Module 12

Advanced Evaluation Designs (IEs)

* Certain slides used with permission from the Abdul Latif Jameel Poverty Action Lab (JPAL) and Center of Evaluation for Global Action (CEGA)
Module Objectives

By the end of this module participants will be able to:

1. Understand experimental and quasi-experimental IE designs
2. Recognize the shortcomings of common ‘counterfeit’ counterfactuals
3. Identify the advantages and challenges of advanced IE designs
4. Plan mitigation strategies for common threats in IEs
5. Propose multiple approaches to contracting IEs
Today’s Roadmap

Defining Impact Evaluation and the Counterfactual

Experimental & Quasi-Experimental Methods
- Randomized Controlled Trials (RCTs)
- Regression Discontinuity
- Differences-in-Differences
- Matching

Planning and Managing Threats to IEs
Why do we do Impact Evaluations? Review

• **Accountability** to stakeholders
  – Measure effectiveness, relevance and efficiency
  – Disclose findings to stakeholders

• **Learning** to improve effectiveness
  – Select evaluation questions to test assumptions underlying project designs
  – Utilize appropriate methods to generate valid findings
  – Share findings
  – Integrate conclusions and recommendations into decision-making
How to Answer Impact Questions

1. Convert your question into a testable hypothesis
   - IE module

2. Develop and implement an evaluation design
   - IE module

3. Collect Data
   - Quantitative and Qualitative Methods Module

4. Test your hypothesis
   - Quantitative and Qualitative Methods Module
Which of the following are typical IE questions?

• Was it the intervention that caused observed results?
• What proportion of women participated in the program?
• Were women affected differently than men?
• Was the program implemented on time and budget?
• Which approach is most effective in achieving the desired result?
• Under what conditions would the intervention produce similar results if it were replicated?

Are these other questions still important for IEs?
Converting Questions to Hypotheses

- How do we answer the evaluation questions?
  - Turn them into hypotheses that can be tested empirically (i.e. through data and evidence)

- Often written as if…then statements
  - If the intervention, then the outcome

- In research terminology:
  - If the independent variable (IV), then the dependent variable (DV)
    - The DV is caused by the IV…it is dependent on the IV
Hypotheses to Test with IEs

• **Examples:**

• **Do microfinance schemes impact suicide rates?**
  – If a small business owner cannot repay a loan, then risk of committing suicide rises

• **Does increased elementary school attendance impact family income?**
  – If a child goes to school, then the family loses labor opportunity

• **Do increased literacy rates impact local government participation?**
  – If a citizen learns to read, then he/she will vote in the next election
USAID Definition of Impact Evaluation - Review

• *Impact Evaluations are based on models of cause and effect* …

  – Requires a causal theory of change
    • E.g. results framework, logical framework, development hypothesis

• *… and require a credible and rigorously defined counterfactual to control for factors other than the intervention that might account for the observed change.*

  – The *counterfactual* … identifies what would have happened to the beneficiaries absent the program.
    • The ‘impact’ of a program is defined as changes relative to this counterfactual
Counterfactual...in Pictures

Primary Outcome

Intervention

Impact

Counterfactual

Time

Counterfactual
The Counterfactual

• Can we directly measure the counterfactual?
  – Unless you have a time machine, you can’t be in two places at one time
  – You can’t both participate and not-participate in the program at the same time

• Since we cannot, how can we estimate it in an unbiased way?

Let’s start by looking at a couple ‘counterfeit’ counterfactuals…
‘Counterfeit’ Counterfactual 1: Before–After

Before – After Evaluation

• What is the ‘impact’?

• What is the counterfactual?
  - Assumes no change

Consider another scenario…
What if participants would have been worse off without the program?

Before – After gave a biased impact estimate

Legend

- Participants
- Impact
- Counterfactual
• We observe that income after food aid is lower than income before food aid

• Did the program fail?

• Suppose that the road got washed away in the floods that triggered the food aid?

The household’s income before the floods is a poor estimate of the counterfactual (i.e. the household’s income without the food aid)
Income effects of food aid

- With food aid income decreased…
- But what would it have been without the food aid?
- Before-after would have given a biased estimate
- What are the DV and IV?

Legend

- With Food-Aid
- Without Food Aid
- Impact

‘Counterfeit’ Counterfactual 1: Example
‘Counterfeit’ Counterfactual 2

- Compare participants to non-participants:

  Those who are selected or choose to participate

  Treatment

  Those who are NOT selected or choose NOT to participate

  Comparison
Why might non-participants be a biased counterfactual?

• Potential Problem:
  – Why do our participants participate?
    ➢ They (or their communities) are often selected because they are poorer or better connected, for example
    ➢ Or may choose to participate because they are highly motivated

• Many times reasons are directly related to the outcome of interest
  – E.g. more motivated=more likely to increase income even without the program

• Selection bias:
  – Differences between treatment and comparison group which influence the outcome of interest
  – Is the comparison group really comparable?
Counterfeit Counterfactual 2: An Example

- Policy Question: Does VCT reduce risky behavior?

  **Treatment Group:**
  - VCT
  - High risk type

  **Comparison Group:**
  - No VCT
  - Low risk type

- What if the treatment group went for VCT because they know they engage in high-risk behaviors?
  - In the absence of VCT, they still have a higher chance of HIV

- **Comparison group not a good counterfactual**
  - They have different expected outcomes even *without* the program
Read through **Exercise 12-1** and answer the following questions:

- How credible are the results of this impact evaluation?
- What is the counterfactual assumed in this impact evaluation?
- **Bonus Question**: Assuming that we could start over, how would you improve the design to generate more credible results?
Good Counterfactuals Exist

• …but they require careful planning

• Important that our ‘stand-ins’ are comparable to beneficiaries
  – Both initially
  – And during the project

• You need to be convinced that the remaining difference is ONLY caused by the intervention
Main Types of Impact Evaluation Designs

1. **Experimental designs:**
   - Units are randomly assigned to the project and control groups
   - Provides the strongest counterfactual

2. **Quasi-experimental designs:**
   - Units are either self-selected or selected by the project
   - A comparison group is selected to match as closely as possible the project group
   - Design quality varies in terms of how closely the two groups are matched (selection bias)

3. **Non-experimental designs:**
   - Used when it is not possible to select a comparison group
   - The analysis is statistically much weaker but the findings may be credible and useful to managers
IE Designs: Experimental Designs

Experimental Designs
(Randomized Evaluations, RCTs)
Eligible individuals, communities, schools etc are randomly assigned to:

- The project group (that receives the services) or
- The control group (that does not have access to the project services)
Why Randomize? Review

• Helps to ensure no systematic differences between the project beneficiaries and the control group at the start of the project.
  – None of the initial characteristics which might influence the outcomes of interest are correlated with participation.

• Removes the following sources of bias that are difficult to control for in non-randomized designs:
  – Self-selection bias: individuals decide to join the project (or not)
  – Agency-selection bias: the implementing agency selects units to participate (or not)
Targeted Assignment

Blue villages don’t make a good counterfactual. These villages are different. They are better off.
Random Sampling and Assignment

Randomly sample from area of interest (select some eligible communities)

Remove ineligible communities from sample

Randomly assign to treatment and control out of sample

Randomly sample for surveys (from both treatment and control)

Income per person before the program

5000

0

Income per person after the program

12,486

10,057

5000

0
How do we Implement Random Assignment?

• What does the term “random” mean here?
  – Equal chance of participation for every unit
    ➢ Units with ‘favorable’ characteristics not more (or less) likely to receive services

• How do we do this?
  – Typically by computer (random number generator)
    ➢ Permits stratification and balance checking
  – Draw numbers out of a hat or public lottery
    ➢ More transparent – and credible to evaluation sample
  – Exploit naturally occurring randomness
    ➢ Veteran status (draft) depended on date of birth
    ➢ Last digit of vehicle number plate odd/even
Esther Duflo TED talk
Random assignment can occur at different levels with different types of ‘units’:

1. Randomizing at the individual level
2. Randomizing at the group level
   - E.g. class, school, clinic, community, district
   - “Cluster Randomized Trial”
Unit of Randomization: Individual?
Unit of Randomization: Class?
Unit of Randomization: Class?
Unit of Randomization: School?
Unit of Randomization: School?
Unit of Randomization: Considerations

• Which level to randomize?
  – What unit does the program target for treatment?
    ➢ If the program must be implemented at the school level, we can’t assign at the individual level
    ➢ Cannot assign at a unit lower than the unit of treatment
  – What is the unit of analysis?
    ➢ If we are looking at school level outcomes, we can’t assign at the individual level
    ➢ Cannot assign at a unit lower than the unit of analysis

• Other considerations
  – Spillovers
  – Power
  – Ethical/Political
Sometimes screening matters

- Suppose there are 2000 applicants
- Screening of applications produces 500 “worthy” candidates
- There are 500 slots
- A simple lottery will not work

- What are our options?
Consider the screening rules

- What are they screening for?
- Which elements are essential?
- Selection procedures may exist only to reduce eligible candidates in order to meet a capacity constraint
- If certain filtering mechanisms appear “arbitrary” (although not random), randomization can serve the purpose of filtering and help us evaluate
Randomization at the margins

- Sometimes a partner may not be willing to randomize among eligible people.
- Partner might be willing to randomize at the margins.
- People “at the margins” are people who are borderline in terms of eligibility.
  - Just above the threshold → not eligible, but almost
- What treatment effect do we measure? What does it mean for external validity?
Randomization at the margins

At the margins, compare treatment to control

Non-participants

participants

Treatment

Control
Quasi-Experimental Designs
Quasi-Experimental Designs

• Quasi-experimental methods offer a wider set of evaluation strategies

• 3 Examples:

  1. Regression Discontinuity Design

  2. Difference-in-differences

  3. Matching
Quasi-Experimental Design 1: Regression Discontinuity
I. Regression Discontinuity – Review

• What is it?
  • Treatment and comparison group assigned based on cut-off score on a quantitative variable that is believed to be correlated with intended outcome

  - Examples of cutoffs?
    ➢ Students must score above 80 on entrance exam
    ➢ Household income below $1,000/yr
    ➢ Farms below 1.5 hectares
I. Regression Discontinuity: An Example

• Agricultural Loan Program
  • Loans provided to farmers to increase farm productivity and income
  • To be eligible, farmers must have a credit rating of at least 100

• What are the IV, DV, and development hypothesis in this example?
Example of RD: Impact of Agricultural Loans

Prior to Agricultural Loan Program

Baseline Farm Income

Assignment variable (Credit Score)
Example of RD: Impact of Agricultural Loans

Comparison group

Cut-off score = 100

Assignment variable (Credit Score)
Defining the Selection Cut-off Point

• The cut-off point is determined by an analysis of the distribution of scores on the scale.
  – Ideally there will be approximately equal numbers of subjects above and below the cut-off point

• However, this if often strongly influenced by resources or implementation concerns
Challenges to RD

• It will sometimes be difficult to construct an ordinal or interval scale.
  – The information may not be available on all subjects.
  – The selection criteria used by management or experts may not be easy to convert into an ordinal scale.

• Generalizability
  – Only measure impact at the cutoff point
Quasi-Experimental Design 2: Difference in Differences (Double Difference)
2. Difference in Differences – Review

• What is it?
  • Compare the change in treatment group with change in a comparison group
  • 2 level comparison:
    1. observed changes in outcomes (Before-After)
    2. for a sample of participants and non-participants (With-Without)

• How are the treatment and comparison groups selected?
  • Treatment group is usually purposefully selected
  • Comparison Group either:
    • Matched
      • Judgmental or Statistical
    • Unmatched
  • Can be used with Randomization (method of analysis for RCT)
2. Difference in Differences

The diagram illustrates the concept of Differences in Differences (DID). The graph shows changes over time between participants (T) and non-participants (C) in terms of an outcome of interest.

- **Intervention** line indicates the change in the outcome for the participants.
- **Non-Participants** line shows the change in the outcome for the non-participants.

The difference between the changes in the two groups is measured by the DID impact, given by:

\[ \Delta = \Delta_T - \Delta_C \]

Where:
- \( \Delta_T \) is the change in the outcome for the participants.
- \( \Delta_C \) is the change in the outcome for the non-participants.

The impact is the difference in these changes, indicating the effect of the intervention on the outcome.
2. Difference in Differences - Example

- Hopetown HIV/AIDS Program (2008-2012)

- Objectives
  - Reduce HIV transmission

- Intervention: Peer education
  - Target group: Youth 15-24

- Indicator: Teen pregnancy rate (proxy for unprotected sex)
2. Difference in Differences

![Graph showing teen pregnancy rates before and after intervention.](image)

- **Before**: 46.37 in 2008, 62.9 in 2012
- **After**: 57.5 in 2012, 66.4 in 2012

- **Participants**
- **Comparison**
2. Difference in Differences

![Graph showing change in teen pregnancy rates before and after intervention]

- **Before**
  - 2008: 62.9
  - 2012: 46.37

- **After**
  - 2008: 66.4
  - 2012: 57.5

**Impact**

Impact = Difference in Differences

\[\text{Impact} = \Delta \text{‘After’} - \Delta \text{‘Before’} = 8.87 - 16.53 = -7.66\]
2. Difference in Differences

![Graph showing teen pregnancy rates with counterfactual (Parallel Trends) and impact calculations.](image)

- Teen Pregnancy (per 1000)
  - 2008: 46.37
  - 2012: 57.5

- Counterfactual (Parallel Trends)
  - 2008: 62.9
  - 2012: 74.0

- Impact: -7.6
### 2. Difference in Differences

<table>
<thead>
<tr>
<th></th>
<th>Average rate of teen pregnancy in</th>
<th></th>
<th>Difference (2008-2012)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2008</td>
<td>2012</td>
<td></td>
</tr>
<tr>
<td>Participants (P)</td>
<td>62.90</td>
<td>66.37</td>
<td>3.47</td>
</tr>
<tr>
<td>Non-participants (NP)</td>
<td>46.37</td>
<td>57.50</td>
<td>11.13</td>
</tr>
<tr>
<td>Difference (P-NP)</td>
<td>16.53</td>
<td>8.87</td>
<td>-7.66</td>
</tr>
</tbody>
</table>
2. Difference in Differences

• Advantages

  – QEDs allow targeted selection of participants
    ➢ Avoids some of the ethical or political objections to random assignment

  – In special circumstances, can be done *retrospectively*
    ➢ Baseline and outcome data must be available for potential comparison units
      • E.g. when secondary data sources are used
2. Difference in Differences

- Challenges
  - Reliance on assumption of **parallel trends**
    - Generally, how comfortable are we with this assumption?
    - Improved when groups are similar at baseline
  - Counterfactual less reliable than other IE designs (RCT, RD, PSM)
Quasi-Experimental Design 3: Matching Designs
3. Matching – Review

• What is it?
  • Construct a comparison group that looks as similar as possible to the treatment group
  • Pair each program participant with one or more non-participants, based on observable characteristics

• How do we construct the comparison group?
  • Selection
    • Judgmental
    • Match on one variable
    • Match on multiple variables (e.g. propensity score matching)
  • Individual vs group level matching
3. Matching

Select Treatment Communities

Identify Comparison Communities Similar to Treatments

Comparison village are matched on observed characteristics but may differ on unobserved.
3. Matching

![Graph showing the change in teen pregnancy rates before and after intervention.](image)

- **Before**: 66.37 (2008)
- **After**: 73.36 (2012)

**Effect**: -7.01
3. Matching: Propensity Score Matching (PSM)

• What is it?
  – Extension of matching technique to efficiently match on multiple variables

• How is it done?
  – After treatment group is selected, uses regression analysis to identify the baseline characteristics correlated with selection
  – Each treatment unit and potential match is given a **propensity score** based on the degree to which it shares these baseline characteristics
    ➢ The propensity or likelihood of being in the treatment group
    ➢ Intuitively: How similar is this participant to the ‘typical’ treatment unit?
  – Select “nearest neighbors” (usually around 3-5) from comparison group who most closely match a participant.
    ➢ Note: Other matching methods exist
### 3. Matching: Propensity Score Matching (PSM)

Estimating project impact on farm income using propensity score matching and DD

<table>
<thead>
<tr>
<th></th>
<th>Project farm 1</th>
<th>Project farm 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Farm 1</td>
<td>Farm 2</td>
</tr>
<tr>
<td></td>
<td>Dollars/month</td>
<td>Dollars/month</td>
</tr>
<tr>
<td></td>
<td>Average for 5</td>
<td>Average for 5</td>
</tr>
<tr>
<td></td>
<td>nearest neighbors (nearest matches)</td>
<td>nearest neighbors (nearest matches)</td>
</tr>
<tr>
<td>Before project</td>
<td>350</td>
<td>410</td>
</tr>
<tr>
<td>After project</td>
<td>500</td>
<td>520</td>
</tr>
<tr>
<td>Change</td>
<td>+150</td>
<td>+110</td>
</tr>
<tr>
<td>Impact estimate:</td>
<td>150 – 70 = + 80</td>
<td>110 – 125 = - 15</td>
</tr>
</tbody>
</table>

Estimated total project impact =

Sum of impacts for each farm/ number of project farms
• Assumption
  • Unobserved characteristics do not affect outcomes of interest
    ➢ Things that we do not or cannot measure
    ➢ Examples?

• Can we test this assumption?
  ➢ By definition…no
  ➢ But we can try to measure proxies
3. Matching

• Advantages…similar to DD
  – Similar to DD but stronger matching techniques are able to control for observed characteristics
  – QEDs allow targeted selection of participants
  – In special circumstances, can be done retrospectively
    ➢ Baseline and outcome data must be available for potential comparison units
    ➢ If done retrospectively without baseline date, rely on ‘covariates of convenience’
      • Variables that are not affected by treatment and can be observed at outcome measurement
        » (age, sex, marital status, race, etc)
      • Leaves lots of unobservables
        » Bias?

Prospective whenever possible
3. Matching

- Challenges
  - Cannot control for unobservables
    - Generally, how comfortable are we with this assumption?
  - Often need large data set to identify appropriate comparisons
  - Relies on finding similar comparison units
    - Can we find comparison units that are similar to all treatment units?
  - Region of common support
Note:

- In order to do PSM well, you will need to find many comparisons, which will require a very large sample size.
  - This is not easy!
What can we say about those outside the region of common support?
Summary of IE Designs - Review

- Gold standard is randomization – minimal assumptions needed, intuitive estimates

- Quasi-experimental approaches require assumptions
  - Hard to defend them sometimes
  - But can be very useful when randomization is not possible

Complete Exercise 12-2 IE Design Options – Scenarios 1-4 and Bonus Questions in groups based on the IE designs you have reviewed this module.
Summary – Which Design Do I Choose?

1. Can a comparison group be identified?
   - Yes
   - No → Non-experimental methods (e.g. before after)

2. Can assignment be randomized?
   - Yes
   - No

3. Is a cutoff used to determine participation?
   - Yes
   - No → Combine DD with: Propensity Score Matching, Other Matching, Unmatched (Quasi-Experimental)

   - Yes → Regression Discontinuity (Quasi-Experimental)

   - No

Experimental Evaluation
Exercise III – Which Design is it? (Exercise 12-3)

• Read each of the executive summaries and write down
  – The key evaluation design being used
  – What concerns you have about the design being used
  – Note:
    ➢ *ex ante*: planned before
    ➢ *ex post*: planned after
Threats in Designing and Managing Impact Evaluations
Types of Threats

• Design Threats
  – Ethical and Political Constraints on Comparison Group
  – Limitations to External Validity
  – Sufficient Sample Size to Yield Significant Results

• Implementation Threats
  – Spillovers
  – Non-Compliance
  – Attrition
  – Other Threats to Internal Validity

• Have you actually analyzed those threats?
Ethical and Political Constraints

What happens when we can’t create a comparison group?
Ethical and Political Constraints

- Is participation offered to all?
  - Usually not...resource constraints
    - If not, we next ask: Is random assignment the ‘best’ way to determine participation?

- When offered to all, we still might have design options:
  - Phased
  - Multiple Treatment

- Some projects by their nature affect whole populations
  - E.g. passing laws at the national level
  - In these cases, we often cannot identify a comparison group and must use other evaluation designs
Randomized-phase in example: Double Shifting in Secondary Schools in Uganda

- Assignment to groups A and B random
- Group B schools serve as controls for Group A in 2011/12…
Randomized-phase in example: Double Shifting in Secondary Schools in Uganda

<table>
<thead>
<tr>
<th>Timing of implementation</th>
<th>2011/12</th>
<th>2012/13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group A: 100 schools</td>
<td>Program</td>
<td>Program</td>
</tr>
<tr>
<td>Group B: 100 schools</td>
<td>Program</td>
<td></td>
</tr>
</tbody>
</table>

...and Group B schools also participate in 2012/3

- Is there a comparison group for 2012/3?
Limitations on External Validity

Is the evaluation sample representative of the target population?

Are the Target Population and Context similar to ‘new project’?

If conditions are truly unique, learning from IE will be limited.
What happens when we are constrained by a ‘small’ sample?

- Might not be able to identify an impact even if there really was one
Spillover

- Program impacts ‘spillover’ to comparison group
- Comparison group benefits indirectly from program
- Special case: donor gap filling
Spillover

- What are the implications of positive spillovers for impact estimates?

- What do we do about spillovers?
  - Measure
    - Construct or use differences in treatment intensity
  - Limit
    - Doesn't mean we have to limit all spillovers…just to comparison units
    - Select unit of assignment to encompass spillovers
  - Example of Deworming project
Non-Compliance

• Some Treatment units don’t participate…and some comparison units do
• How do we analyze results?
  – Based on actual participation, or
  – Original assignment
Attrition

- Some people can’t be found for follow-ups
- Why not?
  - Move and can’t be found
  - Refuse to respond
  - Fatalities
**Attrition**

- What are the implications of attrition for impact estimates?
  - Depends on whether attrition was:
    - Random or Differential

- **Random Attrition** reduces sample size but internal validity is maintained

- **Differential Attrition**: attrition is correlated with:
  - Treatment (e.g. more attrition in comparison group)
  - Other Characteristics (e.g. the poorest drop out)

- What are the implications for differential attrition?
How to handle attrition

- Limiting Attrition
  - Collect good location information
  - Consider (carefully) incentives for responding

- Dealing with Attrition
  - Assess whether attrition was differential
    - Allocate resources for identifying drop-outs
    - Use baseline data
Exercise IV–Kinerja Case Study (Exercise 12-4)

Read and answer the study questions in your groups.
Managing Impact Evaluations
Impact evaluations require sufficient timing to ensure:

• Contracting
• Lead time before implementation
  ➢ Team formation
  ➢ Background research
  ➢ Design
  ➢ Baseline planning
• Collaboration between IE and project partner

IE design should occur during the program design stage
Illustrative Prospective Impact Evaluation Timeline

Project and Impact Evaluation Timelines Must be Coordinated

**PROJECT**
- Project Design
- RFP Issued
- Implementation Signed
- IP Contract
- Start Up
- Intervention Starts
- Intervention Ceases

**EVALUATION**
- IE Decision & Preliminary Design Decisions
- Evaluation RFP Issued
- Evaluation Contract Signed
- Design/Methods Finalized
- Baseline Measures
- Assignment to Groups
- Monitor Delivery, Attrition, Spillovers/Crossovers; Collect Additional Data From Groups
- Endline Measures Analysis Report
LOE and Budgeting Implications

• Evaluation design
• Sample size and geographic dispersion
• Number of survey rounds
• Local costs (enumerators, materials, transport, security)

• Total cost is highly variable
  – Anywhere from <$100k to more than $2M
Triangulation

- Qualitative studies
  - Contracted in tandem with Impact Evaluation to interpret and enrich findings

- Prior studies
  - USAID or third-party evaluations
  - Qualitative or quantitative

- Extant data

High-Quality Evaluation
Contracting Impact Evaluations

• Hierarchical
  – IE component is subcontracted to the IP
  – Evaluation Policy now says this mechanism is NOT to be used
    • Why?

• Parallel
  – IE contractor operates independently of the IP
  – Timing/Planning implications

• Mission-level M&E Partner
  – Approach used by a number of missions
  – Partner contracted for M&E work, including IEs, across multiple projects
Baselines

• Baseline data is critical to IEs

• Baselines allow you to measure the progress of your intervention against a starting point

• Especially useful for midterm measurement
  – Allows you to modify intervention mid-way if necessary
Review Questions

• What is a counterfactual? Why is it important?

• What are the common IE designs? Name one strength and weakness of each.

• Name two common threats to IEs and how to mitigate those threats.

• Based on what you now know, how does the design of your project change? How will you phase in IE?
“Economic Scene: A model for evaluating the use of development dollars, south of the border”

Alan B. Krueger

The Power of Credible Evaluation