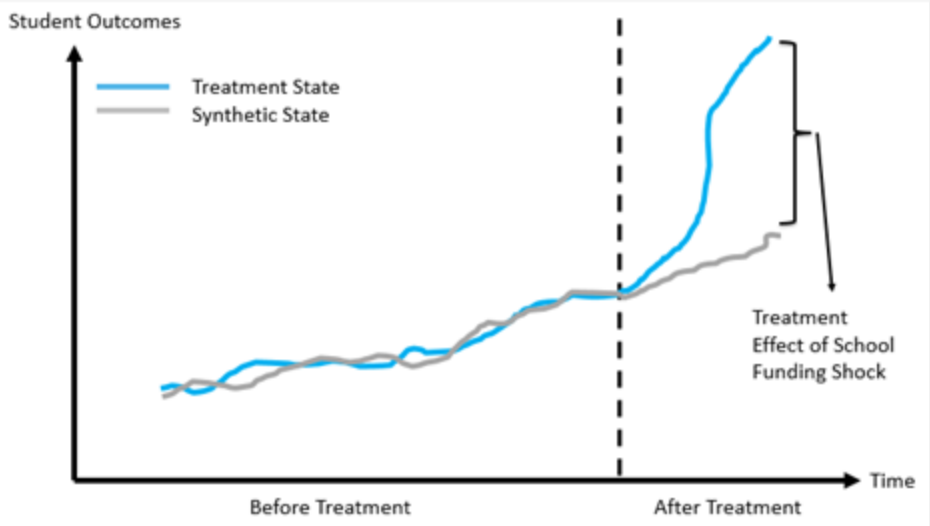


Figure 2: Synthetic Control Method



Propensity Scoring

Propensity Score Matching (PSM) is a statistical matching technique, introduced by Rosenbaum and Rubin, to allow analysis of a nonrandomized study.[5] It estimates the effect of a treatment by reducing or eliminating confounding effects when using observational data by making the groups receiving treatment and not-treatment more comparable. PSM attenuates the bias by comparing outcomes among units that received the treatment versus those matched units that did not.

[5] Rosenbaum, P. and D. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," Biometrika, 70: 41-50.

The basic idea of the matching approach is to find a large group of non-treatment individuals who are like the treatment individuals in all relevant pre-treatment characteristics. Rosenbaum and Rubin suggested using a balancing score (based on the relevant pre-treatment characteristics) that is independent of assignment into treatment for the matching procedure. PSM requires generating matched sets of treated and untreated individuals who share a similar value of the propensity score.

Summit staff used PSM to evaluate the benefits of transferring mortgage servicing portfolios from traditional to high-touch servicers.[6] Following the mortgage crisis, Fannie Mae initiated servicing transfers to improve mortgage servicing performance, and Summit staff (then at Fannie Mae) were charged with evaluating the benefits of the transfers.

At first blush we envisioned comparing loan performance of the transferred loans to a hold-out sample – a sample of randomly selected loans that was not transferred. However, as discussed in our earlier section on RCTs, this would not be appropriate since the servicer from whom loans had been transferred would change its servicing too (whether because they now had additional capacity to service the remaining loans, or if the threat of additional transfers would cause it to change its behavior). Instead, we used PSM to select comparable non-transferred loans (from other servicers in Fannie Mae's book of business) based on a propensity (probability of default) score.[7,8] We then tracked how projected losses diverged over time between the transferred loans and the comparable set of matched loans using Fannie Mae's internal loss forecast models. This formed the basis for our evaluation of servicing improvement.

Conclusion

Federal agencies are increasingly looking to implement rigorous, data-driven evaluations of their programs. RCTs are often considered the gold standard for estimating the impact of a program or service. However, there are multiple circumstances when a RCT would be impractical, infeasible, or prohibitively costly for an agency to implement. In these cases, agencies look to quasi-experimental evaluation methods to estimate impact effects using available observational data. In this white paper, we provided an overview of three types of quasi-experimental methods: Regression Discontinuity Design (RDD), Difference-in-Difference (DiD) and Synthetic Control, and Propensity Score Matching (PSM). We discussed how these methods can reduce selection bias in the estimation of program impacts and provided examples of their implementation from Summit evaluators' experience with our clients. Finally, we provided a table comparing the different quasi-experimental methods and outlining the requirements for implementing the methods.

---

[6] For example, in Walter Investment Management Corporation's 2013 10-K report (available at: https://www.sec.gov/Archives/edgar/data/1040719/000119312514071098/d631621d10k.htm) there is discussion of servicing transfers from Bank of America, NA. Of interest, on page 13, they write: "Our strong servicing performance has been a key driver to our success in winning servicing transfers and sales."

[7] We also note that the propensity score was calculated in a two-step process. Since the score was based on a default model it is straightforward to see why this was needed. Suppose the default model was estimated using current loan-to-value (LTV) and FICO. A high LTV is riskier than a low LTV, and a high FICO is less risky than a low FICO (other things equal). We can thus get the same default probability for a loan with a high LTV and high FICO as with a low LTV and low FICO loan. (We may also get the same score for a loan with a medium value for both variables). As such, in the first stage we constrained matched loans to be in the same interval for all matching variables. For example, a loan with FICO in the range 660-700 could only be matched with loans in the same FICO range, and a loan with LTV between 0.9-1.0 could only be matched with loans in the same LTV range. The propensity score was then used to select loans from the (filtered) list of candidate loans after the first step was completed.

[8] For additional discussion, including stratification and interval matching, see: Caliendo, M. and S. Kopeinig (2005): "Some Practical Guidance for the Implementation of Propensity Score Matching," IZA Discussion Paper No, 1588.

Methodology	Overview	Technical Process	Best when you have:
Regression discontinuity designs	Compares groups that barely qualified for the program with groups the barely missed qualifying for the program. Operates on the assumption that treatment and control groups on the margin of program qualification are virtually randomly assigned.	1. Identify threshold that qualifies participants for the program. 2. Restrict analysis to participants just above and just below that threshold. 3. Compare linear regressions on outcomes of participants around the threshold. 4. If a discontinuity in outcomes at the threshold exists, attribute the discontinuity to the program.	A well-defined (and upheld) threshold that assigns participants to the program Many participants close to the threshold
Difference-in-differences	Compares treatment group outcomes before and after the program relative to control group outcomes. Operates on the assumption that a treatment group and control group that have similar outcome trends *before* the program would have had similar outcome trends *after* the program if not for the program.	1. Identify a control group that has a statistically equal pre-program outcome trend to the treatment group. 2. Test whether that statistical equivalence still holds *after* the program. 3. If the equivalence still holds, conclude that the program had no statistically significant effect on the outcome. 4. If it doesn't, attribute the difference to the program.	Panel (longitudinal) data Many treatment and control groups At least three time periods of pre-program data
Synthetic Control	Compares treatment group with a weighted average of potential control groups that is constrained to be approximately equal to the treatment group. Operates on the assumption that a control group that is approximately equal to the treatment group before the program would have been approximately equal to the treatment group after the program, if not for the treatment.	1. Identify the treatment group and the donor pool of potential control groups. 2. Construct a synthetic control group, which is a weighted combination of several control groups from the donor pool, that is approximately equal to the treatment group in pre-program characteristics. 3. If a difference in outcomes exists between the treatment group and its synthetic control group, attribute the difference to the program.	Only one treatment group and at least 20 possible control groups Panel (longitudinal) data
Propensity score matching	Compares treatment groups with statistically-matched control groups. Operates on the assumption that treatment and control groups with the same likelihood of program participation are virtually randomly assigned.	1. Identify the treatment group and the donor pool of potential control groups. 2. Estimate the probability of being in the treatment group for each subject in the sample, given their observable characteristics. 3. Match treatment groups and control groups based on their predicted likelihood to receive the treatment. 4. If a difference in outcomes exists between the treatment groups and their statistical matches, attribute the difference to the program.	Many treatment and control groups Treatment and control groups with overlapping observable characteristics

About Summit and the Authors

Summit Consulting, LLC (Summit), headquartered in Washington, DC, specializes in applying cutting-edge quantitative techniques to the real-world challenges facing federal agencies and private-sector enterprises. At Summit, we solve complex analytical challenges with unparalleled customer service and extensive client collaboration. The solutions are complete only when our clients understand them and use them to solve their problems.

Summit's Program Evaluation Directorate uses experimental and quasi-experimental evaluation designs to provide reliable estimates of program effectiveness for a wide range of client services. Summit's staff of economists, econometricians, and research scientists uses quantitative techniques to assist our clients as they model risk, evaluate program performance, and predict future performance.

For more information, visit Summit's website (http://www.summitllc.us).

Whitepaper Contributors

Dr. Eddie Seiler is Summit's Chief Housing Economist. He spent a decade at Fannie Mae where he directed servicing research for the National Servicing Organization. He designed controlled experimental pilots to examine the effects of post-modification counseling. He subsequently joined Summit Consulting in early 2014 to head its Mortgage Finance practice, and he accepted the role of Chief Housing Economist in early 2017. (Contact information: eddie.seiler@summitllc.us).

Dr. George Cave, Summit's Senior Research Fellow, has had a rich career as a program evaluator. He has led many field experiments using random assignment and other experimental designs to evaluate the effectiveness of various social programs. (Contact information: george.cave@summitllc.us)

Dr. China Layne is a Manager at Summit. She has extensive experience with statistical analysis and evaluation research and expertise in the analysis of large scale survey data and conducting evaluations using administrative data. She previously served as a Survey Statistician for the U.S. Census Bureau, as well as a Research Scientist for the Center for Human Services Research at the State University of New York at Albany. (Contact information: china.layne@summitllc.us).

Dr. Shane Thompson is a Senior Consultant with Summit. He has core competencies in program evaluation and quantitative methodologies. Dr. Thompson also has experience analyzing large administrative datasets. Currently, he conducts research projects for several agencies within the U.S. Department of Labor. His prior experience includes teaching appointments and research on labor economics and the economics of education. (Contact information: shane.thompson@summitllc.us).