

"Impact Evaluation: Reasons and Methods"



Emmanuel Skoufias
The World Bank
PRMPR

Cuarta Reunion de la Red de Monitoreo y Evaluacion
Bello Horizonte, Brasil
November 24-25, 2008

Outline of presentation

1. Why Evaluate?
2. The Evaluation Problem
3. Impact Evaluation Methods
 - Cross- Sectional Estimator
 - Before and After
 - Double Difference
 - Experimental
 - Quasi-Experimental: RDD

1. Why Evaluate

Why Evaluate?

- Need evidence on what works
 - Limited budget forces choices
 - Bad policies could hurt
- Improve program/policy implementation
 - Design: eligibility, benefits
 - Operations: efficiency & targeting
 - Management tool to improve operations
- Information key to sustainability
 - Budget negotiations
 - Informing public opinion and press
- Results agenda & aid effectiveness

Allocate limited resources

- Benefit-Cost analysis
 - Comparison of choices
 - Highest return investment
- Benefit:
 - Change in outcome indicators
 - Measured through impact evaluation
- Cost:
 - Additional cost of providing benefit
 - Economic versus accounting costs

What kinds of questions does IE answer?

- What is effect of program on outcomes?
- How much better off are beneficiaries because of the intervention?
- How would outcomes change under alternative program designs?
- Does the program impact people differently (e.g. females, poor, minorities)
- Is the program cost-effective?
- Traditional M&E cannot answer these

For Example IE Answers...

- What is effect of scholarships on school attendance & performance (test scores)?
- Does contracting out primary health care lead to an increase in access?
- Does replacing dirt floors with cement reduce parasites & improve child health?
- Do improved roads increase access to labor markets & raise income

Types of Impact Evaluation

➤ Efficacy:

- Proof of Concept
- Pilot under ideal conditions

➤ Effectiveness:

- Normal circumstances & capabilities
- Impact will be lower
- Impact at higher scale will be different
- Costs will be different as there are economies of scale from fixed costs

Use impact evaluation to....

- Scale up pilot-interventions/programs
- Adjust program design
- Kill programs
- Inform (i.e. Finance & Press)
- e.g. PROGRESA/OPORTUNIDADES (Mexico)
 - Transition across presidential terms
 - Expansion to 5 million households
 - Change in benefits
 - Inform the press
 - Worldwide public good (Brazil vs Mexico)

2. Evaluation Problem

How to assess impact

➤ What is beneficiary's test score with program compared to without program?

➤ Formally, program impact is:

$$E(Y \mid T=1) - E(Y \mid T=0)$$

➤ Compare same individual with & without programs at same point in time

➤ So what's the Problem?

Solving the evaluation problem

- **Problem:** we never observe the same individual with and without program at same point in time
 - Observe: $E(Y | T=1)$ & $E(Y | T=0)$ → NO!
- **Solution:** estimate what would have happened if beneficiary had not received benefits
 - Observe: $E(Y | T=1)$ → YES!
 - Estimate: $E(Y | T=0)$ → YES!!

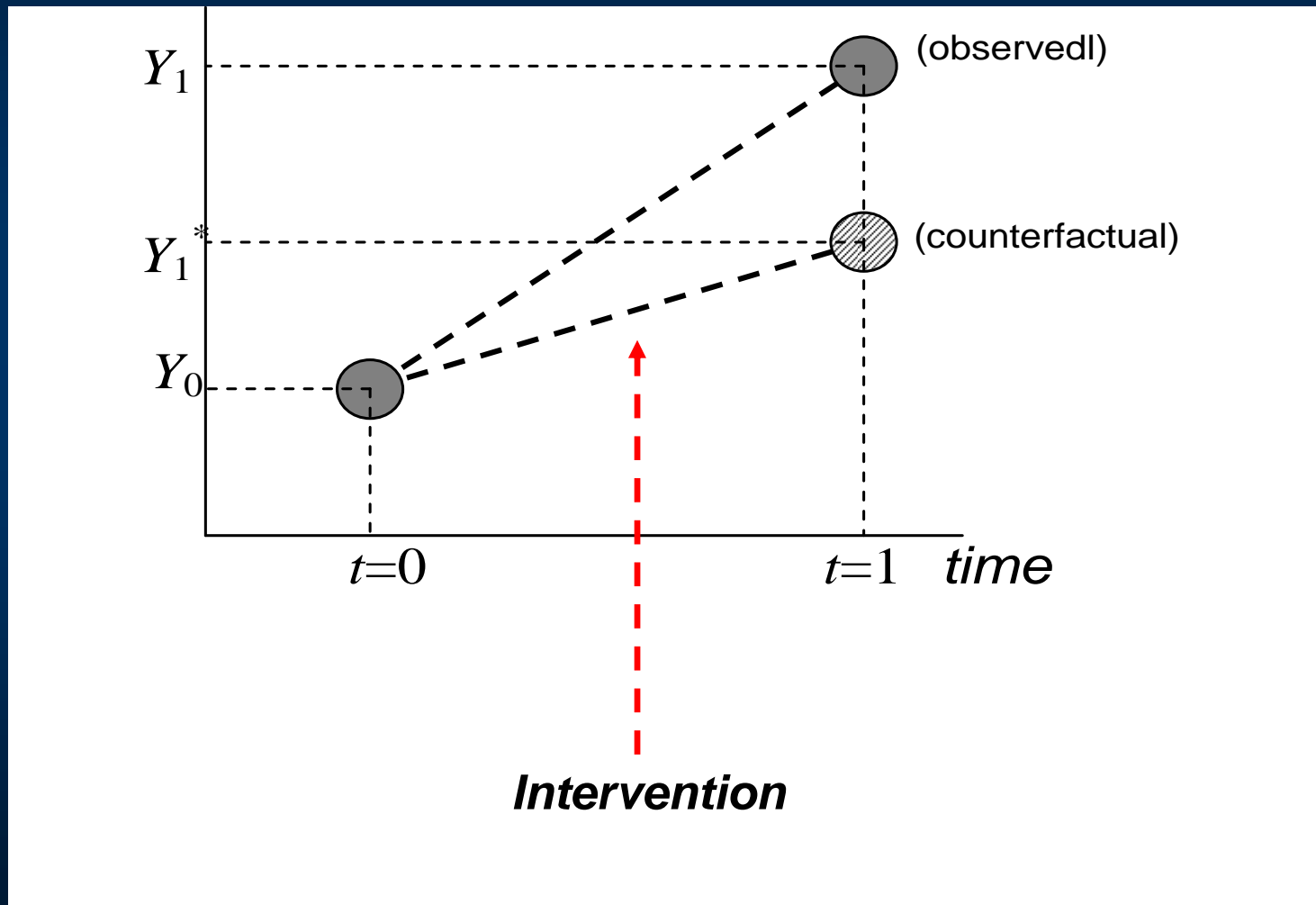
Solving the evaluation problem

- Counterfactual: what would have happened without the program
- Estimated impact is difference between treated observation and counterfactual
- Never observe same individual with and without program at same point in time
- Need to estimate counterfactual
- Counterfactual is key to impact evaluation

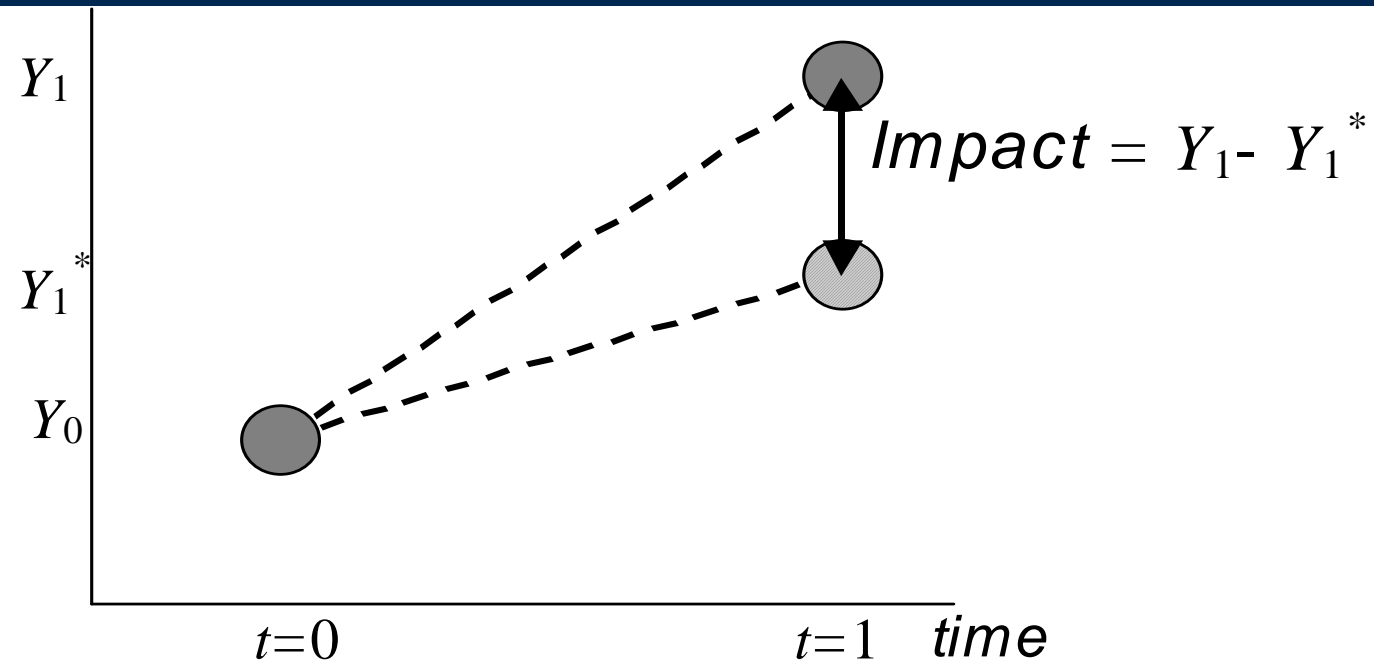
Finding a good counterfactual

- Treated & counterfactual
 - have identical characteristics,
 - except for benefiting from the intervention
- No other reason for differences in outcomes of treated and counterfactual
- Only reason for the difference in outcomes is due to the intervention

Having the "ideal" counterfactual.....



allows us to estimate the true impact



Comparison Group Issues

- Two central problems:
 - Programs are targeted
 - Program areas will differ in observable and unobservable ways precisely because the program intended this
 - Individual participation is (usually) voluntary
 - Participants will differ from non-participants in observable and unobservable ways (selection based on observable variables such as age and education and unobservable variables such as ability, motivation, drive)
- Hence, a comparison of participants and an arbitrary group of non-participants can lead to heavily biased results

The Challenge

- Matching the Evaluation Design to the evaluation's
 - purpose,
 - resources,
 - and timeline to optimize use

Impact Evaluation methods

Differ in how they construct the counterfactual

- Cross sectional Differences
- Before and After (Reflexive comparisons)
- Difference in Difference (Dif in Dif)
- Experimental methods/Randomization
- Quasi-experimental methods
 - Propensity score matching (PSM) (not discussed)
 - Regression discontinuity design (RDD)
- Econometric methods
 - Instrumental variables (not discussed)
 - Encouragement design (not discussed)

Cross-Sectional Estimator

- Counterfactual for participants: Non-participant in the same village or hh in similar villages

- But then:

$$\text{Measured Impact} = E(Y | T=1) - E(Y | T=0) = \text{True Impact} + \text{MSB}$$

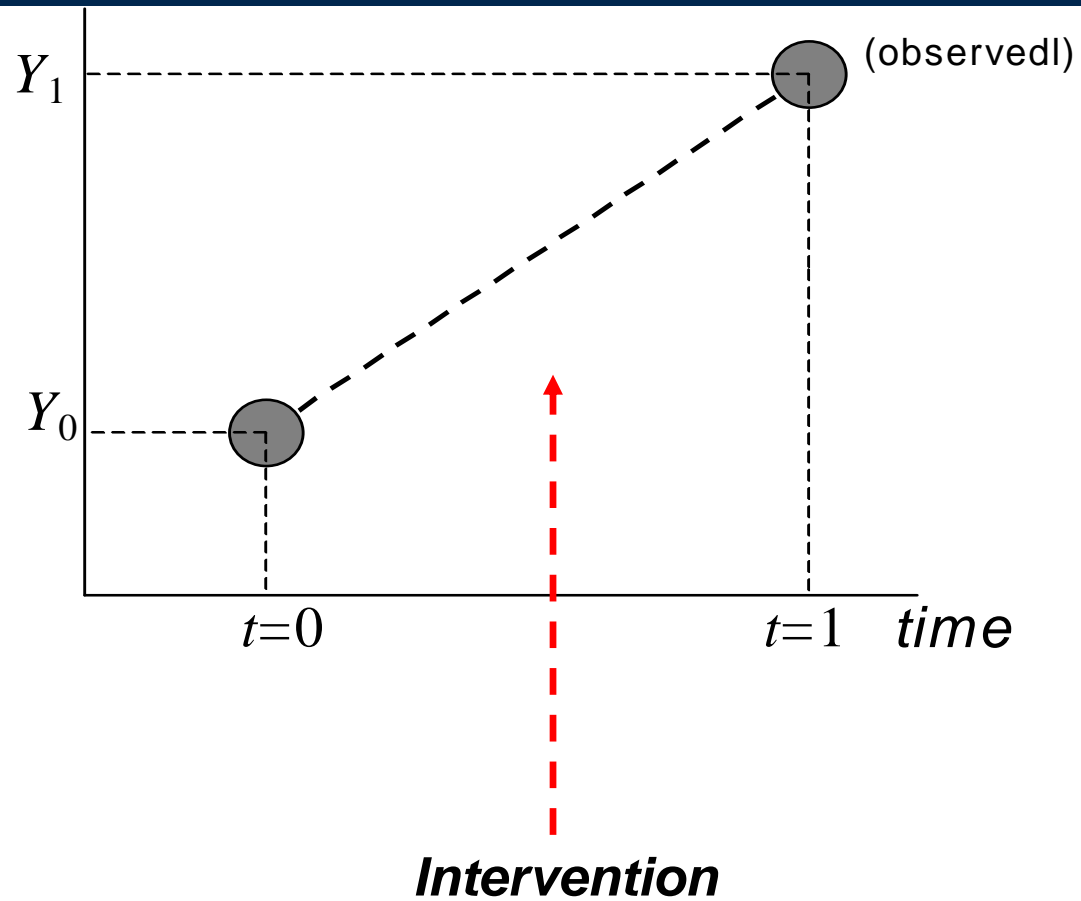
where MSB=Mean Selection Bias = $MA(T=1) - MA(T=0)$

- If $MA(T=1) > MA(T=0)$ then $MSB > 0$ and measured impact $>$ true impact
- **Note: An Experimental or Randomized Design**
 - Assigns individuals into T=1 and T=0 groups randomly.
 - Consequence: $MA(T=1) = MA(T=0) \rightarrow MSB=0$ and
 - Measure Impact = True Impact

Before and After Estimator

- Counterfactual for participants: the participants themselves before the start of the program
- **Steps:**
 - Collect baseline data on potential participants before the program
 - Compare with data on the same individuals (villages) after the program
 - Take the difference (after – before) or use a regression with a dummy variable identifying round 2 obs
- This allows for the presence of selection bias assuming it is time invariant and enters additively in the model

Before and After Estimator



Shortcomings of Before and After (BA) comparisons

- Not different from “Results Based” Monitoring
- over/under-estimates impacts
- $\text{Measured Impact} = \text{True Impact} + \text{Trend}$
 - Attribute all changes over time to the program (i.e. assume that there would have been **no trend**, or no changes in outcomes in the absence of the program)
- Note: Difference in difference may be thought as a method that tries to improve upon the BA method

Difference-in-difference:

- Counterfactual for participants: Observed changes over time for non-participants
- **Steps:**
 - Collect baseline data on non-participants and (probable) participants before the program.
 - **Note: there is no particular assumption about how the non-participants are selected. Could use arbitrary comparison group**
 - Or could use comparison group selected via PSM/RDD
 - Compare with data after the program.
 - Subtract the two differences, or use a regression with a dummy variable for participant.
- This allows for selection bias but it must be time-invariant and additive.

Difference-in-difference: Interpretation 1

- Dif-in-Dif removes the trend effect from the estimate of impact using the BA method
 - True impact= Measured Impact in Treat G (or BA)– Trend
- The change in the control group provides an estimate of the trend. Subtracting the “trend” from the change in the treatment group yields the true impact of the program
 - The above assumes that the trend in the C group is an accurate representation of the trend that would have prevailed in the T group in the absence of the program. That is an assumption that cannot be tested (or very hard to test).
 - What if the trend in the C group is not an accurate representation of the trend that would have prevailed in the T group in the absence of the program?? Need observations on Y one period before the baseline period.

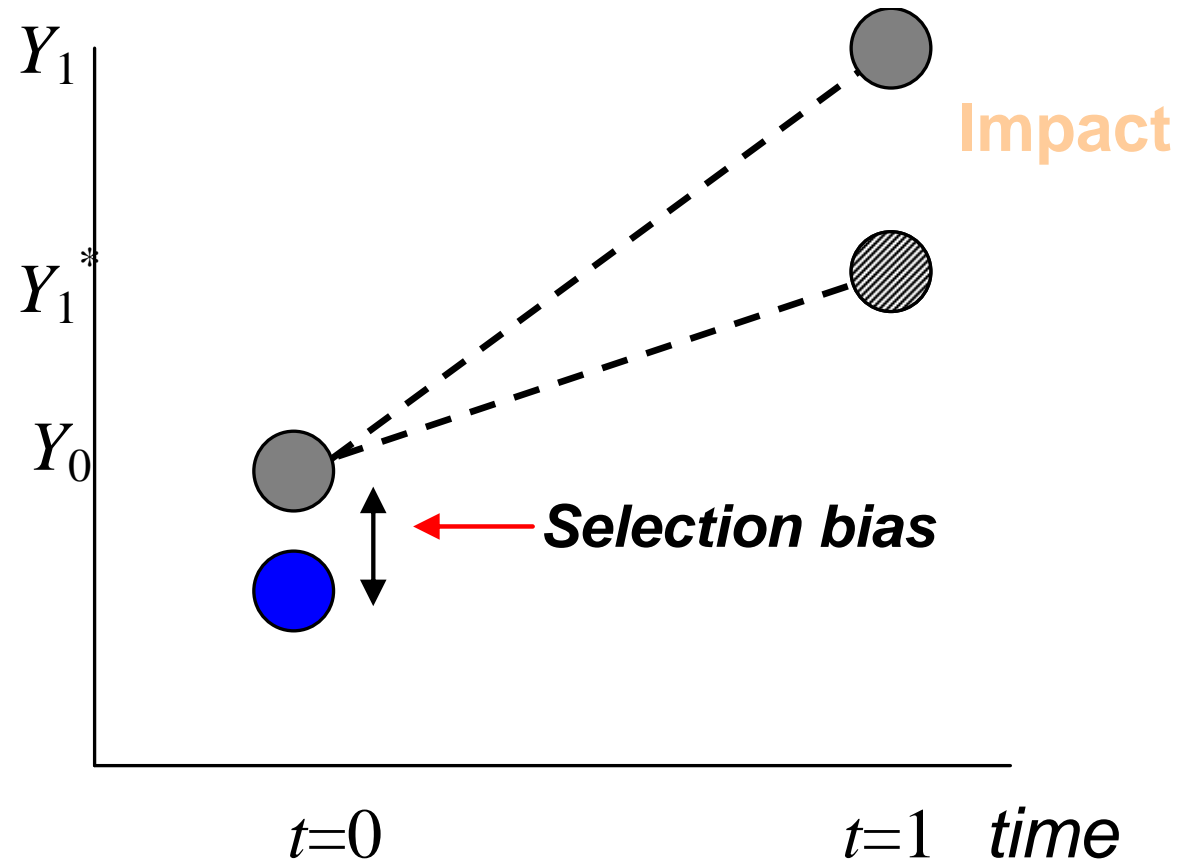
$$(Y^T - Y^C)_{t=1} - (Y^T - Y^C)_{t=0} = (Y_{t=1}^T - Y_{t=0}^T) - (Y_{t=1}^C - Y_{t=0}^C) = \text{Measured Impact} - \text{Trend}$$

Difference-in-difference: Interpretation 2

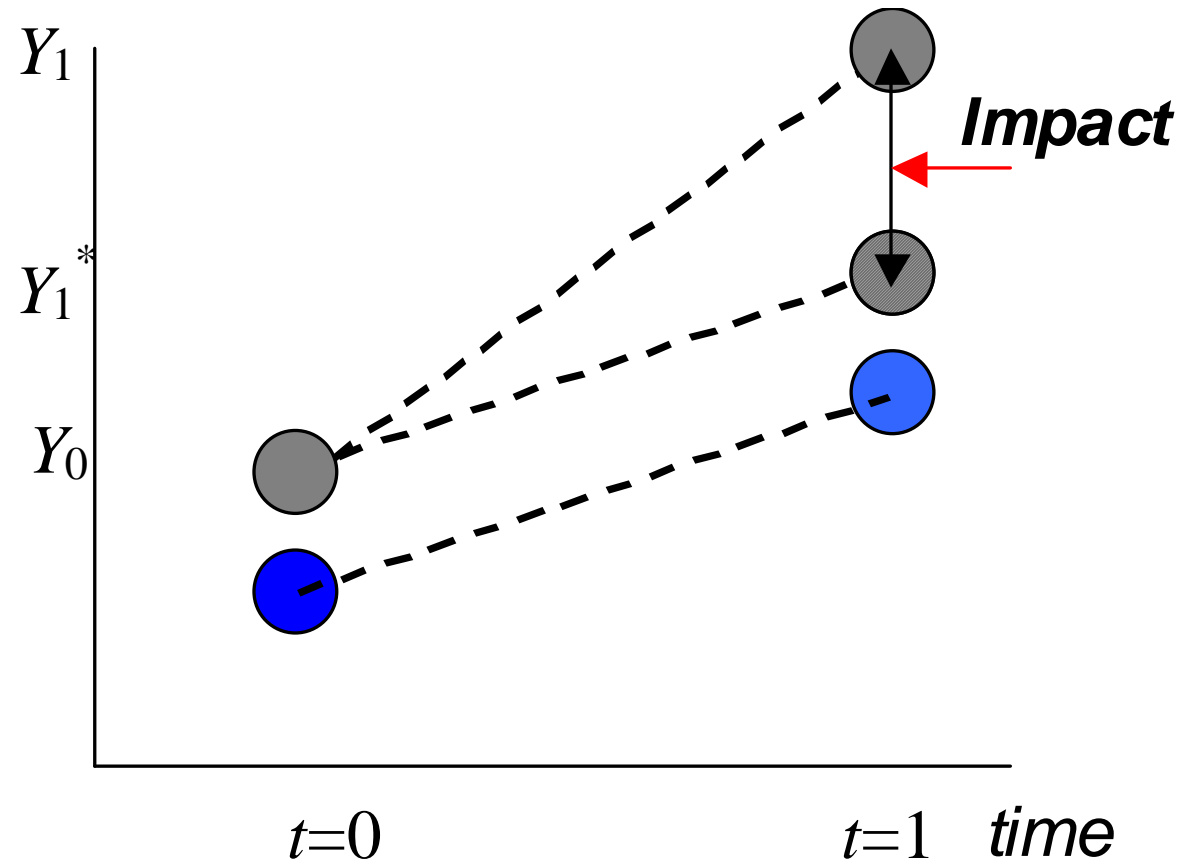
- Dif-in-Dif estimator eliminates selection bias under the assumption that selection bias enters additively and does not change over time

$(Y^T - Y^C)_{t=1} - (Y^T - Y^C)_{t=0} = \text{True impact} - (MSB_{t=1} - MSB_{t=0})$. The latter term drops out if $MSB_{t=1} = MSB_{t=0}$, i.e. MSB is time invariant

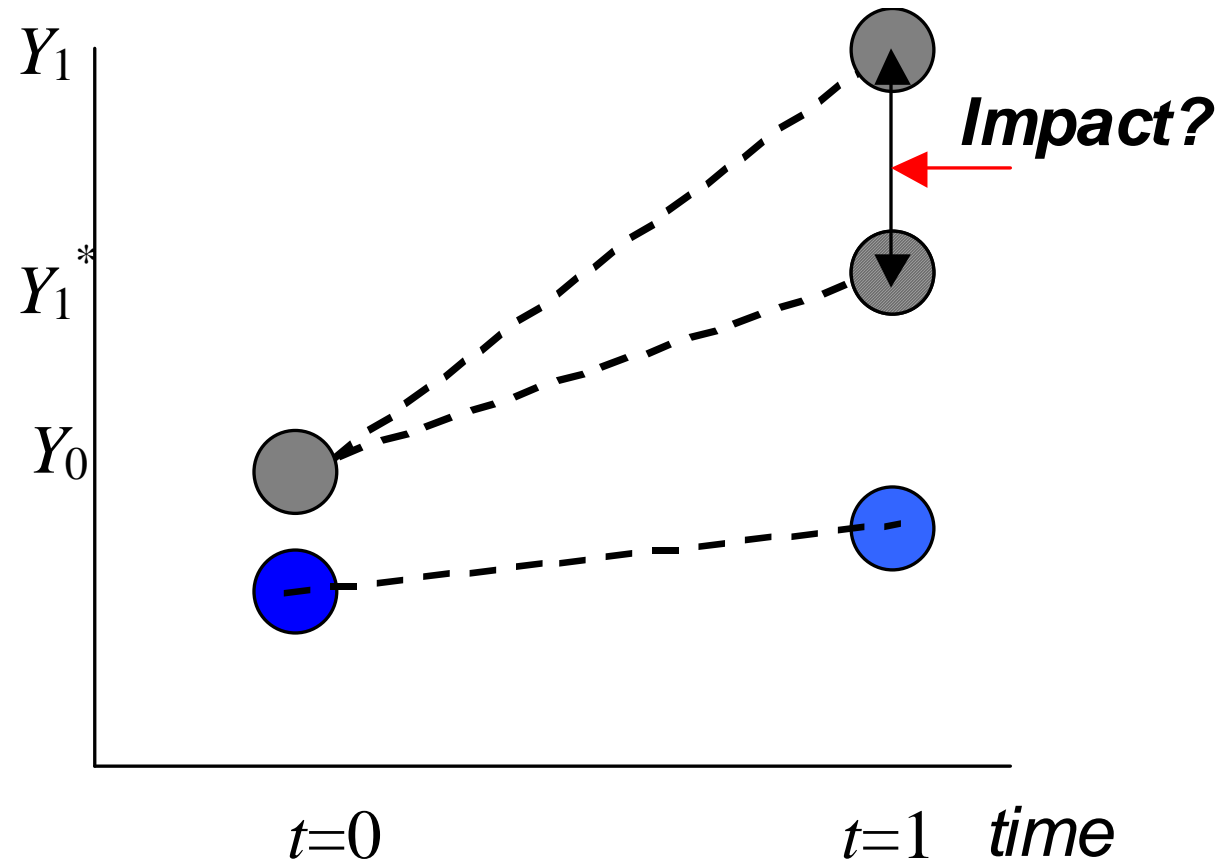
Selection bias



Diff-in-diff requires that the bias is additive and time-invariant



The method fails if the comparison group is on a different trajectory



3a. Experimental Designs

The experimental/randomized design

- In a randomized design the control group (randomly assigned out of the program) provides the counterfactual (what would have happened to the treatment group without the program)
- Can apply CSDIFF estimator (ex-post observations only)
- Or DiD (if have data in baseline and after start of program)
- *Randomization equalizes the mean selection bias between T and C groups*
- **Note: An Experimental or Randomized Design**
 - Assigns individuals into T=1 and T=0 groups randomly.
 - Consequence: $MA(T=1) = MA(T=0) \rightarrow MSB=0$ and
 - **Measured Impact = True Impact**

Examples for developing countries

- PROGRESA in Mexico
 - Conditional cash transfer scheme
 - 1/3 of the original 500 communities selected were retained as control; public access to data
 - Impacts on health, schooling, consumption
- Recently: many other experimental evaluations

Lessons from practice--1

Ethical objections and political sensitivities

- Deliberately denying a program to those who need it and providing the program to some who do not.
 - Yes, too few resources to go around. But is randomization the fairest solution to limited resources?
- Intention-to-treat helps alleviate these concerns
=> randomize assignment, but free to not participate
- But even then, the “randomized out” group may include people in great need.

Lessons from practice--2

Internal validity: Selective compliance

- Some of those assigned the program choose not to participate.
- Impacts may only appear if one corrects for selective take-up.
- Randomized assignment as IV for participation
- *Proempleo example*: impacts of training only appear if one corrects for selective take-up

Lessons from practice--3

External validity: inference for scaling up

- Systematic differences between characteristics of people normally attracted to a program and those randomly assigned ("randomization bias": Heckman-Smith)
 - One ends up evaluating a different program to the one actually implemented
- => Difficult in extrapolating results from a pilot experiment to the whole population

PROGRESA/Oportunidades

- First national CCT program to be evaluated with an experimental design
- IE was a major public good-generated a lot of interest in academic and policy arena
- Generated a wealth of data
- Turned out the program could have been evaluated (equally reliably) with a Quasi-Experimental design-RDD

3b. An Example of a Quasi- Experimental Design

Advantages of RDD for Evaluation

- RDD yields an unbiased estimate of treatment effect at the discontinuity
- Can many times take advantage of a known rule for assigning the benefit that are common in the designs of social policy
 - No need to “exclude” a group of eligible households/individuals from treatment

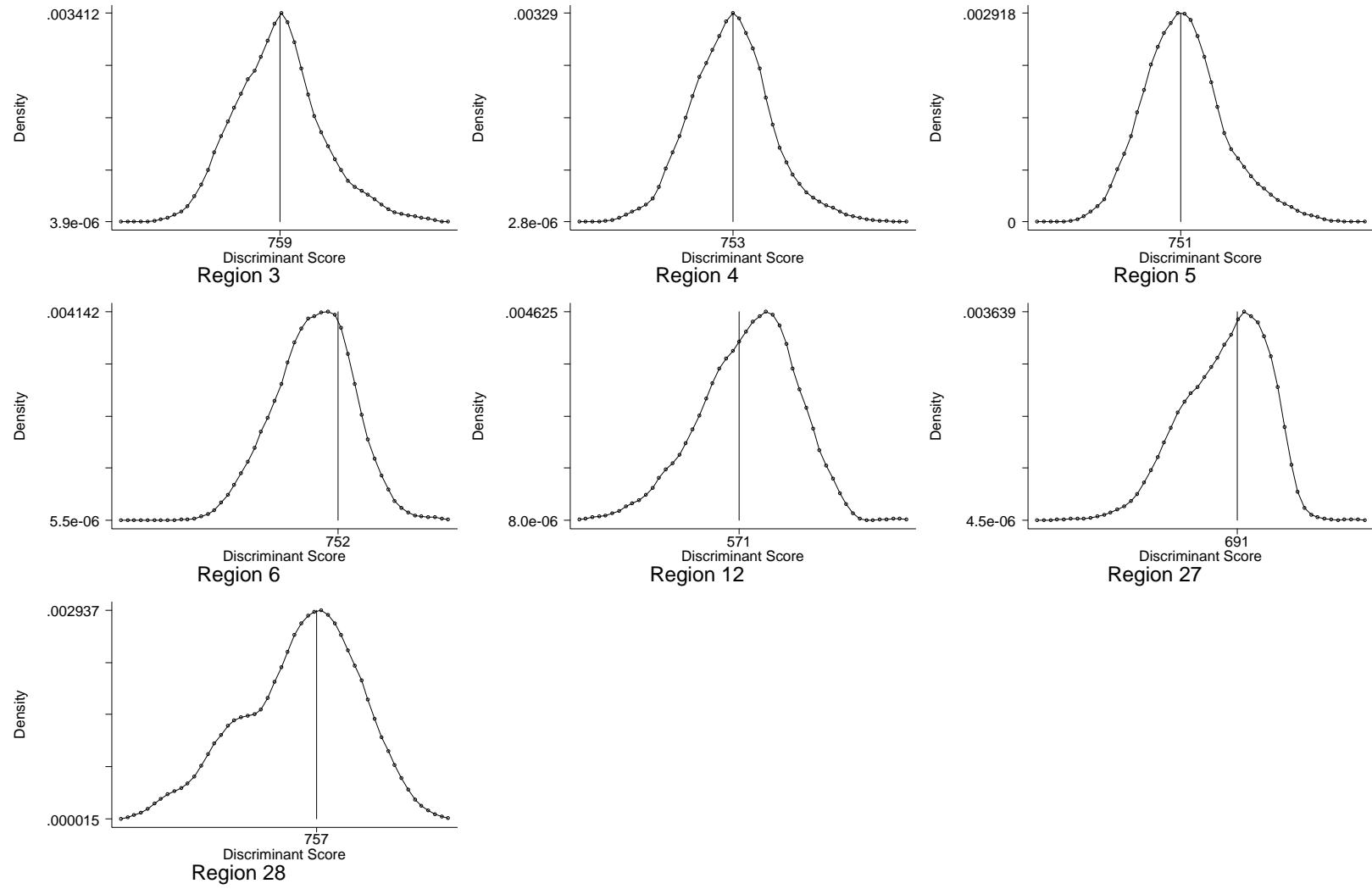
Potential Disadvantages of RD

- Local treatment
 - effects cannot be generalized (especially if there is heterogeneity of impacts)
- Power:
 - effect is estimated at the discontinuity, so we generally have fewer observations than in a randomized experiment with the same sample size
- Specification can be sensitive to functional form: make sure the relationship between the assignment variable and the outcome variable is correctly modeled, including:
 - Nonlinear Relationships
 - Interactions

Some Background on PROGRESA's targeting

- Two-stage Selection process:
 - Geographic targeting (used census data to identify poor localities)
 - Within Village household-level targeting (village household census)
 - Used hh income, assets, and demographic composition to estimate the probability of being poor (Inc per cap < Standard Food basket).
 - Discriminant analysis applied separately by region
 - Discriminant score of each household compared to a threshold value (high DS=Noneligible, low DS=Eligible)
 - Initially 52% eligible, then revised selection process so that 78% eligible. But many of the "new poor" households did not receive benefits

Figure 1: Kernel Densities of Discriminant Scores and Threshold points by region



Main Results

- Overall the performance of the RDD is remarkably good.
 - The RDD estimates of program impact agree with the experimental estimates in 10 out of the 12 possible cases.
 - The two cases in which the RDD method failed to reveal any significant program impact on the school attendance of boys and girls are in the first year of the program (round 3).



Thank you