Impact Evaluation of Development Programs

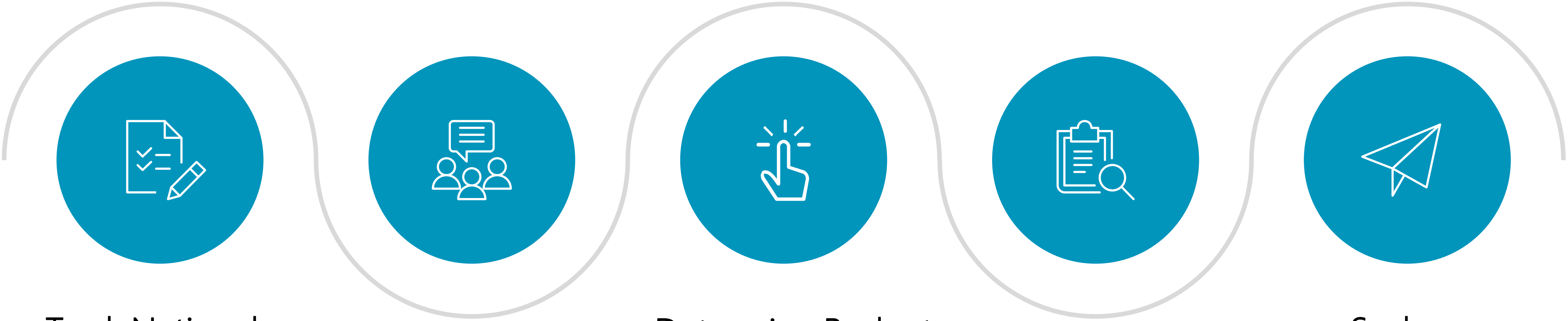
Dr. Pradeep Panda

Professor and Dean Academics

IIHMR Delhi

IIHMR DELHI

Why Impact Evaluation?



Track National and International Targets

Enhance Accountability and Lesson Learning

Determine Budget Allocation (curtailing inefficient programs)

Guide Program Design and Policy Decisions

Scale-up interventions that are successful

Impact Evaluation and Evidence-based Policy Making

- To test effectiveness of a given programme (whether a programme is effective compared to absence of the programme?)

- To test design innovations (whether a particular design innovation can boost programme effectiveness or lower cost?)

- To test effectiveness of program implementation alternatives (which one is most-effective program modality)

- To test heterogeneity in program impact across subgroups (whether a program is more effective for one subgroup?)

Examples of Impact Evaluation Questions

Impact of a Program

01

Did a water and sanitation program increase access to safe water and improve health outcomes?

Impact of a Program Modality

02

Did a new curricula raise test scores among students?

Impact of a Design Innovation

03

Was the innovation of including non-cognitive skills as a part of a youth training program successful in fostering entrepreneurship and rising incomes?

What is Impact Evaluation?

1

Seeks to answer a specific cause-and-effect question (CAUSALITY)

2

“What is the impact (or causal effect) of a program on an outcome of interest?”

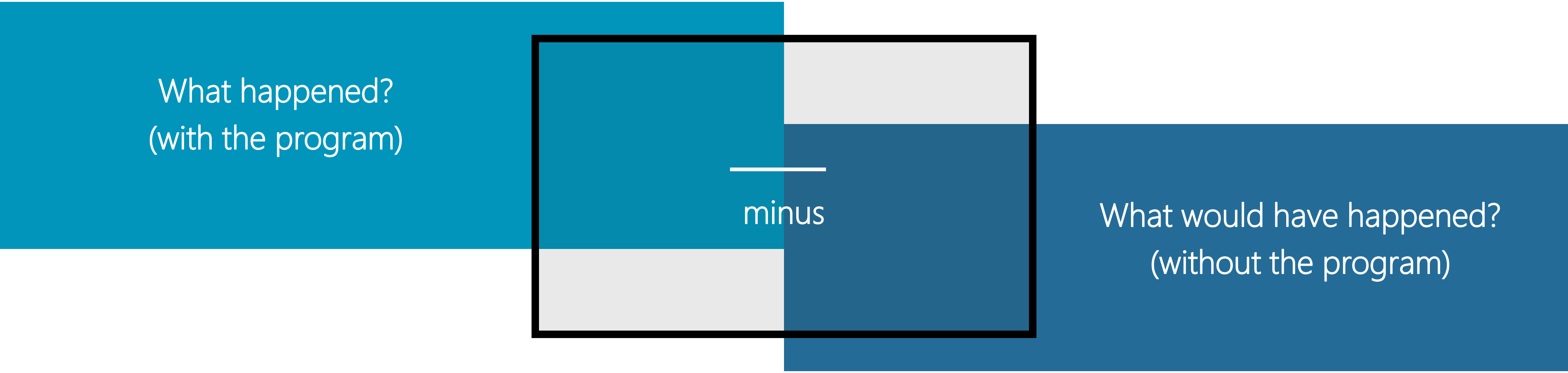
3

The focus is on impact, and the change directly attributable to a program, program modality, or design innovation (ATTRIBUTION)

How to Measure Impact?

What would have happened in the absence of the program?

Take the difference between:

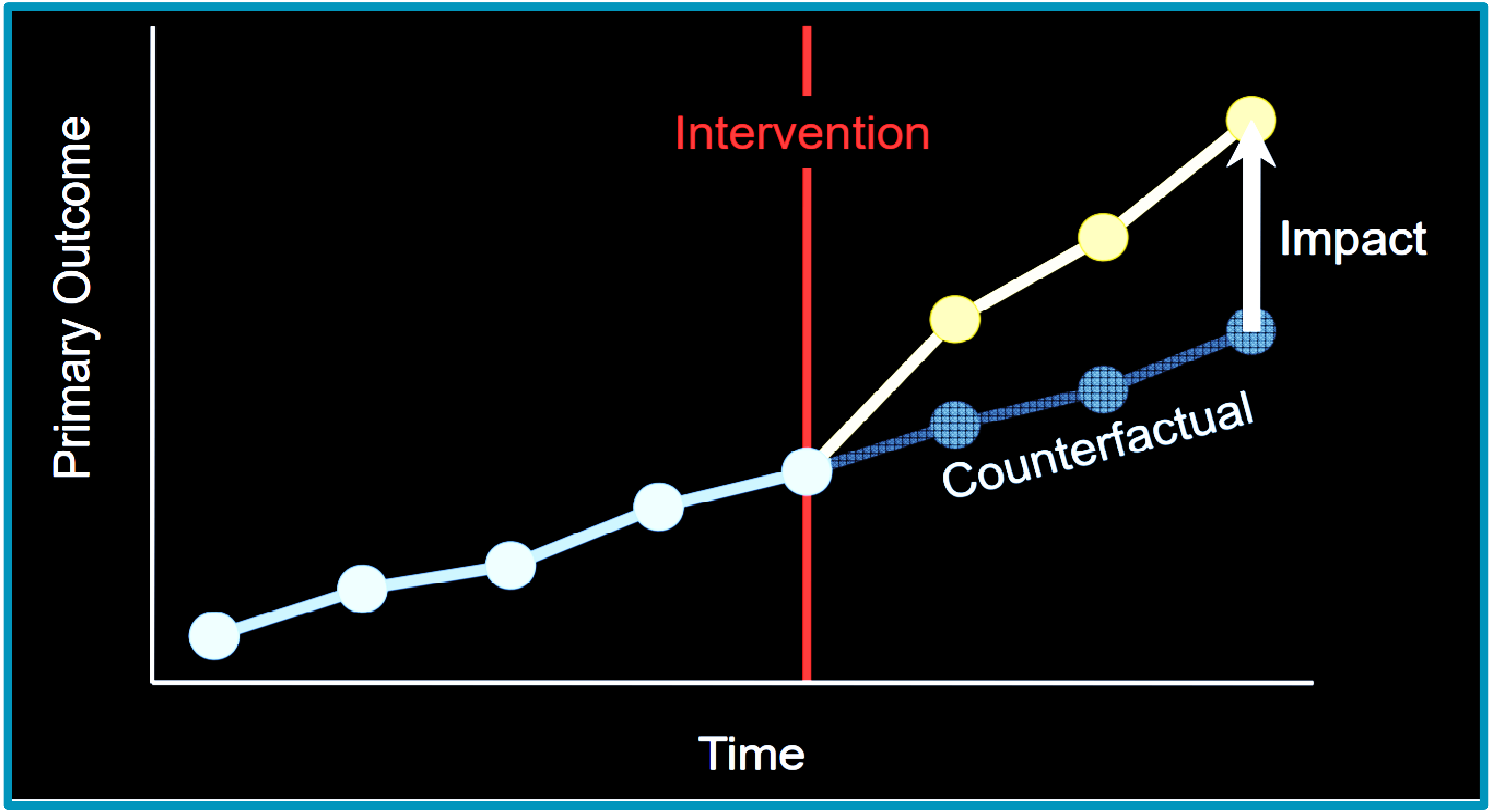


What happened?
(with the program)

minus

What would have happened?
(without the program)

Impact: What is it?



Impact and the Counterfactual

Counterfactual:
What would have happened in the absence of the program?

Since counterfactual is not observable, the key goal of all impact evaluation methods is to construct or “mimic” the counterfactual

Counterfactual can be estimated by selecting- a group not affected by the program: a *comparison* group or *control* group

Constructing the Counterfactual

Counterfactual is often constructed by selecting a group not affected by the program

RANDOMISED

Use random assignment of the program to create a control group which mimics the counterfactual

NON-RANDOMISED

Argue that a certain excluded group mimics the counterfactual

Methodologies in Impact Evaluation

EXPERIMENTAL

- Randomized Evaluations

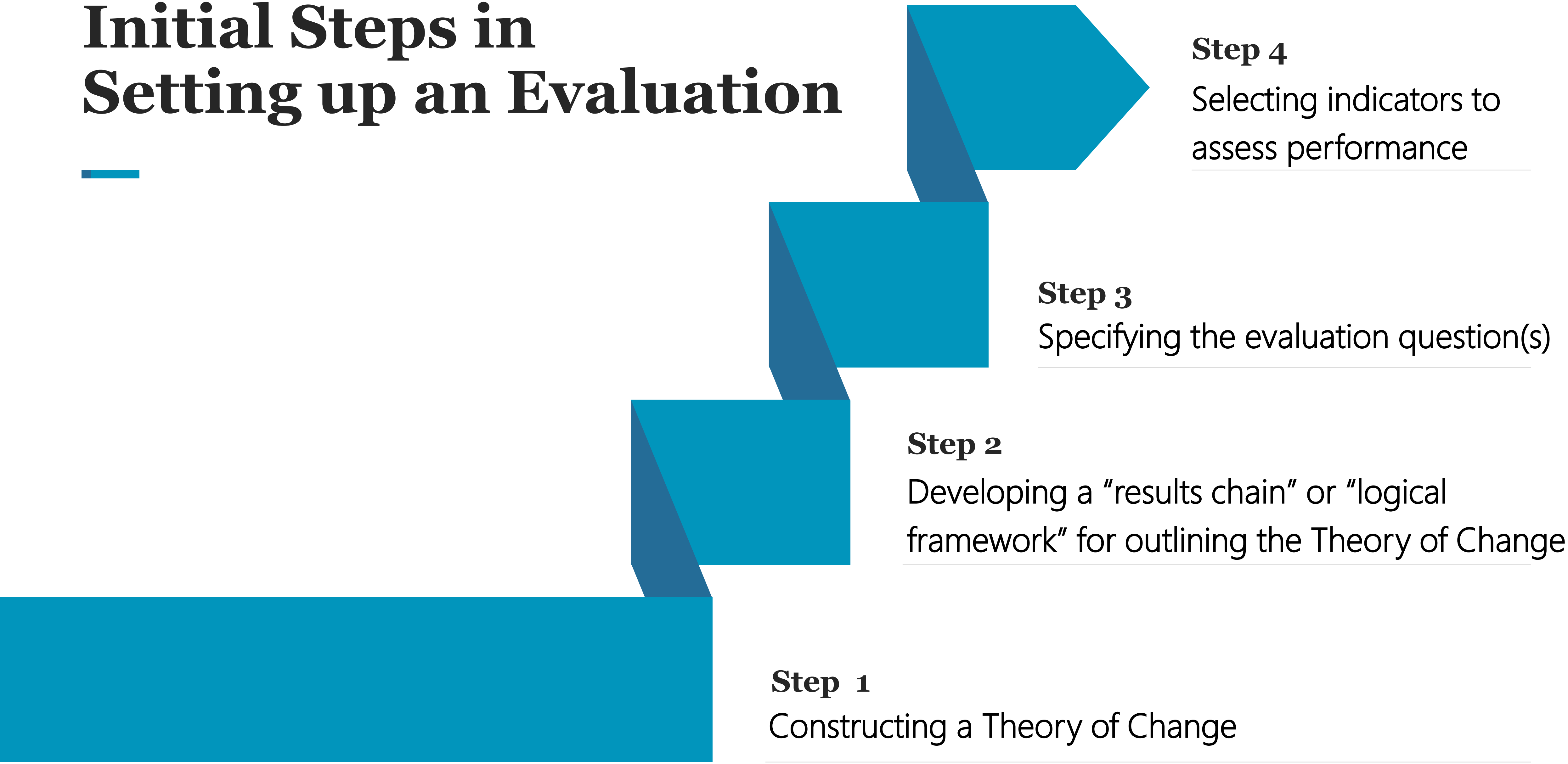
QUASI-EXPERIMENTAL

- Instrumental Variables
- Regression Discontinuity Design

NON-EXPERIMENTAL

- Pre-post
- Difference in Differences
- Regression
- Matching

Initial Steps in Setting up an Evaluation



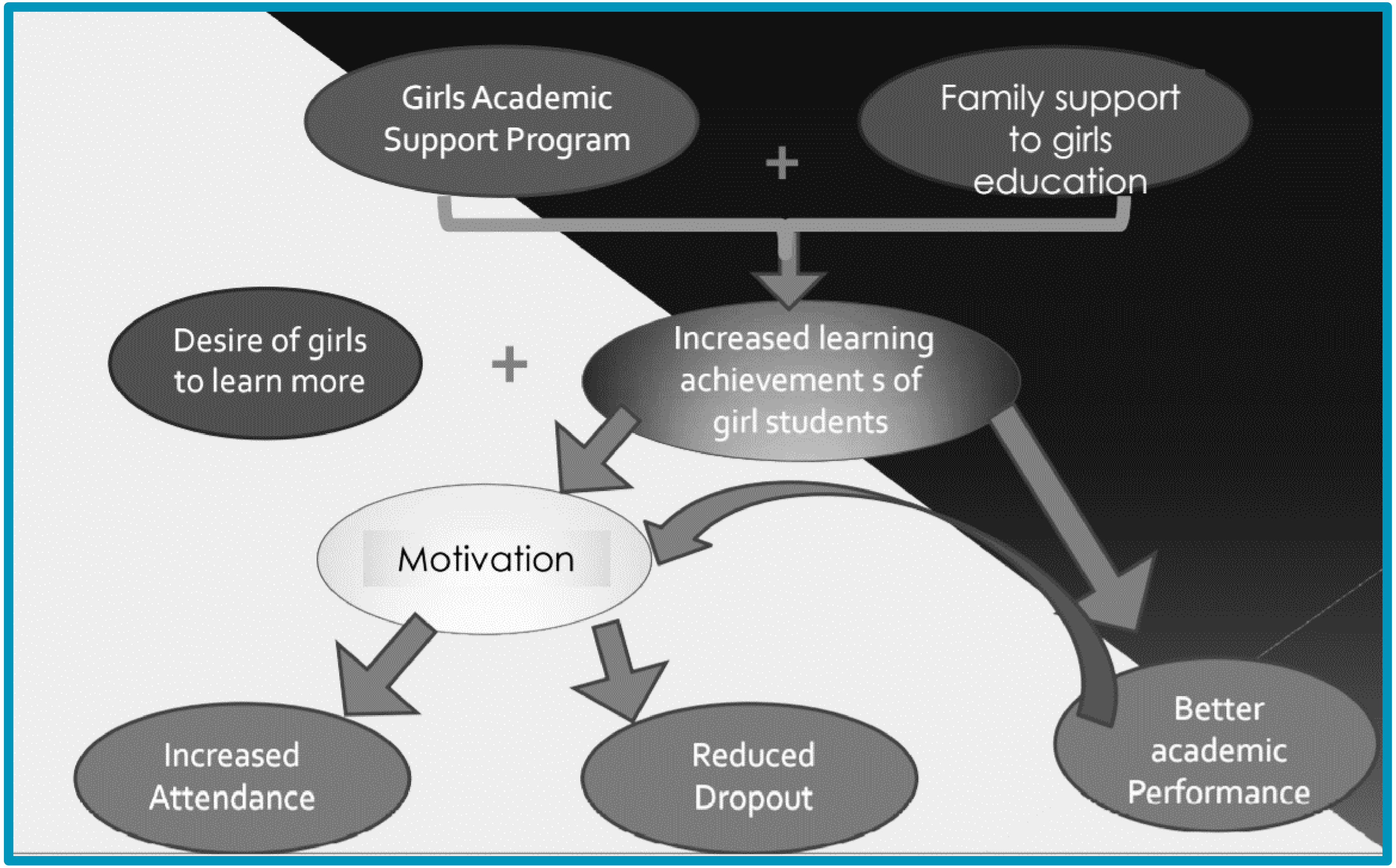
Step 4
Selecting indicators to assess performance

Step 3
Specifying the evaluation question(s)

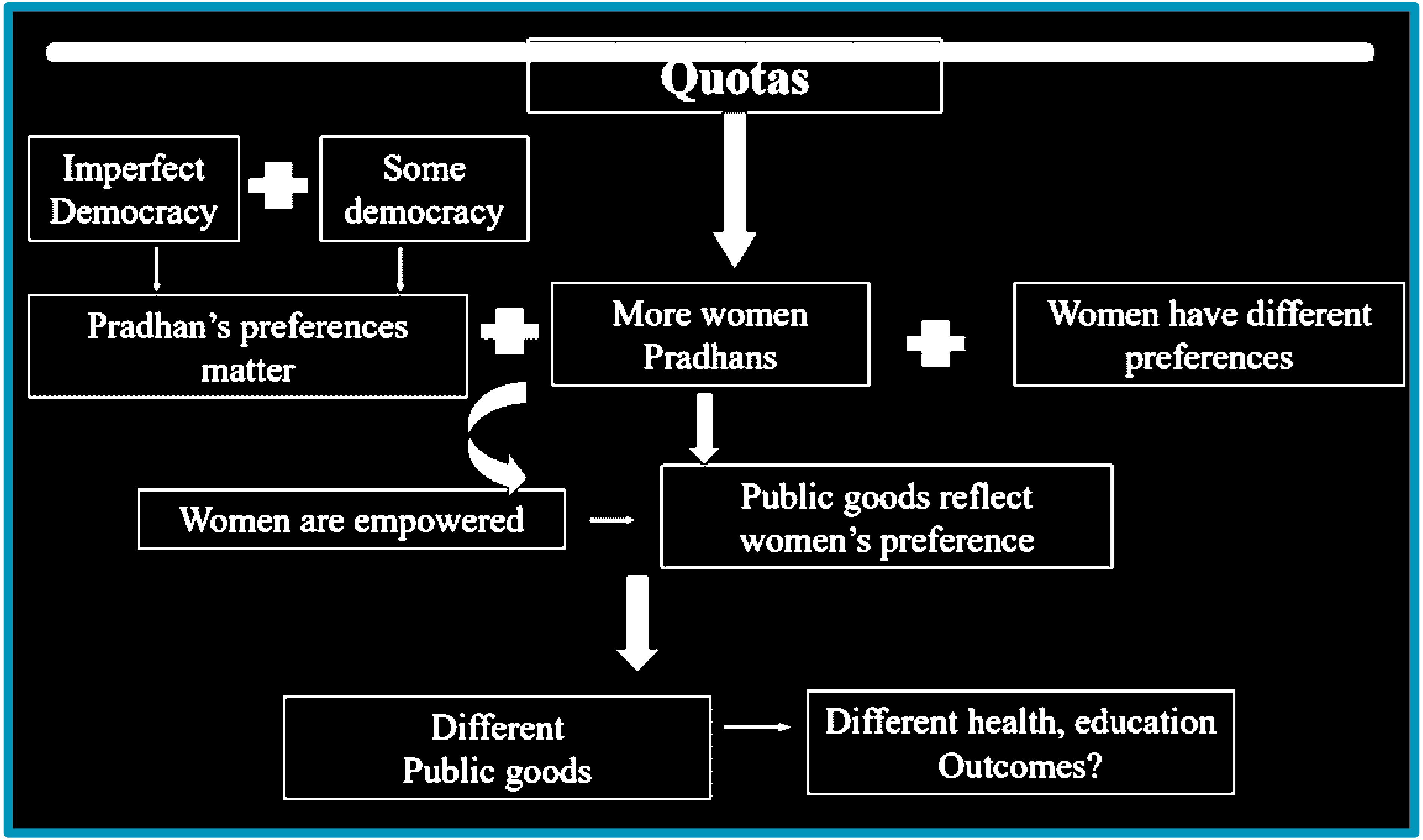
Step 2
Developing a “results chain” or “logical framework” for outlining the Theory of Change

Step 1
Constructing a Theory of Change

Theory of Change (1/2)


Girls Academic Support Program
+
Family support to girls education
Desire of girls to learn more
+
Increased learning achievement s of girl students
Motivation
Increased Attendance
Reduced Dropout
Better academic Performance

Theory of Change (2/2)


Quotas
Imperfect Democracy
Some democracy
Pradhan's preferences matter
More women Pradhans
Women have different preferences
Women are empowered
Public goods reflect women's preference
Different Public goods
Different health, education Outcomes?

Logical Framework (Example 1)

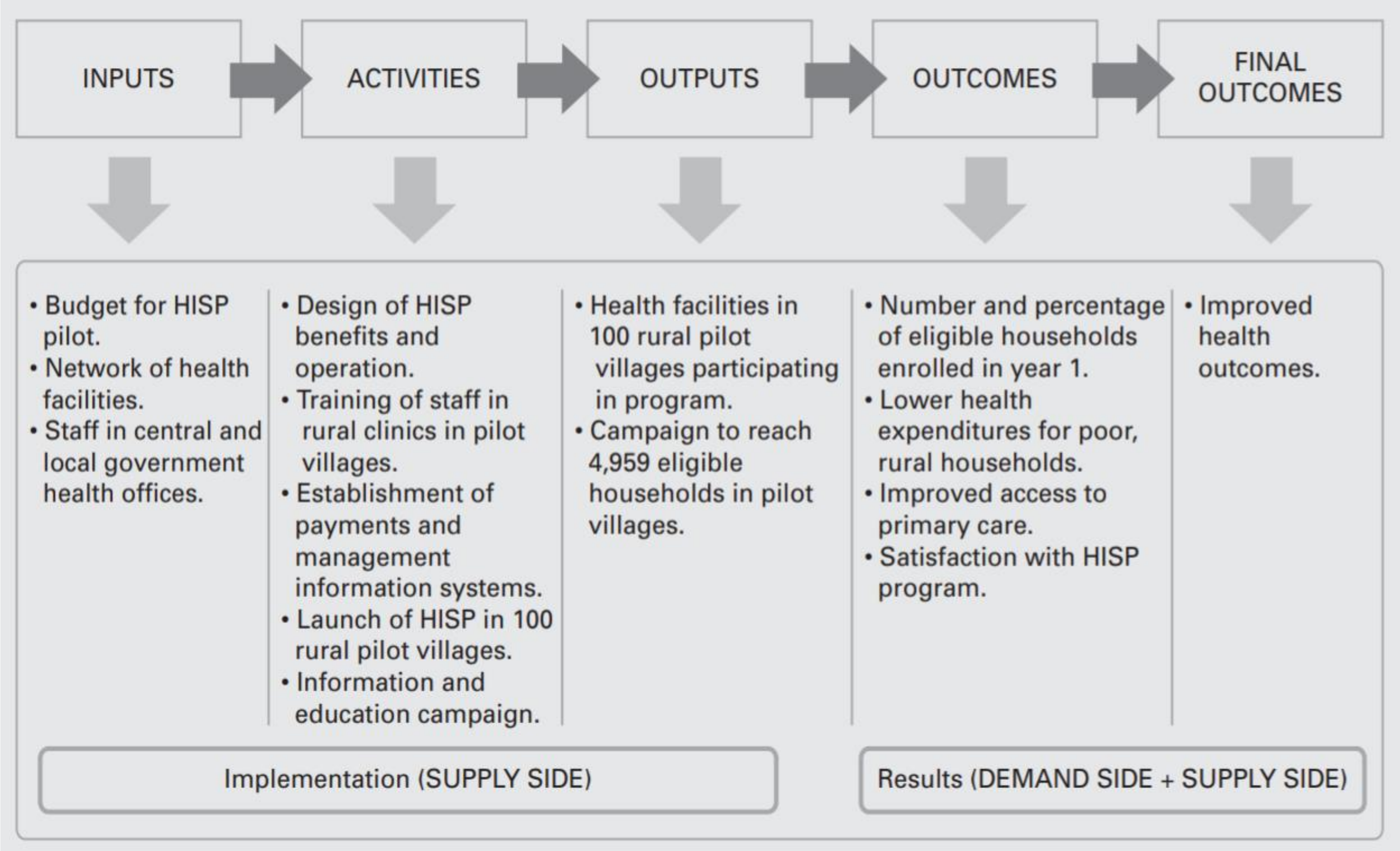
Needs	Input	Output	Outcome	Impact (short-term)	Goal (long-term)
The prevalence of DM and CVD is high; Government is not actively involved in their prevention & management through PHCs. Lack of knowledge	CHW screen for DM, CVD associated risk factors, education and follow up identified cases	At risk and diseased population is identified, given education and followed-up	Improved knowledge about the disease and lifestyle changes required to prevent/ manage them	Behavioral change in diseased and at risk populations. Reduction in risk factors and increased number of well managed cases of DM and CVD	Reduction in the incidence of DM and CVD

Logical Framework (Example 2)

Community Mobilization for Education

Needs	Input	Output	Outcome	Impact (short-term)	Goal (long-term)
Nearly 50% of children in rural India are functionally illiterate, despite being enrolled in school	NGO mobilizes community to monitor teacher attendance and activity	Parents visit schools daily and report teacher absence or failure to teach	Teachers attend school more regularly and teach when in school	Higher rates of literacy among school children	Improved educational outcomes and career opportunities

Developing a Results Chain



INPUTS → ACTIVITIES → OUTPUTS → OUTCOMES → FINAL OUTCOMES

INPUTS
- Budget for HISP pilot.
- Network of health facilities.
- Staff in central and local government health offices.

ACTIVITIES
- Design of HISP benefits and operation.
- Training of staff in rural clinics in pilot villages.
- Establishment of payments and management information systems.
- Launch of HISP in 100 rural pilot villages.
- Information and education campaign.

OUTPUTS
- Health facilities in 100 rural pilot villages participating in program.
- Campaign to reach 4,959 eligible households in pilot villages.

OUTCOMES
- Number and percentage of eligible households enrolled in year 1.
- Lower health expenditures for poor, rural households.
- Improved access to primary care.
- Satisfaction with HISP program.

FINAL OUTCOMES
- Improved health outcomes.

Implementation (SUPPLY SIDE)

Results (DEMAND SIDE + SUPPLY SIDE)

The HISP Results Chain

Specifying the Evaluation Question

Evaluation question is derived from the Theory of Change and formulated as a well-defined, testable hypothesis

What is the effect of a health insurance scheme on poor households' out-of-pocket health expenditures?

What is the effect of a new mathematics curriculum on students' test scores?

Selecting Outcome and Performance Indicators

A clear evaluation question should be accompanied by Outcome measures.

Outcome indicator is used to judge programme success

Outcome indicator forms the basis for the power calculation, used to determine the sample size

Minimum expected effect sizes determine programme success

Effect sizes are the changes expected as a result of the programme (Changes in Test Scores, Changes in Enrolment)

Selecting Outcome and Performance Indicators

If sample size is not large enough to detect changes, they are “underpowered”

Results Chain or Logical Framework informs the selection of indicators:

Indicators to monitor programme implementation and evaluate results

Randomized Controlled Trials

Also known as randomized assignment method

Represents the strongest method (gold standard) in evaluating impact

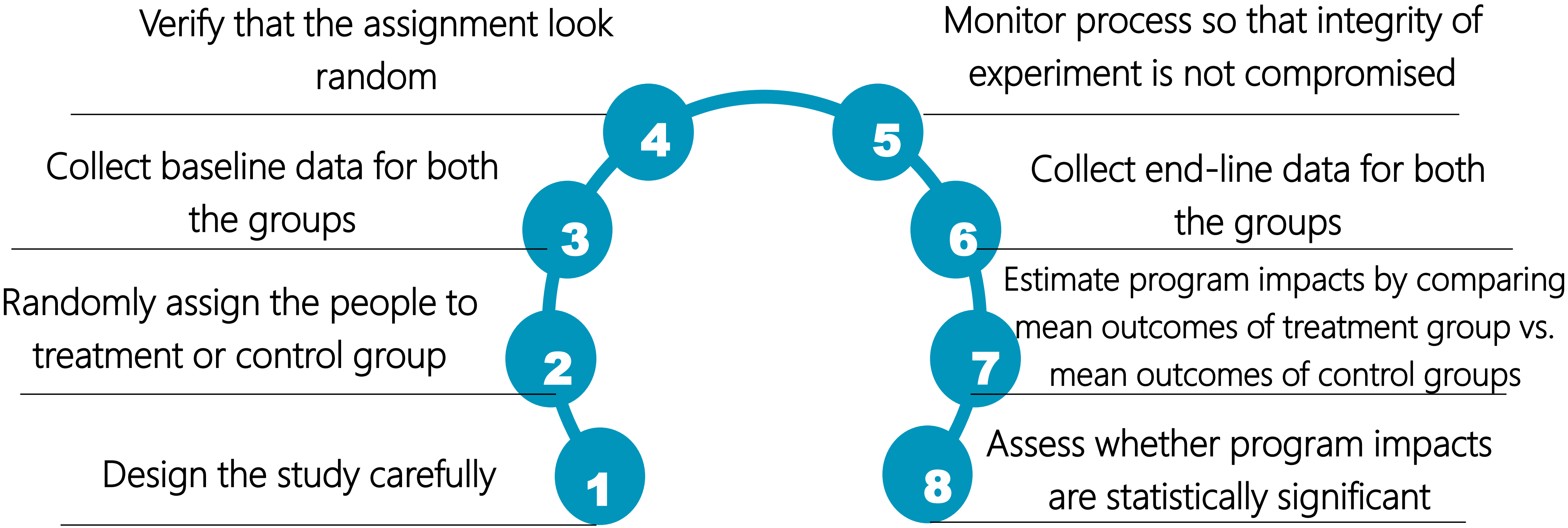
It can be verified from the baseline data

Uses fair and transparent rule for allocating scarce resources among equally deserving population

When a program is assigned at random, we can generate a robust estimate of the counterfactual.

Comparison group will be similar to Treatment group (statistically identical)

Steps in conducting a randomized experiment



1 Design the study carefully

2 Randomly assign the people to treatment or control group

3 Collect baseline data for both the groups

4 Verify that the assignment look random

5 Monitor process so that integrity of experiment is not compromised

6 Collect end-line data for both the groups

7 Estimate program impacts by comparing mean outcomes of treatment group vs. mean outcomes of control groups

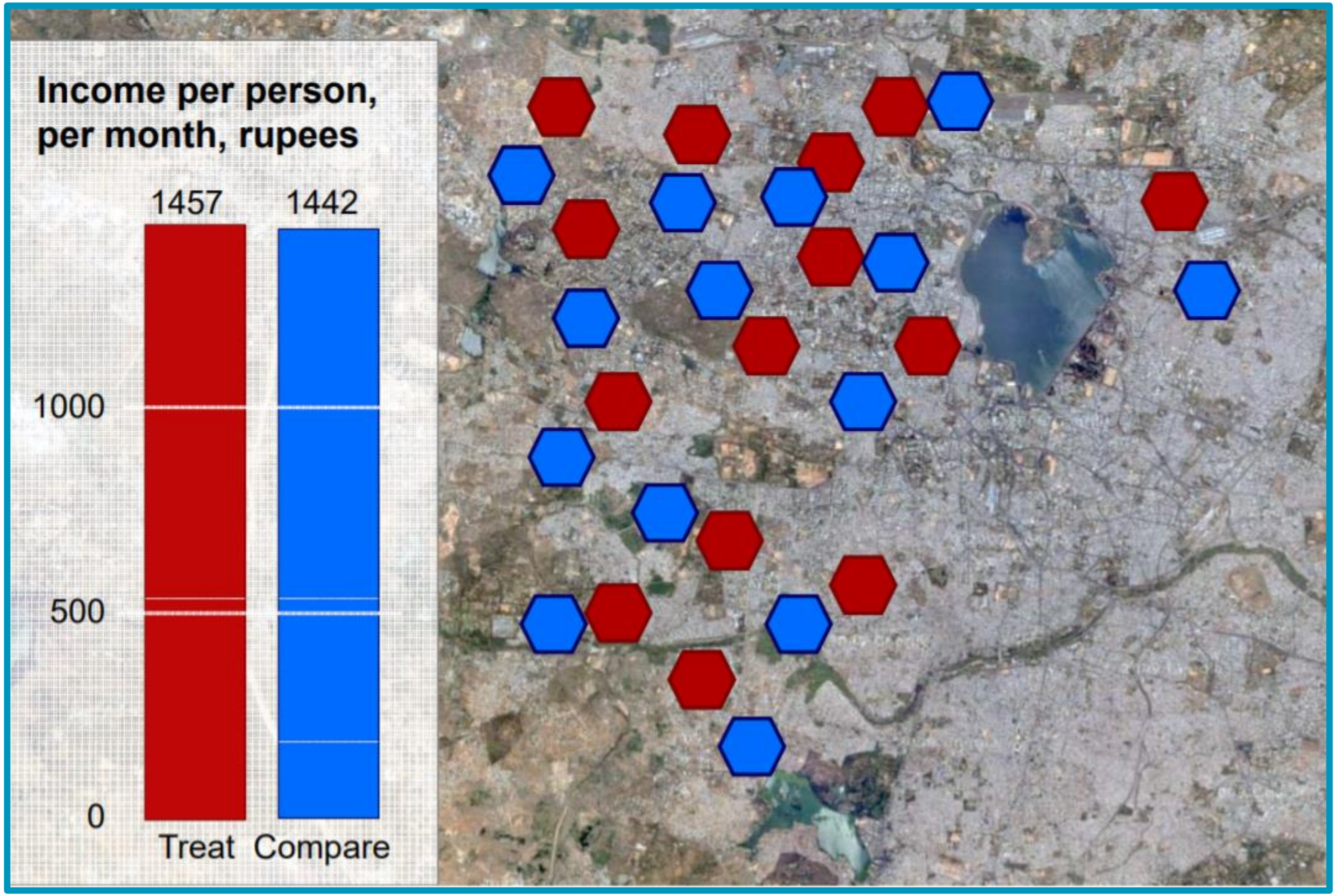
8 Assess whether program impacts are statistically significant

Impact of Balsakhi Program: Summary Results

Method	Impact Estimate
Pre-Post	26.42*
Simple Difference	- 5.05*
Difference-in-Differences	6.82*
Regression	1.92
Randomized Experiment	5.87*

* Statistically significant at the 5% level

Random Assignment



Options for Unit of Randomization

Which level to randomize?

Individual Level

Group Level (Cluster Randomized Trial)

Considerations:

What level at which the treatment is administered?
What is the unit of analysis?

Best to randomize at the level at which the treatment is administered

Sample size for randomized evaluation

When we use a “95% confidence interval”
How frequently we will “detect ” effective programme?

That is Statistical Power

Power: Main considerations

Variance – the more ‘noisy’, the harder it is to measure effects

More precise effect size to be detected – requires larger sample

Sample Size – Larger the sample size, more likely to obtain true difference

Standardized Effect Sizes

How large an effect you can detect with a given sample depends on how variable the outcome is

The standardized effect size is the effect size divided by the standard deviation of the outcome

Modest effect size (0.2), Large (0.5) and very large (0.8)

What affects effect size?

One variable that can affect effect size is take-up

A job training program increases income by 20%

But only 50% of the people in treatment group participate

We need to adjust impact estimate accordingly (from 20% to 10%)

The larger the sample, larger the power

Common power used – 80%, 90%

Clustered Design

Randomization at cluster level, Unit of analysis at individual level

We call r (rho) the correlation between the units within the same cluster

Value of r (rho_ must be between 0 and 1

A lower r is more desirable (0, 0.05, 0.08)

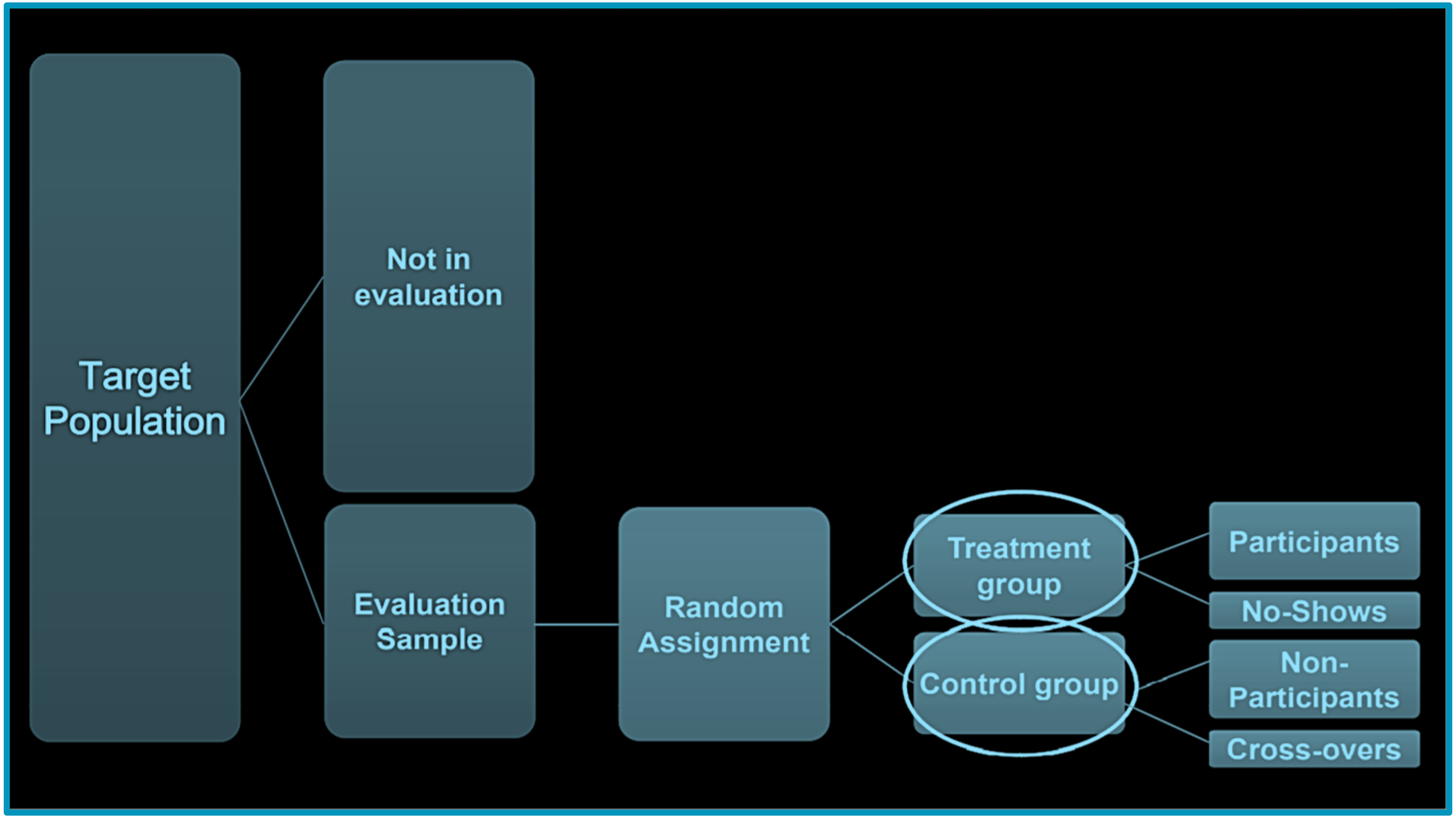
- We need to take clustering into account, when planning sample size
- It is extremely important to randomize adequate number of groups
- Often the number of individuals within groups matter less than the total number of groups

How to measure program impact in the presence of spillovers ?

Design the unit of randomization so that it encompasses the spillovers

If we expect externalities that are all within school, randomization at the level of school allows for estimation of the overall effect.

Basic set up of a randomized evaluation



Sample selection bias (1/2)

Sample selection bias could arise if factors other than random assignment program allocation

Individual assigned to comparison group could attempt to move into treatment group

Alternatively, individuals allocated to treatment group may not receive treatment

Some students in treatment school not treated (22%)

Some students in comparison school treated (5%)

What do you do?

Sample selection bias (2/2)

Use the original assignment

If a child ended up in a treatment school but was from the control, she should be assigned to control when calculating the effect

This gives us the Intention to Treat (ITT)

Intention to Treat (ITT)

What does "intent to treat" measure?

"What happened to the average child who is in a treated school in this population?"

Is this the right number to look for?

Remember: In the deworming case, many children in treatment schools were not treated and some children in comparison schools were.

When is ITT useful?

May relate more to actual programs

For example, we may not be interested in the medical effect of deworming treatment, but what would happen under an actual deworming program?

If students often miss school and therefore don’t get the deworming medicine, the intention to treat estimate may actually be most relevant

Treatment on the treated (TOT)

The effect of the treatment on those who got the treatment:

- Suppose children who got the treatment had a weight gain of A, irrespective of whether they were in a treatment or a control school
- Suppose children who got no treatment had a weight gain of B, again in both kinds of schools
- We want to know A-B, the difference between treated and non-treated students

Then:

$$Y(T) = A * \text{Prob}\,[\,\text{treated} \mid T\,] + B\,(\,1 - \text{Prob}\,[\,\text{treated} \mid T\,]$$

$$Y(C) = A * \text{Prob}\,[\,\text{treated} \mid C\,] + B\,(\,1 - \text{Prob}\,[\,\text{treated} \mid C\,]$$

$$A - B = (\,Y(T) - Y(C)\,) \,/\, (\,\text{Prob}\,[\,\text{treated} \mid T\,] - \text{Prob}\,[\,\text{treated} \mid C\,]\,)$$

= The “treatment on the treated” effect

External Validity

Internal validity is a necessary condition for the results of a randomized experiment to be generalizable

But it’s not sufficient

Threat to External Validity: Behavioral responses to evaluations

One limitation of randomized evaluations is that the evaluation itself may cause the treatment or comparison group to change its behaviour

Treatment group behavior changes: Hawthorne effect

Comparison group behavior changes: John Henry effect

In Addition: a program may generate behavioral responses that would not occur if the program were generalized

Generalizability of results

Depends on three factors:

1

Program Implementation: Can it be replicated at a large (national) scale?

2

Study Sample: Is it representative?

3

Sensitivity of Results: would a similar but slightly different program, have the same impact?

Thank You