REAL-WORLD IMPACT EVALUATION – APPLYING IE METHODS CREATIVELY

7 May 2017

Jo (Jyotsna) Puri – Head of Evaluation, GCF
Anna Henttinen – IE Specialist, WFP
Bidisha Barooah – Evaluation Specialist, 3ie

Q1: Have you worked on an IE before?

Q2: Do you think you might work on an IE in the future?

WELCOME!

UNEG Impact Evaluation Workshop – Real-World Impact Evaluation – Applying IE methods creatively 7 May 2018 – Draft Agenda

Workshop Purpose: To consider options for designing creative impact evaluations in difficult (i.e. real-world) contexts where data may not be available or the context may be shifting. The objective of the workshop is to introduce the audience to main impact evaluation techniques, and share how they have been applied creatively. The morning of the workshop will be more traditional presenations and discussion, while during the afternoon, the participants will have an opportunity to design their own impact evaluation with support from the facilitators.

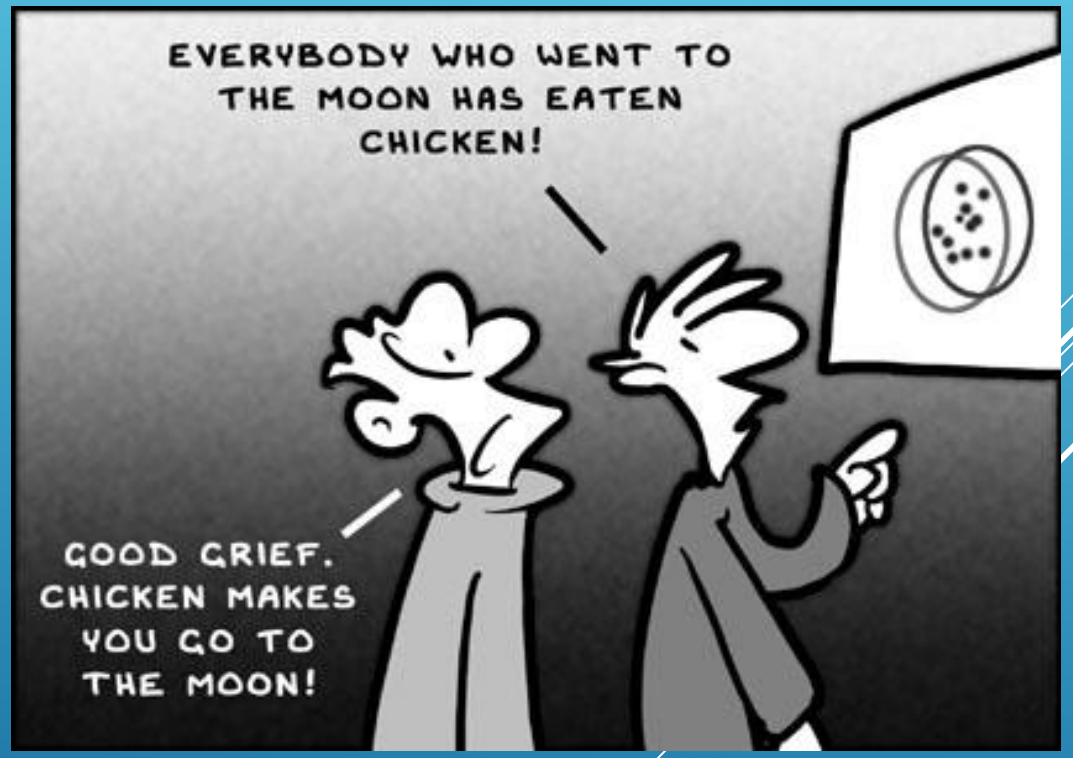
Workshop style: This workshop will be facilitated, and highly participatory, with presentations and discussion in the morning and an interactive impact evaluation design exercise in the afternoon.

Time	Session Description	Activities	Intended Outcomes
09:00-09:15	Introduction and welcome to the day.	• Go through the workshop expectations and agenda, purpose and outline of the day, scope and outcomes.	• Clarity over the purpose and scope of the day and objectives.
09:15-10:30	What is Impact Evaluation and what are the common design options	• Definition of impact evaluation • Basic design frames for undertaking impact evaluations: Experimental and Quasi-Experimental Impact Evaluation designs, with a particular focus on quasi-experiments.	• Shared understanding of the main methods and definition of impact evaluation.
10:30-11:00 Break			
11.00-12.30	Examples of Being Creative with Impact Evaluation designs	• Examples from GCF and WFP on how impact evaluation techniques have been applied creatively in the field	• Exploration of 'real-world' scenarios and application of quasi-experimental methods.
12:30- 14:00 Lunch			
14:00–14:15	Introduction to the Afternoon and the Impact Evaluation Design Game	• Explain the afternoon session and the 'impact evaluation design game'.	• Clarity over the mode and agenda for the afternoon.
14:15-15:30	Planning and designing your impact evaluation	• Audience divided into groups to design their impact evaluation on a specific topic, based on a menu of design choices.	• Participants can apply their own knowledge and expertise, with the help of the facilitators to design impact evaluations.
15:30-16.00 Break			
16:00-17:00	Presenting the different Impact Evaluation Designs	• Feedback from groups on their design choices	• Participants share their work and designs choices to the groups.

SESSION 1 - IMPACT EVALUATION – COMMON DESIGN OPTIONS

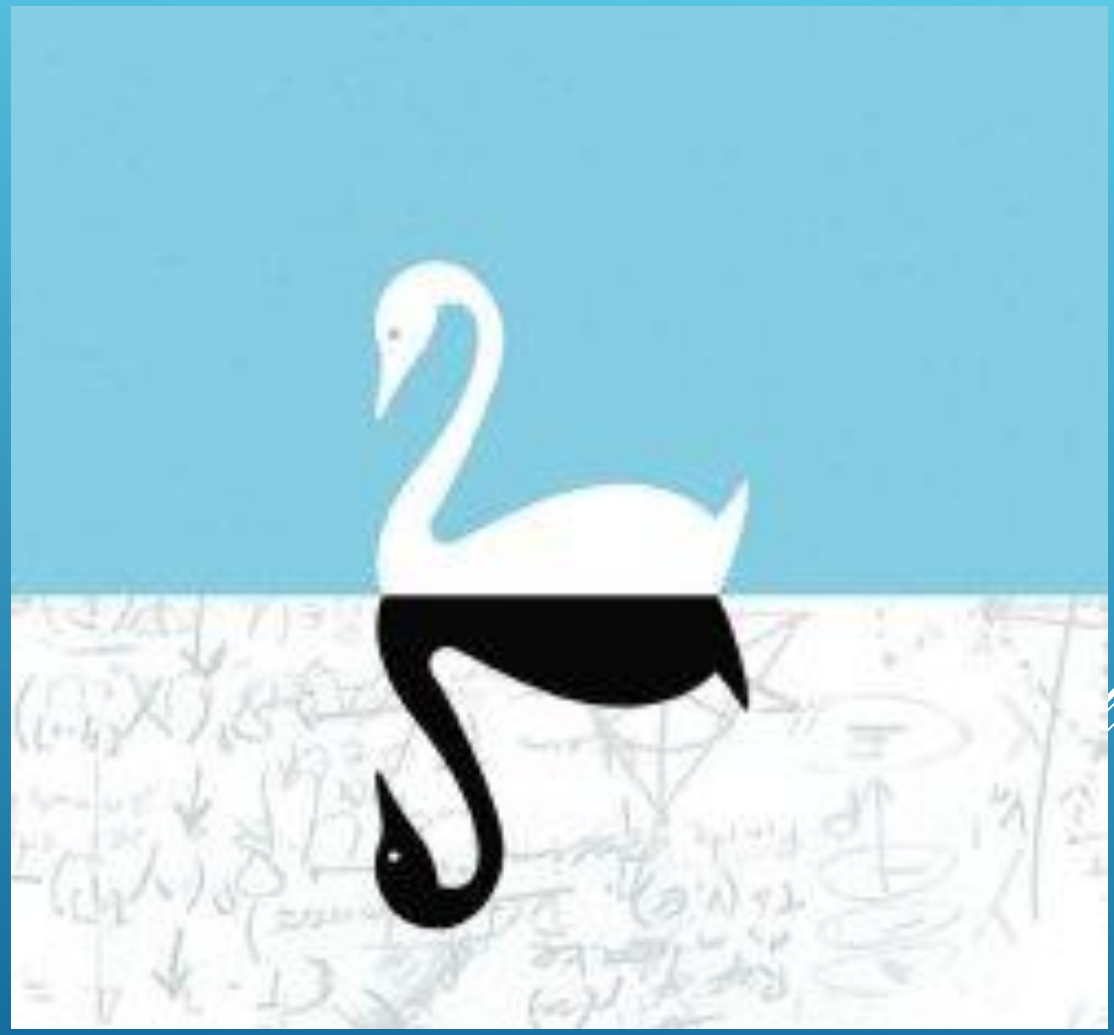
WHAT IS IMPACT EVALUATION?

- Answers **Cause-and-Effect** questions
- Can identify *what* happened and *how* it happened
- *Works at any point* of the results chain;
- Can identify *who* benefitted or if a programme is cost-effective.
- Can measure short, medium or long-term effects
- Can be retrospective or prospective



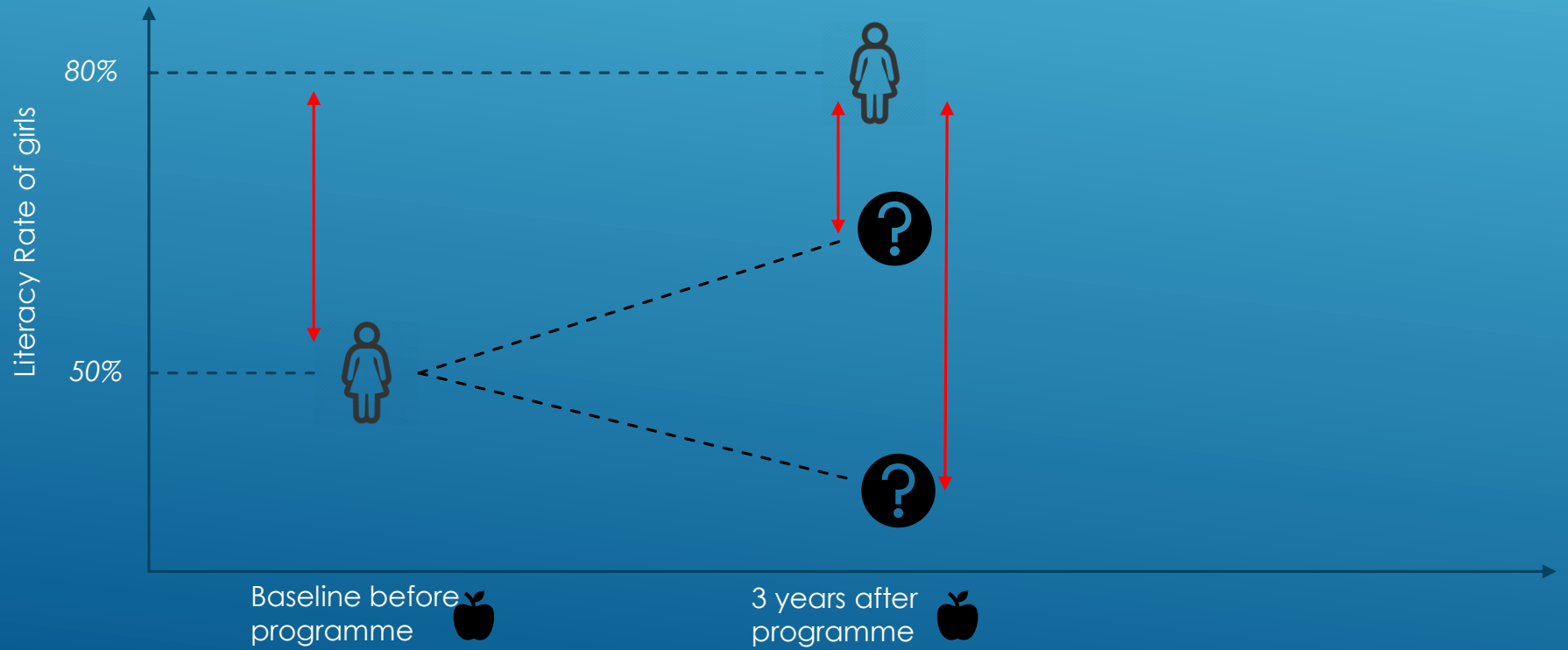
DEFINITION – WHY IT MATTERS

- Most international organisations, and Donors include the following words:
  - Counterfactual
  - Attribution



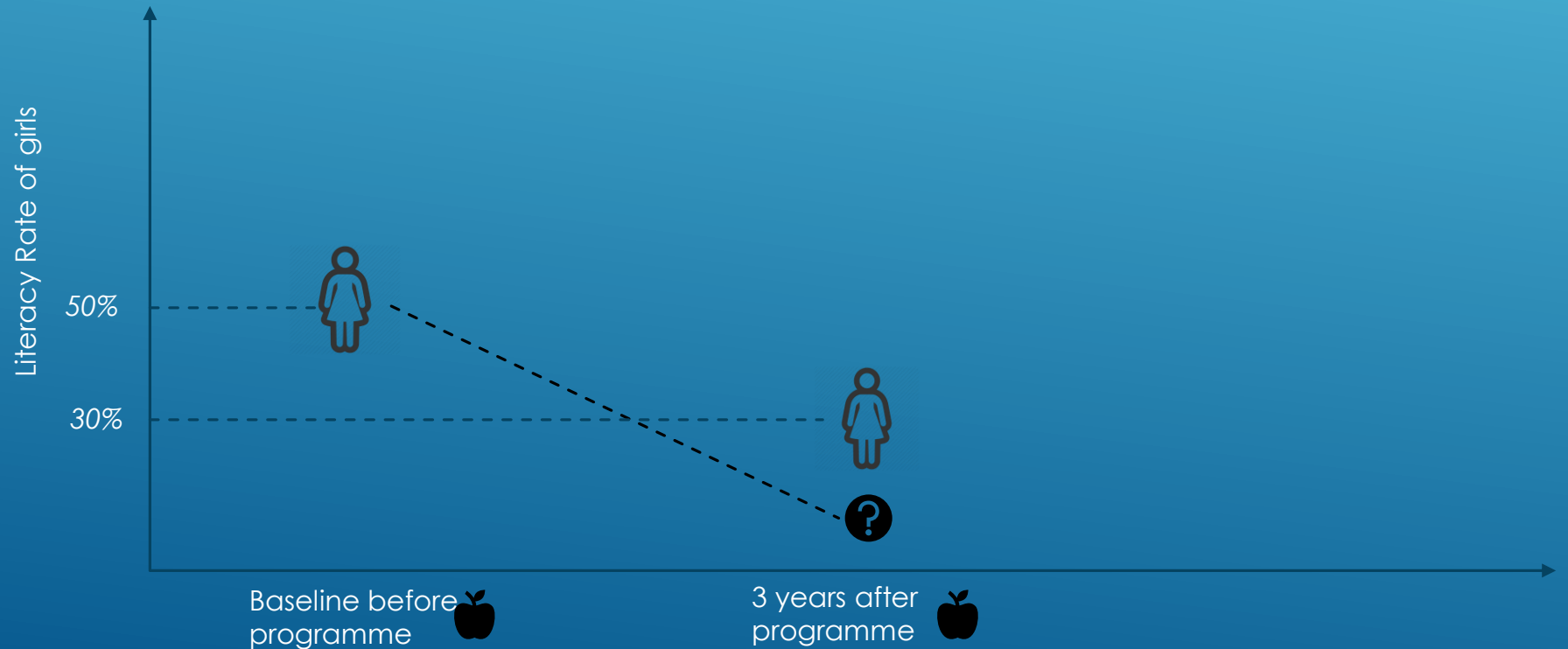
Counterfactual matters

- *Use of a credible counterfactual* to identify **what would have happened in the absence of the intervention**



Why counterfactual matters

- *What if outcomes and impact worsen during operations?*



WHAT IF WE DIDN'T DO THEM

What do we need to measure impact?

PROVIDING CASH TRANSFERS TO THE DISADVANTAGED AND LOW INCOME GROUPS

	Before	After
Project (treatment)		92
comparison		

The majority of evaluations have just this information … which means we can say absolutely nothing about impact

BEFORE VERSUS AFTER SINGLE DIFFERENCE COMPARISON

BEFORE VERSUS AFTER = 92 – 40 = 52

	Before	After
Project (treatment)	40	92
comparison		

"the cash transfer project has led to a higher incomes in a number of villages"

This 'before versus after' approach is outcome monitoring.

Outcome monitoring has its place, but it is not impact evaluation

POST-TREATMENT COMPARISON COMPARISON
SINGLE DIFFERENCE = 92 – 84 = 8

	Before	After
Project (treatment)		92
comparison		84

But we don’t know if they were similar before…

DOUBLE DIFFERENCE = (92-40)-(84-26) = 52-58 = -6

	Before	After
Project (treatment)	40	92
comparison	26	84

Conclusion: Longitudinal (panel) data, with a comparison group, allow for the strongest impact evaluation design (though still need matching).

SO WE NEED BASELINE DATA FROM PROJECT AND COMPARISON AREAS

What do we need to measure impact?

	Before	After
Project		
Comparison		

SO IN FACT

	Before	After
Project		
Comparison		

EXERCISE: 10 MINUTES

EXERCISE: PART 1

Step 1: Think of an intervention you would like to assess the impact of.

Step 2: Define one main impact indicators for your intervention

Step 3: Using hypothetical outcome data for one indicator write down the before/after, comparison/treatment numbers in the table below

	Before	After
Project		
Comparison		

Step 4: Write down the following numbers in the sheet you received:

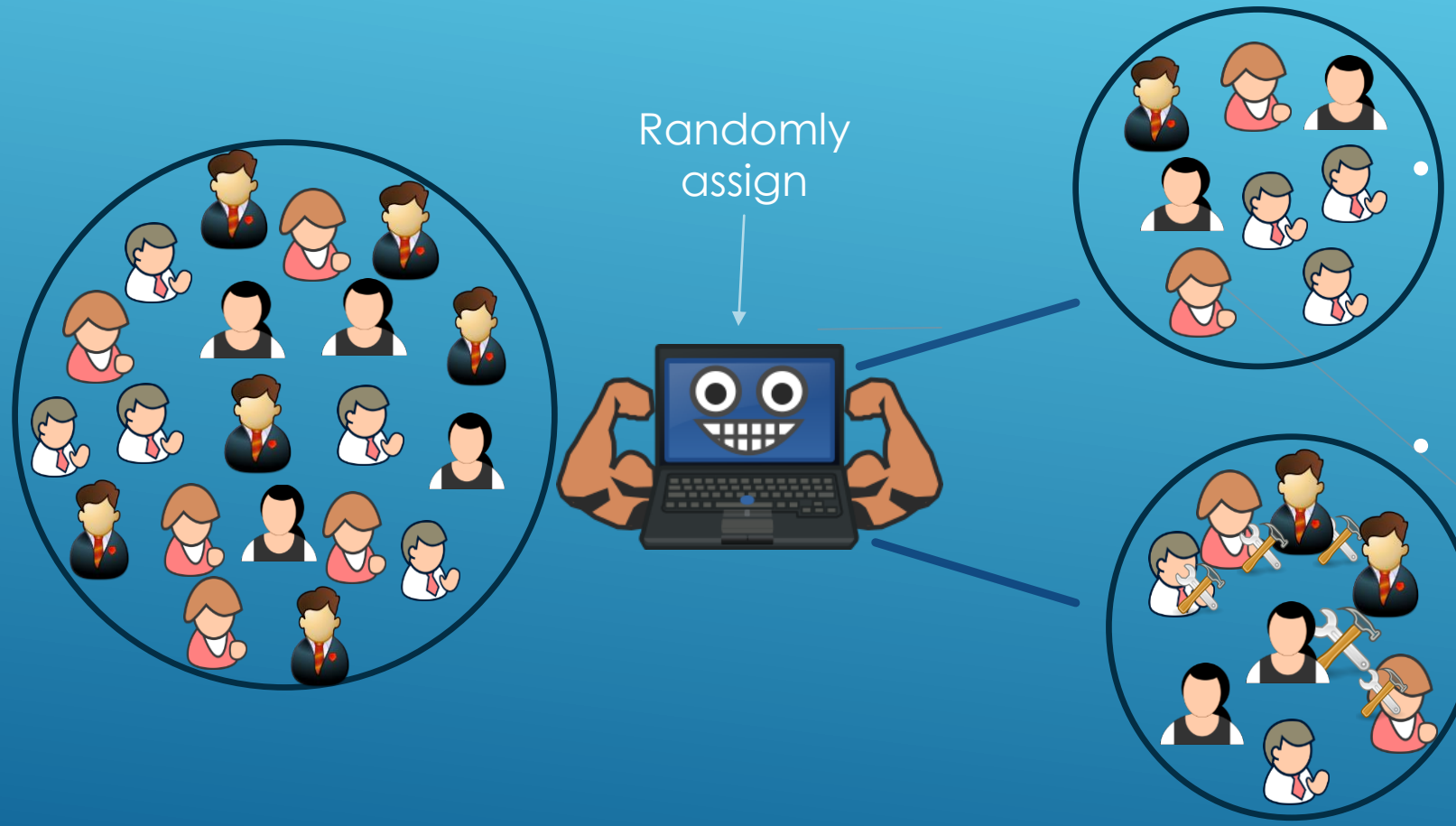
- Ex-post single difference
- Before versus after (single difference)
- Double difference impact estimates

HOW DO YOU CREATE A COUNTERFACTUAL?

METHODS – FEW NOTES BEFORE WE DIVE IN…

- Most development impact evaluations today use different methods and mixed methods.
- Some are 'conventional' RCTs … but increasingly other more creative methods are used in more complex settings.
- What follows is a light taster of a range of methods…

RANDOMIZATION

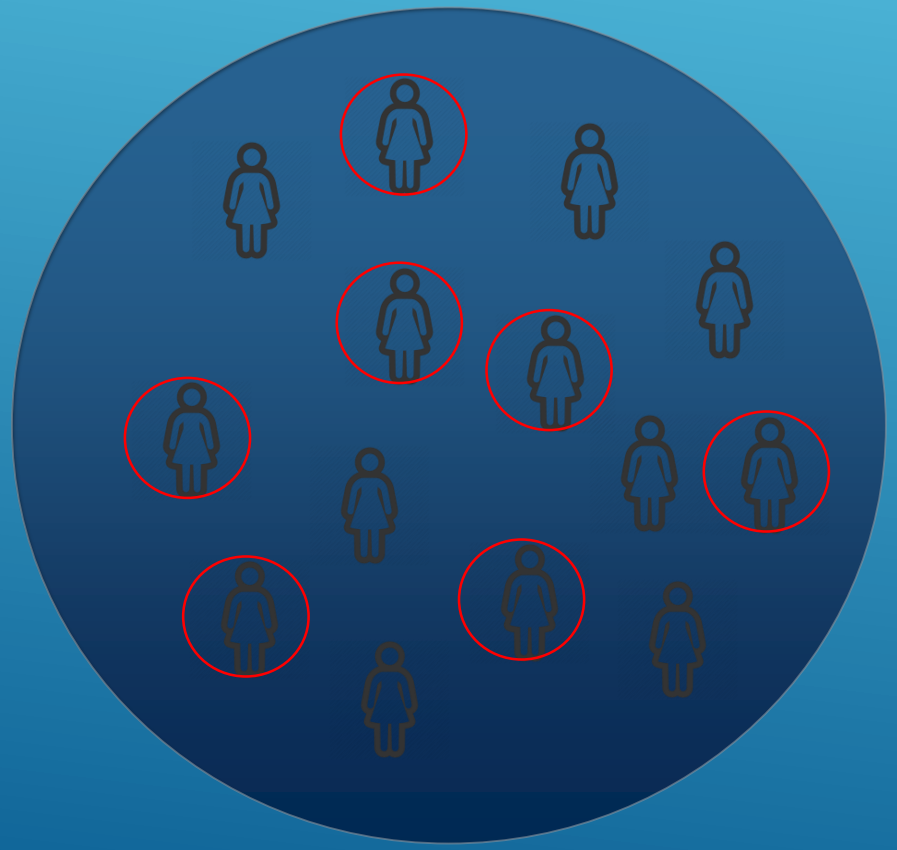


- Two levels of randomization
  - Individual randomization
  - Cluster randomization
- Individual randomization
  - Threats of spillover and contamination
  - Ethics
- Cluster randomization
  - Eg. Schools instead of students
  - Sample size requirements may be bigger

‘BUT I CANNOT RANDOMIZE EVERYONE IN MY PROGRAM…!’

- Pipeline design
  - Most development programs are implemented in phases. Assignment to phases is random
  - Measures duration of program
- Factorial design
  - All groups get a base treatment
- Lottery
  - Oversubscription to a program
- Encouragement design
  - Low sign-up to a program, encourage to increase participation

RCTs – two practical ways to include an RCT



MEASURING IMPACT – THE CHALLENGES

Program placement is hardly ever random.

There is ‘selection’ in who benefits from nearly all interventions.

Need a comparison group which has the same characteristics as those selected for the intervention.

COMMON QUASI-EXPERIMENTS

- Propensity Score Matching
- Difference in Differences
- Regression Discontinuity Design
- Instrumental variable

PROPENSITY SCORE MATCHING

- Prennushi and Gupta (2014)
  - Evaluate a program on woman's empowerment where women are mobilized into self-help groups. Joining a group is voluntary
  - Compare participants to non-participants
    - Not as simple as matching on means
  - Each observation gets a 'score' of its probability of being in the program based on its observable characteristics

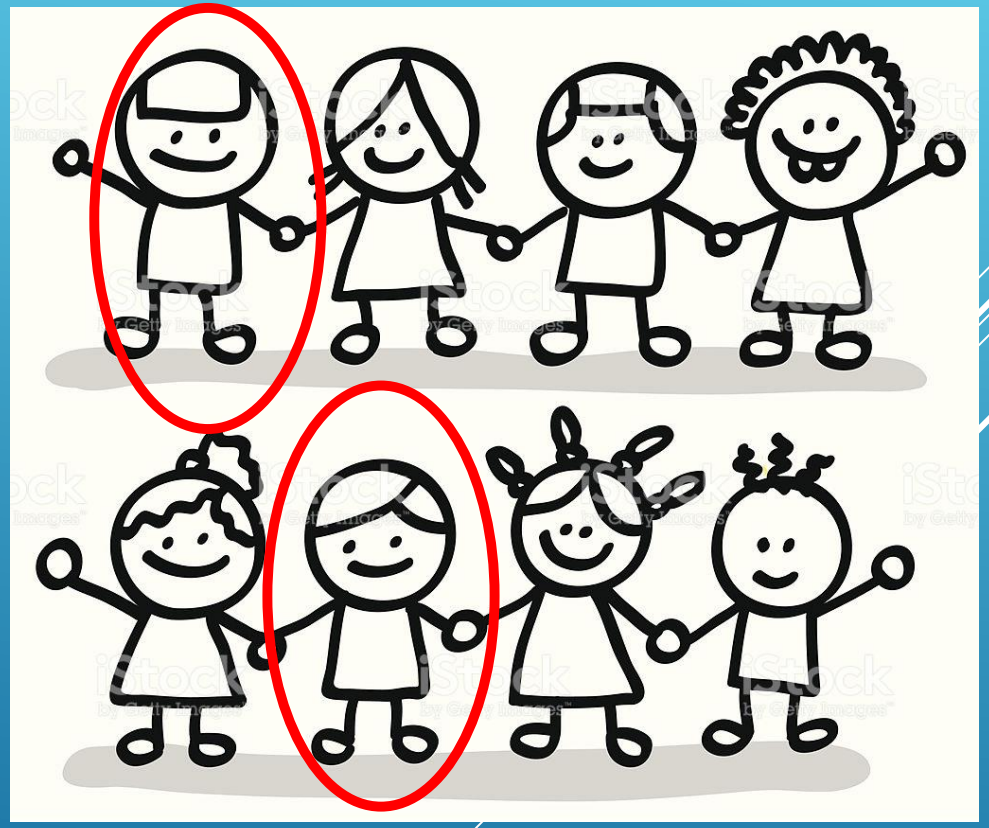
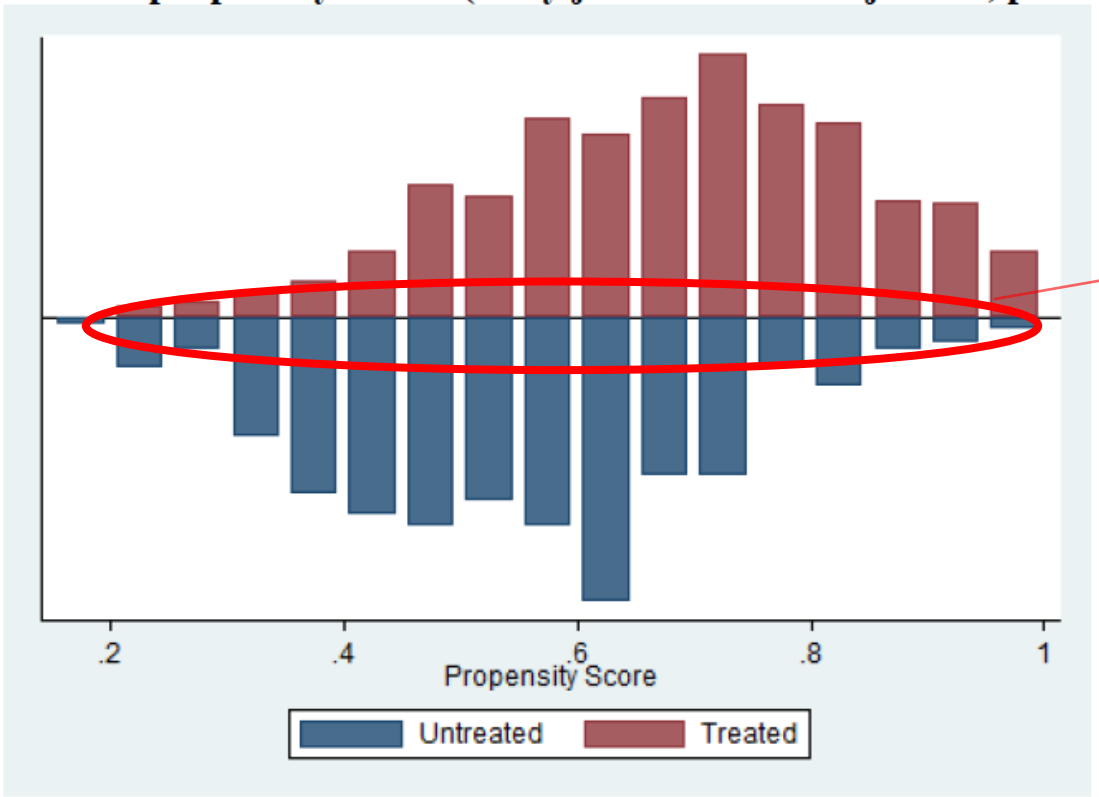


Figure 2: Estimated propensity scores (early joiners vs. never joiners, poor households)



Common support

Table 7: Means of covariates in the unmatched and matched sample (early vs never joiners, poor)

Variable		Mean		%bias	%reduct |bias|
	Unmatched/Matched	Treated	Control		
Household size (*zhhsize*)	Unmatched	4.3395	3.8141	32.3	
	Matched	4.3395	4.4917	-9.4	71
Female head (*zfemalehead*)	Unmatched	.07792	.09295	-5.4	
	Matched	.07792	.11503	-13.3	-147
Highest year of schooling in the family (*zeducyears*)	Unmatched	6.7328	4.9071	43.7	
	Matched	6.7328	6.5974	3.2	92.6
No of members that can write in the household (*znowriters*)	Unmatched	1.833	1.1923	47.5	
	Matched	1.833	1.9221	-6.6	86
Total expenditure 2004 Rs. (*ztotexpb*)	Unmatched	2094.3	1744.7	36.9	
	Matched	2094.3	2168.7	-7.8	79
Household owned any land in 2004? (*zanylandowned*)	Unmatched	.50649	.48718	3.9	
	Matched	.50649	.48980	-1.3	91
Household owned any livestock assets in 2004? (*zanylivestock*)	Unmatched	.30241	.23718	14.7	
	Matched	.30241	.30798	-4.2	71.6
Household owned any farm assets in 2004? (zanyfarmassets)	Unmatched	.83117	.80769	6.1	
	Matched	.83117	.82931	0.5	92.1
SC/ST (*zcaste2_1*)	Unmatched	.45455	.30769	30.6	
	Matched	.45455	.43785	3.5	89
Other Castes (*zcaste2_3*)	Unmatched	.11317	.21795	-28.4	
	Matched	.11317	.12987	-4.5	84

DIFFERENCE IN DIFFERENCES

- Afridi, Barooah and Somanathan (in progress)
- School meals were started in urban public schools of Delhi in 2003
- Phased implementation with 410 in first phase (2003) and the rest in phase 2 (2004)

TABLE 2: AVERAGE ATTENDANCE LEVELS AND CHANGES.

	Control	Treatment	Difference
	(1)	(2)	(3)=(2)-(1)
(A) Δ 2002	0.06	0.06	0
	(0.008)	(0.010)	(0.013)
Mean attendance in April 2002	0.81	0.79	
	(0.073)	(0.087)	
(B) Δ 2003	0.07	0.11	0.04***
	(0.008)	(0.009)	(0.012)
Mean attendance in April 2003	0.80	0.78	
	(0.086)	(0.063)	
Difference (B)-(A)	0.01	0.05***	0.04***
	(0.011)	(0.014)	(0.018)

Control experiment

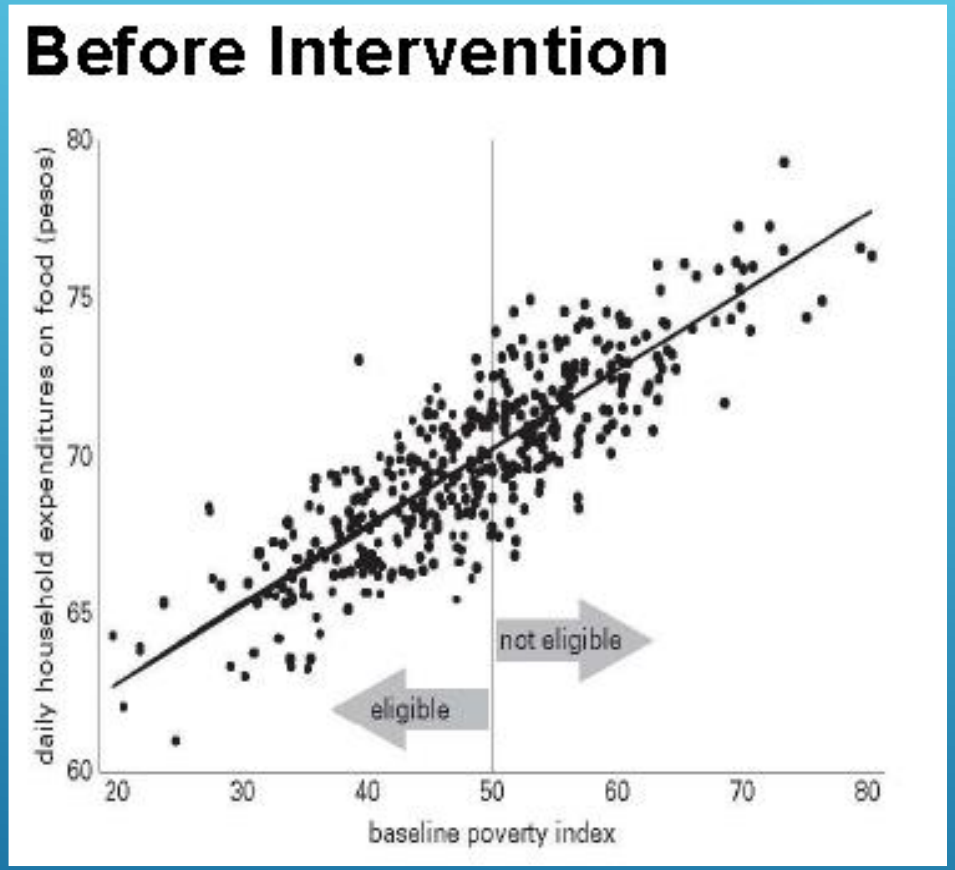
Treatment experiment

REGRESSION DISCONTINUITY DESIGN

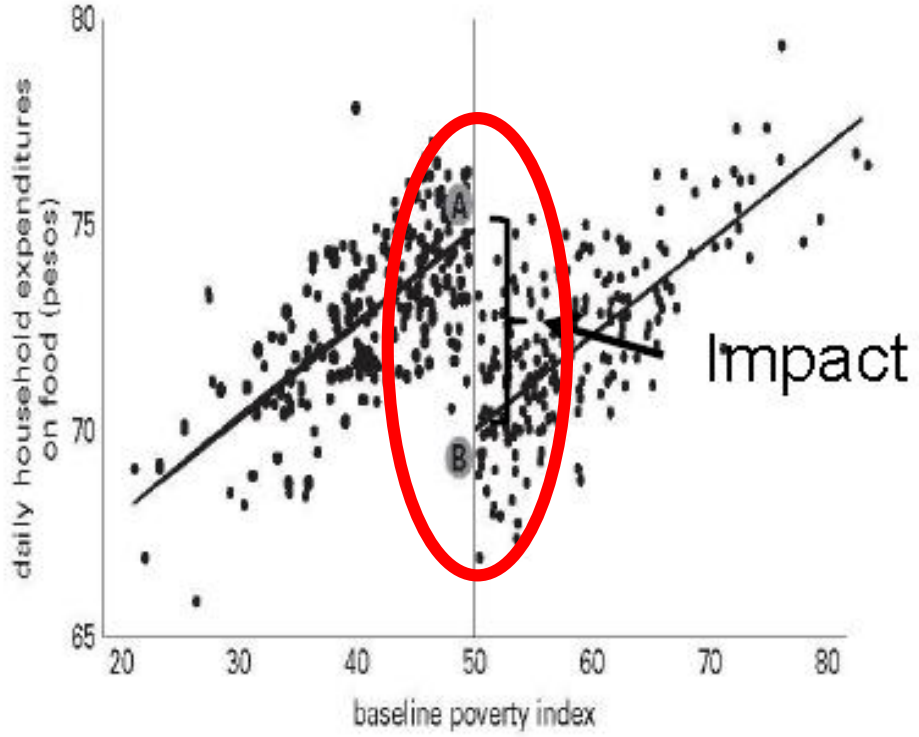
There is a programme allocation 'threshold rule' dividing participants and non-participants

Variable	Threshold rule
Poverty index	Impact of development projects to households below a poverty incidence threshold (eg BPL cards)
Age	Impacts on subsidies for senior citizens (above 60 y.o.)
Date	Impact of introduction of a reform after a certain time

Before Intervention



After Intervention

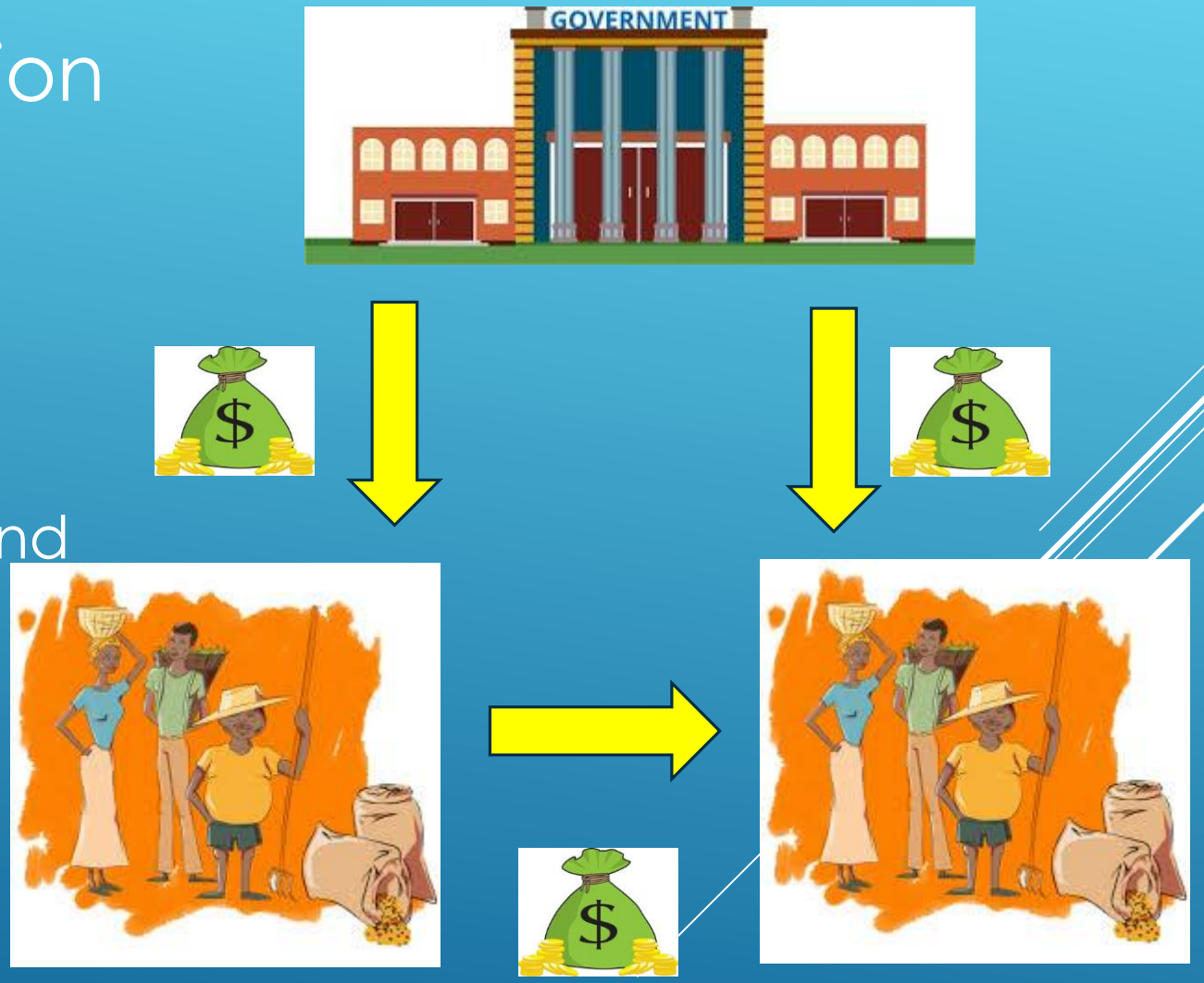


SESSION 2:

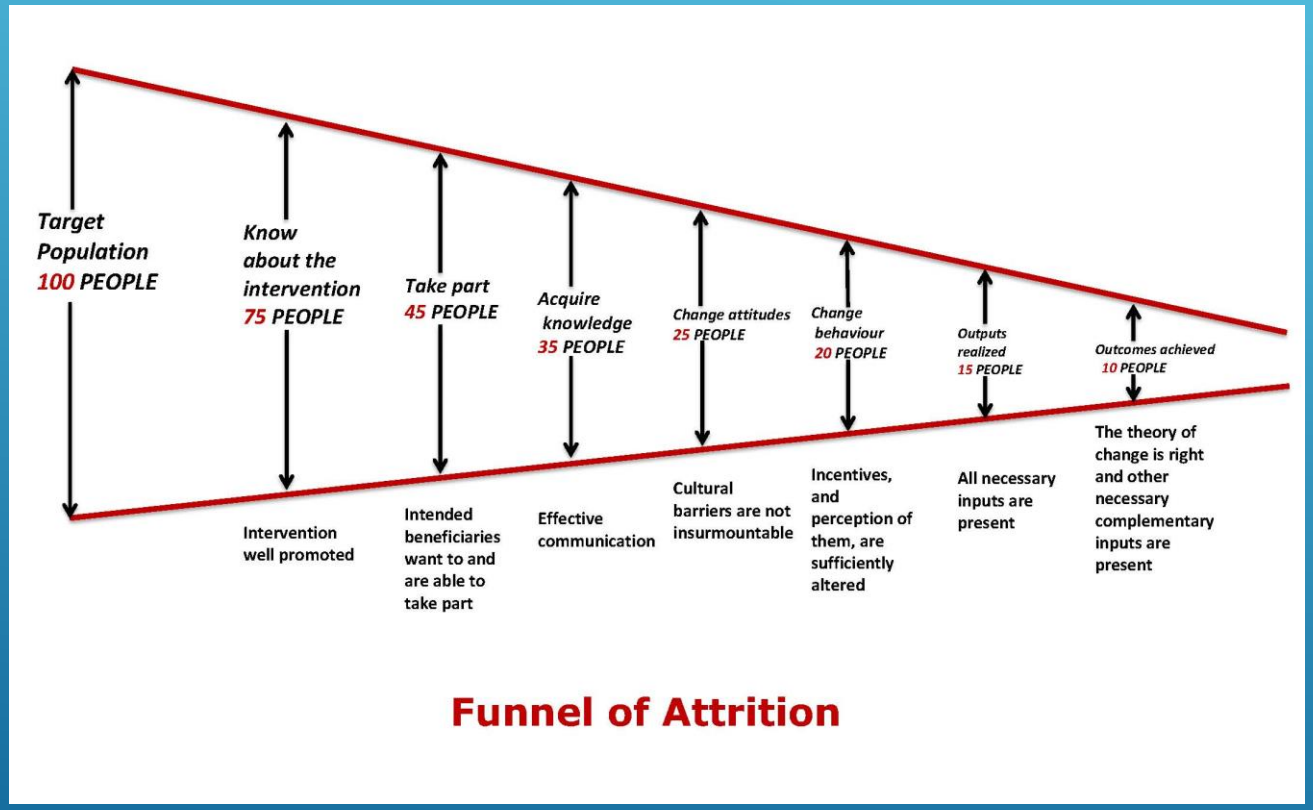
HOW TO BE CREATIVE WITH IMPACT EVALUATIONS

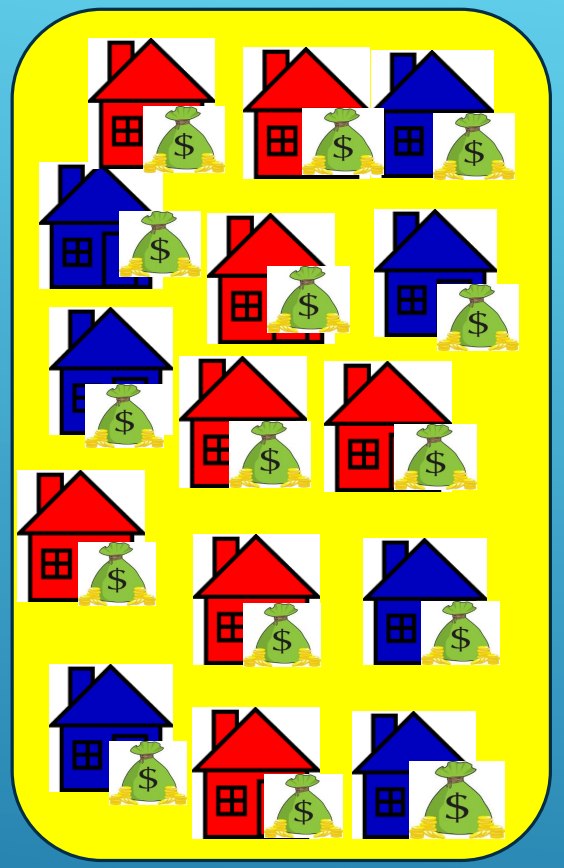
BIASES IN IMPACT EVALUATIONS

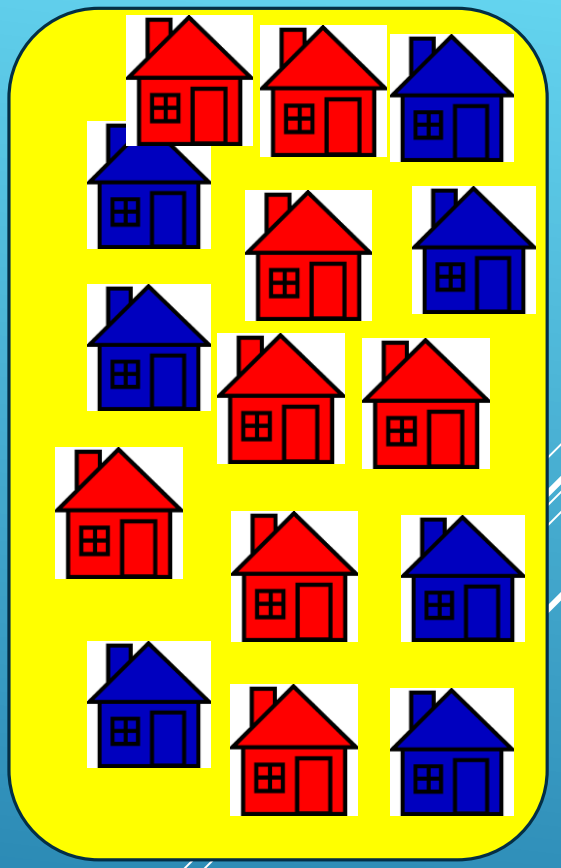
- Spillover and Contamination
  - Threatens the validity of IEs
- What to do? Examples
  - Design
    - Unit of treatment of a village and not some villagers
  - Intervention
    - Non-transferrable vouchers
  - Monitoring



ATTRITION

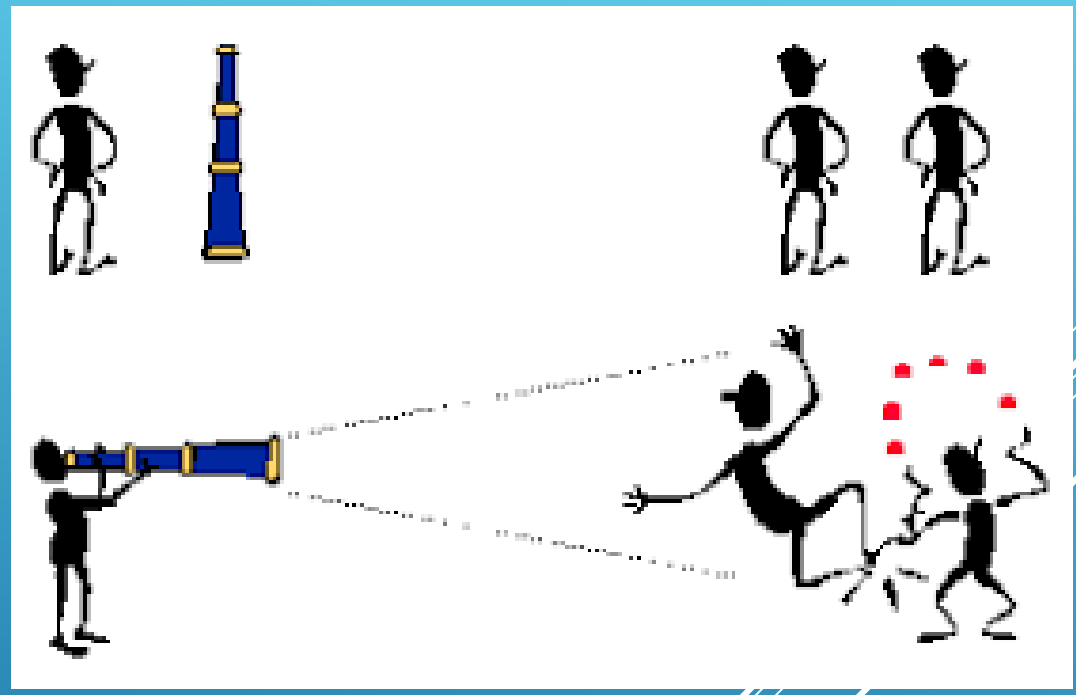

Target Population 100 PEOPLE
Know about the intervention 75 PEOPLE
Take part 45 PEOPLE
Acquire knowledge 35 PEOPLE
Change attitudes 25 PEOPLE
Change behaviour 20 PEOPLE
Outputs realized 15 PEOPLE
Outcomes achieved 10 PEOPLE
Intervention well promoted
Intended beneficiaries want to and are able to take part
Effective communication
Cultural barriers are not insurmountable
Incentives, and perception of them, are sufficiently altered
All necessary inputs are present
The theory of change is right and other necessary complementary inputs are present
Funnel of Attrition





OTHER BIASES

- Hawthorne Effect
  - Treatment group modifies behavior not because of the treatment but being observed
- John Henry effect
  - Control groups change behavior
- What to do? Examples
- ‘ Sensitive survey and monitoring systems



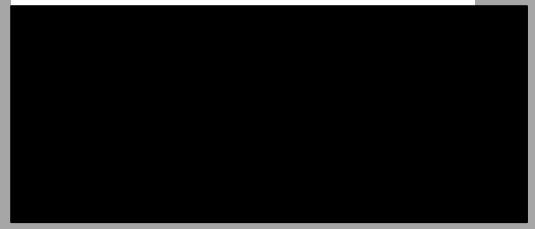
Assessing the robustness of our methods

SAMPLE SIZE AND POWER

Sample Size Calculations

Larger sample → more likely that treatment and control are comparable

	Years of education	
	Treatment	Control
n=2	12.0	9.0
n=20	6.4	5.8
n=50	5.8	5.3



LESSONS LEARNT: GENERAL RULES

1. You need a minimum sample size to make good estimations.
2. You need a sample that is diverse enough to represent the population studied.
3. The larger the sample the better and more accurate are your estimations.
4. If you increase sample size, you are likely to increase power.

You need a large sample size!

SOME SAMPLING BASICS

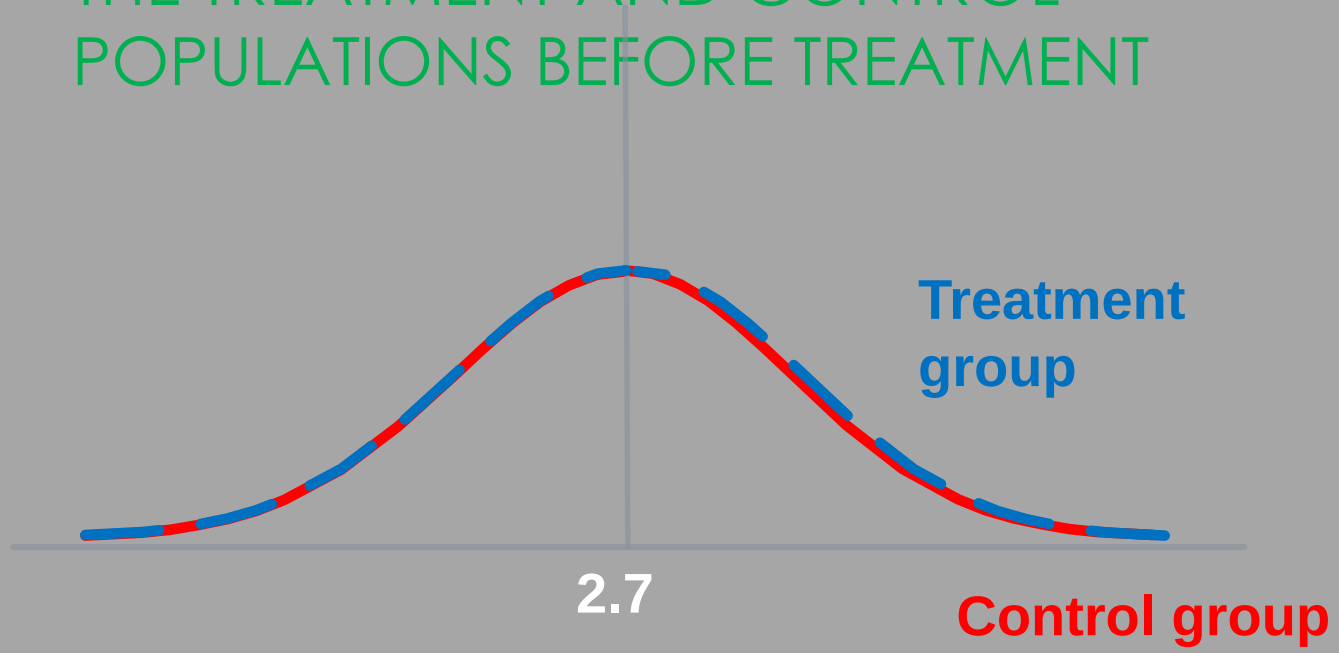
Population mean: the true value of a parameter, i.e. the average weight for age of all children aged under in the region of interest.

Sample mean: the average weight for age in a sample drawn from the population.

The larger the sample the more likely it is that the sample mean is close to the population mean (provided our sample is a random sample)

What do impact evaluators do?

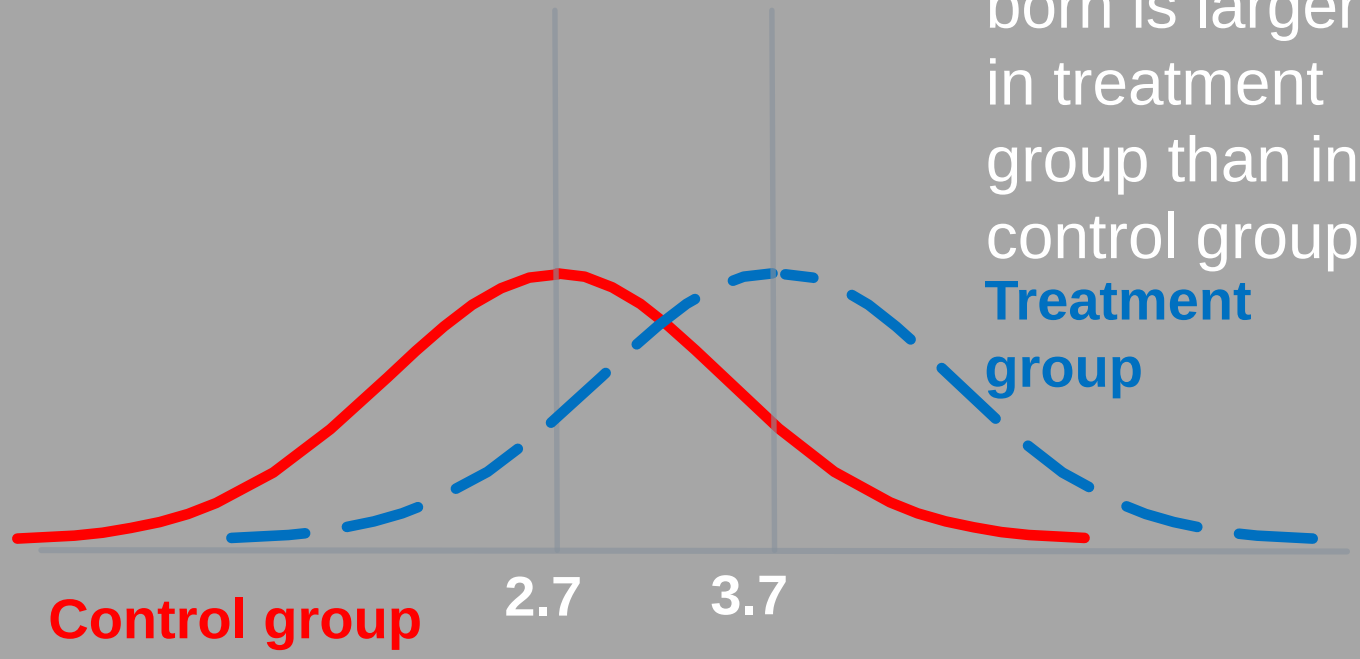
DISTRIBUTION OF NEWBORN WEIGHT IN THE TREATMENT AND CONTROL POPULATIONS BEFORE TREATMENT



What do impact evaluators do?

AND AFTER TREATMENT

On average weight of new born is larger in treatment group than in control group.



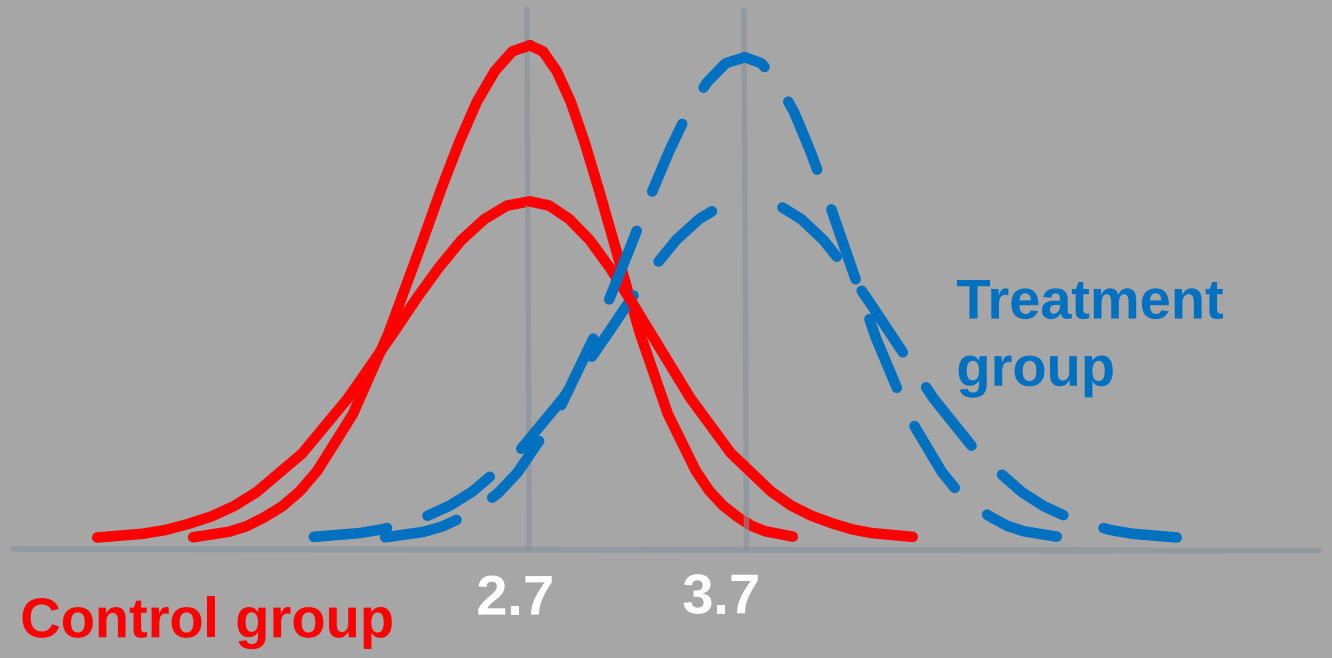
Power Calculation and sample size

- *Power (or statistical power)* of an impact evaluation is the likelihood that it will detect a difference between the treatment and comparison groups, when in fact one exists.

  - Power calculation indicate the smallest sample size required for an evaluation to detect a meaningful difference in outcomes between the treatment and comparison groups.

SAMPLE SIZE AND STANDARD ERROR

A larger sample



WHY POWER CALCULATION?

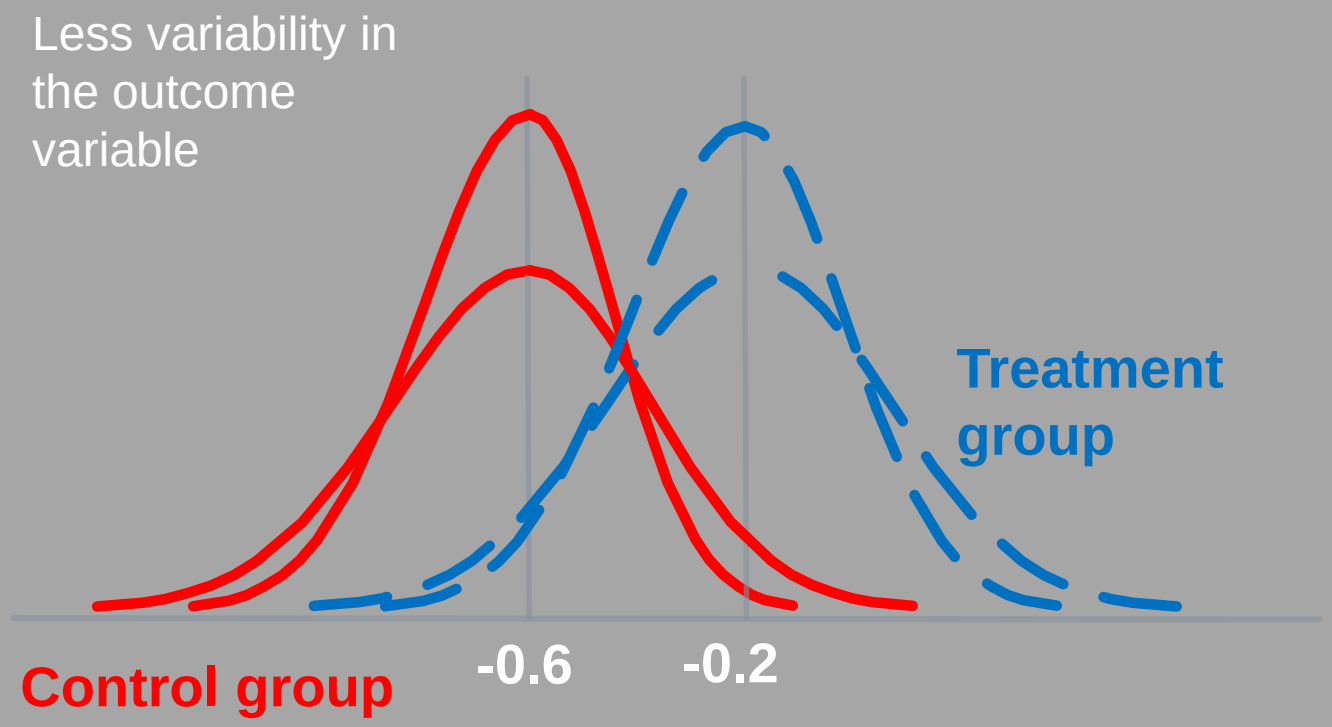
1. Not acceptable to conduct a study that would not be stringent enough to detect a real effect due to a lack of statistical power.

2. Not acceptable to conduct a study by recruiting 1000s of participants when sufficient data could be obtained with 100s of participants instead.

3. Avoid misleading policy recommendations

So how large a sample do we need?

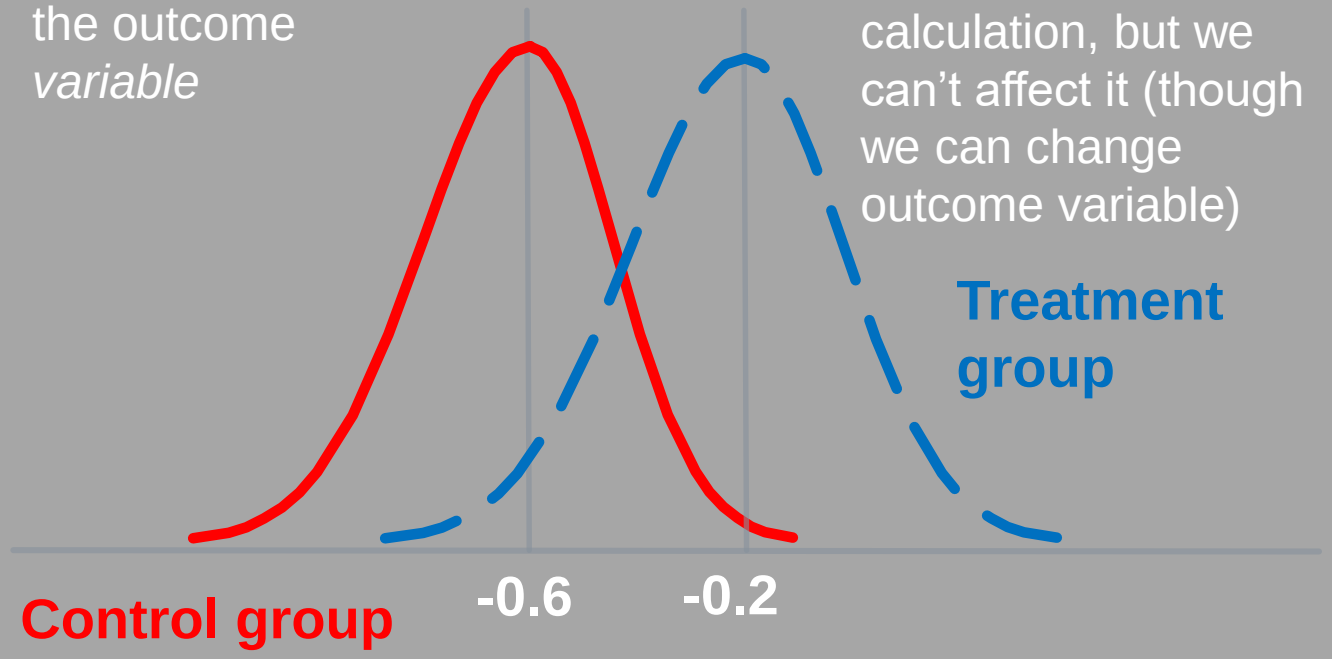
WHAT MAKES IT EASIER TO DETECT PROGRAMME IMPACT?



WHAT MAKES IT EASIER TO DETECT PROGRAMME IMPACT?

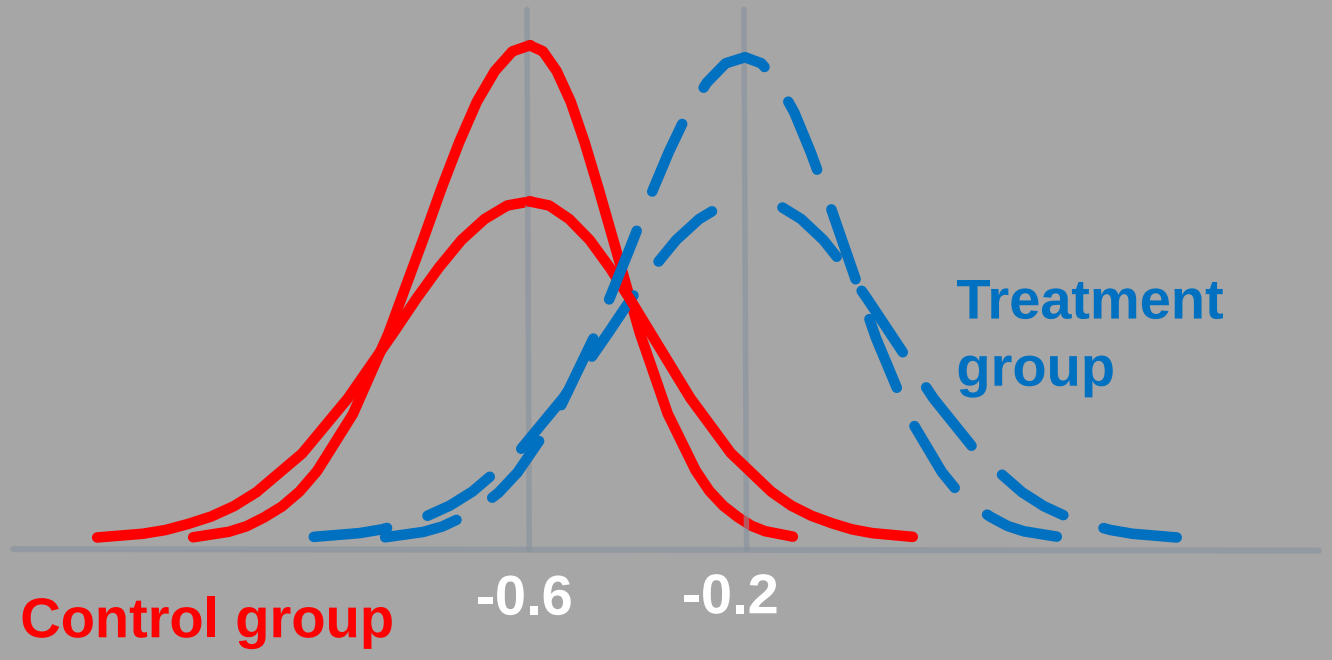
Less variability in the outcome *variable*

So we need to know that for our power calculation, but we can't affect it (though we can change outcome variable)



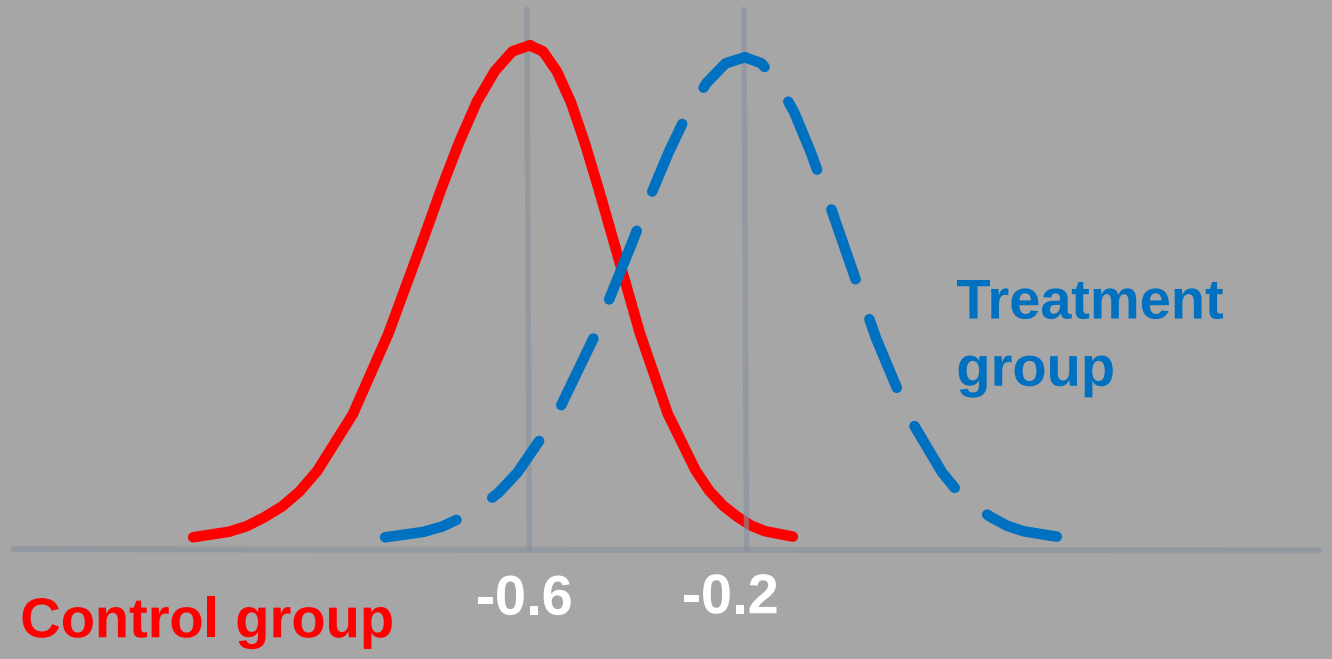
WHAT MAKES IT EASIER TO DETECT PROGRAMME IMPACT?

A larger sample



WHAT MAKES IT EASIER TO DETECT PROGRAMME IMPACT?

A larger sample



More formally

EQUAL TREATMENT AND CONTROL SAMPLES

$$MDE = (t_{\alpha} + t_{1-\beta})\,\sigma_y \sqrt{\frac{1}{P(1-P)n}}$$

MDE = f[1/P(1-P)]

And obviously increasing n helps

$\delta(MDE)/\delta P = (1-P) - P = 1 - 2P = 0 \Rightarrow P = ½$

$\delta^2(MDE)/\delta P^2 = -2$ so maximize MDE

Formative/Process Evaluation

Understand the context and the program, ground realities

Develop the program theory of change

Set out research questions- what can the IE address and not?

Design the impact evaluation

1. Sample size
2. Data requirements
3. Costs
4. Methodology
5. Biases
6. Monitoring of implementation
7. Plan data collection

EVALUATION OF THE NATIONAL RURAL LIVELIHOODS MISSION IN INDIA

- Large scale program on group-based livelihoods support
- The government had conducted baseline surveys in 13 states of India before the roll-out of the program
- There were matched treatment and control areas

- **Our initial scope of work**

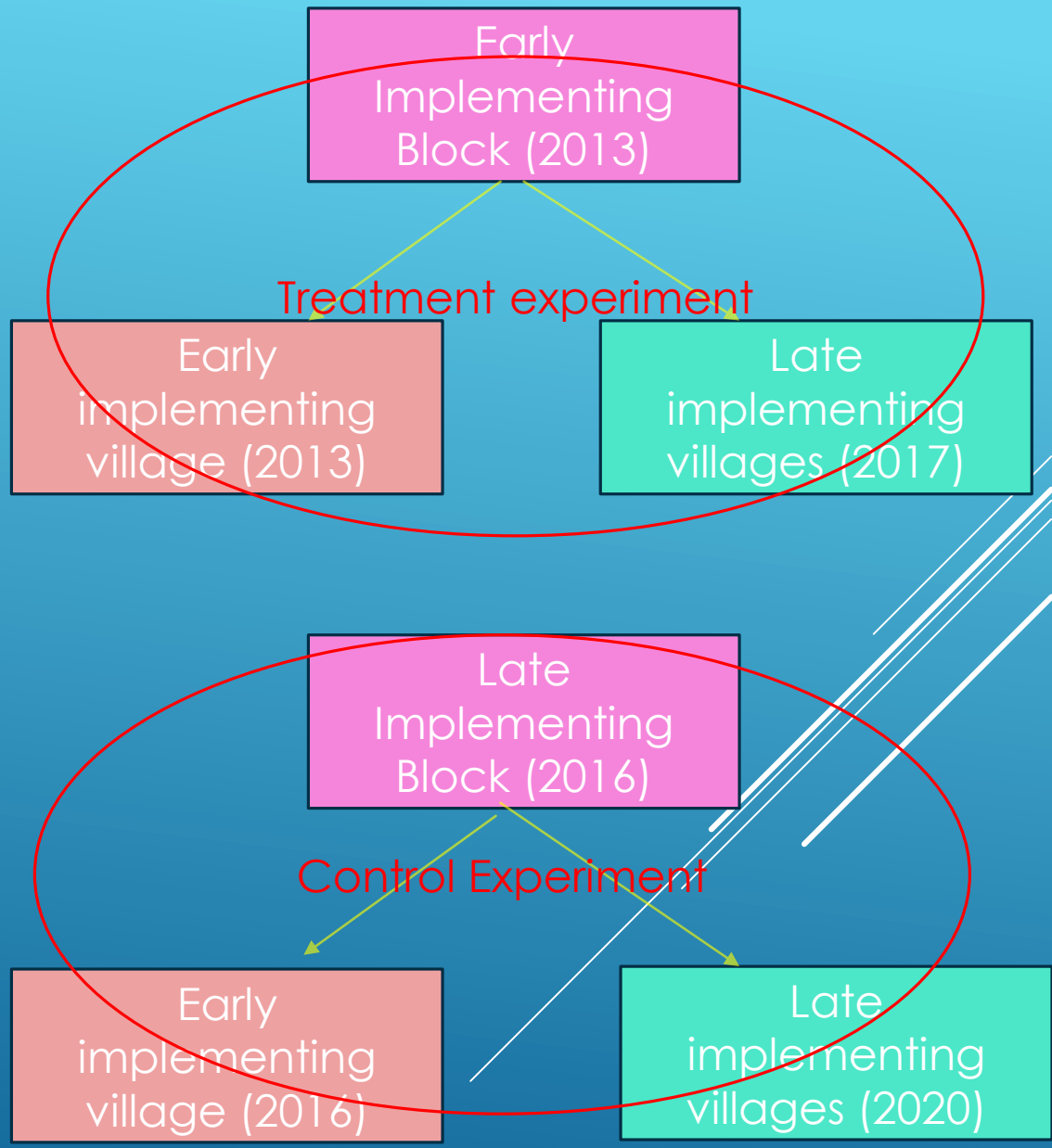
- Design an endline survey and report on findings

BUT

- The baseline data was not usable
- The program was rolled out in control areas

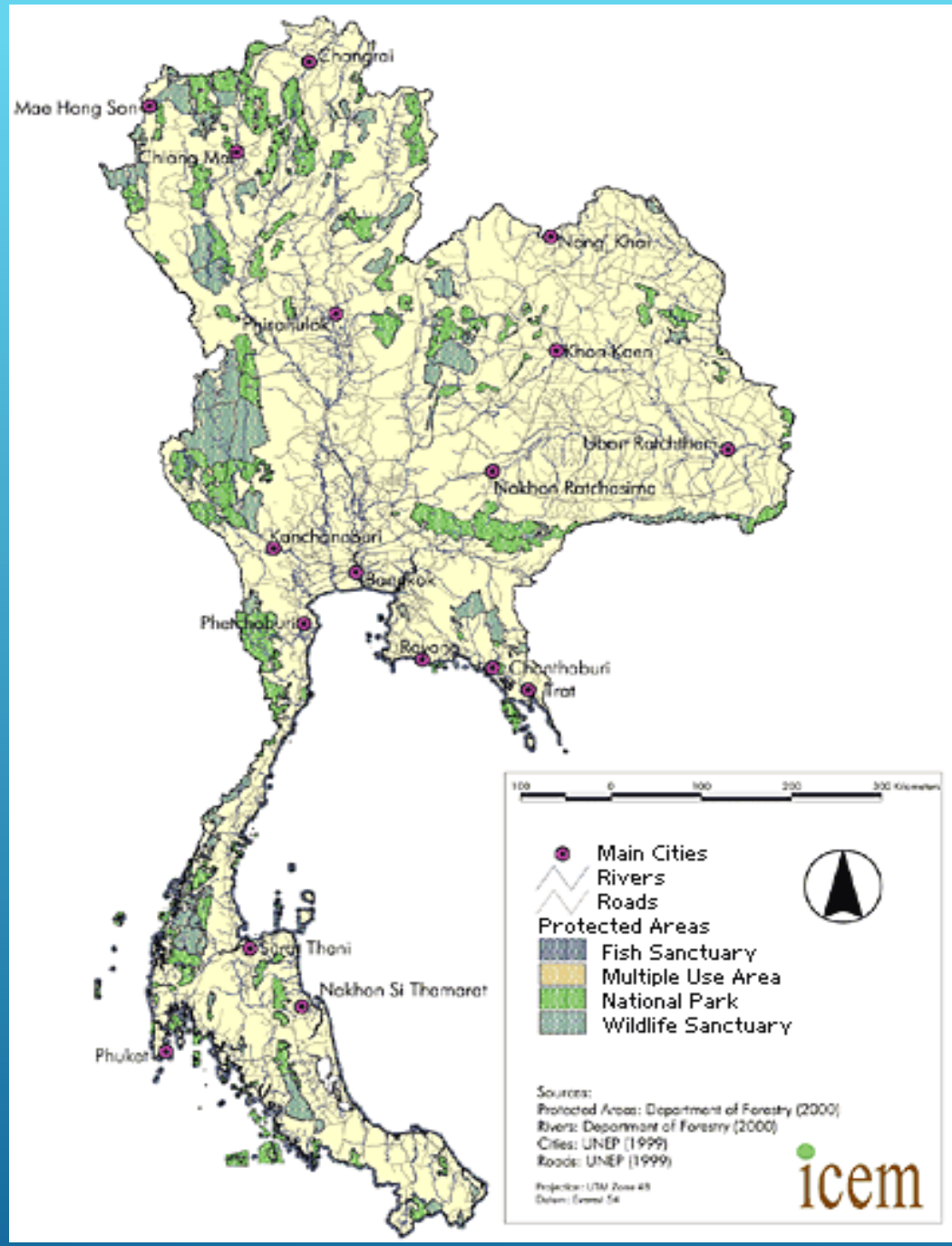
What we did

- Examined program records and MIS
- Intensive ground work
- Conversations with field teams
- Proposed a Difference-in-Difference strategy



USING IMPACT EVALUATION: TO ESTIMATE THE IMPACT OF PROTECTED AREAS AND ROADS

Most protected areas and forest reserves in Thailand are in the north.



SELECTION BIAS

Areas with high elevation and slopes and bad soils are where protected aeas are located.

So do protected areas and forestry programmes really help?

ed are those

gricultural

ally bring

vity.

THE ECONOMETRIC MODEL

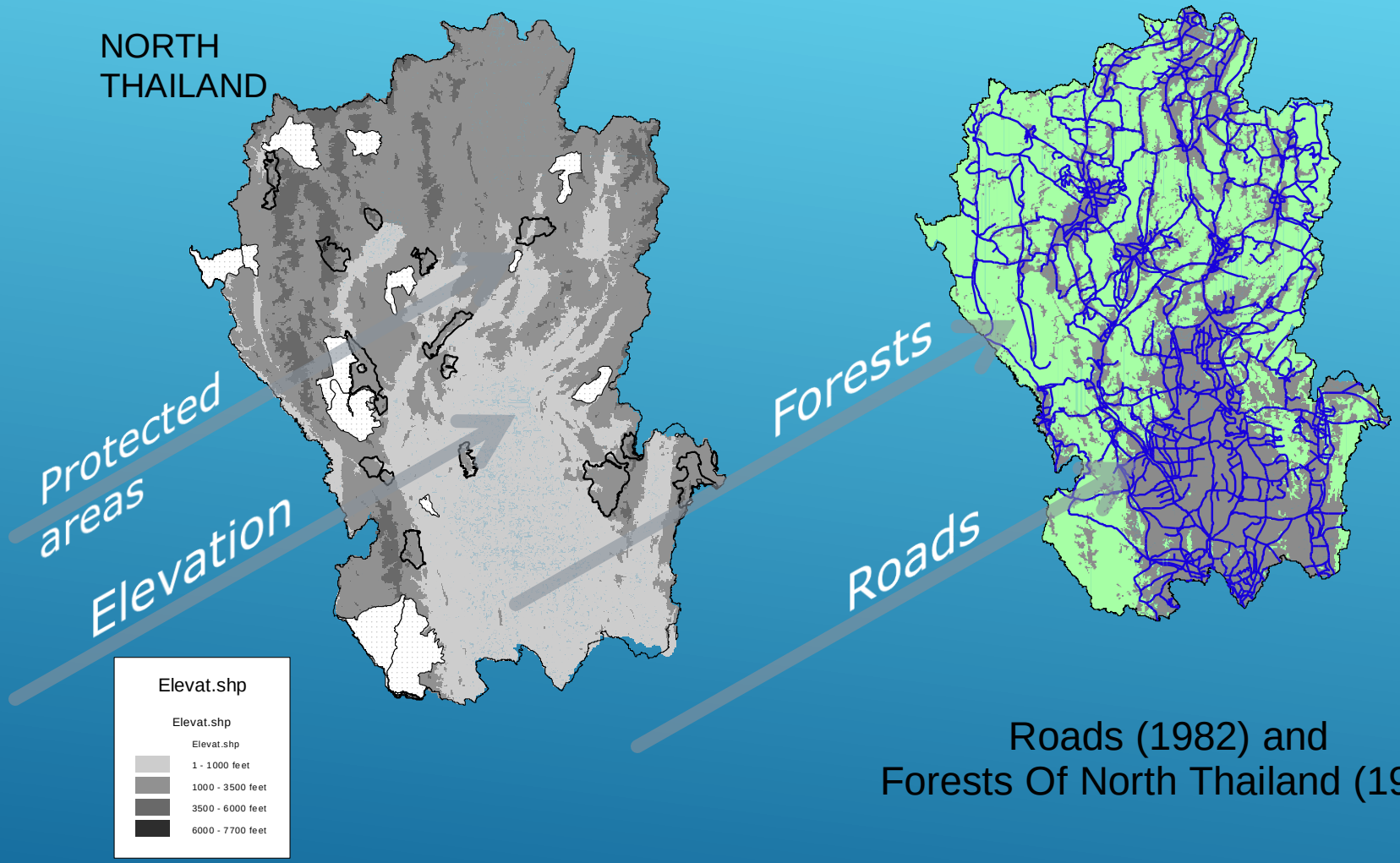
- The econometric model that we estimate is thus given by

- **Zi : Plot attributes (Slope, Elevation, Impedance weighted travel time, Soil Dummy, Population density)**
- **Y1i*: Net profit from clearing**
- **Y2i*: Net utility from protecting a plot**

$$Y_{1i}^{*} = Z_i B_1 + \gamma Y_{2i} + e_{1i} \qquad Y_{1i} = 1 \text{ if } Y_{1i}^{*} > 0; = 0 \text{ otherwise}$$

$$Y_{2i}^{*} = Z_i B_2 + \alpha W_i + e_{2i} \qquad Y_{2i} = 1 \text{ if } Y_{2i}^{*} > 0; = 0 \text{ otherwise}$$

THAILAND PROTECTED AREAS

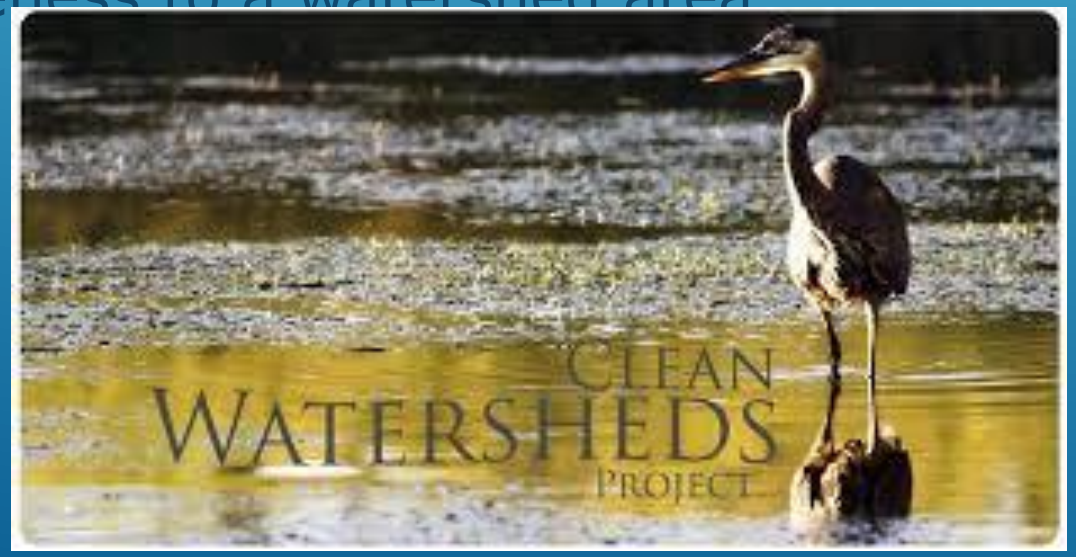


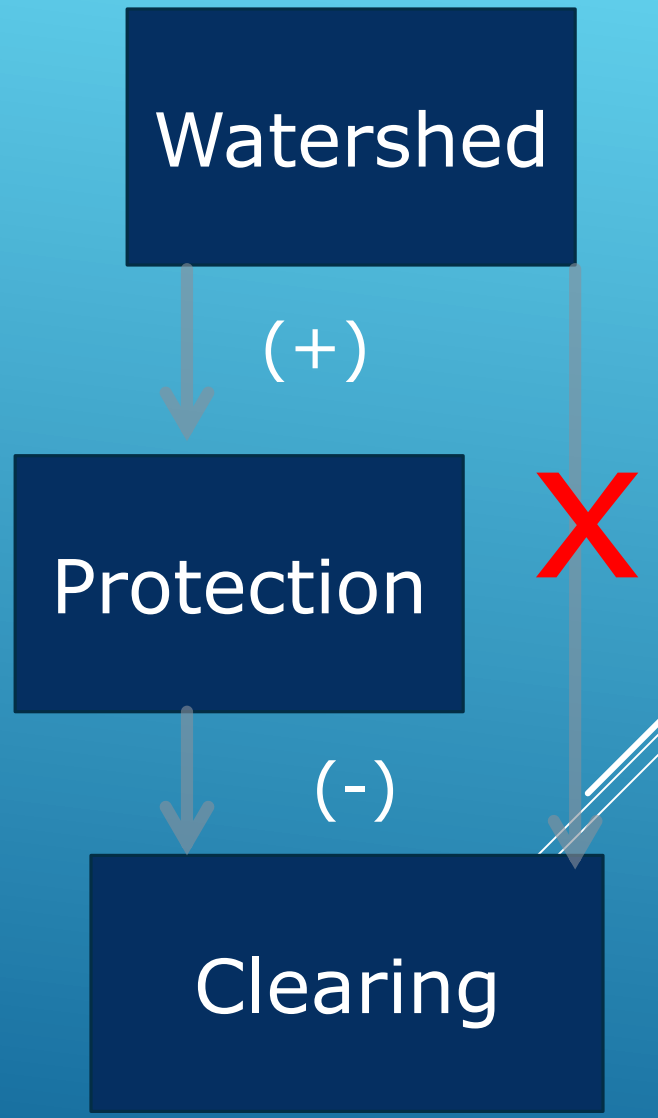
Roads (1982) and
Forests Of North Thailand (1986)

METHOD: INSTRUMENTAL VARIABLES

Probability of land getting cleared = determined by soil fertility, slope, elevation, distance to the market, administrative factors, population pressure etc.

Probability of land being protected = determined by some of the same factors AND closeness to a watershed area





Cleared Land (Y1 = 1)		T- Stats
Slope (degrees)	-0.088	-10.652
Elevation (ms.)	-0.001	-8.095
Population density1990 (people/km^2)	0.003	4.532
Log(cost) (1982)**	-0.191	-9.729
Soil and Province Dummies	Not	Shown
Protected Area dummy (1986)	-6.28	-10.332
Constant	1.295	8.870
No. of observations	4946	

Cleared Land (Y1 = 1)		T- Stats
Slope (degrees)	-0.088	-10.652
Elevation (ms.)	-0.001	-8.095
Population density1990 (people/km^2)	0.003	4.532
Log(cost) (1982)**	-0.191	-9.729
Soil and Province Dummies	Not	Shown
Protected Area dummy (1986)	-0.077	-0.332
Constant	1.295	8.870
Protected Area (Y2 = 1)		**Equation**
Slope (Degrees)	0.034	5.297
Elevation (ms.)	0.001	9.058
Population density1990 (people/km^2)	0.001	2.297
Log(cost) (1982)	0.192	7.477
Soil and Province Dummies	Not	Shown
Watershed dummy	0.188	3.543
Constant	-4.098	-14.010
Log Likelihood	-3714.7	
No. of observations	4946	

THAILAND PROTECTED AREAS: RESULTS

- Naïve model: Protection has a large effect on preventing deforestation.
- After you account for selection bias, in the static model, there is no effect. Protected lands would not have been cleared even if they had not been protected.

LUNCH!

The WFP Moderate Acute Malnutrition Impact Evaluation Series – 4 Impact Evaluations + Synthesis

- All examine aspects of WFP's food security and moderate acute malnutrition (MAM) prevention and treatment programmes, and their **impact on nutrition and food security outcomes**.
- Commissioned by WFP's OEV and managed by the International Initiative for Impact Evaluation's (3ie).
- All 4 Impact Evaluations implemented by different teams

Background – Where?

Chad, Mali, Niger and Sudan

- Selection criteria : number of beneficiaries, countries with both prevention and treatment interventions, malnutrition figures, mix of programme categories and geographic representation.

- Short-list refined by feasibility for Country Office engagement and timeliness.

Overview - What is given to whom and why?

Blanket Supplementary Feeding

- Nutritional supplements and transferring children to treatment

Targeted Supplementary Feeding

- Treatment programme

+

In-kind and cash transfers in all

School feeding in **Mali**

Behaviour Change Communications in **Sudan**

Food for Assets in **Niger**

Things to note: Questions

- IEs appear to ask similar questions on similar outcomes. The detail underneath is wildly different – different things measured.
- Questions were tailored, due to local contextual and data quality issues.

Country	Primary Questions
Niger	What is the impact of receiving different combinations of the components within WFP's Protracted Relief and Recovery Operation?
Sudan	What is the impact of different MAM treatment and prevention interventions on the incidence and prevalence of MAM and SAM?
Chad	What is the difference in impact of MAM prevention on the incidence and prevalence of MAM, when access to MAM treatment is good or poor?
Mali	What is the impact of conflict and food assistance on child malnutrition and other developmental outcomes?

Challenges? Of course not – it was easy!

Niger
- Baseline not designed for follow up + security risks. Result - high attrition (75%).
- Everybody in baseline received something – no control.
- Too small a sample to answer the initial study questions.

Sudan
- No baseline
- One intervention did not reach the beneficiaries

Chad
- No maps – identification plan had to change
- Targeting agreed so comparison groups from different areas
- If a malnourished child was identified – she/he should be referred

Mali

Things to note: Creative Methodology

- All use **different** methods and **mixed** methods

- None are 'conventional' RCTs

  - but all have a way to identify the impact

  - all methods are 'complicated'

Country	Methodology
Niger	• Difference-in-differences • Instrumental Variables • Qualitative analysis • Selection correction models
Sudan	• Stepped wedge cluster controlled trial design • Qualitative analysis
Chad	• Analysis of covariates and propensity score matching • Use of qualitative data to inform and interpret results
Mali	• Qual. and Quant. data used to characterise exposure to conflict and humanitarian aid. • Natural experiment, Difference-in-differences and propensity score matching

Findings?

Niger

- Food for Assets with Prevention or Treatment has an impact on child nutrition and Food for Assets programme is well targeted
- Prevention and treatment programmes less well targeted

Sudan

- No impact on the prevalence, but impact on children-at-risk.
- No change in feeding behaviours and practices as a result of the behavioural intervention.

Chad

- Prevention programme lowers incidence in under-2s.
- Prevention is more effective in reducing malnutrition for those with poor access to Treatment.

Mali

- Impact on caloric intake and zinc consumption, and increase in vitamin A availability
- Households receiving two forms of assistance had improved nutrition outcomes.

Technical difficulties that were resolved creatively

- Difficult to identify a counterfactual? Can often be done creatively.
- High level of attrition – complicates things but can be corrected for.
- Low sample sizes – change in design?
- No baselines – several techniques exist to constructing it either ex-post or artificially.

Lessons from creative IEs: 1. Evaluation Management

Robust management always important but with complex methods even more so:

- Regular comms between evaluation team
  - Changes in evaluation questions
  - Changes in evaluability
  - Unforeseen challenges

Lessons from Creative IEs: 2. Balance of skills

Need a range of skills:

- Rigorous impact evaluation
- Understanding of context and programmes
- Presenting the results and communicating
- Working to timelines

Lessons from creative IEs: 3. Define quality carefully

Agree a common understanding and expectation of "quality"

- High quality methodology
- Integration of gender dimension
- Ethical approvals and management of ethics
- High quality report drafting
- Bespoke communication products

Key takeaways

- Creativity is a must!
- IEs work in ‘real-life’ and complex settings
- Quasi-experiments are a friend of complexity
- Ethics is important but not an obstacle
- Planning with programme/implementers is crucial
- Extra focus on comms is key

AFTERNOON SESSION

DESIGN YOUR OWN IMPACT EVALUATION!

TWO TASKS:

1.) WHAT ARE YOUR EVALUATION QUESTIONS?

2.) WHAT IS YOUR IMPACT EVALUATION DESIGN?

- Jyotsna Puri: jpuri@gcfund.org
- Anna Hentinnen: anna.hentinnen@wfp.org
- Bidisha Barooah: bbarooah@3ieimpact.org