



USAID
FROM THE AMERICAN PEOPLE

Module 7b

Introduction to Impact Evaluation

Design:

Planning and Managing Impact Evaluations

Module Objectives

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

1. • Distinguish performance and impact questions.
2. • Identify the counterfactual in IE designs.
3. • Identify the strengths, weaknesses, and uses of 4 IE designs.
4. • Anticipate common challenges in applying IE designs.
5. • Define strategies for strengthening IE designs and IE management when working under real-world constraints.

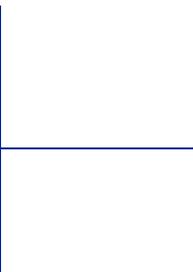
ADS references for this module

| Reference | Topic or Issue |
|------------------|--|
| ADS 203.3.1.1 | Definition of impact evaluation. |
| ADS 203.3.1.3 | Mission leadership will identify at least one opportunity for an impact evaluation for each Development Objective. |
| ADS 203.3.1.4 | Any activity within a project involving untested hypotheses or demonstrating new approaches that are anticipated to be expanded in scale or scope through USG assistance or other funding sources will, if feasible, undergo an impact evaluation. This also applies to pilot and “proof of concept” projects. |
| ADS 203.3.1 | A parallel contractual or grant agreement will be established at the inception to accompany implementation of impact evaluations. |
| ADS 203.3.1.7 | Implementing partners encouraged to identify opportunities for IEs. |
| ADS 203.3.1.6 | Experimental methods generate the strongest evidence for IEs. Alternative methods should be utilized only when random assignment strategies are not feasible. |
| ADS 203.3.3.1 | Rigorous impact evaluations may be undertaken for Development Objectives. |

Impact Evaluation (IE) Overview

What do **you** need to do?

1. Understand the definition of IEs according to the ADS
2. Identify the hypothesis that will be tested through the IE
3. Get help in defining a sample frame
4. Define a treatment and comparison group. If you're doing a randomized control trial, ensure the program is set up accordingly
5. During implementation, mediate between the evaluator and implementer
6. Ensure reports are written, disseminated, etc...



Contract the evaluator and implementer

Key Impact Evaluation Questions

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?

Why do we do IEs?

- Avoid continuing programs that do not have a real impact
- Avoid terminating programs that do have an impact
- Improve design, cost-effectiveness and efficacy
- Provide correct guidance on whether the program should be replicated



Village representatives show the cards they drew during the Kpomasse land tenure lottery as part of the MCC's program in Benin.

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

USAID Definition of Impact Evaluation

“ Impact evaluations measure the change in a development outcome that is attributable to a defined intervention. Impact evaluations are based on models of cause and effect and require a credible and rigorously defined counterfactual to control for factors other than the intervention that might account for the observed change.”

USAID Definition of Impact Evaluation (Part 1)

- *Impact evaluations measure the change in a development outcome...*
 - Focused on higher level results (outcomes) rather than inputs and outputs
 - E.g. health, income, school enrollment, agricultural productivity
 - To measure change, need data from at least two points in time
 - Baseline
- *...that is attributable to a defined intervention*
 - Intervention must be clearly defined and implemented
 - Must rule out other potential causes

Does measuring change tell us about attribution?

Change Does Not Prove Causality

An observed
change

≠

Causality
[impact]

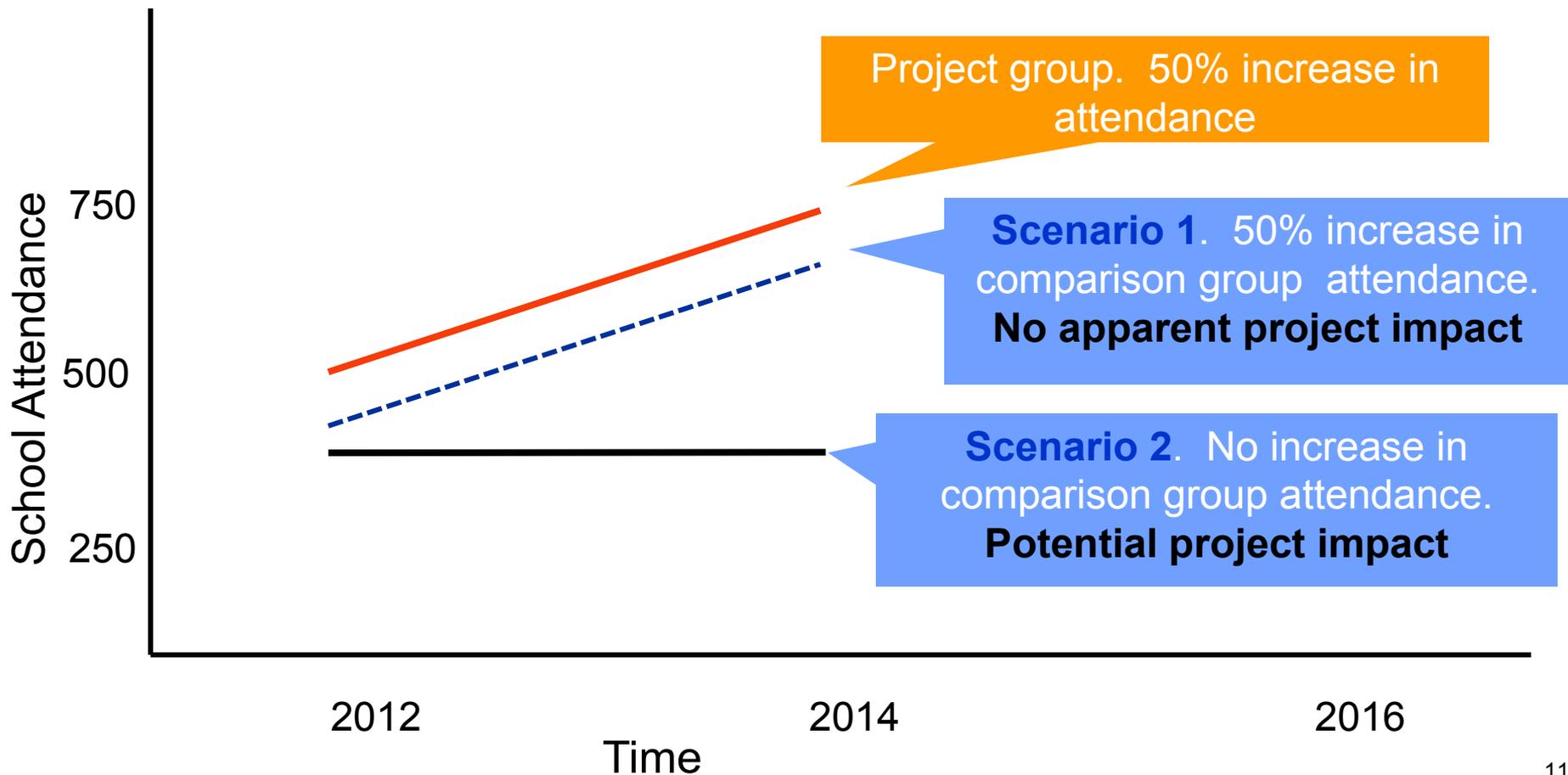
More children
attending
school in project
areas

Does **NOT**
prove

Project
interventions
caused the
increase

Change Does Not Prove Causality

Consider two scenarios:



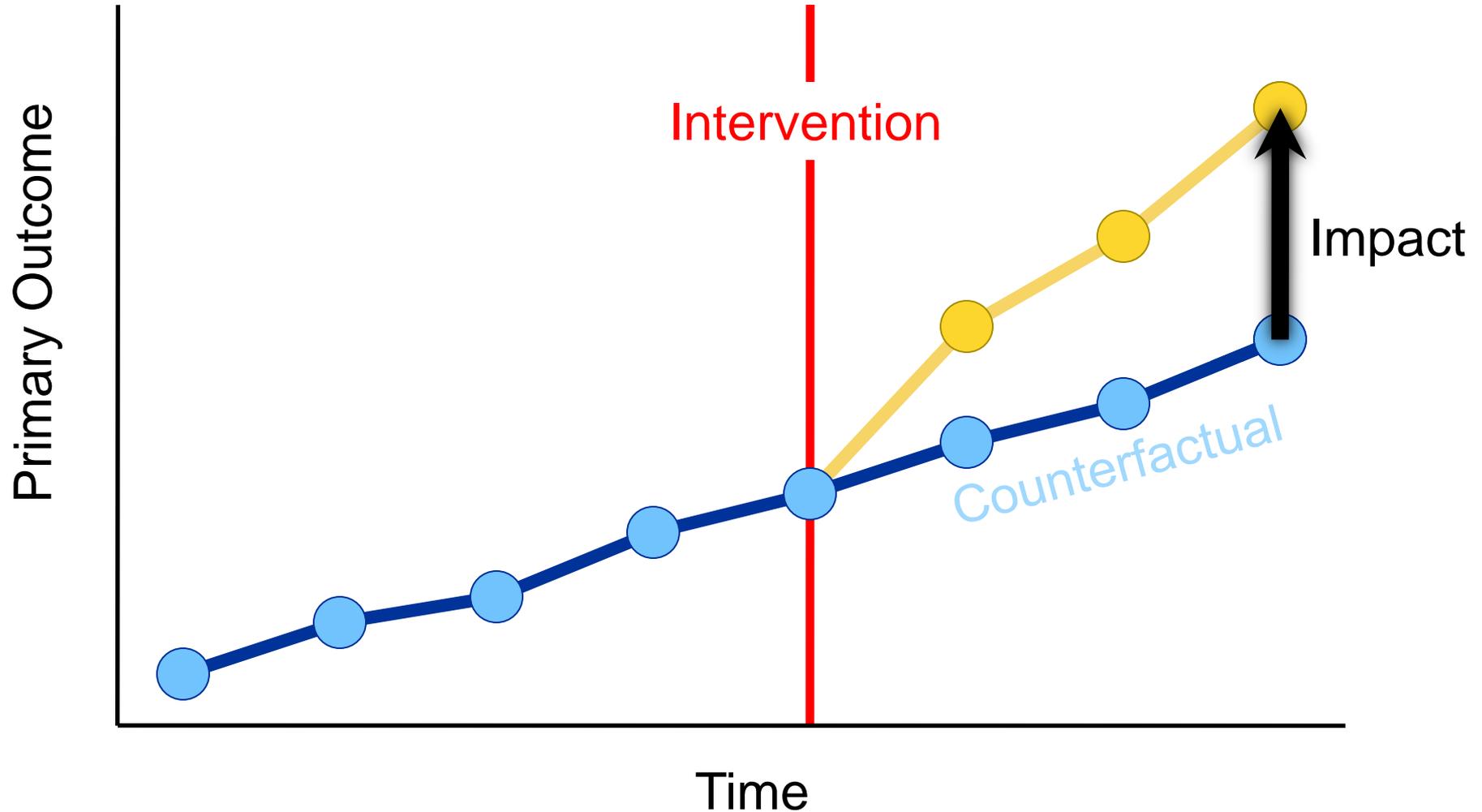
USAID Definition of Impact Evaluation (cont)

- *Impact Evaluations are based on models of cause and effect...*
 - Require a causal theory of change
 - E.g. results framework, logical framework, development hypothesis
- **Quick exercise.** Convert the following evaluation questions into testable hypotheses:
 - **Example:** Did the agribusiness loan program impact farm productivity?
 - **If** a farmer receives a loan, **then** they will have larger harvests the following year.
 - Does the capacity building program have impacts on CSO sustainability?
 - Did certain types of CSOs benefit more from the capacity building program than others?
 - Did boys or girls benefit differently from the school rehabilitation program?

USAID Definition of Impact Evaluation (cont.)

- *Impact Evaluations are based on models of cause and effect...*
- *... and require a credible and rigorously defined **counterfactual** to control for factors other than the intervention that might account for the observed change.*
 - The ‘impact’ of a program is defined as changes relative to what would have happened without the program.
 - The **counterfactual**... *identifies what would have happened to the beneficiaries absent the program.*
 - Why is the counterfactual so challenging?

Counterfactual...in pictures



Estimating the Counterfactual

- Can we directly measure the counterfactual?
 - No...we can't know what participants would have been like if they had not participated
- Since we cannot, how can we estimate it in an unbiased way?
 - Compare outcomes with **participants' pre-program** status?
 - Compare outcomes with **non-participant outcomes**?
 - Compare changes in outcomes with **non-participant changes**?

Estimating the Counterfactual

- Compare outcomes with **participants' pre-program status?**
 - Before-After evaluation design
 - Demonstrates change, but can't attribute that change to project
- Compare outcomes with **non-participant outcomes?**
 - Post comparison group evaluation design
 - What if they were different to begin with?
- Compare changes in outcomes with **non-participant changes?**
 - Pre-post comparison group evaluation design
 - Relies on comparison group being similar to participants

➤ **IE designs focus on this last structure**

Main Types of Impact Evaluation Designs

1. Experimental Designs (Randomized Control Trials):

- Units are randomly assigned to the project and control groups
- This is the strongest statistical design as it provides unbiased estimates of project impacts

2. Quasi-Experimental Designs:

- Units are either self-selected or selected by the project agency
- 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:

- It is not possible to select a comparison group
- Subject to more sources of bias but the findings may be credible and useful to managers
- Not considered true IE designs

IE Designs: Experimental Designs

Experimental Designs **(Randomized Evaluations, RCTs)**

IE Designs: Experimental Designs

- Eligible individuals, communities, schools, etc are **randomly assigned** to:
- The treatment group (that receives the services) or
- The control group (that does not have access to the project services)

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

How to randomize: Lottery + Phase-in

Round 1

Treatment: 1/3

Control: 2/3

Round 2

Treatment: 2/3

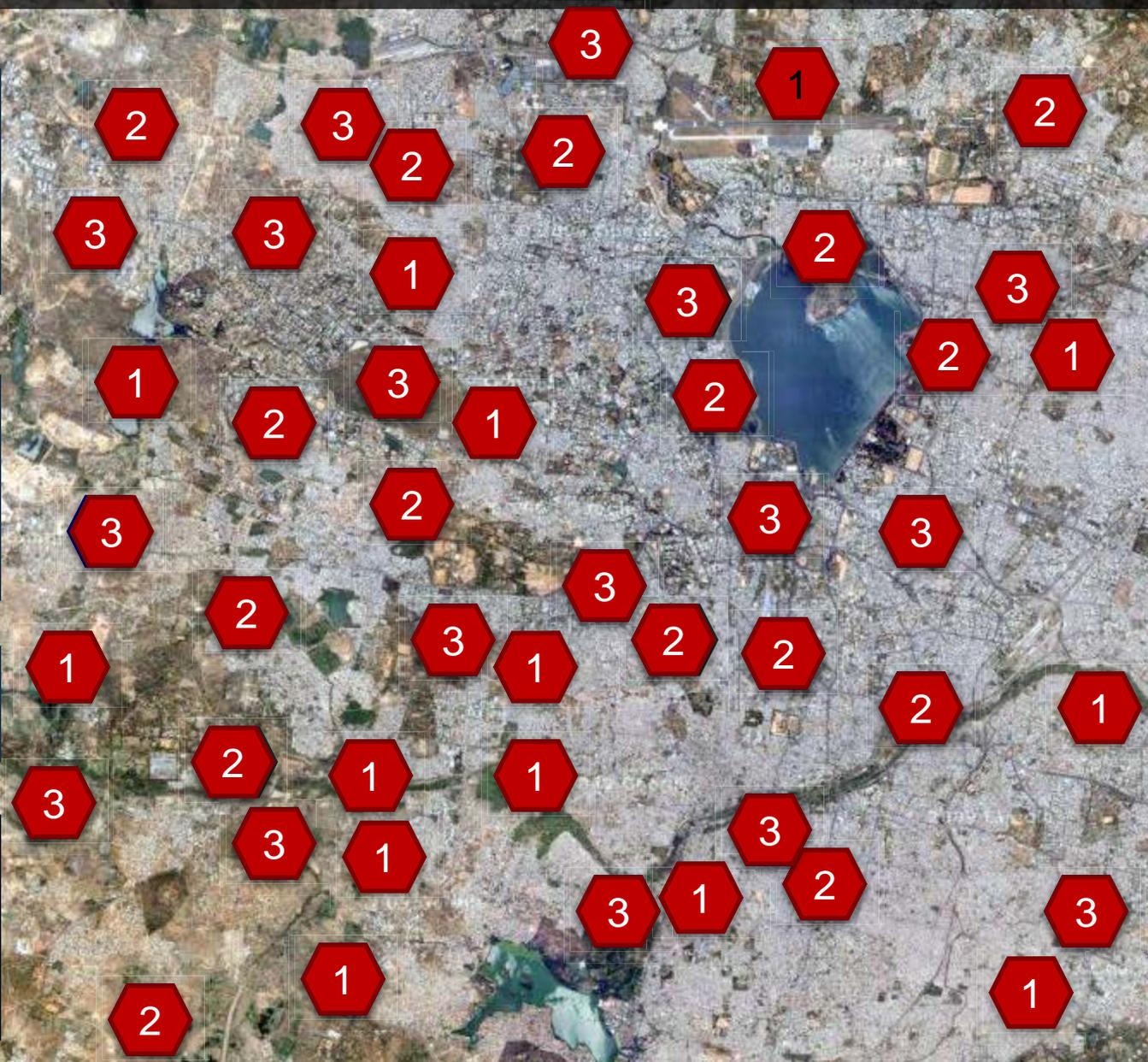
Control: 1/3

Randomized
evaluation ends

Round 3

Treatment: 3/3

Control: 0



Why Randomize?

- 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)

Randomized Control Trials

- The gold standard for constructing counterfactual
 - ... but not a 'silver bullet'
 - Sometimes not possible or recommended

When to Use Randomization?

- Randomization might not be the best IE method when...
 - For logistical, ethical or political reasons we feel that we must work with a select group of participants (RCT can work around this if the number is limited)
 - The project started before we could randomly assign participation (RCT only possible when planned in advance)
- On the other hand, RCT often used when the # of eligible individuals is larger than available resources
 - Fair and transparent way to assign benefits
 - Gives an equal chance to everyone in the sample

QUASI-EXPERIMENTAL DESIGNS

Quasi Experimental Designs

Objective

- Find a plausible counterfactual

Reality check

- Every method is associated with an assumption
 - Namely, that the treatment and comparison groups are equivalent
- The weaker the assumption, the more we need to worry about whether we are measuring the true causal effect



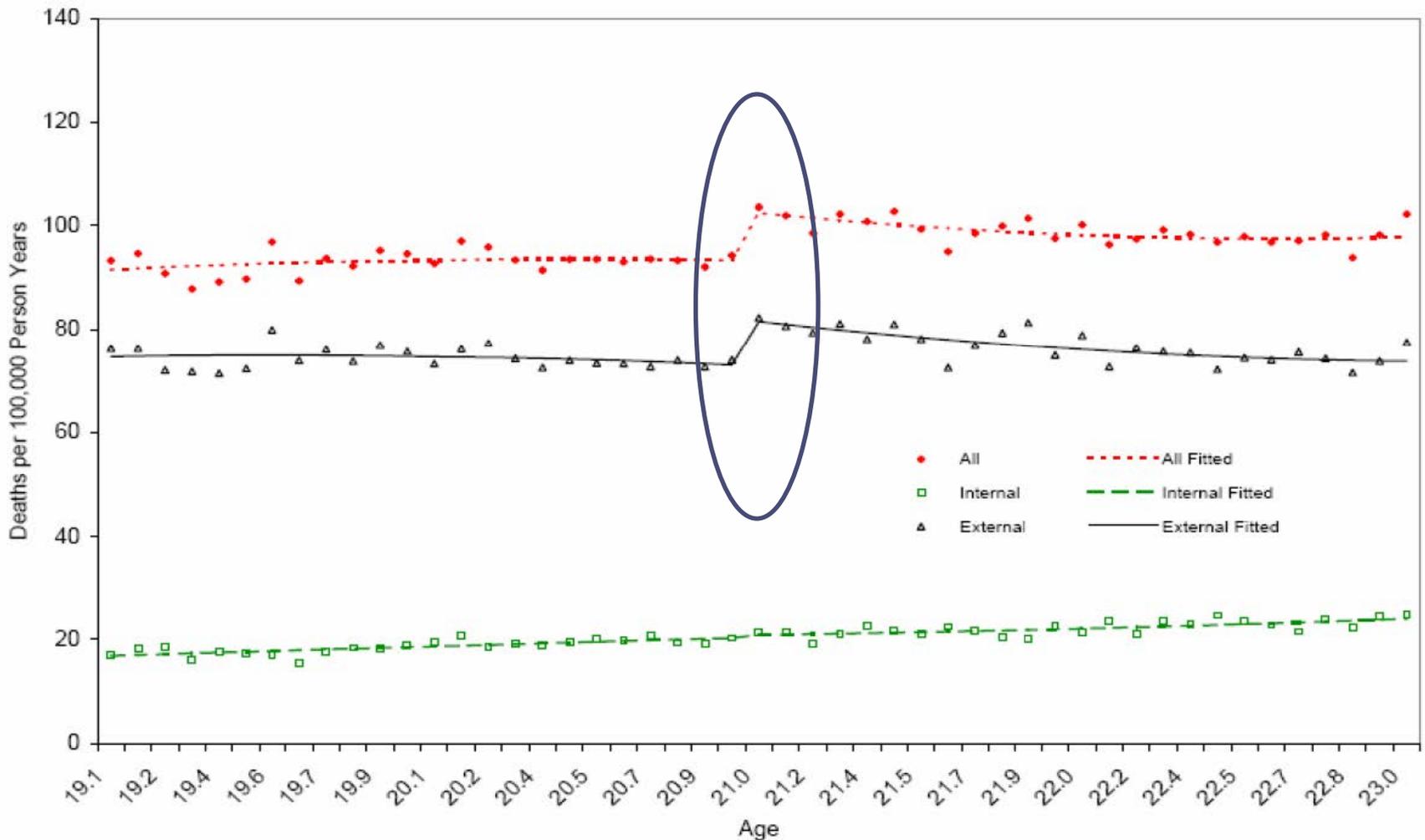
» Question your assumptions

1. Regression Discontinuity (RD)

- What is it?
 - Treatment and comparison group assigned based on cut-off score on a quantitative variable
- Examples?
 - Students must score above 80 on entrance exam
 - Household income below 1,000/yr
 - Farms below 1.5 hectares

1. Regression Discontinuity (RD)

Figure 5: Death Rates by Age



1. Regression Discontinuity

- Counterfactual:
 - Non-participants just below the eligibility cutoff are the comparison for participants just above the eligibility cutoff
- Assumption:
 - The small difference between units around the cutoff does not represent significant differences in expected outcomes

The Advantages of the RD Design

- Provides relatively unbiased impact estimates
- Allows for strict targeting or selection criteria
 - Can use desired selection criteria (for example selecting the poorest families or those considered most likely to succeed) rather than random assignment
 - Avoids some of the ethical or political objections to random assignment

Challenges using RDs

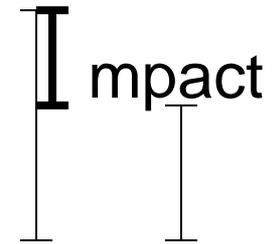
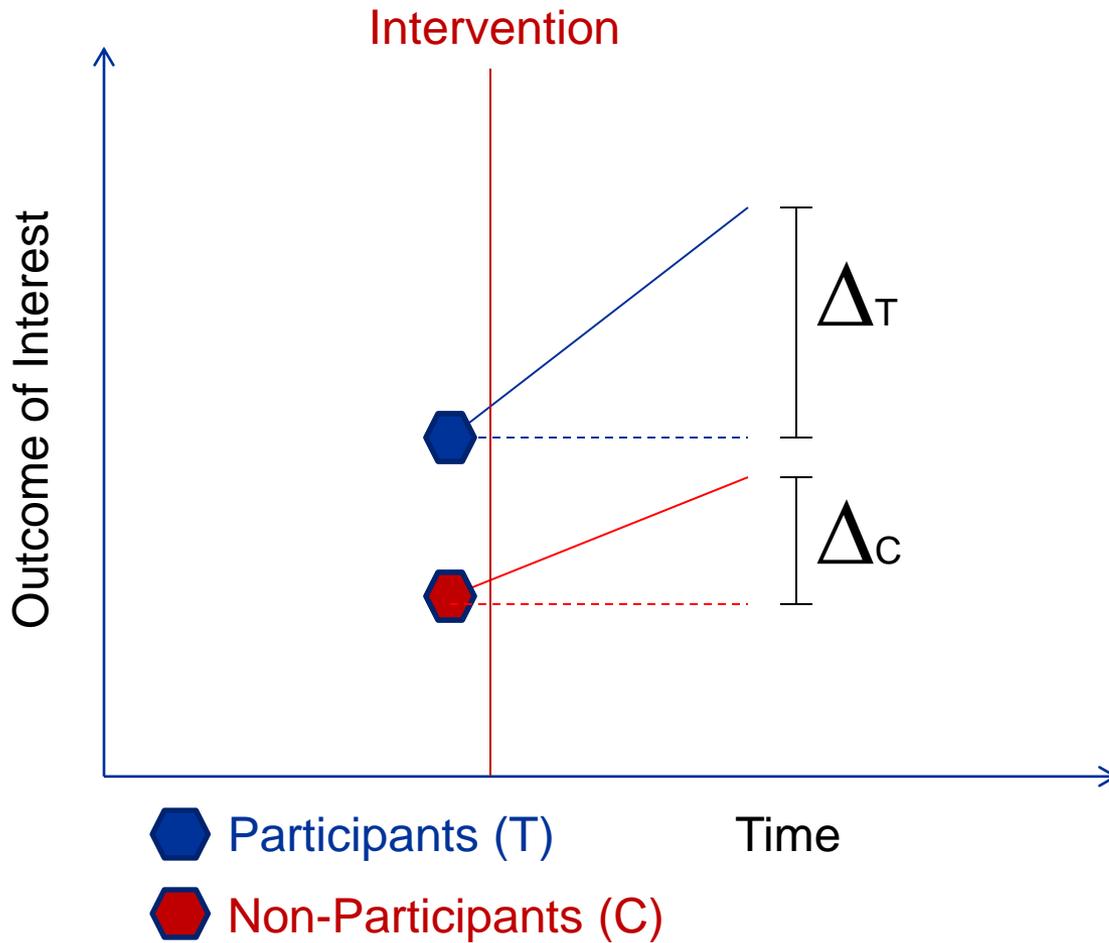
- Ensuring the cut-off point is strictly adhered to
 - Political or other pressures to allow people below the cut-off to participate.
 - Sometimes agencies do not have the administrative capacity to follow the cut-off
 - E.g. they may not know how much land a farmer owns
- A relatively large sample size is required in order to get sufficient numbers around the cutoff
- **External validity is limited to those around the cutoff.**

2. Difference in Differences

- What is it?
 - Compare the change in treatment group with change in a comparison group
 - Two 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

Differences in Differences



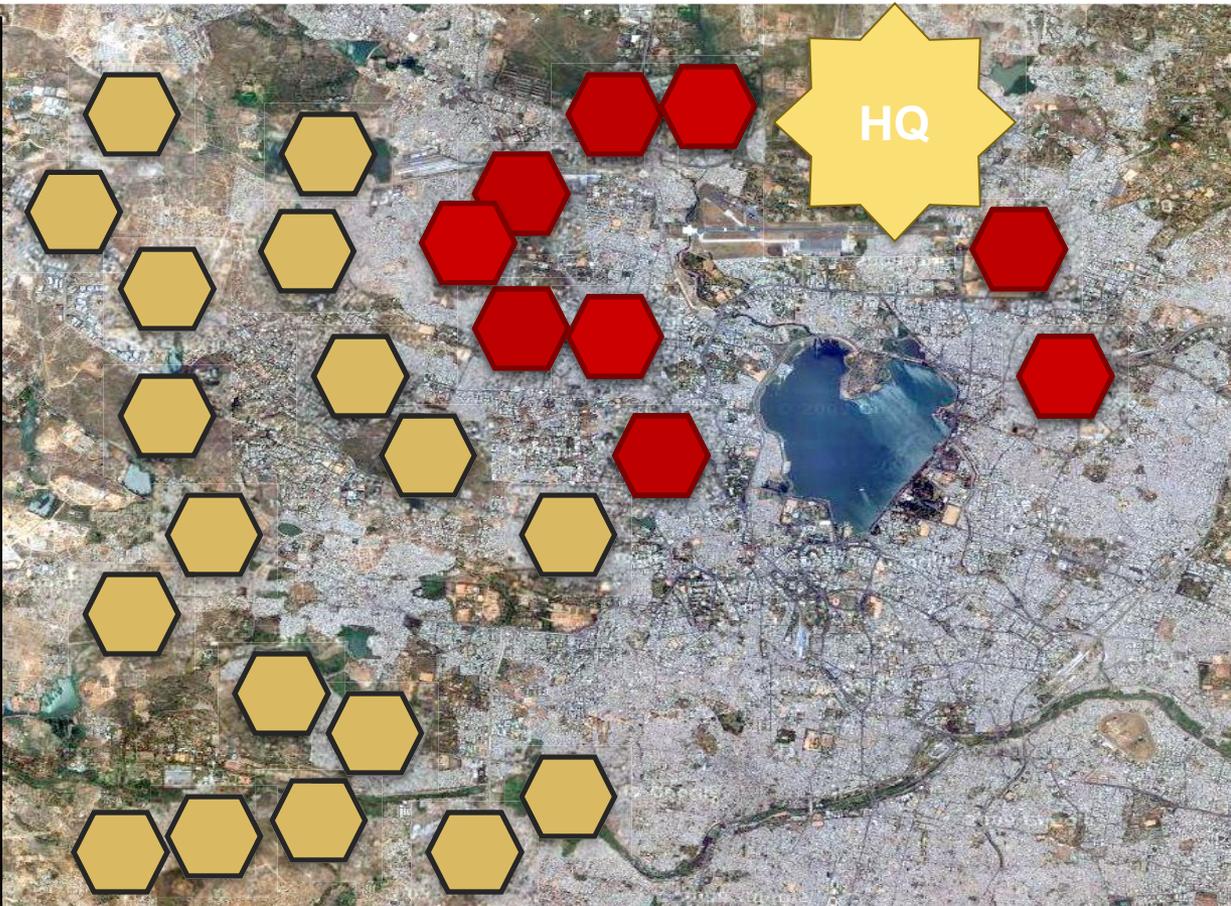
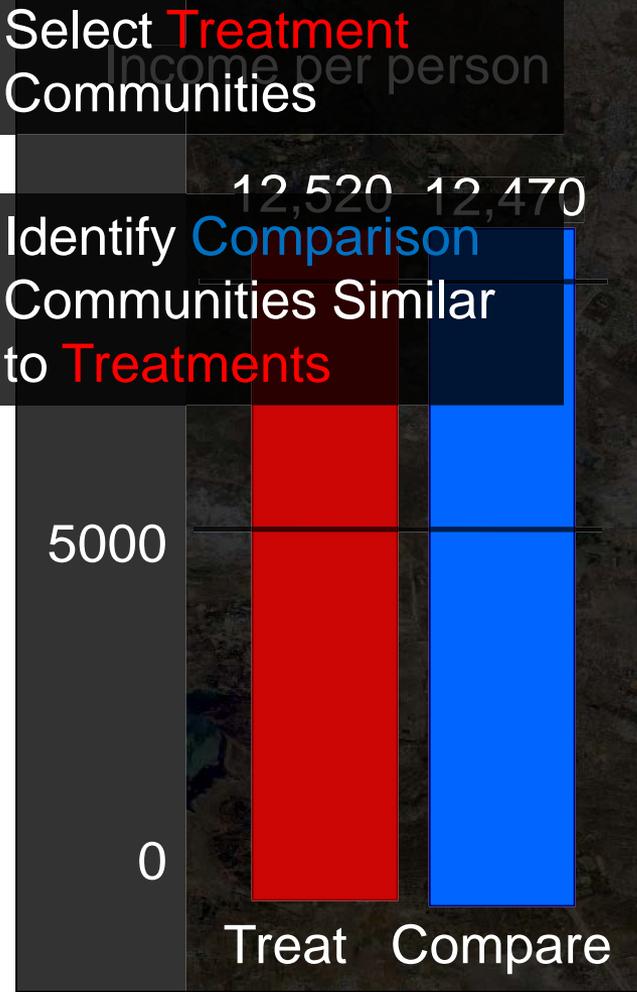
2. Difference in Differences

- 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
- Disadvantages
 - Reliance on assumption of **parallel trends**
 - Generally, how comfortable are we with this assumption?
 - Improved when groups are similar at baseline
 - Possibility of selection bias
 - Counterfactual less reliable than other IE designs (RCT, RD, PSM)

3. Matching

- 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



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

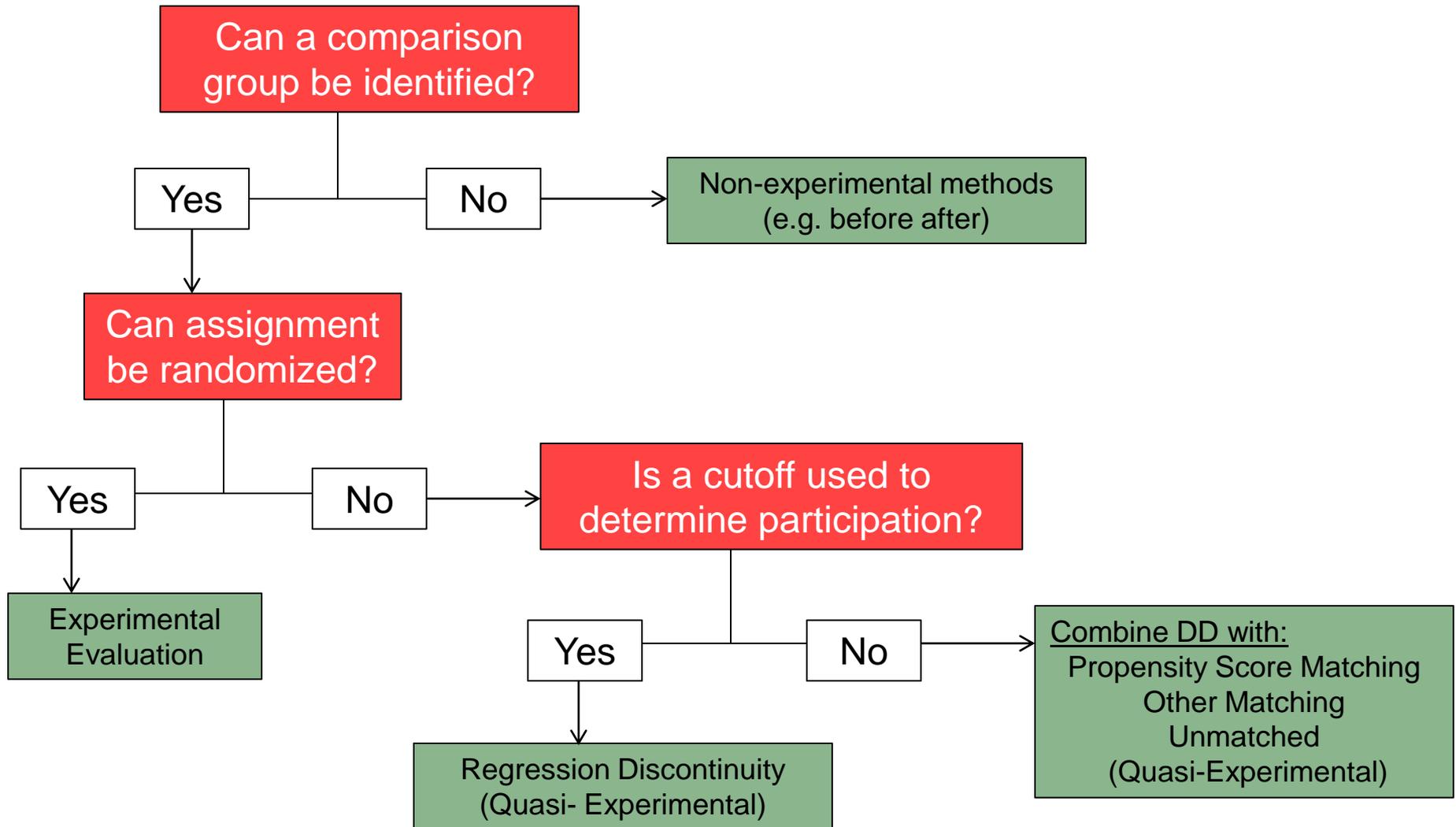
3. Matching

- 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
- Disadvantages
 - Cannot control for unobservables
 - 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?

Summary of IE Designs

- Gold standard is randomization – minimal assumptions needed, intuitive estimates
- Quasi-experimental approaches require assumptions
 - Hard to defend them sometimes

Summary – Which Design Do I Choose?



Exercise 7-1b: What Kind of Design Is It

Working in groups, use [Exercise 7-1b](#) to determine the best evaluation designs for each of the scenarios.

What Can Program Managers Do To Strengthen Impact Evaluations?

Deciding to evaluate stage

1. Be very selective about impact evaluation
2. Begin impact evaluation design as early as possible in the strategy and program design phase
3. Enlist IE specialists for scoping and developing the SOW for the impact evaluation
4. Begin to build a base of support for the impact evaluation through training of Mission and partner staff

What Can Program Managers Do To Strengthen Impact Evaluations?

Planning stage

1. If RCT, design intervention such that randomization is logistically possible
2. If QED, work with IE specialists to ensure that the comparison group is matched as closely as possible to the project group
3. Understand and try to reduce sample selection bias and assess its potential effects in the interpretation of the evaluation findings
4. Identify and test for alternative explanations of the differences between the project and comparison group outcomes

What Can Program Managers Do To Strengthen Impact Evaluations?

Managing stage

1. Continue to build support for the impact evaluation in the Mission and with IP's during the program launch phase
2. Modifying the program design requires costly modifications to the IE design—be prepared to manage changes to both
3. As data become available, look for negative cases that challenge the hypothesis of project impact and explore these carefully
4. Avoid accepting findings that support your hypothesis when the evidence is weak
5. Remember: Anecdotes and individual cases are not solid evidence. Avoid selective presentation of findings

What Can Program Managers Do To Strengthen Impact Evaluations?

Managing stage (cont)

6. Use triangulation to obtain two or more independent estimates of key findings
7. Review threats to the validity and conclusions based on internal and external threats identified in planning stage:
 - **Threats to external validity:** do the findings justify a recommendation that the program would produce similar results if replicated in a different setting?
 - **Threats to internal validity:** does the evaluation design support the conclusion that it was the program treatment that caused the observed outputs/impacts?

What Can Program Managers Do To Strengthen Impact Evaluations?

Using stage

1. Given the large investment in the impact evaluation ensure a strong communication strategy throughout the program life
2. Be sure to disseminate findings widely—they are intended to build the broader base of knowledge about development effectiveness and to inform future investment decisions
 - USAID and implementing partners
 - Host country governments
 - Other donors working in the sector
 - OECD/DAC
3. Publish findings in peer reviewed journals to build the knowledge base

Impact Evaluation Case Study

Steps

1. Using your case study as a base, choose an evaluation design you could do that would answer a cause-and-effect question.
2. Be prepared to explain which question you chose, and why you chose a particular IE design!



Review Questions

- What are some key impact evaluation questions?
- What are some tactics for defining the counterfactual?
- What are common impact evaluation designs?
- What can program managers do to strengthen impact evaluations?