Capacity Building workshop on Impact Evaluation of Employment Programs

Measuring Impact: Part 1

Maciej Jakubowski, Gdańsk, February 21, 2017
Outline

• Objective
• Perfect counterfactuals
• Bad counterfactuals
  • Before and after
  • Enrolled vs not enrolled
• Randomized Control Trials (RCTs)
Impact Evaluation

What is our objective?
Our Objective

Estimate the causal effect \((\text{impact})\) of program \((P)\) on outcome \((Y)\).

\[
\text{Impact} = Y^T - Y^C
\]

\begin{align*}
P &= \text{Program or Treatment} \\
Y &= \text{Indicator, Measure of Success} \\
Y^T &= \text{Outcome with the program} \\
Y^C &= \text{Outcome without the program (control)}
\end{align*}
Research Question

What is the Impact of....

...a youth training program...

... on probability of being employed?
Challenge – No counterfactual

Impact = $Y_T - Y_C$

We do not observe what would have happened to the jobseekers if they did not receive any training (the counterfactual)?

The key to a good impact evaluation is a valid estimate of the counterfactual!
Perfect Counterfactual

CLONE/PARALLEL UNIVERSE
Perfect Experiment

First, identify the target beneficiaries ...
Perfect Experiment

... and then clone the target beneficiaries
Perfect Experiment

... and then give the training to the target beneficiaries
Perfect Experiment - Clones

... and compare their probability of being employed some time later

• Because the people who received the training are exactly the same as those who did not receive the training, we can truly attribute the difference to the program
Bad counterfactuals

Case 1: Before & After
Case 1: Before & After

(1) Observe only participants

(2) Two observations in time:
   Employment status 12 months before the program
   Employment status 12 months after the program

ESTIMATE OF IMPACT = A - B = +13 percentage points
Case 1: Before & After

Problem: we don’t know what would have happened without the program

Economic Boom:
- Real Impact = A - C
- A - B over-estimates impact

Economic Recession:
- Real Impact = A - D
- A - B under-estimates impact

Before & After problem: other things that matter also change over time!
Case 1: Before & After

Example of Labor Market Program: Youth training program
Probability of being employed after 6 months of training

(1) Observe only participants in the program

(2) Two observations in time
1997/8
1999/9

ESTIMATE OF IMPACT = A - B = -10%
Bad counterfactuals

Case 2: Comparing those enrolled with those who don’t
Case 2: Some people enroll, others don’t

<table>
<thead>
<tr>
<th>Ineligibles (Inactive)</th>
<th>Eligibles (Unemployed youth = Target Population)</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Ineligibles" /></td>
<td><img src="image2" alt="Eligibles" /></td>
</tr>
<tr>
<td><img src="image3" alt="Not Enrolled" /></td>
<td><img src="image4" alt="Enrolled" /></td>
</tr>
</tbody>
</table>
Case 2: Some people enroll, others don’t
Case 2: Some people enroll, others don’t

Problem of Selection Bias

What if those who choose not to enroll are different?
Case 2: Some people enroll, others don’t

Problem of Selection Bias

And, what if these differences influence outcomes?

Are the factors that determine enrolment correlated with the probability of being employed?
Case 2: Some people enroll, others don’t
Example of Public Works Program (PWP) on first-come first-serve basis

Do you think enrolled jobseekers and those who came to register but were too late are similar?

Jobseekers who arrived early and were enrolled in the program

Jobseekers who arrived later, when no more public works jobs were available

What may be the problems in comparing these two groups?
Keep in Mind…

**Before-After Comparison**

**Problem:** other factors that matter also change over time.

**Compare those who enroll with those who don’t**

**Problem:** Selection bias. The enrolled may be different, and we don’t observe these differences.

Both comparison groups may lead to biased estimates of the program impact.
Good counterfactuals

Randomized Control Trial (RCT)
Example: Effect of Youth Job Guarantee on employment [Denmark]

Program
WHAT? Youth employment program (<30) with intensified activation
WHERE? 2009 to 2010, Denmark
WHY? Increase employment among long-term unemployed youth

Method/Eligibility
Researchers conducted a randomized evaluation (RCT) to test the impact of intensified ALMPs on youth employment and educational attainment. The 32-week program targeted job seekers under 30 who became or were already unemployed in the period from November 2009 to February 2010. Out of the 3,380 participants, researchers randomly assigned 1,683 job seekers to the treatment group and 1,697 to the comparison group.

Results
The intensified program reduced employment for uneducated youth and had no impact for educated youth. In addition, job seekers in the treatment group spent more time receiving sickness benefits. The program was not implemented as designed, and in practice, the only difference between the treatment and the comparison group was that job seekers in the treatment group met with a caseworker more frequently.
Basic setup of a randomized evaluation

1. Target Population
2. Random Sampling
3. Evaluation Sample
4. Baseline Survey
5. Random Assignment
6. Treatment group
7. Control group
8. Endline Survey

Key Concepts:
- Randomization methods
- Individuals/Clusters
- Multiple/Cross-cutting treatments
- Random Assignment
- Balance Check
- Measure Impact
- External Validity
- Internal Validity
Randomly sample from area of interest

Random sampling and random assignment
Randomly sample from area of interest

Random sampling and random assignment

Randomly assign to treatment and control

Randomly sample from both treatment and control
Opportunities to Randomize when Resources are Limited
Lottery when Oversubscription

More are eligible than can be served by the program
Lottery when oversubscription:

Oversubscription of a training in nursing – California, USA
Randomly vary *when* to receive program

Before

During

After

Time

evaluation period
Randomly vary program strength
Summary: opportunities to randomize

Consider

1. **Targeting rules**: Who is eligible for the program? Is program participation restricted, widely open, some combination?

2. **Resources**: Sufficient program resources to reach all eligible beneficiaries?

3. **Timing**: How are potential beneficiaries enrolled in the program – all at once or in phases over time?
Testing multiple program features

- Total Population
- Target Population
- Not in evaluation
- Evaluation Sample
- Random Assignment

Groups:
- Treatment Group 1
- Treatment Group 2
- Treatment Group 3
- Control Group
Cross-cutting program features

- Target Population
- Evaluation Sample
- Not in evaluation

Random Assignment

- T1
- C1
- T2
- T1*T2
- T2
- C2
- T1
- C
Which level of randomization?

Individual, or labor agency, poviat, voivodeship?
Which level of randomization?

Individual
Which level of randomization?

Employment agency
Which level of randomization?

Factors that matter

1. Unit of Implementation
   Unit of randomization should be at least at the level of the unit of program intervention

2. Spillovers
   Individuals in the control also benefits from the program (if trainees talk to non-participants)

3. Sampling unit and statistical power
   Randomize at the smallest possible level

4. Ethics and Fairness
Question: Is there risk of spillovers in the following programs?

• Example 1: **Job training program**

  • Job seekers in the control are *less* likely to get jobs *as a result* of the program, if the available jobs are given to those who received training

  • Measured program impact = jobs in treatment – jobs in comparison

  • As a result, *measured* program impact is higher than the *real* program impact
Why do we care about sample size?

• General question:
  How large does the sample need to be to credibly detect a given effect size?

• What does “Credibly” mean here?
  It means we can be reasonably sure that the difference between the group that received the program and the group that did not is due to the program

• Randomization removes bias, but it does not remove noise: it works because of the law of large numbers… how large must large be?
Some of the usual ethical concerns

• We can not run “experiments” on development issues
• We cannot leave people aside for the sake of the IE

but...
• Experimenting is part of the way to learn what is working (and what is not)
• We can not intervene with everybody anyway under budget/logistical constraints ➔ the evaluation may be a fair assignation

Lessons Learned:
• Work with counterparts at the beginning on identifying and addressing their concerns (mix political and research constraints)
• Be clear and explain everything
• Offer the evaluation as a solution rather than an extra layer for complications
Some of the usual political concerns

• There is no interest in showing (potential) bad news
• Long durations of the evaluations do not reconcile with political timelines

but...
• It is worse to do something bad and to hide it
• IEs may be designed as a tool to find “areas of improvement”
• The evaluation can be designed to pilot different options
• A good design can go beyond a political cycle

Lessons Learned:
• Understand the political concerns in order to design accordingly
• Work in phase
• Evaluate a subset of the program (in terms of geographic coverage)
• Show results soon (even with some limitations in the analysis) as it keeps politicians interested and engaged with the IE
Some of the usual technical concerns

- We already “know” what works… there is no need for evaluation
- The project is already complicated and we don’t want to add more complexity
- The concept of the project is already agreed upon
- The evaluations are too expensive, we cannot afford it

but…
- It may be complicated, but if we are not evaluating we are not learning
- Correlation and causality are not the same
- In most of the cases, the so called “agreements” are just basic features in the project’s concept

Lessons Learned:
- Projects in the preparation stage are better candidates than those under implementation
- Be pragmatic
The Danish Case

- The Danish National Labor Market Board: Program for obtaining empirical evidence on effectiveness of ALMPs
  - RCTs – on average 2 per year
  - Knowledge-bank – database of empirical studies conducted worldwide on ALMP effectiveness
  - Works actively with research community, providing data, participating in conferences, interactive debates, etc.

- Status so far (2011)
  - 8 RCTs are completed
  - 2 ongoing
  - 1-2 in planning process
  - 30 small-scale ‘pilot projects’ planned

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Content</th>
<th>Region</th>
<th>Jobcentres</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Group meeting each week</td>
<td>Northern</td>
<td>Frederikshavn, Brønderslev, Hjørring</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Jutland</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>Early activation (after 13 weeks)</td>
<td>Mid Jutland</td>
<td>Aarhus</td>
</tr>
<tr>
<td>C</td>
<td>Early activation and group meeting each week</td>
<td>Southern</td>
<td>Esbjerg, Vejle</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Denmark</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>Individual meeting w. case worker every week</td>
<td>Copenhagen &amp; Seeland</td>
<td>Holbæk, Roskilde, Gribskov, Ishøj-Vallensbæk, Vordingborg</td>
</tr>
</tbody>
</table>
Things to have in mind when considering a project

• What is the level at which the program is administered? (individual, village, group, district, etc.)

• If clustered, how many clusters can the implementing organization reach through existing budget?

• If clustered, how many clusters can be identified in addition to the program clusters?

• What level of compliance with the program has been observed in pilots and other settings?

• Are there large spillovers expected?
Things to have in mind when considering a project

• Is the program expected to have a different effect across people/groups with different characteristics?
• Can the program be phased in over time?
• Are there multiple programs (treatments) to be evaluated?
• Are there complementary programs to be evaluated?
Thank you!