Impact Evaluation

Session 2, June 22nd, 2010

Microsoft Research India Summer School on Computing for Socioeconomic Development

Discussion led by Aishwarya Ratan

Types of evaluation

- Usability evaluation
- Intervention impact evaluation
  - Non-experimental methods
  - Randomised Controlled Trials (RCTs)

Impact evaluation

- Our goal is not just to design good technology
- The goal is to design technology that enables socio-economic development
- So need to know if our well-designed technology intervention impacts our target development outcome

Requirements for a valid Impact Evaluation

- Need to establish that
  - The outcome changed after the programme
  - The change occurred among those who received the programme and not among those who did not receive the programme
  - There was nothing else that caused this differential change

Need to establish beyond reasonable doubt that the intervention caused the change in outcome!

So, need a counterfactual

- If the program had not been implemented, we would not have seen the observed change in outcomes.
- 'The "counterfactual" is the imaginary state of the world that program participants would have experienced if they had not participated in the program.'
- Can never implement and not implement the program for the same individuals.
- So need to mimick the counterfactual.

THIS IS TOUGH TO DO!

‘Get out the Vote’ Case - Discussion

- What is the problem?
- What is the suggested intervention/programme?
- How is the intervention/programme operationalised?
- How is the programme implemented?
- How many people were targeted?
- How many ended up receiving the programme?
- What impact did they expect? What outcome were they measuring?

Simple difference

	Voter turnout by group		Impact Estimate
	Reached	Not reached	
Method1: Simple difference	64.5%	53.6%	10.8 pp*

Controlling for other characteristics

Characteristics of Reached and Not-Reached Groups			
	Reached	***Not Reached***	***Difference***
Household Size	1.56	1.50	0.06
Average age	55.8	51.0	4.8
Percent female	56.2%	53.8%	2.4 pp*
Percent newly registered	7.3%	9.6%	-2.3 pp*
Percent from a competitive district	50.3%	49.8%	0.5 pp
Percent from Iowa	54.7%	46.7%	8.0 pp*
Sample Size	25,043	34,929	

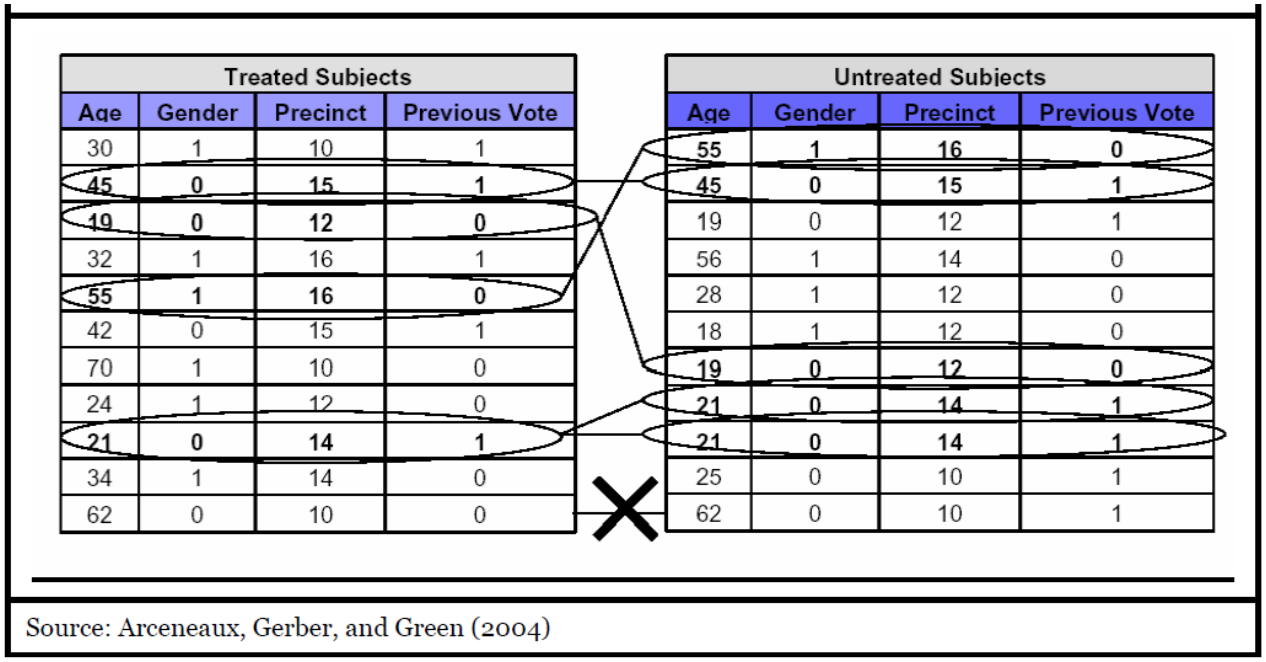
Impact measure of 6.1 pp*

Use panel data

Voter turnout in 1998 and 2000 elections between the reached and not-reached			
	2002 Reached	***2002 Not Reached***	***Difference***
Voted in 2000	71.7%	63.3%	8.3 pp*
Voted in 1998	46.6%	37.6%	9.0 pp*

Impact measure of 4.5 pp*

Matching



Treated Subjects			
Age	Gender	Precinct	Previous Vote
30	1	10	1
45	0	15	1
19	0	12	0
32	1	16	1
55	1	16	0
42	0	15	1
70	1	10	0
24	1	12	0
21	0	14	1
34	1	14	0
62	0	10	0

Untreated Subjects			
Age	Gender	Precinct	Previous Vote
55	1	16	0
45	0	15	1
19	0	12	1
56	1	14	0
28	1	12	0
18	1	12	0
19	0	12	0
21	0	14	1
21	0	14	1
25	0	10	1
62	0	10	1

Source: Arceneaux, Gerber, and Green (2004)

Matching Analysis			
Number of Covariates matched on:	**Subset of Matched Reached**	**Subset of Matched Not-Reached Individuals**	***Impact***
4 (HH size, age, newly registered, state)	64.5%	60.8%	3.7 pp*
6 (HH size, age, newly registered, state in a competitive district, voted in 2000)	64.5%	61.5%	3.0 pp*
All	65.9%	63.2%	2.8 pp*

Randomised experiment

Compares the treatment and control groups on observable characteristics			
	Treatment	**Control**	***Difference***
Voted in 2000	56.7%	56.4%	0.4 pp
Voted in 1998	22.7%	23.1%	-0.5 pp
Household Size	1.50	1.50	0.0
Average age	52.0	52.2	-0.2
% Female	54.6%	55.2%	-0.6 pp
% Newly registered	11.6%	11.7%	0.0 pp
Total people in group	14,972	1,153,072	

	Treatment (60,000 called)	**Control (2M not called)**	***Impact***
Simple Difference	58.2%	58.0%	0.2 pp
Difference after controlling for observable characteristics (multivariate regression)			0.2 pp

Summary

Method	*Estimated impact*
Simple Difference	10.8 pp*
Multivariate Regression	6.1 pp*
Multivariate Regression with Panel Data	4.5 pp*
Matching (All Covariates)	2.8 pp*
Randomized experiment with adjustment to reflect that only 25,000 of 60,000 in the treatment were treated	0.4 pp

NOTES: pp means "percentage points" and * indicates statistically significant at the 5% level

Other scenarios… ?

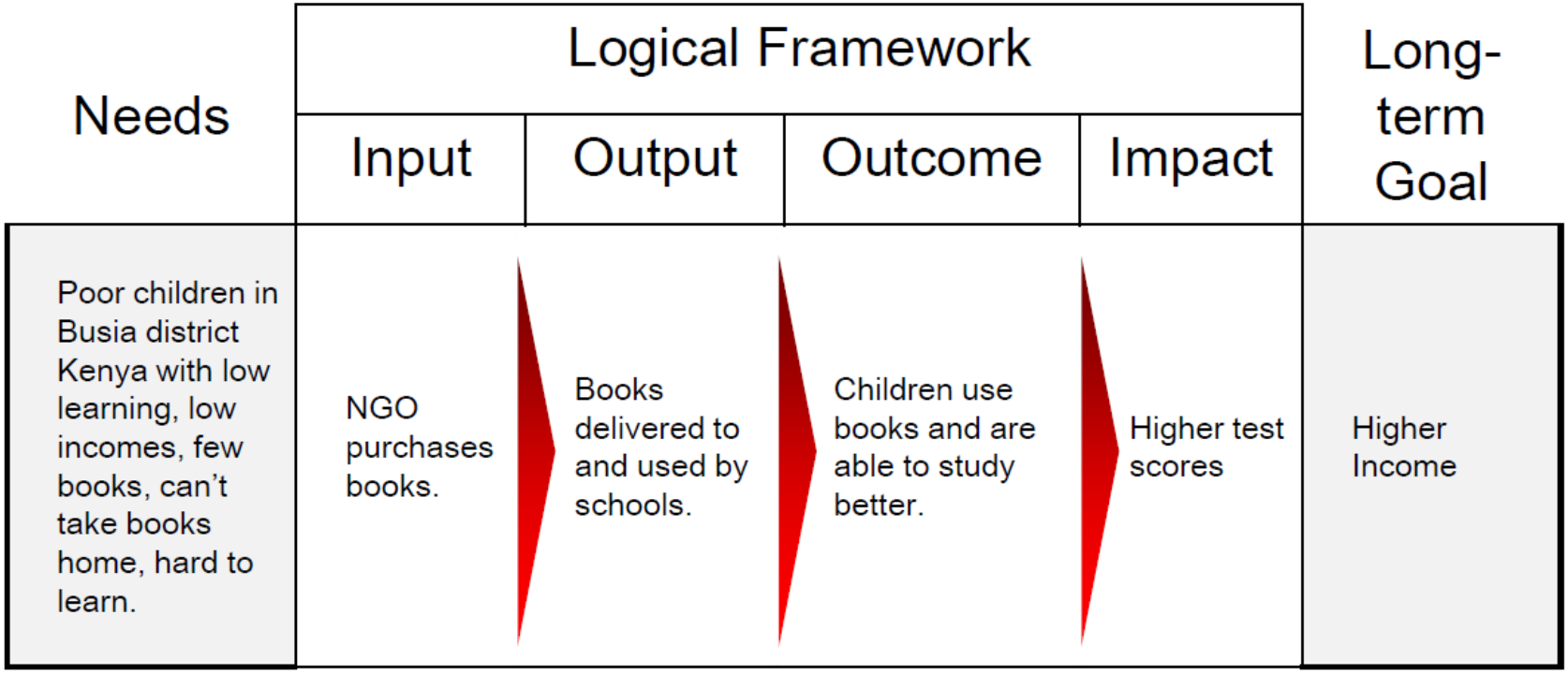
Strength of Randomised Controlled Trials

- Remove bias in the estimation of an intervention's impact
  - Increase internal validity
- For this, rely on random assignment

- Different from random sampling
  - A requirement to increase external validity

1. Constructing a reliable Theory of Change

Example of model: provision of textbooks



2. A lot needs to be done before conducting an impact evaluation

- What are the needs of this population?
- What might be a target need that we can feasibly address?
- What is the right design for this intervention in this context?
- Does the intervention work in test?
- Does the intervention work in deployment - process?
- Does the intervention have an impact? How much?

3. Steps involved in conducting an RCT

Key steps in conducting an experiment

1. Design the study carefully
2. Randomly assign people to treatment or control
3. Collect baseline data
4. Verify that assignment looks random
5. Monitor process so that integrity of experiment is not compromised

4. Steps involved in conducting an RCT

Key steps in conducting an experiment (cont.)

6. Collect follow-up data for both the treatment and control groups in identical ways.
7. Estimate program impacts by comparing mean outcomes of treatment group vs. mean outcomes of control group.
8. Assess whether program impacts are statistically significant and practically significant.

5. Practical considerations

- How to randomise?
  - Lottery
  - Randomised phase-in
  - Rotation
  - Multiple treatments
- Unit of randomisation
  - If spill-over effects expected, then higher-level units required
- Operational constraints

6. How big does the RCT need to be?

- Power (sample size)
  - More is better
  - Power of 80-90% considered normal at 95% sig level; what does this mean?
- Effect size
  - Smaller effect requires larger sample
  - Standardised effect size = effect size/std. dev (Std effect size of 0.2 usually low; 0.4 medium)
- Variability
  - Higher variability, more difficult to measure effect

- Good tool to use for power calculations: Optimal Design Software http://sitemaker.umich.edu/group-based/optimal_design_software

7. Analysis

- If randomised designed and implemented correctly, only need a simple difference & means-testing

- Impact of intervention on the participants we intended to treat
  - Average impact of intervention
  - Difference between average levels of Treatment group and Control group

- When partial compliance, estimate impact of intervention on those within the Treatment group who actually received the Treatment
  - Impact of 'Treatment on the Treated'

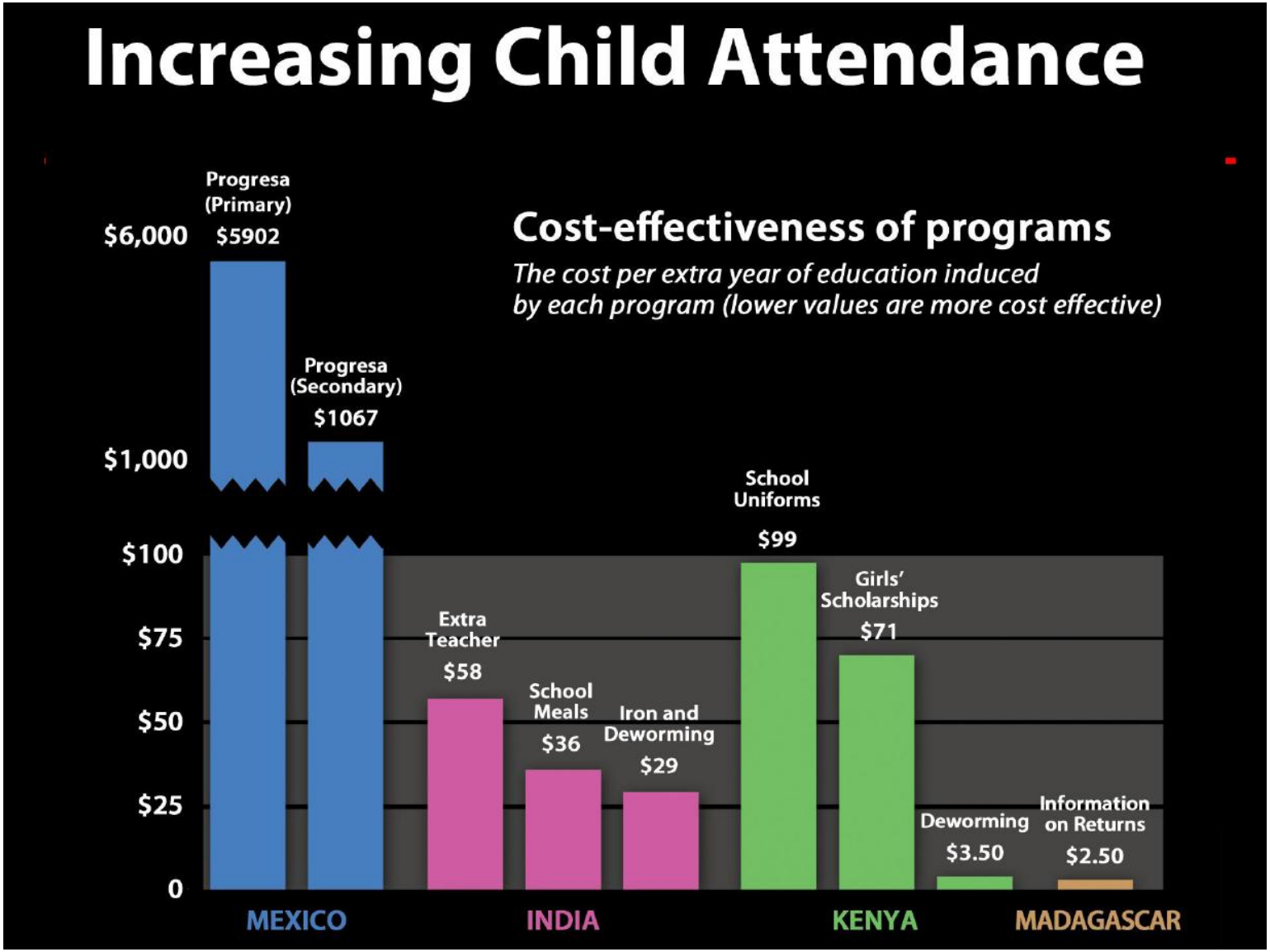
8. Limitations

- Despite great methodological advantage of experiments, they are also potentially subject to threats to their validity. For example,
  - Internal Validity

    (e.g. Hawthorne Effects, survey non-response, no-shows, crossovers, duration bias, etc.)
  - External Validity

    (e.g. are the results generalizable to other populations?)
- It is important to realize that some of these threats also affect the validity of non-experimental studies

9. Aggregating towards external validity and cross-programme comparisons



“The Digital Provide” paper - Discussion

- What is the target problem?
- What is the intervention?
- What is the Theory of Change?
- On what assumptions is it based?
- Do these assumptions hold in the study context?
- Who receives the intervention?
- Who serves as the control group? Do they provide a valid counterfactual?
- How was data collected?
- Does the analysis convince you? Is this impact estimate true? (Internal validity, external validity)

Reference: Jensen, R. (2007) “The Digital Provide: Information (Technology), Market Performance, and Welfare in the South Indian Fisheries Sector.” Quarterly Journal of Economics, 122(3), 879-924.

References/ Resources

- Duflo, Esther, Rachel Glennerster, and Abhijit Banerjee, *RES.14-001 Abdul Latif Jameel Poverty Action Lab Executive Training: Evaluating Social Programs, Spring 2009*. (Massachusetts Institute of Technology: MIT OpenCourseWare), http://ocw.mit.edu

- Optimal Design Software http://sitemaker.umich.edu/group-based/optimal_design_software

- Jensen, R. (2007) "The Digital Provide: Information (Technology), Market Performance, and Welfare in the South Indian Fisheries Sector." Quarterly Journal of Economics, 122(3), 879-924.