Outline

• What is Impact Evaluation
• Causal Inference
• Counterfactuals
• Experimental Methods
• Quasi Experimental Methods
• Case Study
Monitoring and Evaluation (M&E)

A continuous process of collecting information

• to compare how well a project, strategy of policy is performing against expected results, and
• to inform implementation and program management.
Impact Evaluation (IE)

The assessment of the causal effect of a project, program, or policy on beneficiary outcomes, by estimating the change in outcomes attributable to the intervention.
Step 1: Challenge

We start with a development problem......

• More than 10% of children in Paraguay suffer delays in their physical, emotional or cognitive development

• Diarrhea is the second leading cause of child mortality

• School enrolment in Afghanistan is 54%
Step 2: Outcome

We determine a desired outcome we want to achieve….

• Decrease cognitive development delays by half
• Reduce diarrhea incidence by 20%
• Achieve universal primary enrollment
Step 3. Intervention

We propose a potential solution, in the form of a program or policy....

• Develop an early child development curriculum for caretakers

• Provide access to improved water and sanitation

• Establish a conditional cash transfer (linked to school enrolment)
Step 4. Evaluating impact

Determine the impact of the intervention....

• Did the inclusion of an early child development curriculum increase cognitive development?
• What was the effect of providing access to water and sanitation on children’s health?
• Are conditional cash transfers effective in increasing school enrolment?
More impact evaluation questions

• What is the effect of information on risky sexual behavior and HIV prevalence?
• Does contracting out primary health care lead to an increase in access and quality?
• Do bonuses to sales people generate more revenue than consumer price discounts?
• Do micro loans increase the productivity of small entrepreneurs?
Additional questions answered

- What is the effect of different sub-components of a program on specific outcomes?
- What is the right level of subsidy for a service?
- How would outcomes be different if the program design changed?
- Is the program cost-effective?

Traditional M&E cannot answer these.
Objetives of impact evaluation

- Determine if a program had *impact*, by measuring the *causal effect* between an intervention and an outcome of interest
- Estimate the *level* of impact
- Compare *real impact* with the *expected impact* at the time of designing the intervention
- Determine *adequate intensity* of intervention
- Compare *differential impact* among geographical areas, communities, or interventions
Estimating Causal Inference: How Do We Evaluate?
How do we evaluate?

Estimate the causal effect ($\alpha$) of an intervention ($T$) on an outcome ($Y$).

($\alpha$) = Effect, Impact  
($T$) = Program, Policy, Intervention, Treatment  
($Y$) = Outcome, Measure, Indicator

Example: What is the effect of an ECD program ($T$) on cognitive development in kids under 5 ($Y$)?
Formally

Question:
What is the impact of (T) on (Y)?

Answer:
\[ \alpha = [Y_i \mid T] - [Y_i \mid C] \]
Problem of incomplete information

\[ \alpha = [Y_i | T] - [Y_i | C] \]

For a program beneficiary \( i \):
- Observe:
  \[ [Y_i | T] \]: diarrhea incidence (Y) of beneficiary \( i \) having participated in the program (T)
- But do NOT observe:
  \[ [Y_i | C] \]: diarrhea incidence (Y) of beneficiary \( i \) not having participated in the program
Solution

Estimate what would have happened with $Y$ in the absence of $T$.

We call that ….. **Counterfactual**

The key for a good impact evaluation is a valid **counterfactual**!
Counterfactuals: How Do We Construct a Valid Counterfactual?
Counterfactual criteria

- Treated and Counterfactual
  1. Have identical characteristics
  2. Only difference is 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.
Example: What is the impact of... giving Mr. Nice

ECD

Cognitive development (Batelle & Denver)

\((T)\)

\((Y)?\)
The perfect “Clon”

Mr. Nice

“Clon”

\[ E[Y_i|T] = 620 \]

\[ E[Y_i|C] = 600 \]

IMPACT = 620 − 600 = 20 points
Using statistics

Treatment

Comparison

\[ \hat{E}[Y_i|T] = 620 \text{ puntos} \]

\[ \hat{E}[Y_i|C] = 600 \text{ puntos} \]

**IMPACT** = 620 – 600 = 20 points
Estimating impact

\[ \alpha = E[Y_i \mid T] - E[Y_i \mid C] \]

Observe \( E[Y_i \mid T] \)
Outcome \((Y)\) under treatment \((T)\)

Estimate \( E[Y_i \mid C] \)
Counterfactual using Control group

IMPACT = Outcome Treated Group \((T)\) - Outcome Control Group \((C)\)
Two counterfeit counterfactuals

- Before and After
  - Same individual or group before and after the treatment

- Those not enrolled
  - Those who did not enroll in (were not offered) the program versus those who did (were)
Case 1: Before and After

What is the impact of ECD (T) on scores (Y)?

**IMPACT** = A – B = 33
Case 1: What is the problem?

Recibe nutritional complement = C
- “Real” impact = A - C
- A-B *overestimates* impact

Start working = D
- “Real” impact = A - D
- A-B *underestimates* impact

Pre-program condition does NOT control for external factors that vary over time
Case 2. Participants vs Non-Participants

Compare versus non-elegible beneficiaries of a programa

- Treatment group: Registered / Participants
- Control group: Non-registered / Non-participants

*They chose NOT to participate*

*Are NOT eligible to participate*

Self-selection bias

- Characteristics of the population are correlated with de their condition to participate ($T$) and their outcomes ($Y$)

*We can control for observable characteristics*

*But we can’t control for unobservables*
Remember

**Pre-post**

**Compares:** Same unity of observation before and after receiving $T$.  
**Problem:** Other factors can occur over time that may affect the final outcome

**Self-selected**

**Compares:** Different unity of observation that opts to whether receive $T$.  
**Problem:** Self-selection bias makes groups not comparable

Both counterfactuals can result in a biased estimate of impact
Summary

• Impact evaluation measures the causal effect between an intervention (T) and an outcome (Y).

• The counterfactual is the theoretical concept of a program beneficiary without the benefit or treatment. Can NOT be observed!

• A counterfactual is estimated using a control group

• Program impact is the difference between the average outcome under treatment, and the counterfactual estimate, measured as the average outcome of the control group:

\[ \alpha = E[Y_i|T] - E[Y_i|C] \]
Experimental Methods: Randomized Control Trials (RCTs)
Experiments/Random Assignment/RCTs

Allocation of intervention (T) through lottery or another random process...

- Generates two groups statistically identical

When do we randomize?

- **Oversuscription**: # elegibles > available resources
- **Innovation**: need rigorous evidence about the efficacy of the program / intervention

Advantages

- Gives all eligible units the same probability to receive the intervention (T)
- Selection criteria is ethical, quantitative, fair and transparent
- Produces the most accurate counterfactual and is intuitive
But how do we randomize?

1. Population
2. Sample
3. Treatment

- Externa validity
- Internal validity

[CEGA logo]
Unit of randomization

- Select depending on the program
  - Individual/Household
  - School/Health Center
  - Street/Block
  - Town/Community
  - District/Municipality/Region

- Keep in mind:
  - It’s necessary to select a “big enough” number of units to detect a minimum detectable effect: Statistical power
  - Spillovers / contamination
  - Operational and survey costs

Rule of thumb: always randomize to the smallest implementation unit possible.
Case: RCT

- ECD Program in Paraguay
- Unit of randomization: Student
- 500 Elegibles students

Randomization in two phases:
  - 250 students in pilot schools receive an ECD intervention starting in 2015
    - Treatment group
  - 250 students in the rest of the country continue with national strategy
    - Grupo Control
Case: RCT

|                  | Treatment group (Randomly assigned to treatment) | Counterfactual (Randomly assigned to comparison) | Impacto $\hat{E}[Y_i|T] - \hat{E}[Y_i|C]$ |
|------------------|-------------------------------------------------|-------------------------------------------------|---------------------------------------------|
| **Baseline (2014)** Score (Y) | 587                                             | 586.9                                           | 0.1                                         |
| **Follow-up (2015)** Score (Y)  | 620                                             | 600                                             | 20                                          |

Note: **statistically significant at 1%**
Comparing different treatments

- Conventional evaluation question:
  - What is the impact of a program (T) on an outcome (Y)?

- Other interesting questions:
  - How to optimize a program?
    - What is the optimal level of a benefit or treatment?
  - Why does a program work?
    - What is the impact of different sub-components of a program?

- Random assignment with different treatments:

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Low benefit</th>
<th>High benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>1 Hour of training per week</td>
<td>3 Hours of training per week</td>
</tr>
</tbody>
</table>
Random assignment for two levels of benefits

1. Eligible population
2. Sample for evaluation
3. Random assignment (2 levels)

= Ineligible
= Eligible
Combined impact for two levels of benefits

- How do two benefits complement?
- Random assignment of a package of interventions:

<table>
<thead>
<tr>
<th>Intervention 1</th>
<th>Comparison</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervention 2</td>
<td>Group A X</td>
<td>Group C</td>
</tr>
<tr>
<td>Comparison</td>
<td>Group B Training</td>
<td>Group D Training</td>
</tr>
<tr>
<td>Treatment</td>
<td>Group B Training</td>
<td>Group D Training</td>
</tr>
</tbody>
</table>

CEGA
Random assignment with multiple levels of intervention

1. Eligible population
2. Evaluation sample
3. Random assignment 1
4. Random assignment 2

= Ineligible
= Eligible

Training
Random Assignment

Random assignment of treatment to a large sample produces two groups statistically equivalents

Feasible in prospective evaluations when demand exceeds the supply of services or resources

We have a perfect “clon”!

Randomly assigned to treatment

Randomly assigned to comparison

Ideal for new or innovative interventions to test efficacy

Many programs comply with these condition (pilot programs)
Quasi Experimental Methods: Difference in Differences (DiD)
Diff-in-Diff (DiD)

Y = Girls exam score (percentage of correct answers)
P = Tutoring

<table>
<thead>
<tr>
<th></th>
<th>Enrolled/With tutoring</th>
<th>Not Enrolled/No tutoring</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>After</strong></td>
<td>0.74</td>
<td>0.81</td>
</tr>
<tr>
<td><strong>Before</strong></td>
<td>0.60</td>
<td>0.78</td>
</tr>
<tr>
<td><strong>Difference</strong></td>
<td>+0.14</td>
<td>+0.03 = 0.11</td>
</tr>
</tbody>
</table>

\[ \text{Diff-in-Diff: Impact} = (Y_{t1} - Y_{t0}) - (Y_{c1} - Y_{c0}) \]
**Diff-in-Diff**

Y = Girls exam score (percentage of correct answers)
P = Tutoring

\[ \text{Diff-in-Diff: Impact} = (Y_{t1} - Y_{c1}) - (Y_{t0} - Y_{c0}) \]

<table>
<thead>
<tr>
<th></th>
<th>Enrolled/With tutoring</th>
<th>Not Enrolled/No tutoring</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>After</strong></td>
<td>0.74</td>
<td>0.81</td>
<td>-0.07</td>
</tr>
<tr>
<td><strong>Before</strong></td>
<td>0.60</td>
<td>0.78</td>
<td>-0.18</td>
</tr>
</tbody>
</table>

0.11
\[ \text{Impact} = (A - B) - (C - D) = (A - C) - (B - D) \]

\begin{align*}
\text{Exam score} \\
\text{B} = 0.60 \\
\text{C} = 0.81 \\
\text{D} = 0.78 \\
\text{T} = 0 \\
\text{T} = 1 \quad \text{Before} \\
\text{Enrolled} \\
\text{Impact} = 0.11 \\
\text{A} = 0.74 \\
\text{Not enrolled} \end{align*}
Impact = \((A-B)-(C-D) = (A-C)-(B-D)\)

- Exam score
  - Before: \(B = 0.60\)
  - After: \(C = 0.81\)
  - Enrolled: \(A = 0.74\)
  - Not enrolled: \(D = 0.78\)

- Time
  - \(T = 0\) Before
  - \(T = 1\) After

Impact = ?
Summary DiD

Difference-in-Differences

Difference-in-Differences combines *Enrolled & Not Enrolled* with *Before & After*.

Generate counterfactual for change in outcomes over time

Trends –slopes– are the same in treatments and comparisons (*Fundamental assumption*)

To test this, at least 3 observations in time are needed:
- 2 observations before
- 1 observation after.
Quasi Experimental Methods: Regression Discontinuity Design (RDD)
Regression Discontinuity Design (RDD)

We have a continuous eligibility index with a defined cut-off

- Households with a score \( \leq \) cutoff are eligible
- Households with a score \( > \) cutoff are not eligible
- Or vice-versa

Intuitive explanation of the method:

- Units just above the cut-off point are very similar to units just below it – *good comparison*.
- Compare outcomes \( Y \) for units just *above and below* the cut-off point.

For a discontinuity design, you need:
1) Continuous eligibility index
2) Clearly defines eligibility cut-off.
Examples of Eligibility Index/Score

Many social programs select beneficiaries using an index or score:

- **Anti-poverty Programs**: Targeted to households below a given poverty index/income
- **Pensions**: Targeted to population above a certain age
- **Education**: Scholarships targeted to students with high scores on standardized test
- **Agriculture**: Fertilizer program targeted to small farms less than given number of hectares
Case: Effect of fertilizer program on agriculture production

Goal
Improve agriculture production (rice yields) for small farmers

Method
• Farms with a score (Ha) of land ≤50 are small
• Farms with a score (Ha) of land >50 are not small

Intervention
Small farmers receive subsidies to purchase fertilizer
RDD Baseline

- Eligible
- Not eligible
RDD Post Intervention

Impact

Outcome

Score
Case: Discontinuity Design

- Eligibility for Progresa is based on national poverty index

- Household is poor if score $\leq 750$

- Eligibility for Progresa:
  - Eligible=1 if score $\leq 750$
  - Eligible=0 if score $> 750$
Case: Discontinuity Design
Score vs. consumption at Baseline–No treatment

![Graph showing consumption vs. fitted values for poverty index.](graph.png)
Case: Discontinuity Design

Score vs. consumption at Baseline—No treatment

Estimated impact on consumption (Y) **30.58**

(**) Significant at 1%
Summary RDD

Discontinuity Design requires continuous eligibility criteria with clear cut-off.

- Gives unbiased estimate of the treatment effect, but produces a local estimate:
  - Effect of the program around the cut-off point/discontinuity.
  - This is not always generalizable.

Power:
- Need many observations around the cut-off point.

- No need to exclude a group of eligible households/individuals from treatment.

- Can sometimes use it for programs that already ongoing.
Case Study: Handwashing with soap in Peru
Original intervention

ILM

Activities:
• Mass media campaign
• Promotional events
• Capacity building of agents
• Educational sessions
• HW promotion as a school curriculum
Revised intervention

Component: Mass Media

Activities:
- Mass media campaign
- Promotional events

Component: School and Community

Activities:
- Mass media campaign
- Capacity building of agents
- Educational sessions
- HW promotion as a school curriculum
Intervention design

195 provinces (universe)

40 provinces

40 districts

T1 (700 HH)

40 communities

40 districts

40 provinces

40 districts

40 communities

T2 (700 HH)

T2-Esc. (700 HH)

40 communities

C (700 HH)

C-Esc. (700 HH)

40 communities

40 districts
Results – Mass Media

Improved children’s health

Improved HW behavior among mothers and caretakers

Changes in beliefs, knowledge and availability

Exposure to HW with soap promotions

ILM (treatment)
Results – Community and School

- Improved children’s health
- Improved HW behavior among mothers and caretakers
- Changes in beliefs, knowledge and availability
- Exposure to HW with soap promotions
- ILM (treatment)
Advantages of randomization

Diarrhea (previous 7 days)
BASELINE

Control: 16.2%
Treatment: 16.7%

Diarrhea (previous 7 days)
ENDLINE

Control: 6.9%
Treatment: 6.4%
References


Alexandra Orsola-Vidal | aorsolavidal@berkeley.edu
Center for Effective Global Action | www.cega.org
University of California, Berkeley