Energy Efficiency Program Evaluation

Maximilian Auffhammer
George Pardee Jr. Professor
University of California at Berkeley
National Bureau of Economic Research
CESifo Fellow

NARUC Webinar
May 6th 2015

What is E2E?

Berkeley UNIVERSITY OF CALIFORNIA · THE UNIVERSITY OF CHICAGO · MIT Massachusetts Institute of Technology

Why?

The E2e Mission

E2e's mission is to solve one of the most perplexing energy puzzles of our time—the efficiency gap. Study after study has shown potential for huge savings and hefty cuts in carbon emissions from widespread adoption of energy efficient technologies. Yet these gains have failed to fully materialize. Scholars can't say why. Theories abound. Solutions remain elusive.

How?

The E2e Strategies

E2e unites top researchers in economics, engineering and other fields to close the energy efficiency gap using transparent and state-of-the-art analytical techniques.

In addition to conducting research, E2e communicates results to policymakers and the public to maximize impact. Infusing the creation of knowledge with a commitment to non-partisan outreach, E2e aims to create a

Google™ Custom Search | Search

September 8, 2014:

Randomized Control Trials for Energy Efficiency Evaluation, Berlin, Germany

Part 1: Impact Evaluation Overview

Maximilian Auffhammer
George Pardee Jr. Professor
University of California at Berkeley
National Bureau of Economic Research
CESifo Fellow

NARUC Webinar
May 6th 2015

Some questions we should know the answers to…

How much energy do rebates actually save?

EMERGENCY | Safety | About Search GO ENGLISH

MY HOME | MY BUSINESS | BUSINESS TO BUSINESS

My Bill & Account | Service Requests | Outages | Find Ways to Save | Additional Services

My Account

- My Rate Plan
- Usage Analysis & Alerts
- Home Money Saver
- Solar and Renewables
- Savings & Rebate Programs
- Assistance Programs
- Plug In Electric Vehicles
- Energy Savings Tips
- Other Resources

Energy Savings Programs

It pays to save energy, check out the latest rebates

Learn about programs available for your energy-efficient products and appliances. These tools and more will help you save on your energy bill.

Home and Business Area Networking (HAN)
Home and Business Area Networking (HAN) is a wireless technology that allows customers to view their electricity consumption in near real-time, via their SmartMeter™.

SmartAC™
Take control and help your community prevent power interruptions by enrolling in PG&E's SmartAC program. It's free and easy, join now.

SmartRate™
Enroll in the SmartRate Summer Pricing Plan where you can save money, gain more control on your summer electric bills and help the environment.

Energy Upgrade California™
Develop a comprehensive plan for making energy-efficient upgrades, and start seeing greater energy savings sooner.

Contact Us

For general questions, please contact PG&E's Customer Service Center at 1-800-743-5000.

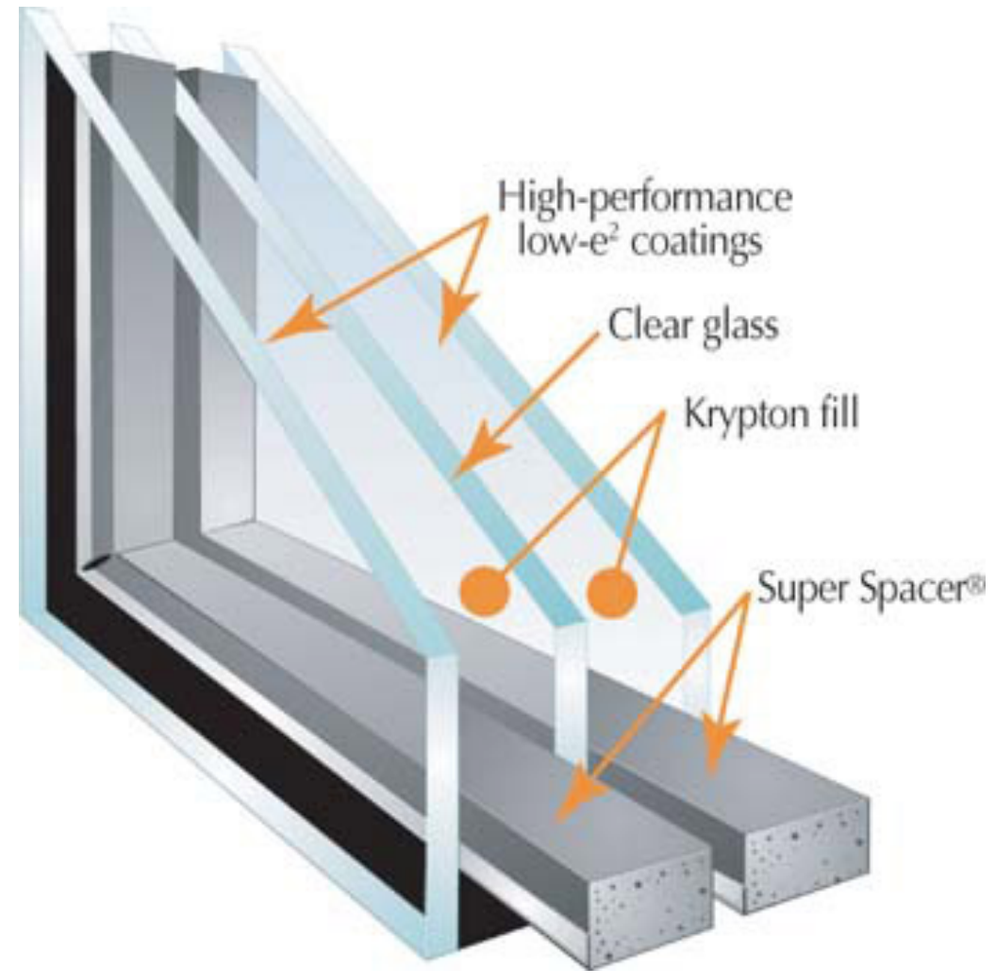
Solar Water Heater Rebate

Conserve resources and save energy by installing a solar water heating system.
Learn more

How much energy do building retrofits save?

How much fossil fuel is actually displaced by Solar PV?

How much energy is saved by EE window retrofits?



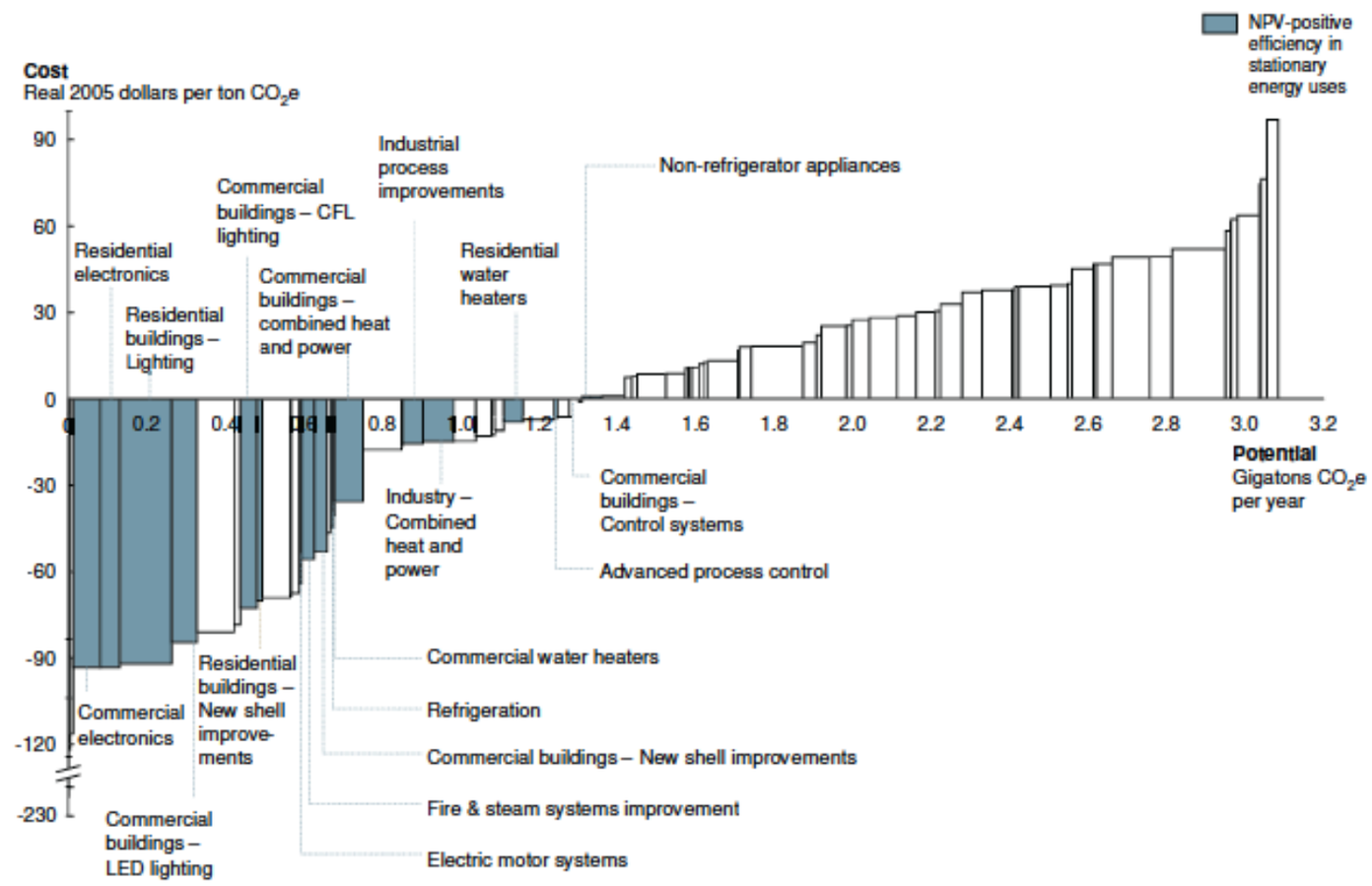
How much energy is saved by EE water heaters?

And the economists ask: What are the costs?

How good are the estimates we have?

Exhibit G: U.S. mid-range greenhouse gas abatement curve – 2030

This exhibit shows greenhouse gas abatement potential as depicted in the mid-range case in McKinsey's greenhouse gas report (2007), with energy efficiency opportunities associated with stationary uses of energy highlighted. The height of each bar represents the incremental cost in dollars to abate one ton of carbon dioxide (or its equivalent); the width shows the gigatons of such emissions that could be abated per year.



Source: McKinsey & Company

The challenge

"The central problem [...] is that of evaluating the effect of the exposure of a set of units to a program or treatment on some outcome."

Imbens and Wooldridge, 2009

The importance of causal effects in policy evaluation

- Programs and policies are typically designed to **change outcomes**.
- Whether or not these changes are **actually achieved** is a crucial public policy question, which is often not examined.
- More commonly, program managers and policy makers focus on controlling and **measuring the inputs and immediate outputs** of a program
- The key question is whether programs have **achieved their intended goals**.
- **Evidence based policy making** requires identifying the **causal relationship** between the project, program, or policy and the outcomes of interest.

Causality versus correlation

- **Causality / Causation:** A change in X causes a change in Y
  - A rebate (X) increases efficient AC adoption (Y)
  - A free home retrofit (X) reduces energy consumption (Y)
- **Correlation:** X and Y move together and are somehow related to each other

- We want to show causality and rule out simple correlation

Evaluate the outcome, not the input or output.

Problem	Logical Framework				
	Inputs	**Output**	**Intermediate Result**	**Strategic Objective**	**Goal**
Electricity consumption in California's households spikes during extreme heat events leading to dispatch of high cost generating units and increased risk of outages.	Utility purchases smart residential meters and in-home displays for households	Smart meters and in-home displays are delivered and installed	Households "use"/interact with the in home displays	Households become more aware of their electricity consumption patterns.	Households will decrease their electricity consumption overall and during peak times.

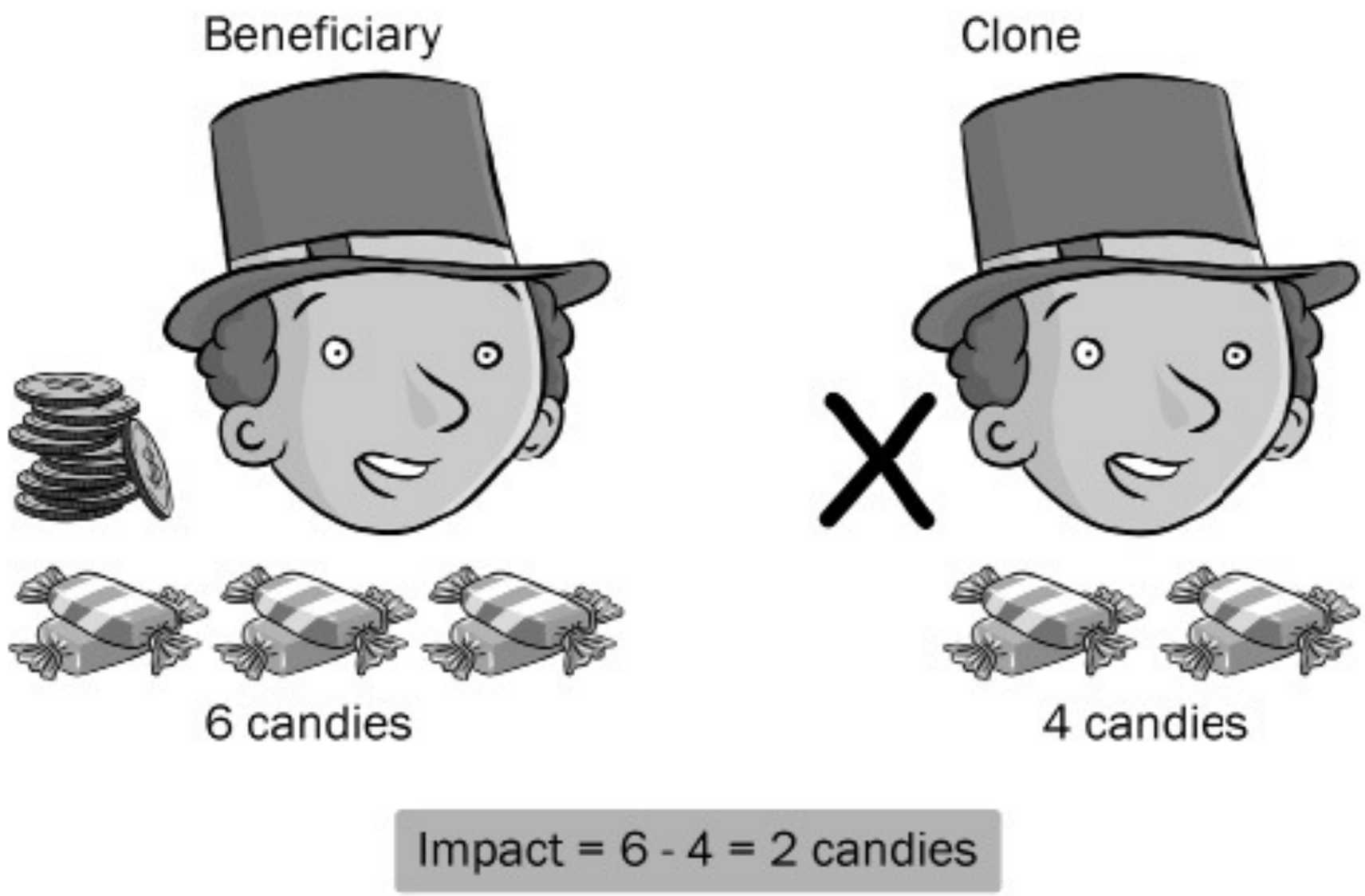
The problem in a nutshell

We would like to know how **individuals** who have **participated in a program** would have **fared in the absence of a program** *or*

We would like to know how individuals who did not participate in a program would have fared had they participated.

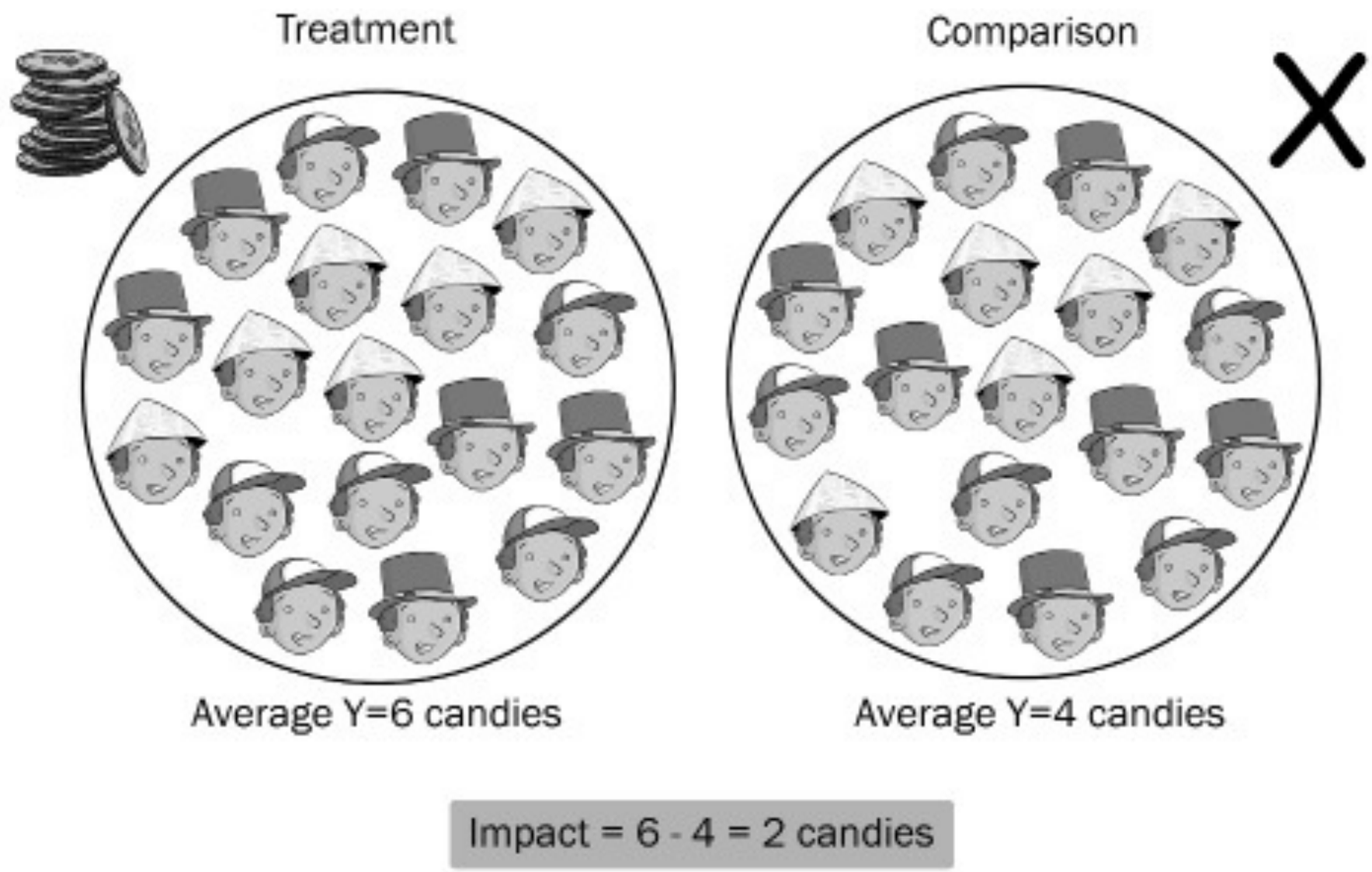
- You **never observe** the necessary **counterfactual**.
- At any point in time **a person is either "treated" or not**.
- The **solution** may lie in **comparing** outcomes for a group of treated to the outcome for a group of untreated individuals.
- There is a **right way** of doing this (**RCTs and quasi-experimental approaches**) and there are wrong ways of doing this.

The comparison we wish we had



Source: Gertler et al. 2009

Choosing a comparison group



Source: Gertler et al. 2009

The trouble with comparison groups

A “valid” **comparison group** is a group of subjects, who absent treatment, **would have had similar outcomes to those of the treatment group**.

This breaks down if individuals who are exposed to treatment are different from the individuals in the control group in observable or unobservable ways.

- Treatment happens in specific areas (e.g. München vs. Nürnberg)
- Participants are screened (e.g. income thresholds)
- Participation in a program is voluntary (e.g. solar PV installation)

Hence, one very often **cannot separate** the **effect of treatment** from that of **preexisting differences**.

The spectrum of evaluation approaches

1. **Ex-Ante Studies**
   - McKinsey Curve
   - Bottom Up Studies
2. **Stated Preference Studies**
3. **Quasi-Experimental Studies**
   - Before and After Comparison
   - Cross Sectional Comparisons
   - Difference-in-Difference Comparisons
4. **Randomized Control Trials (RCTS)**

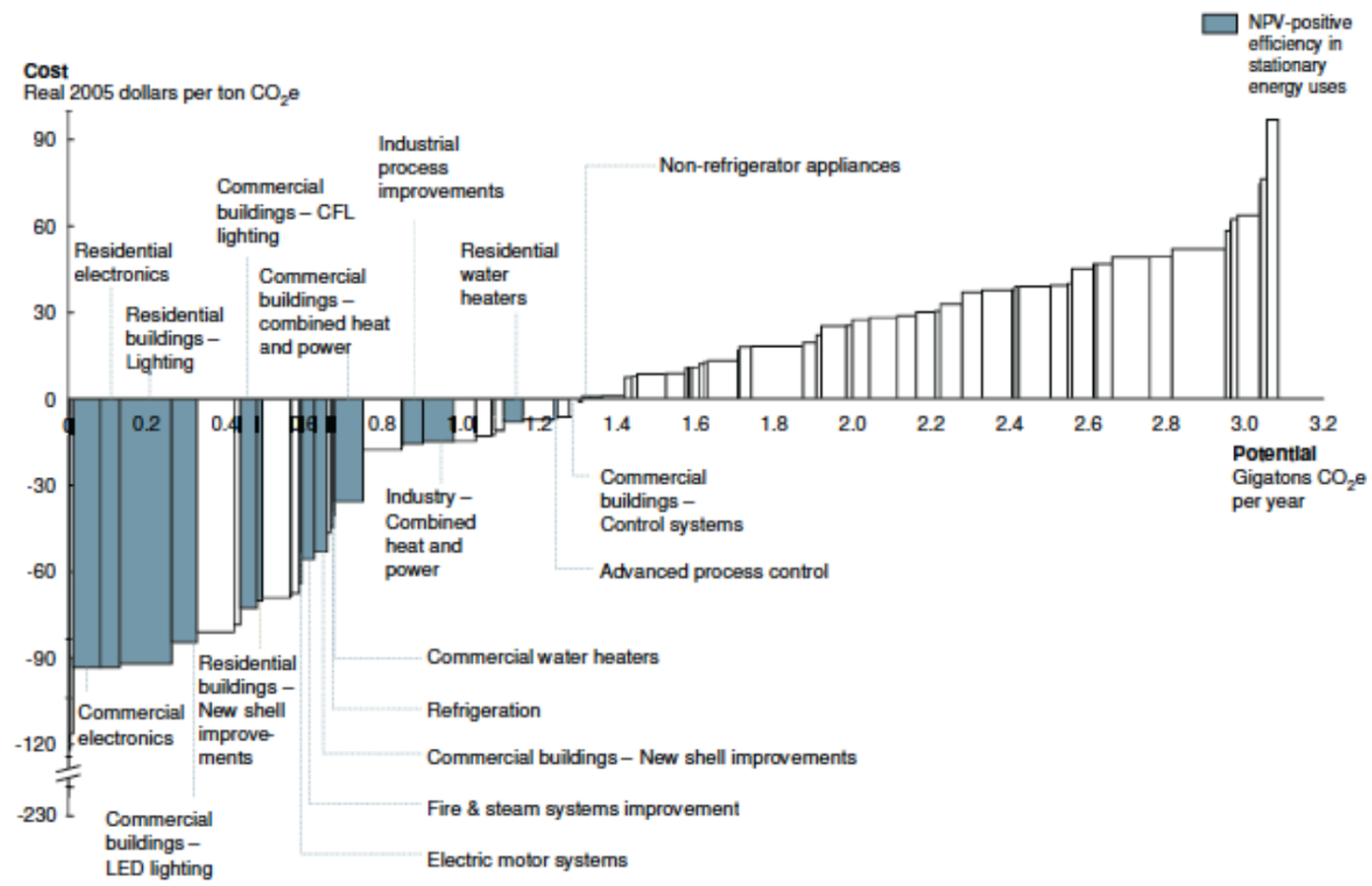
Example 1: Ex Ante Studies

- Ex Ante studies are very **good** at **quantifying potential savings**.
- They have to make **heroic assumptions** about human and firm **behavior**.
  - Have trouble with transactions costs
  - Principal Agent issues....
- Have to make assumptions about how technology performs **in practice.**
- Have trouble quantifying free riding.

Example 1: Ex-Ante Studies

Exhibit G: U.S. mid-range greenhouse gas abatement curve – 2030

This exhibit shows greenhouse gas abatement potential as depicted in the mid-range case in McKinsey's greenhouse gas report (2007), with energy efficiency opportunities associated with stationary uses of energy highlighted. The height of each bar represents the incremental cost in dollars to abate one ton of carbon dioxide (or its equivalent); the width shows the gigatons of such emissions that could be abated per year.



Example 2: Stated Preference Approaches

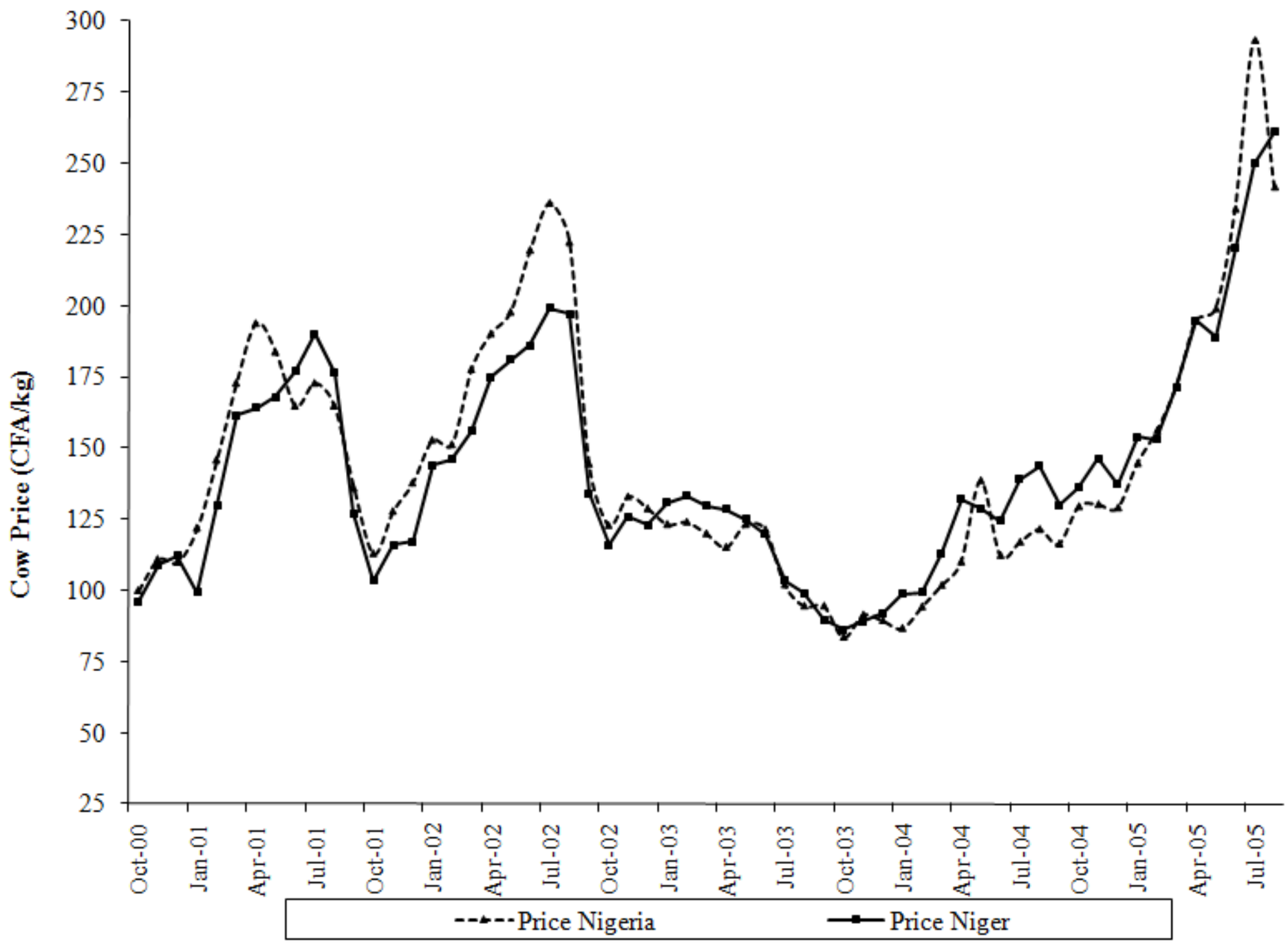
One solution to the counterfactual issue may be asking people what they would have done in the absence of a program.

- Recall bias.
- People are terrible at estimating their own behavior in an alternative version of the universe.
- Response rates are small
- Sample selection issues.
- Incentives to answer truthfully are often misaligned.

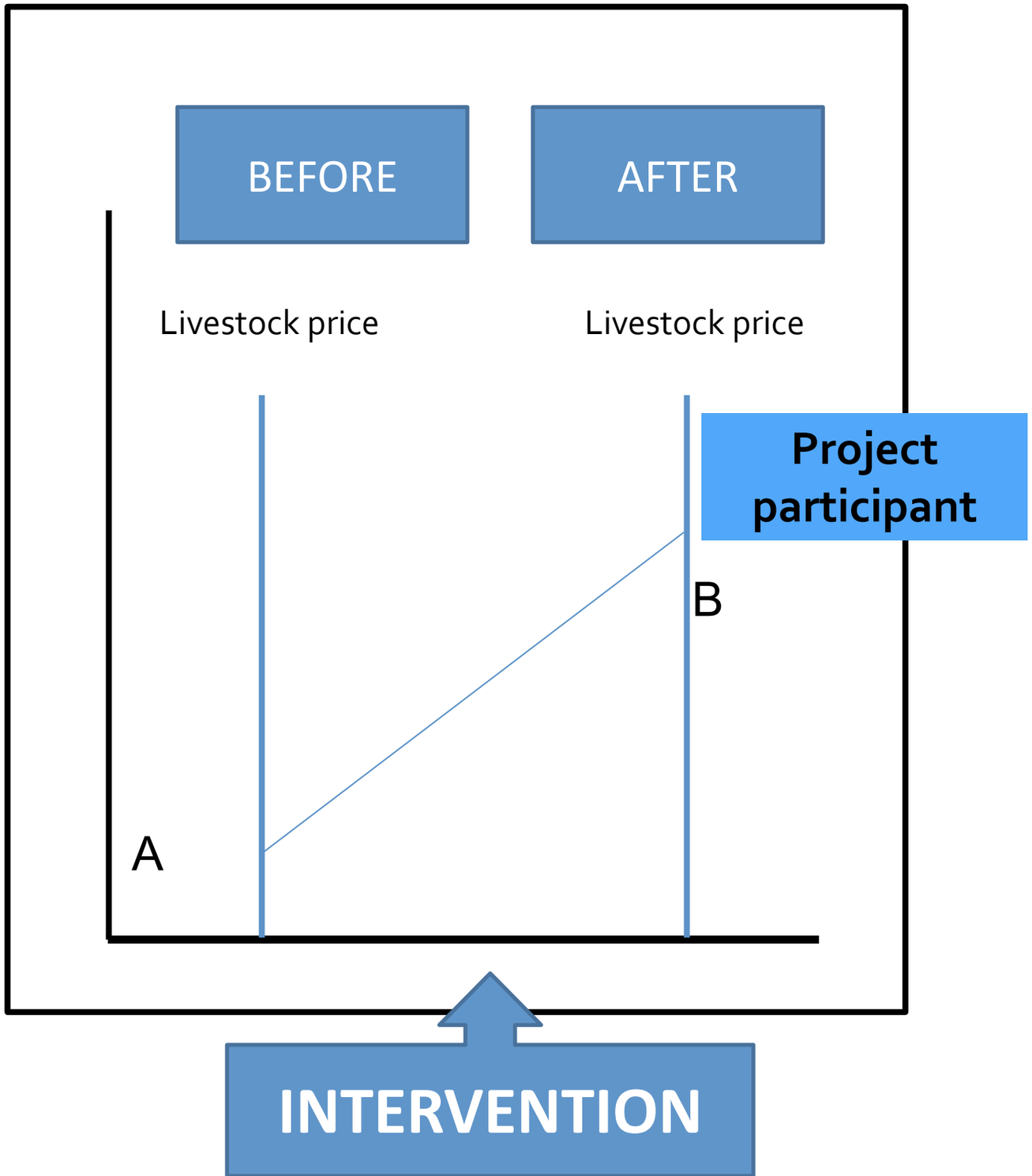
Example 3: Before and after comparison

- The Ministry of Livestock in Niger has a project to improve livestock marketing between Niger and Nigeria.
- The stated objective of the project is to increase prices received by herders.
- The project gives mobile phones to herders so that they can receive livestock prices by SMS from different markets.
- A before and after evaluation compares the prices received by the herders before and after the project. The indicators show that prices are 500 CFA (about USD 1) higher after the project.

Example 3: Before and after comparison

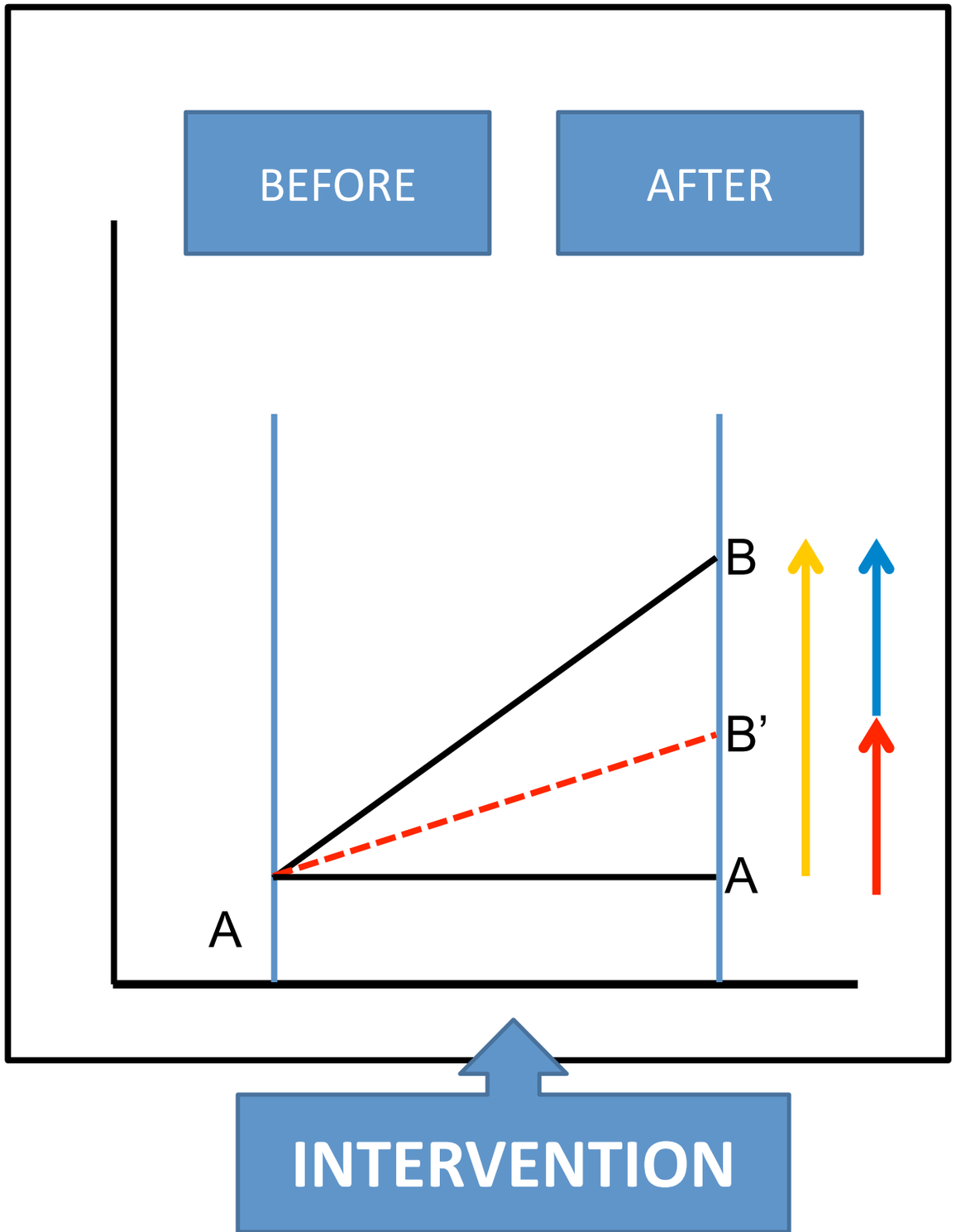


Example 3: Before and after comparison



- Did the program succeed?
- We don't know – we can't distinguish the impact of the program from the trend in livestock prices
- How can we show that the project was successful in increasing herders' prices more than the increase in livestock prices?

Example 3: Before and after comparison



- We want to know what part of the impact or change in program participants' lives is due to the program
- B-A is the overall change
- B'-A is the trend
- B-B' is the **program impact**

Example 4: Cross-Sectional Comparisons.

- We are interested in estimating the causal impact of "$/ student spent on after school programs" on performance in state level achievement tests.
- We have measured data on test scores at the school level and average test scores for each school district for one year.
- Using regression analysis, one can compare outcomes for schools with higher levels of program funding to outcomes for schools with lower levels of funding.

Example 4: Cross-Sectional Comparisons.

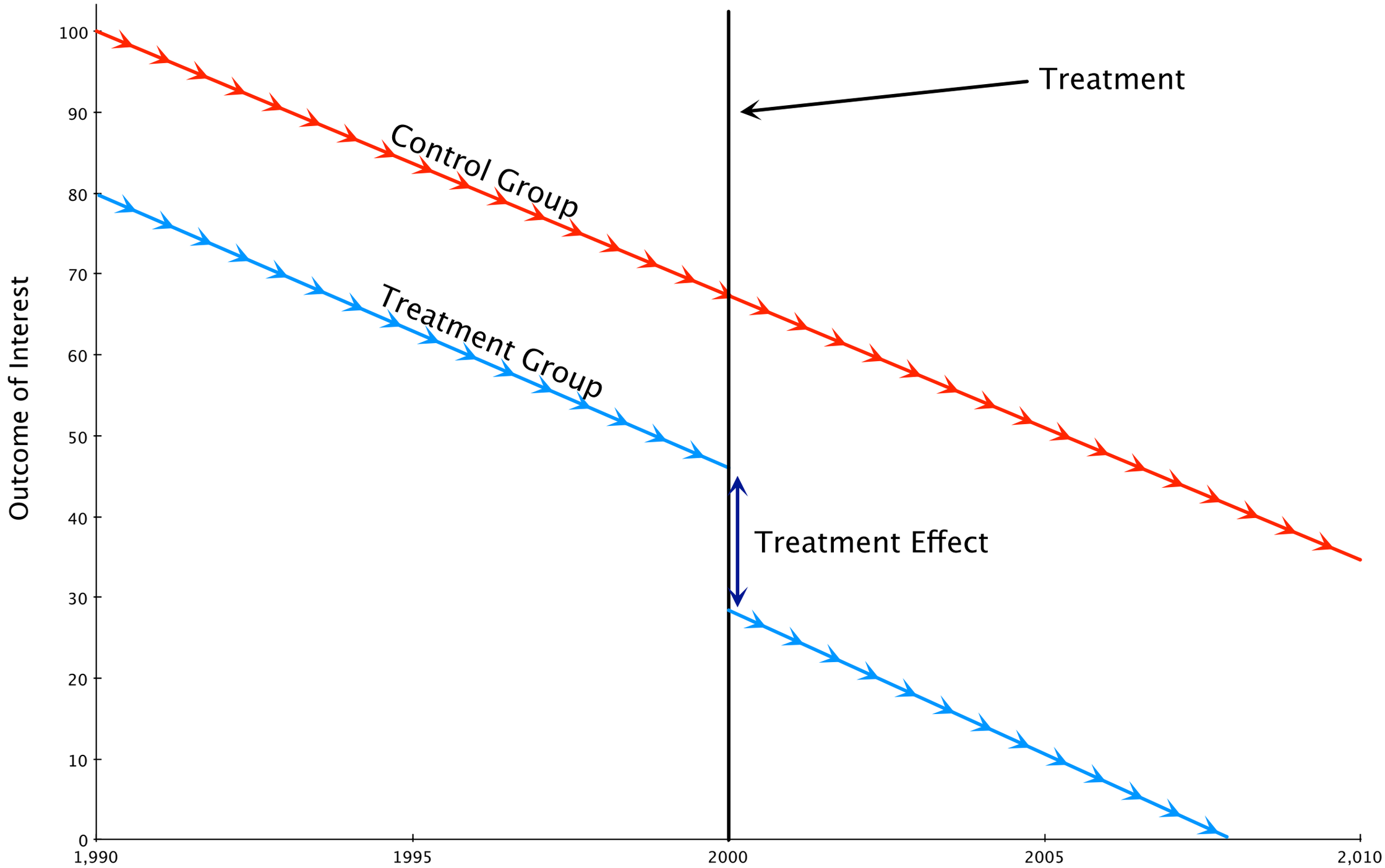
- Schools are not randomly assigned to households.
- You choose what school your kid(s) go to by choosing a house.
- Individuals differ in **observable ways** (e.g. income, education, number of kids)
- Individuals differ in **unobservable ways** (e.g. strength of preference for better schools, ability to relocate).
- These **differences** are likely **correlated with school funding** of after school programs.
- Estimated **difference is not causal**.

Example 5: Difference in Differences Analysis

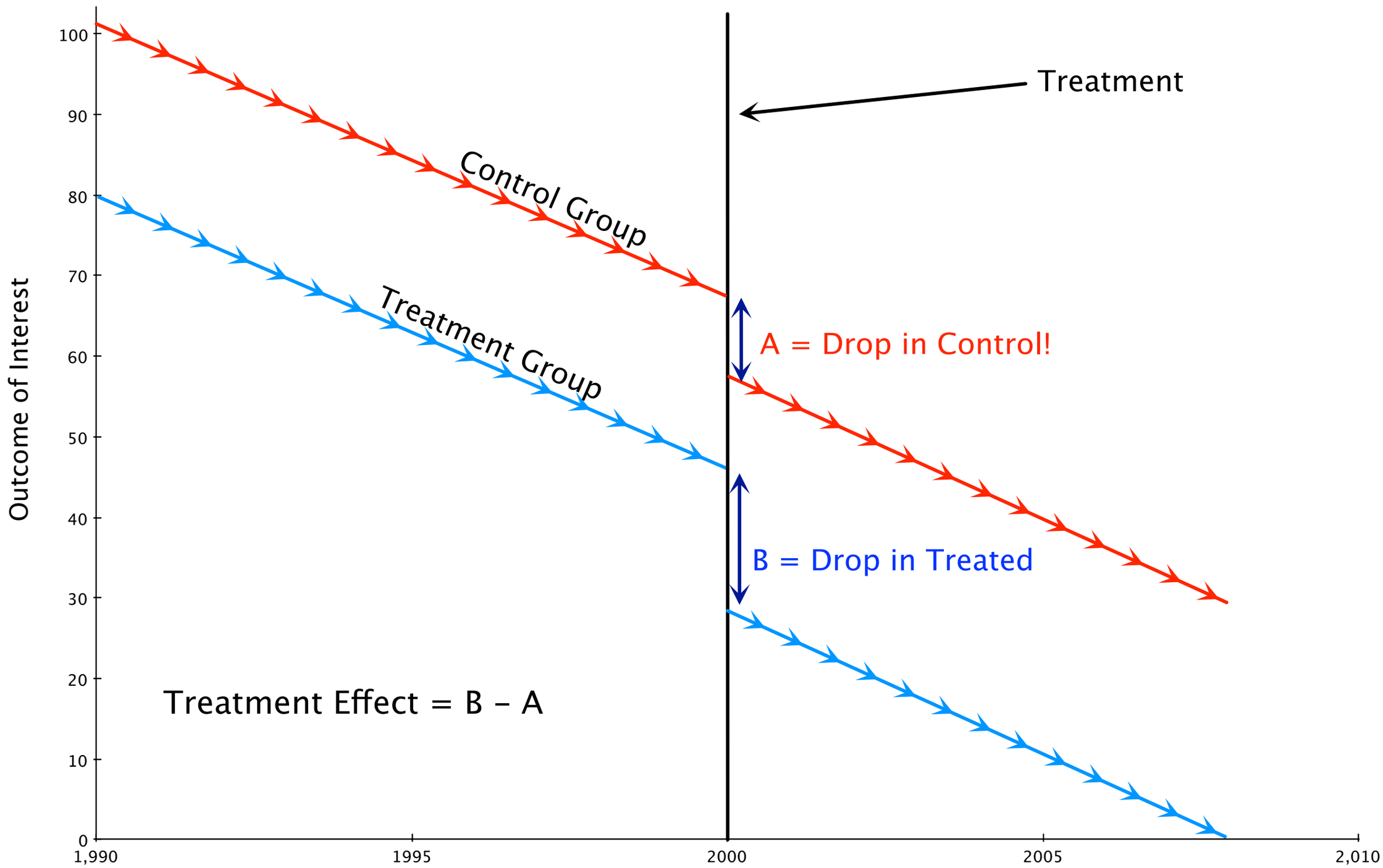
Imagine that you observe two groups over two time periods:

- Neither group receives the treatment in the first period, and only one group receives it in the second period.
- The idea is to calculate the change in outcomes among the treated group between the two periods and then subtract the change in outcomes among the untreated group.

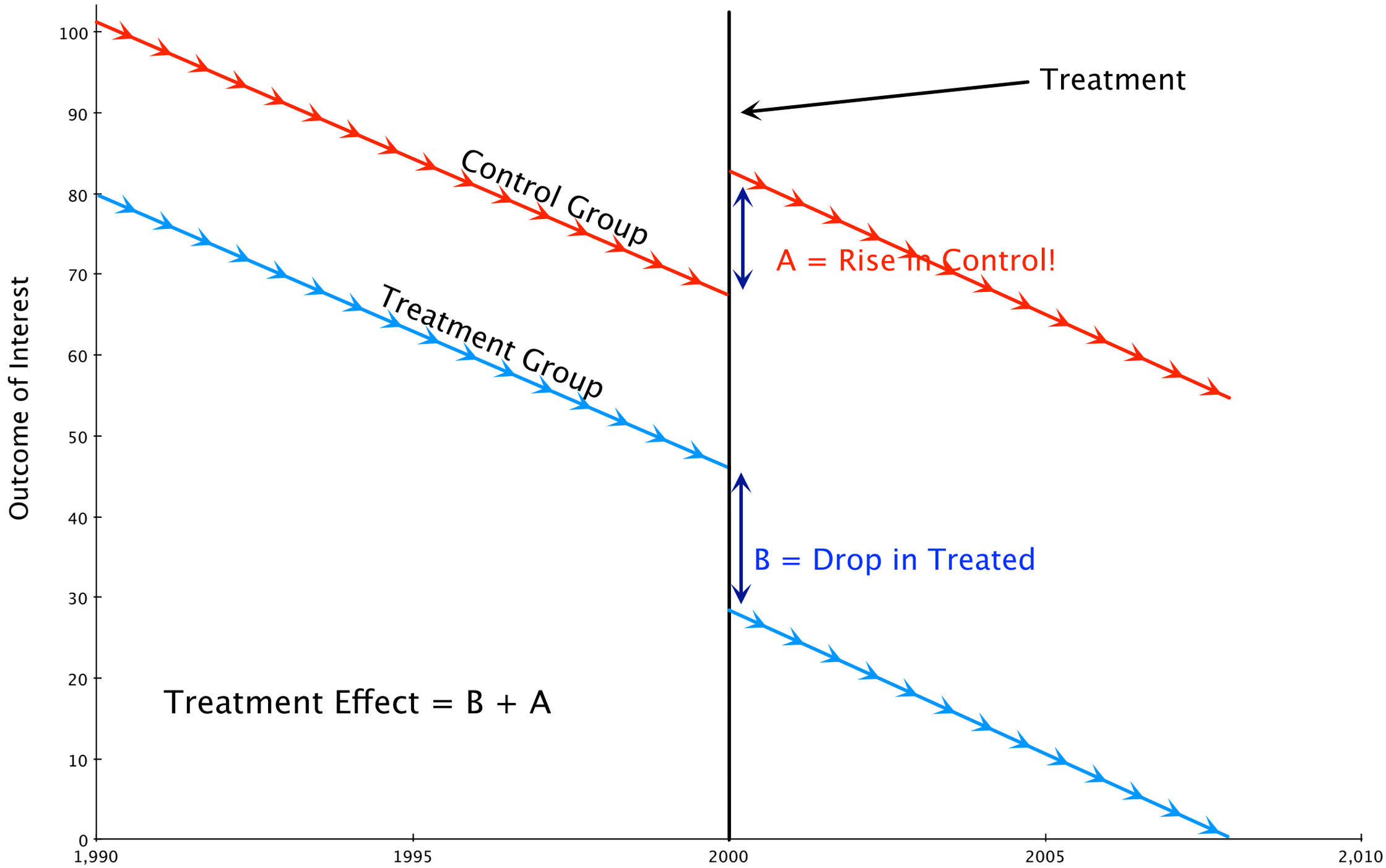
Example 5: Difference in Differences Analysis



Example 5: Difference in Differences Analysis



Example 5: Difference in Differences Analysis



Part 1: Concluding thoughts

- Program evaluation is important (the room is filled with you who have a high opportunity cost of time).
- First and foremost you have to think about what outcome the policy is designed to affect.
- You should then evaluate whether the policy had an impact on said outcome, not an intermediate program input or output.
- There are many flawed comparisons, which do not allow you to estimate a causal effect.
- There are a number of non-RCT approaches, which will allow you to estimate a causal effect if carefully executed.

Part 2: (Quasi) Randomized Experiments

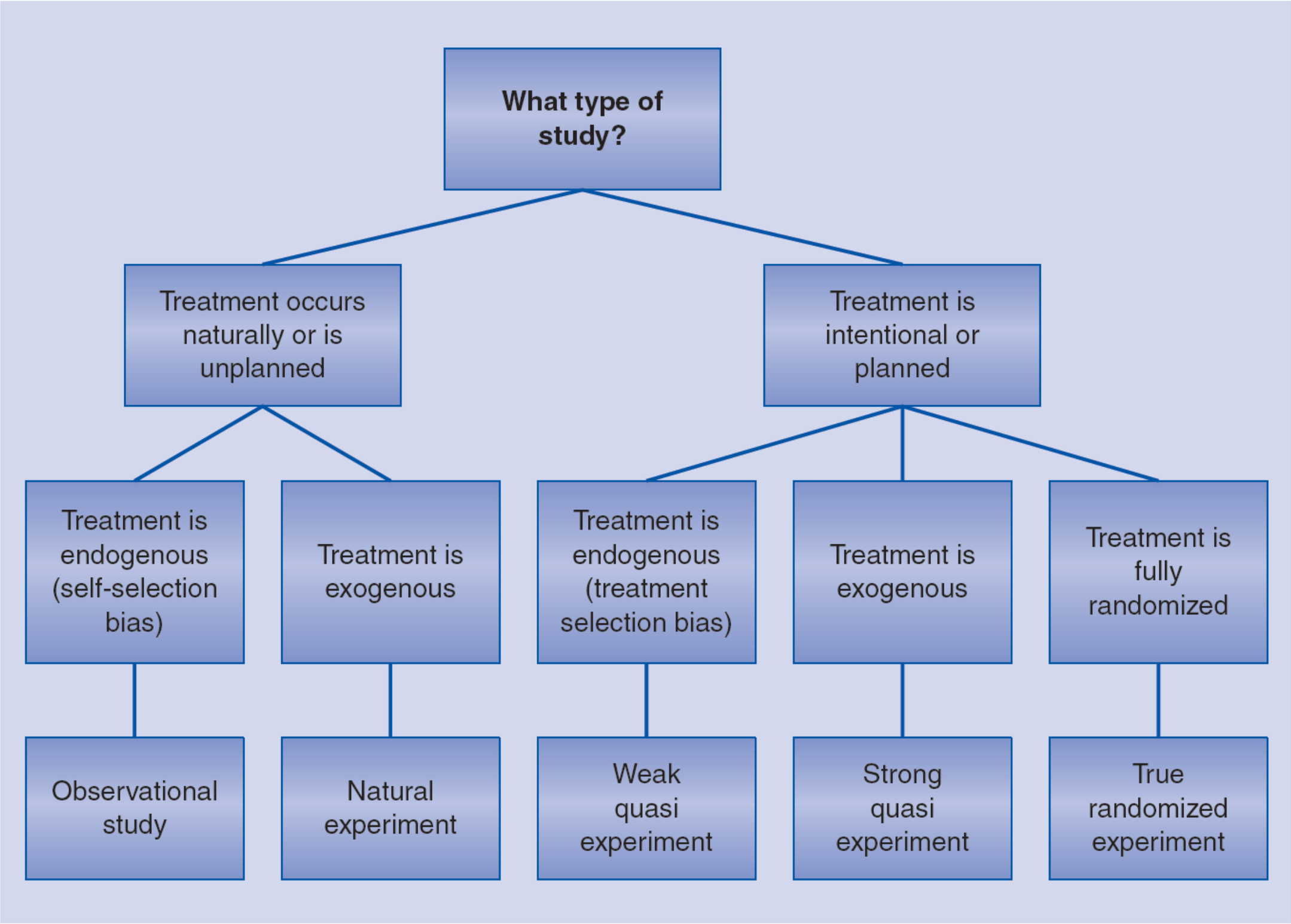
Maximilian Auffhammer
George Pardee Jr. Professor
University of California at Berkeley
National Bureau of Economic Research
CESifo Fellow

NARUC Webinar
May 6th 2015

Motivation: Why should you learn about RCTs?

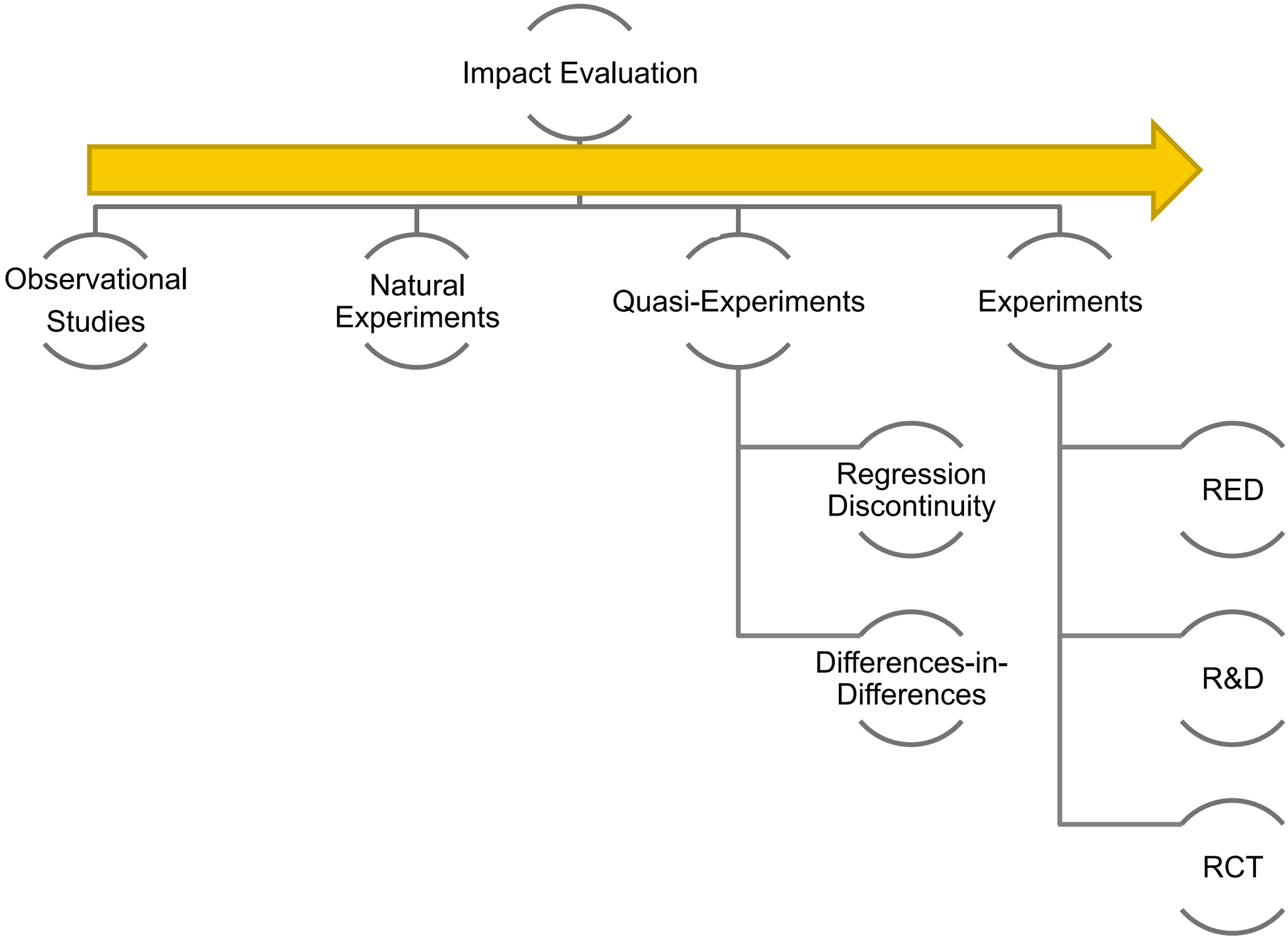
- Because they offer a gold standard to measure the impact of a program
  - Measure the real impact of a program in a simple and clean way
  - Enable you to identify what really works
  - Reveal what's most cost-effective
- The use of experiments is becoming more and more widespread
  - Academia
  - Governments
  - Private Sector
- However, RCTs should be done carefully, as mistakes will lead to meaningless results

Categorizing Studies



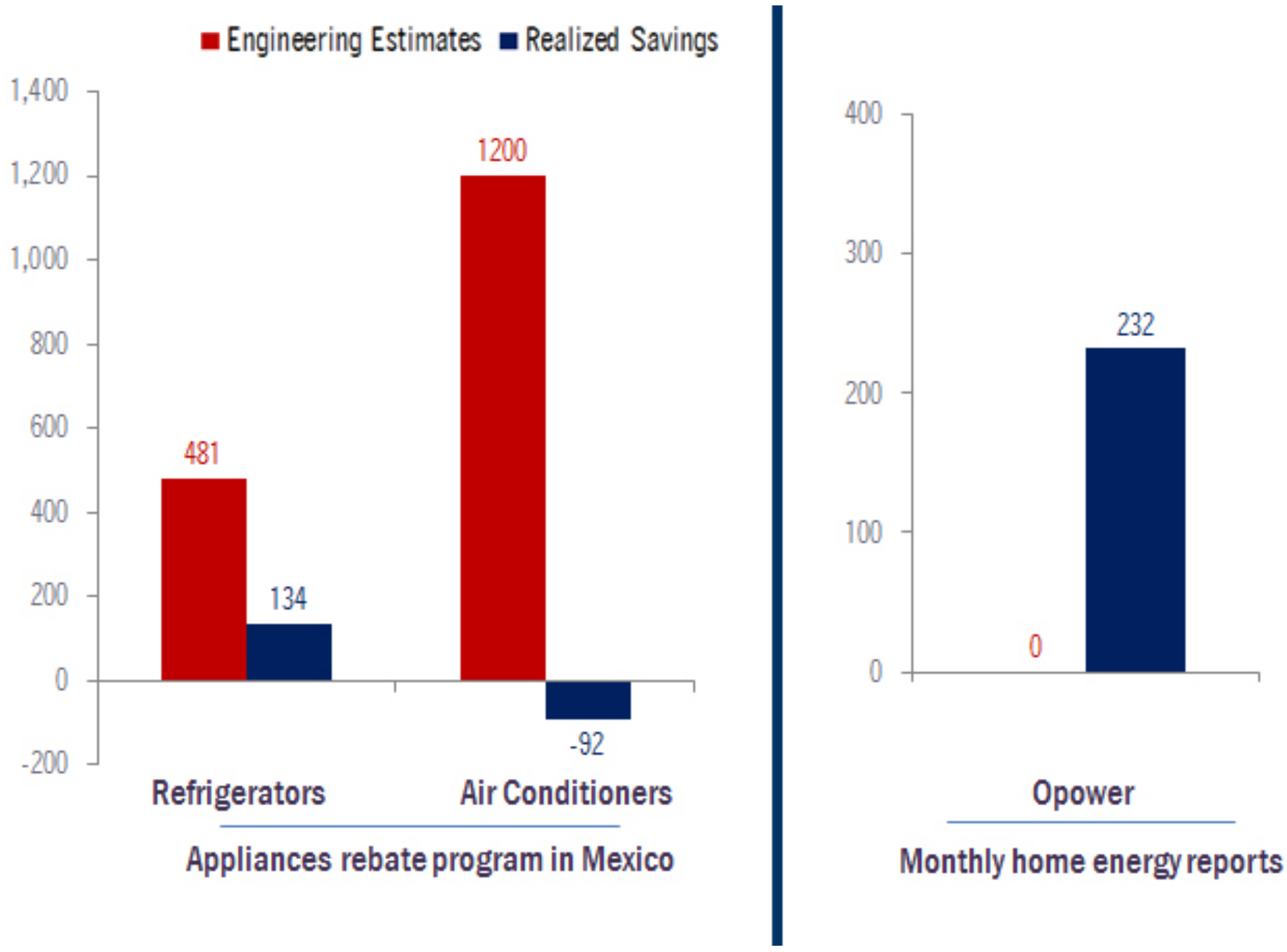
Source: Sage.

Motivation: Credible evaluation is about forming a counterfactual



Motivation: Rigorous measurement can have surprising results

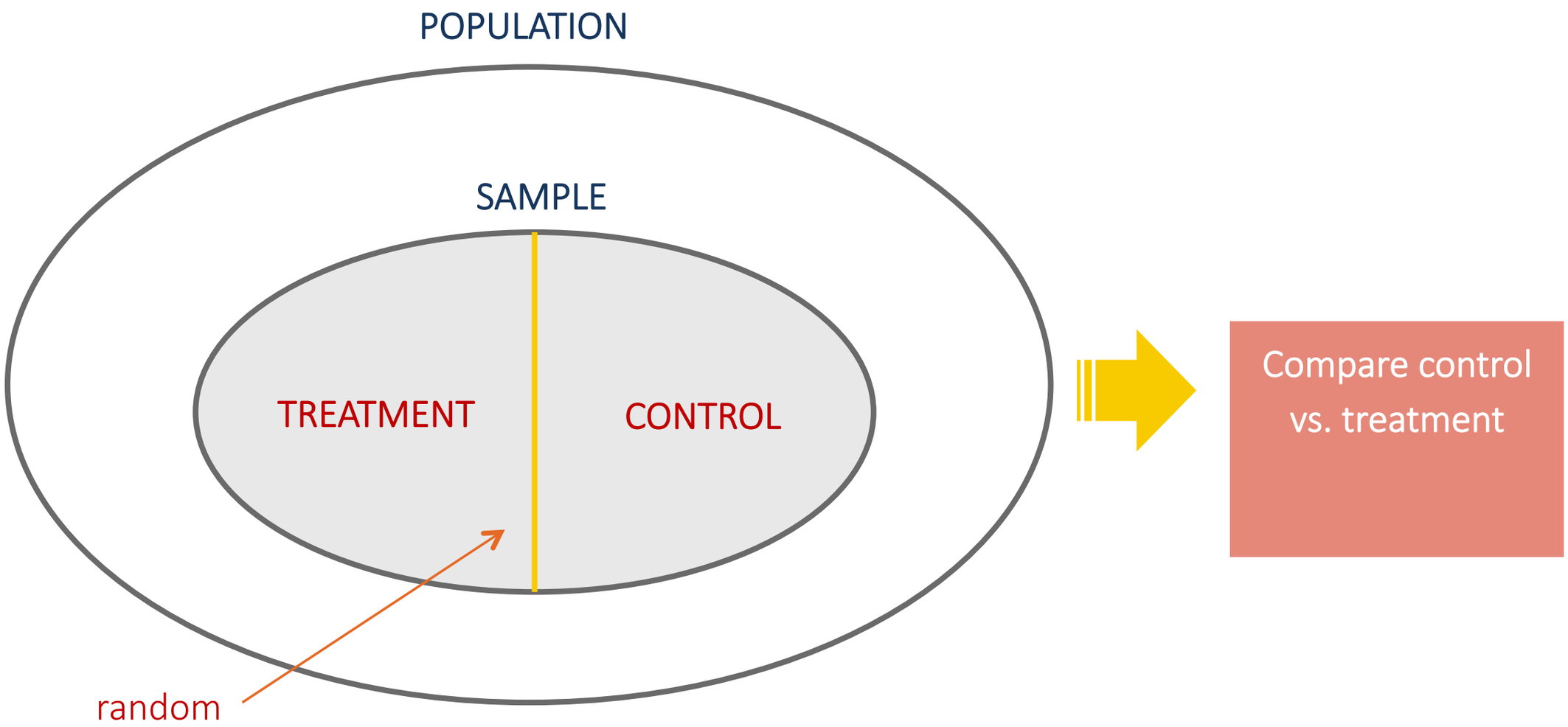
Annual actual vs. estimated savings
kilowatt hours saved per household



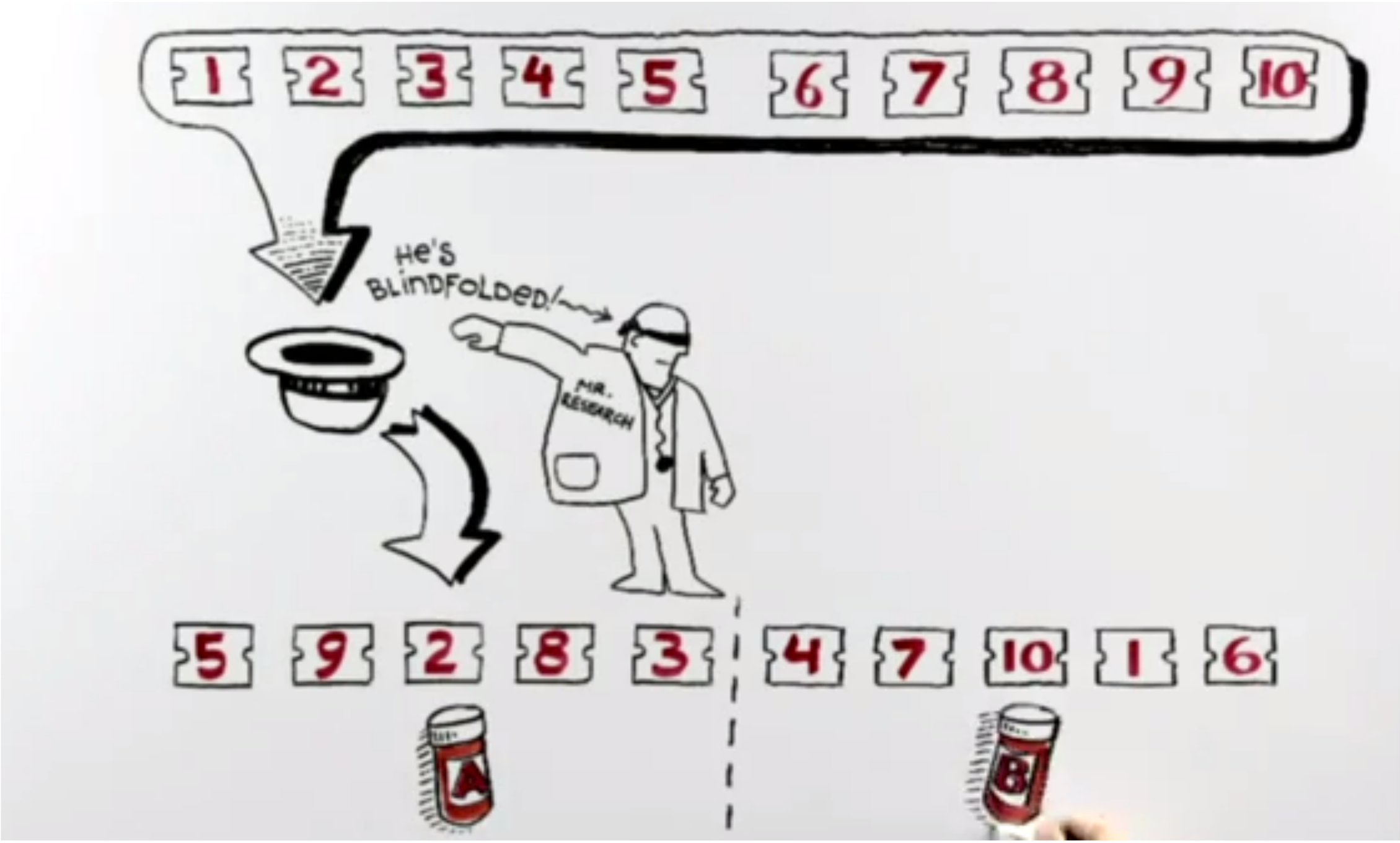
Source: Davis; Alcott.

The framework: What are RCTs all about?

RCT in a single picture



How does randomization work in practice?



Some important things to think about when designing an RCT

Aspects to be considered.

- You need to think about the evaluation ahead of time
- The selection needs to be random
  - Avoid selection bias
- The larger the impact you are expecting, the smaller the sample you need
  - This is called “power”

Aspects to be considered.

- You don't need to have a 50/50 split
  - Can estimate upfront the size of the group you need
  - Especially if treatment is costly

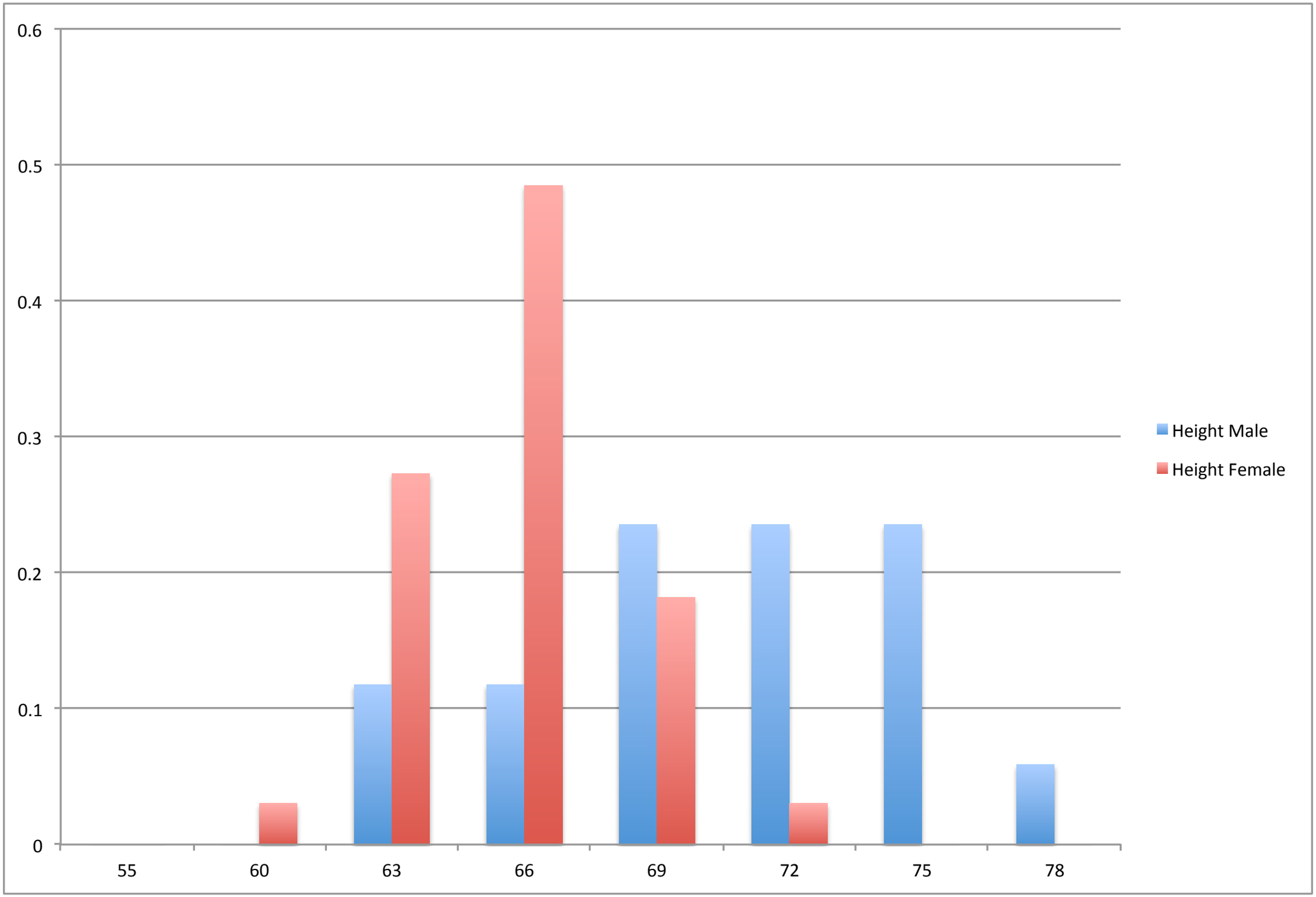
- You can only design treatment around variables that can be manipulated
  - Gender, age, height cannot be changed!

- In many instances, you just can't randomize at the individual level (and that's ok)
  - Impact of teachers with masters degrees on student performance – classroom

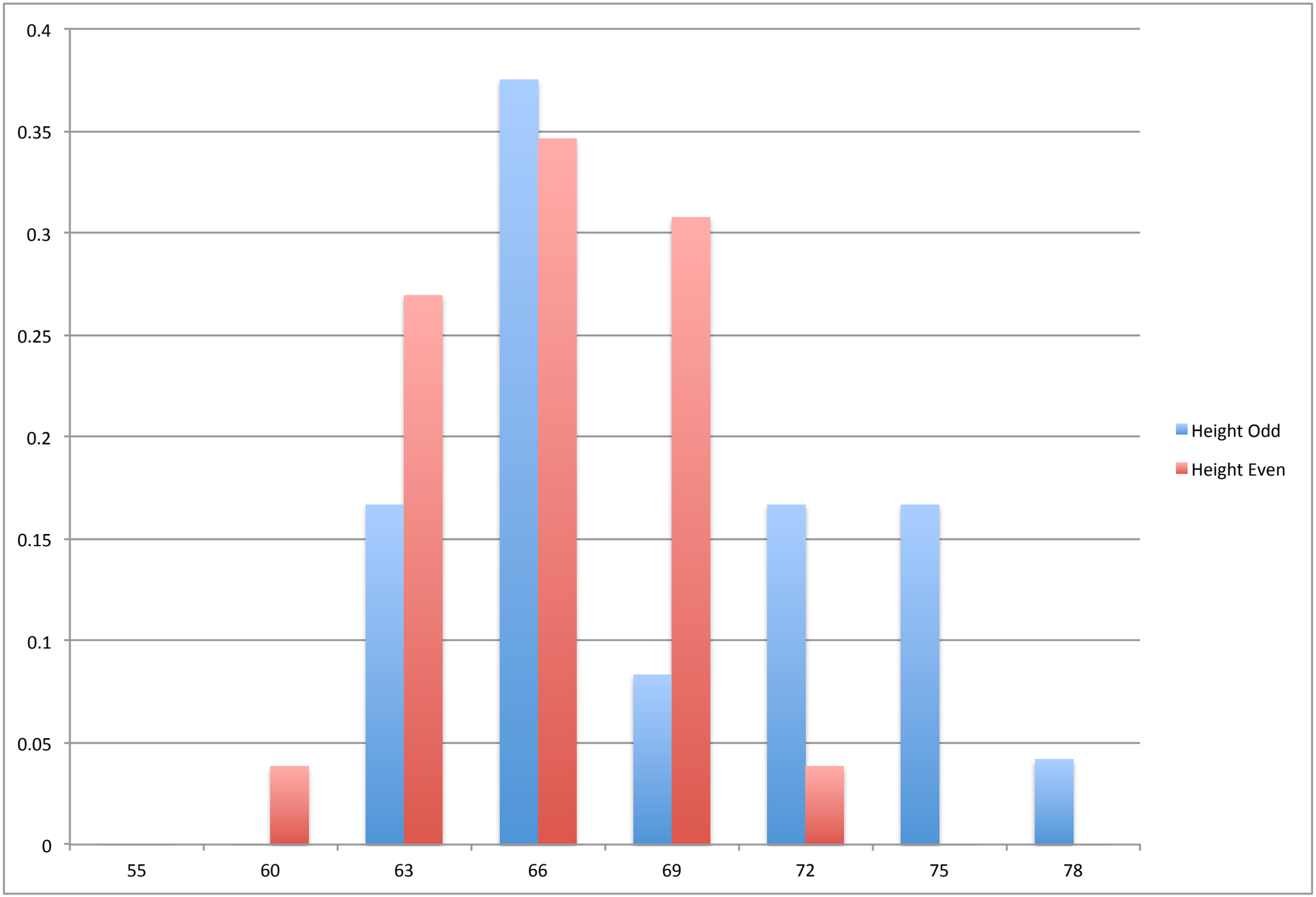
You think some things are random...

- At previous workshop, collected data on height of participants.
- Also collected data on gender and last digit of phone number.

Height distribution by gender



Height distribution by last digit of phone number

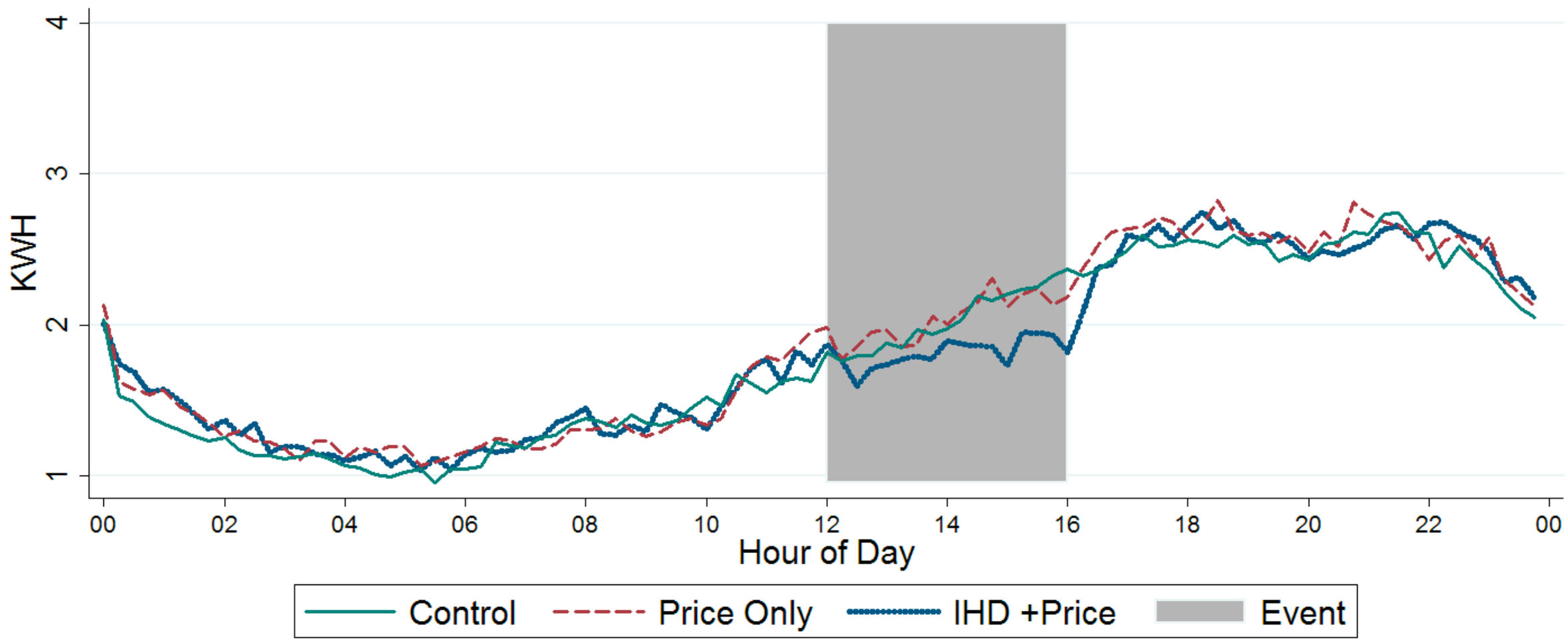


A example of RCT: Impact of TOU pricing and information on consumption

Sometimes treatment cannot be randomized

- Bills appear once a month
- Strange pricing structure
- We do not consume electricity directly
- Even on real time pricing schedules people do not respond much.
- Question: Does information make people more price sensitive
- Identification: RCT of monitors combined with price treatment.

One pictures tells all.



If it only were that simple! Customizing your experiment in more complex situations.

Sometimes treatment cannot be randomized

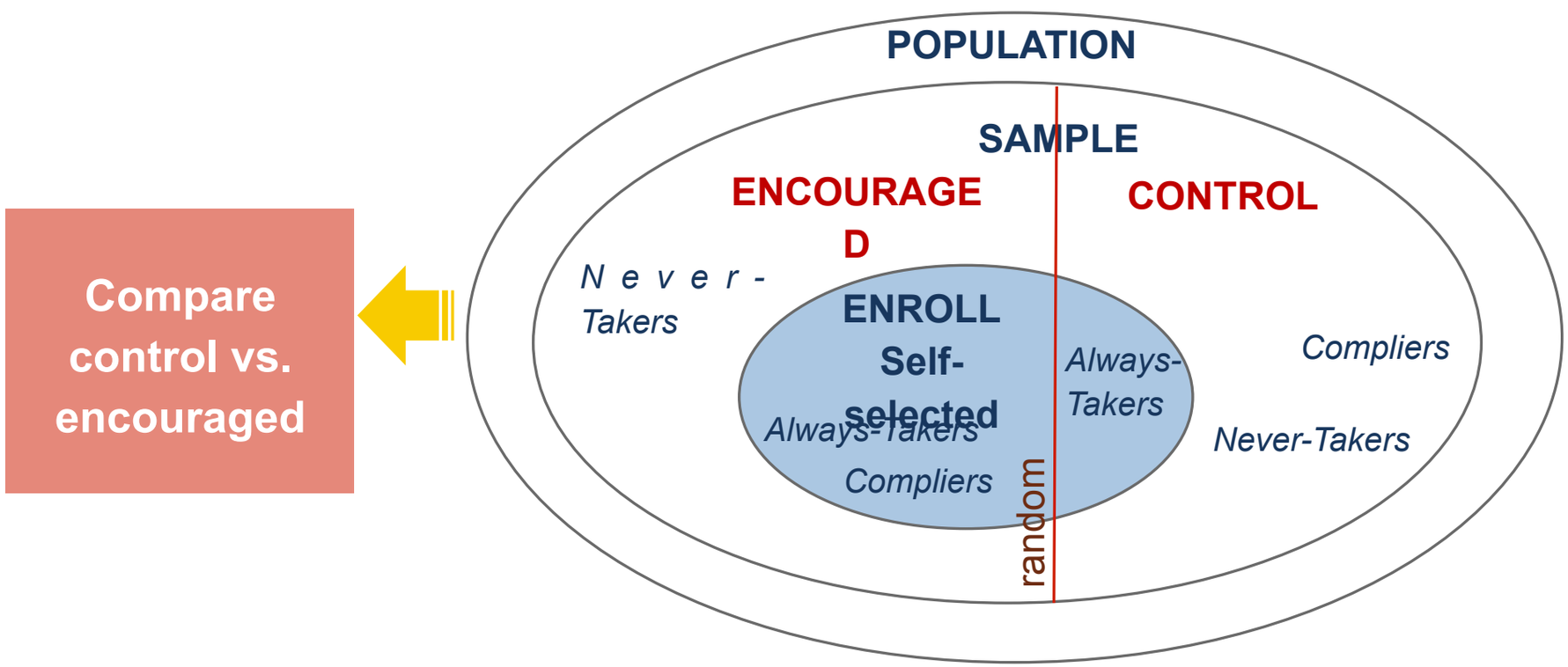
- Mandatory treatment assignment may be infeasible – or inappropriate – in some evaluation contexts.
  - It is available to everyone in the target population (e.g., admission to public school, access to energy efficiency assessments)
  - It is inappropriate to (ultimately) deny treatment to everyone that is interested (e.g., drug trial)

- You still have experimental options
  - Randomized encouragement design
  - Recruit-and-deny (or delay) – over-subscriptions

Randomized Encouragement Design

- Rather than randomize over the intervention itself, we randomly encourage individuals to take the intervention (e.g., a letter informing subject about an available tax benefit)
- Extra advantage of measuring take-up or "cost of acquisition"

Randomized Encouragement Design (RED)



RED Example: Wolfram, Greenstone and Fowlie (2014)

- Partnership with Federal Weatherization Assistance Program (WAP)
- Free energy audits and retrofits up to $6500 to low-income HH
- Seems like "free money" and no brainer to decision for HH

Encouragement Effort	
Encouraged group (households)	8,648
House visits/canvassing	6,694
Number of robo-calls	23,500
Number of personal calls	9,171
Follow up in-person appointments	2,720
Average cost/hh	$55.10

- Experiment Design
- (Randomly selected) treatment HHs were "encouraged" at various steps of the WAP process via outreach about the program
- Outreach is expected to increase probability of WAP take-up in the treatment group

RED Example: Wolfram, Greenstone and Fowlie (2014)

	Control group (20,795)	Encouraged group (7,445)
Submitted an application	535 hh (2.1%)	1517 hh (20 %)
Household inspection/audit	238 hh (0.1%)	494 hh (7%)
Weatherization assistance received!	177 hh (<0.1%)	448 hh (6%)

Note: this table conditions on match with energy consumption data

It's a hard sell

- Even if it is free!
- Suggests large unmeasured costs

Very few free riders

- However, few would have acted in the absence of the encouragements, indicating their value

Low Interest

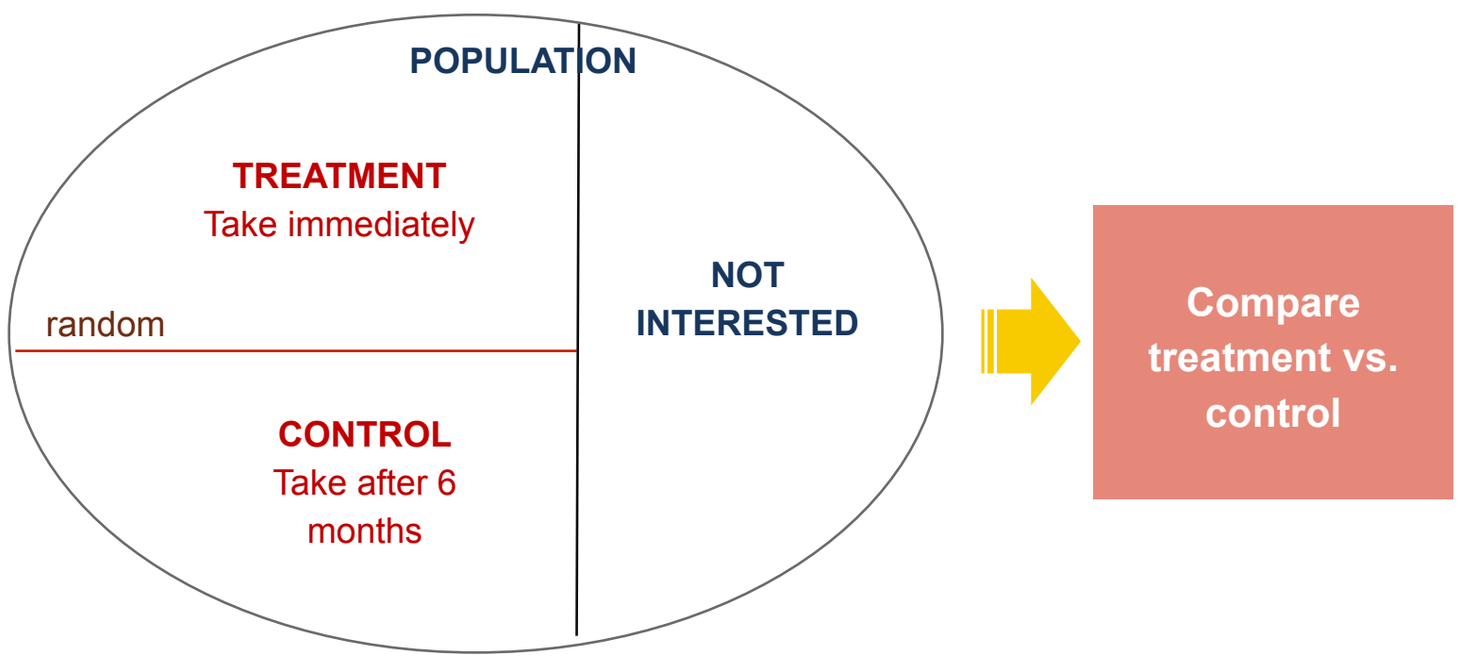
- Vast majority of HH remain uninterested despite rich incentives

Low Energy Savings

- Preliminary estimated suggest ~20-25%, savings, lower than ex ante engineering estimates.

Recruit and Deny/Delay (Oversubscription designs)

- Often used when it is unacceptable to deny treatment to anyone in the target population.
- Researcher randomizes the phasing-in of the treatment
- The (randomly selected) group getting treated later serves as a control group



Oversubscription approaches

PROS

- Results can be analyzed and interpreted in a very simple way;
- We need a smaller sample to detect the impact of the program;
- The sample used will be a true representation of the "usual" clients

CONS

- Results should be interpreted more carefully
- "The impact of the program on the people who are interested in the program"
- May not be generalizable to other states/countries
- Sometimes delaying/ denying is not feasible

What to do when you cannot randomize upfront?

Quasi Experiments

When to use quasi experiments?

- Often randomizations are simply not feasible
  - The program is already concluded
  - Results need to be presented asap
  - The program is running and the evaluation was not planned upfront

- In these cases, quasi-experimental research designs can sometimes be an effective substitute.
  - Data is already being collected (billing data, IRS data)
  - The design of the program might offer interesting opportunities: Thresholds are great!

- The effectiveness of these techniques depends on the exact context and are often found by chance
  - Also a matter of finding a credible counterfactual

- 3 methods:
  - Natural experiments
  - Regression Discontinuity
  - Difference-in-Differences

Natural Experiments - Freedman (1991)

- Sometimes randomization happens without being designed upfront

- London, 1850s Cholera outbreak
  - Postulated that unsanitary water caused cholera not poisonous particles in the air.
  - But how to prove it?
  - Randomly expose people to unsanitary water and measure cholera patients!

- Two water supply companies,
  - One company (Lambeth) drew water upstream of the sewage discharge points in the River Thames
  - The other (Southwark and Vauxhall) drew water downstream of the discharge points.
  - Plausible random assignment to houses

Natural Experiments - Freedman (1991)

Table 3: Snow's Table IX

	Number of Houses	Deaths from Cholera	Deaths per 10,000 Houses
Southwark and Vauxhall	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

Notes: Source: Freedman (1991).

Regression Discontinuity

- Regression discontinuity (RD) is a method of constructing a counterfactual that exploits circumstances created by a threshold
- For people close to the threshold the probability of assignment to the treatment group is a discontinuous function of one or more observed variables.

- Example: impact of subsidies for appliance replacement in Mexico on electricity consumption
- Clear threshold to determine which households were eligible

Regression Discontinuity

- Almond, Doyle, Kowalski, and Williams (QJE 2010) are interested in estimating the returns to medical care on infant health
  - Medical care is not randomly assigned
  - Observational studies are likely biased for this reason

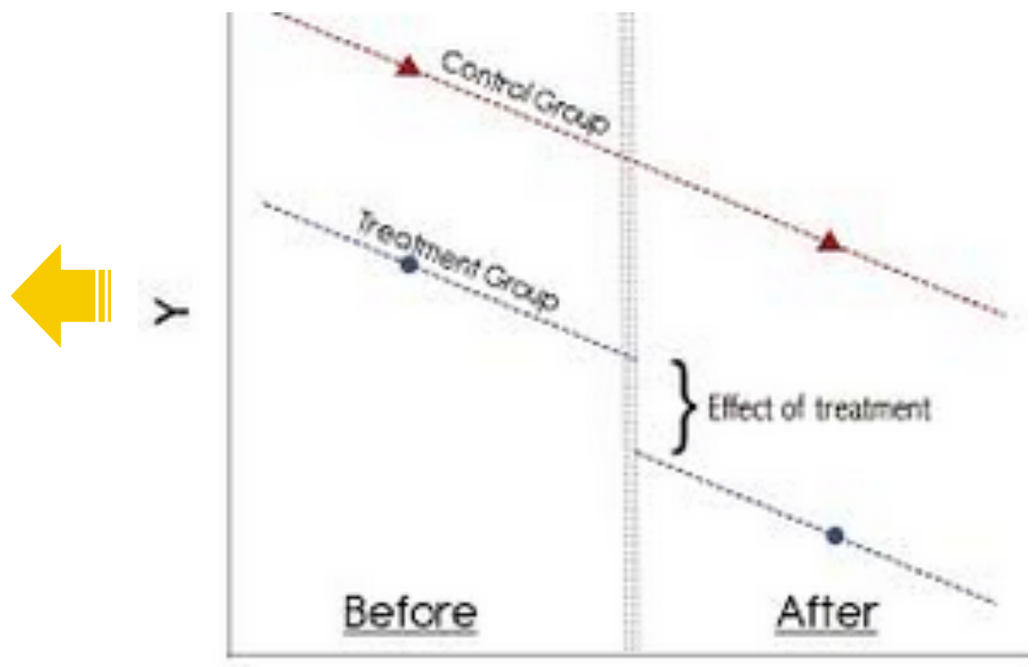
- However, there is a discontinuity in treatment at birth weight of 1500 grams
  - Those born weighing less are provided more treatment
  - If there is nothing "magic" about 1500 grams, those just below and those just above will be similar on observables and un-observables

Differences in differences

- In the standard example, there are two groups.
- Neither group receives the treatment in the first period, and only one group receives it in the second period.
- The idea is to calculate the change in outcomes among the treated group between the two periods and then subtract the change in outcomes among the untreated group.

Differences-in-Differences (DD)

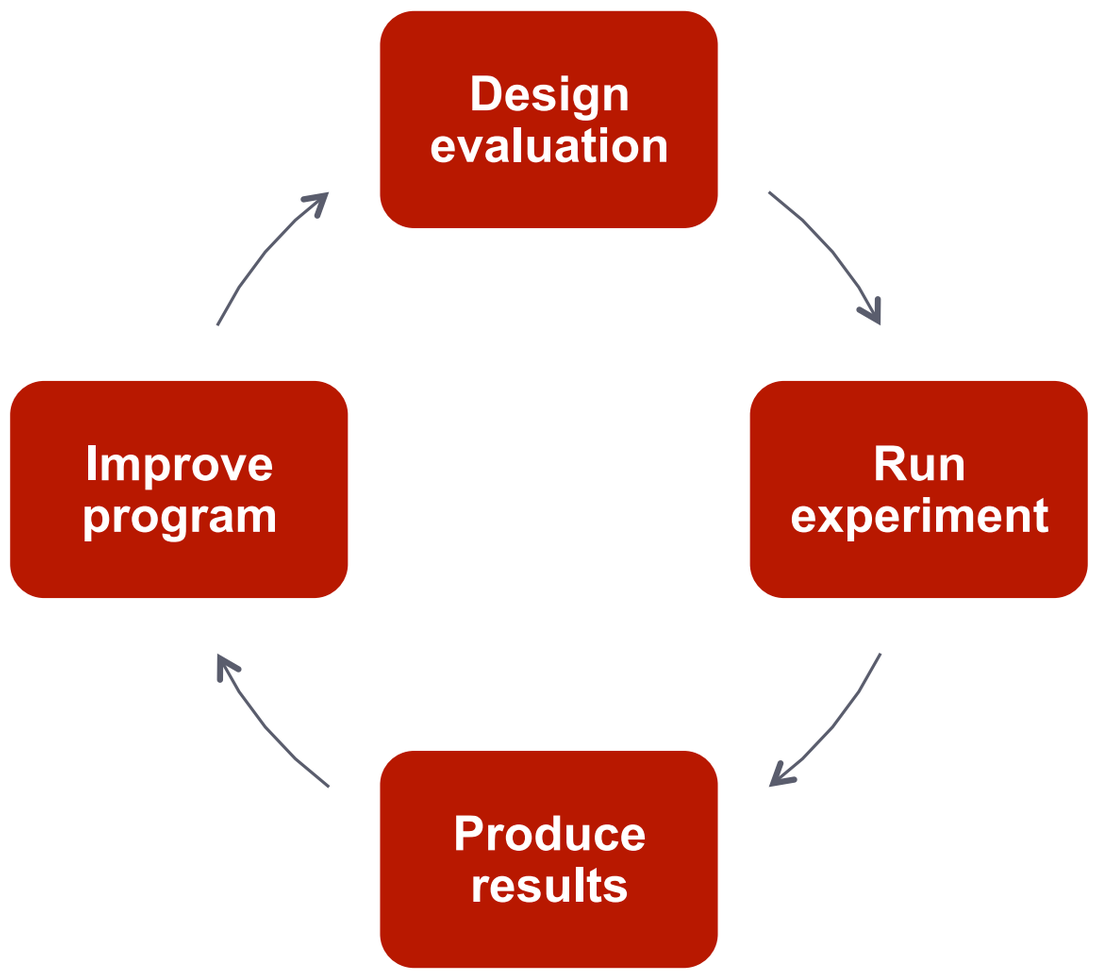
As long as pre-trends are the same



Concluding remarks: What makes a successful RCT?

Let's start at the very beginning

- Building an RCT into the program from the beginning can foster a virtuous cycle of evaluation



Pilot programs are a great opportunity for RCTs

- Why randomize your next pilot?
  - Because you want the pilot to be a representative sample of your population in order to discover the true potential impact
  - Because it's a fair way to select the participants for your pilot – everyone has an equal chance of getting in
  - To test the logistics of implementation and refine
  - To get a first causal assessment of treatment impacts and understand costs and benefits of scaling up

Thanks for listening!

auffhammer@berkeley.edu