

An introduction to Impact Evaluation

Markus Goldstein

PRMPR

Outline

- Why do impact evaluation
- Why we need a comparison group
- Methods for constructing the comparison group
- Practical considerations
- Funding
- Resources

Impact evaluation

- Many names (e.g. Rossi et al call this impact assessment) so need to know the concept.
- *Impact* is the difference between outcomes with the program and without it
- The goal of impact evaluation is to measure this difference in a way that can attribute the difference to the program, and only the program

Why it matters

- We want to know if the program had an impact and the average size of that impact
 - Understand if policies work
 - Justification for program
 - Scale up
 - Meta-analyses
 - (with cost data) understand the net benefits of the program
 - Understand the distribution of gains and losses

What we need

- The difference in outcomes with the program versus without the program – for the *same* unit of analysis (e.g. individual)
- Problem: individuals only have one existence
- Hence, we have a problem of a missing counter-factual, a problem of missing data

Thinking about the counterfactual

- Why not compare individuals before and after (the reflexive)?
 - The rest of the world moves on and you are not sure what was caused by the program and what by the rest of the world
- We need a control/comparison group that will allow us to attribute any change in the “treatment” group to the program (causality)

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
- Hence, a comparison of participants and an arbitrary group of non-participants can lead to heavily biased results

Example: providing fertilizer to farmers

- The intervention: provide fertilizer to farmers in a poor region of a country (call it region A)
 - Program targets poor areas
 - Farmers have to enroll at the local extension office to receive the fertilizer
 - Starts in 2002, ends in 2004, we have data on yields for farmers in the poor region and another region (region B) for both years
- We observe that the farmers we provide fertilizer to have a *decrease* in yields from 2002 to 2004

Did the program not work?

- Further study reveals there was a national drought, and everyone's yields went down (failure of the reflexive comparison)
- We compare the farmers in the program region to those in another region. We find that our "treatment" farmers have a larger decline than those in region B. *Did the program have a negative impact?*
 - Not necessarily (program placement)
 - Farmers in region B have better quality soil (unobservable)
 - Farmers in the other region have more irrigation, which is key in this drought year (observable)

OK, so let's compare the farmers in region A

- We compare “treatment” farmers with their neighbors. We think the soil is roughly the same.
- Let's say we observe that treatment farmers' yields decline by less than comparison farmers. *Did the program work?*
 - Not necessarily. Farmers who went to register with the program may have more ability, and thus could manage the drought better than their neighbors, but the fertilizer was irrelevant. (individual unobservables)
- Let's say we observe no difference between the two groups. *Did the program not work?*
 - Not necessarily. What little rain there was caused the fertilizer to run off onto the neighbors' fields. (spillover/contamination)

The comparison group

- In the end, with these naïve comparisons, we cannot tell if the program had an impact
 - We need a comparison group that is as identical in observable and unobservable dimensions as possible, to those receiving the program, and a comparison group that will not receive spillover benefits.

How to construct a comparison group – building the counterfactual

1. Randomization
2. Matching
3. Difference-in-Difference
4. Instrumental variables
5. Regression discontinuity

1. Randomization

- Individuals/communities/firms are randomly assigned into participation
- **Counterfactual: randomized-out group**
- *Advantages:*
 - Often addressed to as the “gold standard”: by design: selection bias is zero on average and mean impact is revealed
 - Perceived as a fair process of allocation with limited resources
- *Disadvantages:*
 - Ethical issues, political constraints
 - Internal validity (exogeneity): people might not comply with the assignment (selective non-compliance)
 - Unable to estimate entry effect
 - External validity (generalizability): usually run controlled experiment on a pilot, small scale. Difficult to extrapolate the results to a larger population.

2. Matching

- Match participants with non-participants from a larger survey
- Counterfactual: matched comparison group
- Each program participant is paired with one or more non-participant that are similar based on observable characteristics
- Assumes that, conditional on the set of observables, there is no selection bias based on unobserved heterogeneity
- When the set of variables to match is large, often match on a summary statistics: the probability of participation as a function of the observables (the propensity score)

2. Matching

- *Advantages:*
 - Does not require randomization, nor baseline (pre-intervention data)
- *Disadvantages:*
 - Strong identification assumptions
 - Requires very good quality data: need to control for all factors that influence program placement
 - Requires significantly large sample size to generate comparison group

3. Difference-in-difference

- Observations over time: compare observed changes in the outcomes for a sample of participants and non-participants
- Identification assumption: the selection bias is time-invariant ('parallel trends' in the absence of the program)
- Counter-factual: changes over time for the non-participants

Constraint. Requires at least two cross-sections of data, pre-program and post-program on participants and non-participants

- Need to think about the evaluation ex-ante, before the program
- Can be in principle combined with matching to adjust for pre-treatment differences that affect the growth rate

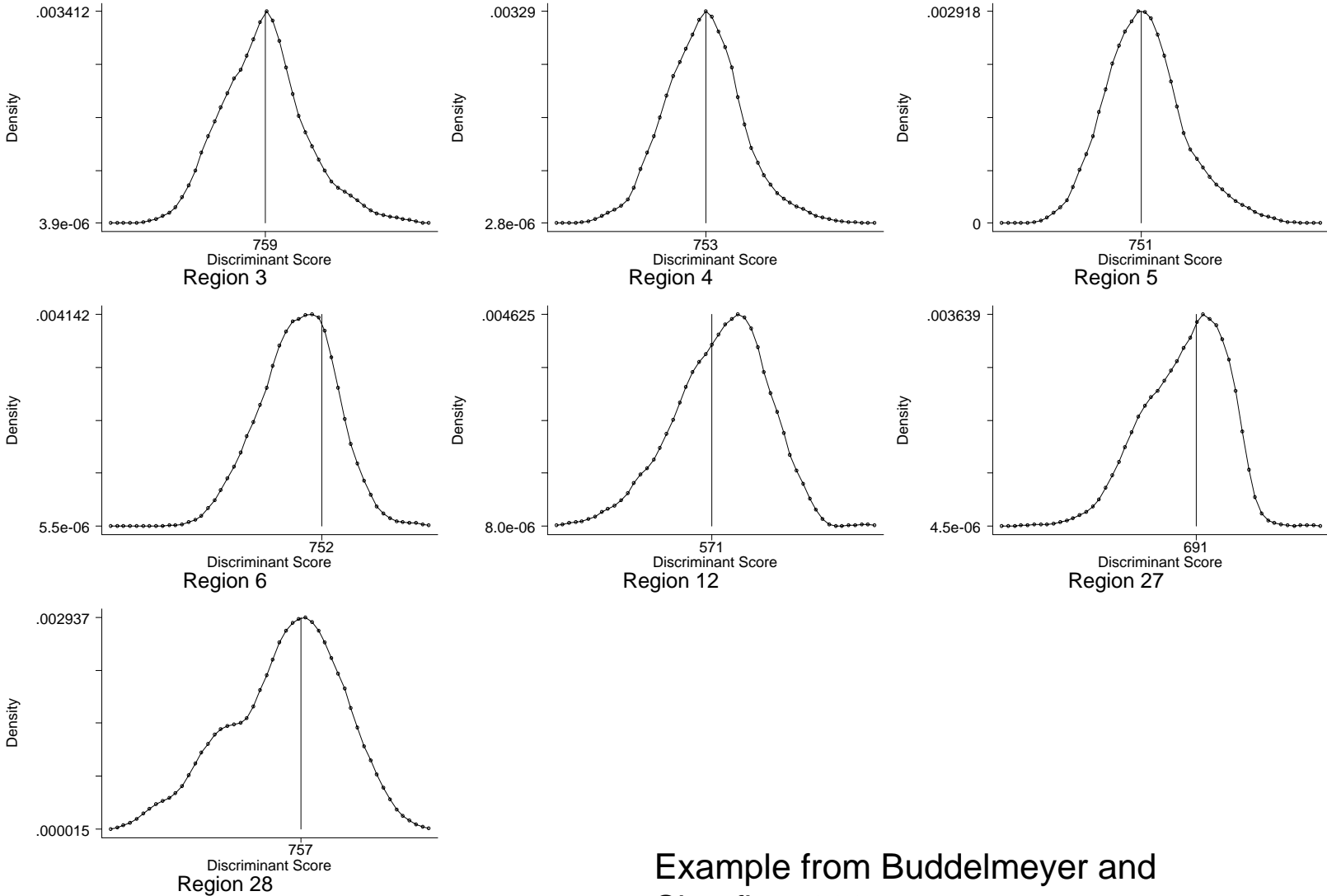
4. Instrumental Variables

- Identify variables that affects participation in the program, but not outcomes conditional on participation (exclusion restriction)
- Counterfactual: The causal effect is identified out of the exogenous variation of the instrument
- *Advantages:*
 - Does not require the exogeneity assumption of matching
- *Disadvantages:*
 - The estimated effect is local: IV identifies the effect of the program only for the sub-population of those induced to take-up the program by the instrument
 - Therefore different instruments identify different parameters. End up with different magnitudes of the estimated effects
 - Validity of the instrument can be questioned, cannot be tested.

5. Regression discontinuity design

- Exploit the rule generating assignment into a program given to individuals only above a given threshold – Assume that discontinuity in participation but not in counterfactual outcomes
- Counterfactual: individuals just below the cut-off who did not participate
- *Advantages:*
 - Identification built in the program design
 - Delivers marginal gains from the program around the eligibility cut-off point. Important for program expansion
- *Disadvantages:*
 - Threshold has to be applied in practice, and individuals should not be able manipulate the score used in the program to become eligible.

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



Example from Buddelmeyer and Skoufias, 2005

New and other directions

- *Estimating heterogeneity of impact*: go beyond average impacts and look at the entire distribution of gains/losses
- *Economy-wide reforms*: by construction do not have a comparison group. Structural modeling
- *Ex-ante simulations*: simulate the effect of the program (and of alternative design features) before it is in place (this is not impact evaluation)

Practical considerations

- Impact evaluation is not for every intervention
- The “gold standard” is plausible causality, not a single impact evaluation method
 - Recognize constraints
 - Be flexible, be creative

Timing the evaluation

- 2E's:
 - Early: start early, work IE into the design of the program
 - Evolve: as project changes, evaluation needs to – it is generally not a one-off exercise
- The evaluation should not subvert project design, BUT thinking about an evaluation early will let you change implementation to improve the evaluation without undermining the objectives

More practical considerations

- Think hard about benefits (what impacts to measure)
 - Link to project objectives
 - Careful choice of indicators
 - Understand time frame for outcomes to materialize
 - Identify logical axes of disaggregation (e.g. income groups, gender) and plan sample accordingly

Thinking about the data collection

- If baseline – need to time it with roll-out of intervention
- Know what other data sources are out there – type data into your browser
 - List of surveys by country
 - Some data online
 - DECDG attempt to coordinate/identify
- Maybe piggyback on existing survey

More practical considerations

- Monitor implementation of program – policy does not always equal reality (know what you are evaluating). The same holds true for data collection.
- Mix methods – qualitative and quantitative
- Watch for contamination of the comparison group

Funding – data collection

- BB project prep funds aren't enough for a baseline
- Options
 - Gov't funds, advanced
 - Convince a bilateral/other donor (e.g. DFID has done this)
 - Bank research committee, 2 windows – but need a research question
 - Trust funds – none earmarked explicitly for evaluation, but...

Trust funds

- Choose a strategic DiME meta-evaluation topic
 - Slum upgrading, CCTs, School based management, AIDS, ECD (pending)
- Opportunistic search of trust funds
 - PHRD (Japanese)
 - ESSD (Nordic)
 - BNPP

Ongoing funding: staffing

- There is a new budget task code for impact evaluation (ref J Adams, kiosk July 14 2005)
- An impact evaluation can be a product – a form of AAA but not ESW
- This becomes part of the work program agreement
- CDs may resist – but this is an emerging product

Staffing considerations

- Add an evaluation resource person to the team
- Places to look: anchors, regional/sector focal people, DEC
- These folks are subsidized
- BB for bank staff
- Project funds for consultant (roster)

Resources

- Informal evaluation network, including regional/sectoral focal individuals
- DiME
 - Meta-evaluations
 - Outreach
- DECRG course, WBI course (for counterparts), other training options can be set up

Thematic Group on Poverty Analysis, Monitoring and Evaluation

- Sectoral methods (nutrition, urban transport, rural roads, water, land reform, HIV/AIDS)
- Clinics – started by HD (HD contact Paul Gertler/Barbara Bruns, non-HD contact me)
- Library of resources (in progress)
- website


[Overview](#)
[Methods & Techniques](#)
[Selected Evaluations](#)
[Data and Tools](#)
[Training Events and Materials](#)
[Key Readings](#)
[Related Links](#)
[Contact Us](#)

Impact Evaluation

Last Updated: August 11, 2005

An **impact evaluation** assesses the changes in the well-being of individuals that can be attributed to a particular project, program or policy. This website aims at disseminating information and providing resources for people and organizations working to assess and improve the effectiveness of interventions aimed at reducing poverty. Information is organized in the following categories:

- [Overview](#) describes impact evaluation, why it is important, when it should be done and how to do it.
- [Methods & Techniques](#) describes different evaluation designs and estimation methods used to construct a counterfactual and measure welfare levels in order to evaluate the impact of an intervention. This section also presents a variety of implementation issues.
- [Selected Evaluations](#) provides a collection of impact evaluations organized by sector and by country.
- [Data & Data Sources](#) presents a range of poverty indicators and trends, datasets and a guide to data instruments and surveys.
- [Training Events and Materials](#) includes presentations and events for poverty impact evaluation.
- [Key Readings](#) presents a list of articles and impact evaluations.
- [Related Links](#) provides a collection of links relating to poverty impact evaluation for reference and research purposes.

Impact Evaluation Handbook



Methods and tools for project managers and policy analysts. Available in [English](#), [French](#), [Russian](#) and [Spanish](#).

Impact Evaluation Experts

A [database of "external" experts](#) who have analytical skills in the design of and/or the analysis for impact evaluations. *(For World Bank users only.)*

PRSP Sourcebook: Monitoring and Evaluation

A [guide to assist countries](#) in evaluating the effectiveness of a poverty reduction strategy.

Related Topics

Poverty analysis, monitoring, and evaluation and PSIA play an important role in a country's strategy to fight poverty.

- [Poverty Reduction Strategies](#)
- [Poverty Analysis](#)
- [Poverty Monitoring](#)
- [Poverty and Social](#)