ECONOMIC EVALUATION
IN DEVELOPMENT

Autumn 2015

Michael King
A plan to end poverty…..

• “I have identified the specific investments that are needed [to end poverty]; found ways to plan and implement them; [and] shown that they can be affordable.”

  Jeffrey Sachs

  End of Poverty
How much money….

• "After $2.3 trillion over 5 decades, why are the desperate needs of the world's poor still so tragically unmet? Isn't it finally time for an end to the impunity of foreign aid?"

*Bill Easterly*

*The White Man’s Burden*
Types of evaluation: A framework

- Program Evaluation
- Monitoring & Evaluation
- Impact Evaluation
- Randomized Evaluation
Different levels of program evaluation

1. Needs Assessment
2. Program Theory Assessment
3. Process evaluation
4. Impact evaluation
5. Cost-benefit/Cost-effectiveness analysis
Different levels of program evaluation

1. Needs Assessment
2. Program Theory Assessment
3. Process evaluation
4. **Impact evaluation** [Our focus]
5. Cost-benefit/Cost-effectiveness analysis
Why focus on impact evaluation?

• Surprisingly little hard evidence on what works
• Can do more with given budget with better evidence
• If people knew money was going to programs that worked, could help increase funds
• Instead of asking “do aid/development programs work?” should be asking:
  – Which work best, why and when?
  – How can we scale up what works?
What makes a good evaluation?

• Ask the right questions
  – For accountability
  – For lesson learning

• Answers those questions in unbiased and definitive way

• To do that you need a model: logical framework/theory of change
  – Who is the target?
  – What are their needs?
  – What is the program seeking to change?
  – What is the precise program or part of program being evaluated?
**EXAMPLE 1: WATER AND SANITATION PROJECT**

<table>
<thead>
<tr>
<th>OBJECTIVES (What you want to achieve)</th>
<th>INDICATORS (How to measure change)</th>
<th>MEANS OF VERIFICATION (Where &amp; how to get information)</th>
<th>ASSUMPTIONS (What else to be aware of)</th>
</tr>
</thead>
</table>
| **Goal:** Reduce death and illness related to Water and Sanitation related diseases in the targeted communities | G1 % (percentage) reduction in water and sanitation related diseases among target population  
G2 % of children under 36 months with diarrhoea in the last two weeks                           | Ministry of Health / WHO statistics  
Records from village clinics                                                                   | Civil war / hostilities do not return  
Improved access to clinical health facilities                                                   |
| **Outcome 1** Improved access to and use of sustainable sources of safe water in target communities   | 1a % of people in the target communities using minimum 25L of safe water per day  
1b % of targeted households with access to an functional water source  
1c % of water points managed by local WatSan committees  
1d # hours spent by women in fetching water daily                                                | 1a,b,d Household survey  
1c Key informant interviews with WatSan committee members                                      | Low rainfall does not limit overall water supply.                                             |
| **Outputs**                                                                                           | 1.1a # (number) of water points constructed to national standard (140)  
1.1ab% of water handpumps rehabilitated to national standard (35)                               | “Community Facility Inspection” field report                                                   | No major disputes or conflicts within the community                                         |
|                                                                                                      | 1.2a # of communities with a WatSan committee established  
1.2b # of WatSan committees with technicians trained to perform basic maintenance on water points  
1.2c % of WatSan committees collecting adequate charges to maintain the water points            | 1.2a Household survey  
Key informant interviews with WatSan committee members                                          |                                                                                               |
| **Outcome 2** Improved access to and use of sustainable sanitation facilities among targeted communities | 2a % of people in the target communities using latrines on a daily basis  
2b % of targeted households with access to functional latrines meeting national standard  
2c % of latrines managed by local WatSan committees                                               | 2a,b Household survey  
2c Key informant interviews with WatSan committee members                                         | Civil war / hostilities do not return                                                          |
| **Outputs**                                                                                           | 2.1a # of fully functioning household latrines constructed (3,500)                                | “Community Facility Inspection” field report                                                   | Flooding or other environmental problems do not affect sanitation facilities                |
### EXAMPLE 1: WATER AND SANITATION PROJECT

<table>
<thead>
<tr>
<th>OBJECTIVES</th>
<th>INDICATORS</th>
<th>MEANS OF VERIFICATION</th>
<th>ASSUMPTIONS</th>
</tr>
</thead>
</table>
| **Goal:** Reduce death and illness related to Water and Sanitation related diseases in the targeted communities | G1 % (percentage) reduction in water and sanitation related diseases among target population  
G2 % of children under 36 months with diarrhoea in the last two weeks | Ministry of Health / WHO statistics  
Records from village clinics | |
| **Outcome 1** Improved access to and use of sustainable sources of safe water in target communities | 1a % of people in the target communities using minimum 25L of safe water per day  
1b % of targeted households with access to an functional water source  
1c % of water points managed by local WatSan committees  
1d # hours spent by women in fetching water daily | 1a,b,d Household survey  
1c Key informant interviews with WatSan committee members | Civil war / hostilities do not return  
Improved access to clinical health facilities |
| **Outputs** | 1.1 Community water points constructed or rehabilitated | “Community Facility Inspection” field report | Low rainfall does not limit overall water supply. |
| 1.1a # (number) of water points constructed to national standard (140)  
1.1ab % of water handpumps rehabilitated to national standard (35) | | |
| 1.2 Community management of water points is improved | 1.2a # of communities with a WatSan committee established  
1.2b # of WatSan committees with technicians trained to perform basic maintenance on water points  
1.2c % of WatSan committees collecting adequate charges to maintain the water points | 1.2a Household survey  
Key informant interviews with WatSan committee members | No major disputes or conflicts within the community |
| 1.2c % of WatSan committees collecting adequate charges to maintain the water points | | |
| **Outcome 2** Improved access to and use of sustainable sanitation facilities among targeted communities | 2a % of people in the target communities using latrines on a daily basis  
2b % of targeted households with access to functional latrines meeting national standard  
2c % of latrines managed by local WatSan committees | 2a,b Household survey  
2c Key informant interviews with WatSan committee members | Civil war / hostilities do not return |
| **Outputs** | 2.1 Sanitation facilities constructed | “Community Facility Inspection” field report | Flooding or other environmental problems do not affect sanitation facilities |
| 2.1a # of fully functioning household latrines constructed (3,500) | | |
How to measure impact?

• What would have happened in the absence of the program?

• Take the difference between
  – what happened (with the program) ...and
  – what would have happened (without the program)

  =IMPACT of the program
Counterfactual

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

- Randomized: Use random assignment of the program to create a control group which mimics the counterfactual.

- Non-randomized: Argue that a certain excluded group mimics the counterfactual.
Methodologies in impact evaluation

1. Experimental
   - Randomized Evaluations

2. Quasi-experimental
   - Regression Discontinuity Design
   - Instrumental Variables

3. Non-experimental
   - Matching
   - Regression
   - Difference in differences
   - Pre v post

Increasing sophistication
Cost-benefit analysis

1. Needs assessment gives you the metric for measuring impact
2. Process evaluation gives you the costs of all the inputs
3. Impact evaluation gives you the quantified benefits
4. Identifying alternatives allows for comparative cost benefit
Figure 1: J-PAL COST EFFECTIVENESS: additional years of student attendance per $100 spent
www.povertyactionlab.org

- Information: 28.6 years
- Health Interventions: 3.4 years
- Incentives/Reduced Costs: 2.8 years
- Multiple Outcomes: 1.4 years

When to do a randomized evaluation?

- When there is an important question you want/need to know the answer to
- Program is representative not gold plated
- Time, expertise, and money to do it right
When NOT to do an R.E.

- When the program is premature and still requires considerable “tinkering” to work well
- When the project is on too small a scale to randomize into two “representative groups”
- If a positive impact has been proven using rigorous methodology and resources are sufficient to cover everyone
- After the program has already begun and you are not expanding elsewhere
Drawing the chain of causality

- We want to draw the chain of causality
  - How effective is the intervention?
- We also want to answer:
  - Why is it effective?
- Defining and measuring intermediate outcomes will enrich our understanding of the program, reinforce our conclusions, and make it easier to draw general lessons.

Diagram:

```
inputs --intermediary outcomes-- primary outcome
```
The basics of R.E.

• Start with simple case:
• Take a sample of program applicants
• *Randomly* assign them to either:
  – **Treatment Group** – is offered treatment
  – **Control Group** - not allowed to receive treatment (during the evaluation period)
Key advantage of experiments

• Because members of the groups (treatment and control) do not differ systematically at the outset of the experiment

➢ any difference that subsequently arises between them can be attributed to the program rather than to other factors.
Example 1: De-worming in Western Kenya

- About 30,000 children in 75 primary schools in rural Kenya were treated en masse in schools with drugs for hookworm, whipworm, roundworm, and schistosomiasis (bilharzia).
- Randomly phased into schools over several years
- Randomisation by school and individual
- **Key Impacts**
  - Reduced the incidence of moderate-to-heavy infections by 25 percentage points.
  - Reduced school absenteeism by 25 percent, with the largest gains among the youngest pupils.
  - School participation in the area increased by at least 0.14 years of schooling per treated child.
  - There was no evidence that de-worming increased test scores.

- By far the most cost-effective method for improving school participation
Example 2: Teacher Absenteeism

- Chaudhury et al. (2006): results of almost 70,000 surprise visits to a representative sample of primary schools and clinics across 9 poor countries.
  - Teachers were absent 19 percent of the time on average, and health care workers 35 percent of the time.
  - In general, the poorer the country the higher the absence rates. On an average day, 27 percent of teachers are not at work in Uganda.
Project to Combat Absenteeism

• Re-orientating Pay Structure
  – Ordinarily, teachers were paid a salary of Rs. 1,000 (about $22) per month, for 21 days of teaching
  – In the treatment schools, guaranteed base pay of Rs. 500.
  – Rewarded with Rs. 50 for each valid day taught
  – Under new system, monthly pay ranged from Rs. 500 to Rs. 1,300

• Verification by Camera
  – Student take a picture of teacher and other students at the beginning and end of day
  – Teacher present if 5 hour difference and min number of students present
  – Cameras collected a few days before the end of a pay period
Results: Project to Combat Absenteeism

- Teacher attendance:
  - Control: 58%
  - Treatment: 79%

- Student improvement from midpoint to final test:
  - Control: 24%
  - Treatment: 35%

- Student graduation to government school:
  - Control: 16%
  - Treatment: 26%
Eight 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
Eight key steps in conducting an experiment

6. Collect follow-up data for both the treatment and control groups.

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.
Detailed Example: Balsakhi Program
Balsakhi Program: Background

- Implemented by Pratham, an NGO from India
- Program provided tutors (Balsakhi) to help at-risk children with school work
- Previously teachers decided which children would get the balsakhi
- Suppose we evaluated the balsakhi program using a randomized experiment: What would this entail? How would we do it?
Methods to estimate impacts

Let’s look at different ways of estimating the impacts using the data from the schools that got a balsakhi

1. Pre — Post (Before vs. After)
2. Simple difference
3. Difference-in-difference
4. Other non-experimental methods
5. Randomized Experiment
Pre-post (Before vs. After)

- Look at average change in test scores over the school year for the balsakhi children
- Can this difference (26.42) be interpreted as the impact of the balsakhi program?

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average <strong>post-test</strong> score for children with a balsakhi</td>
<td>51.22</td>
</tr>
<tr>
<td>Average <strong>pretest</strong> score for children with a balsakhi</td>
<td>24.80</td>
</tr>
<tr>
<td><strong>Difference</strong></td>
<td>26.42</td>
</tr>
</tbody>
</table>
2 - Simple difference

- Compare test scores of...
  - Children that got balsakhi with
  - Children that did not
- Can this difference (-5.05) be interpreted as the impact of the balsakhi program?

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average score for children with a balsakhi</td>
<td>51.22</td>
</tr>
<tr>
<td>Average score for children without a balsakhi</td>
<td>56.27</td>
</tr>
<tr>
<td>Difference</td>
<td>-5.05</td>
</tr>
</tbody>
</table>
3 – Difference-in-Differences

- Compare gains in test scores of...
  - Children who got balsakhi
- With gains in test scores of...
  - Children who did not get balsakhi
- Can 6.82 be interpreted as the impact of the balsakhi program?

<table>
<thead>
<tr>
<th></th>
<th>Pretest</th>
<th>Post-test</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average score for children <strong>with</strong> a balsakhi</td>
<td>24.80</td>
<td>51.22</td>
<td>26.42</td>
</tr>
<tr>
<td>Average score for children <strong>without</strong> a balsakhi</td>
<td>36.67</td>
<td>56.27</td>
<td>19.60</td>
</tr>
<tr>
<td>Difference</td>
<td></td>
<td></td>
<td><strong>6.82</strong></td>
</tr>
</tbody>
</table>
4 – Other Methods

• There are more sophisticated non-experimental methods to estimate program impacts:
  – Regression
  – Matching
  – Instrumental Variables
  – Regression Discontinuity

• These methods rely on being able to “mimic” the counterfactual under certain assumptions

• Problem: Assumptions are not testable
5 – Randomized Experiment

• Suppose we evaluated the balsakhi program using a randomized experiment
  – What would be the advantage of using this method to evaluate the impact of the balsakhi program?
# Impact of Balsakhi - Summary

<table>
<thead>
<tr>
<th>Method</th>
<th>Impact Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Pre-post</td>
<td>26.42*</td>
</tr>
<tr>
<td>(2) Simple Difference</td>
<td>-5.05*</td>
</tr>
<tr>
<td>(3) Difference-in-Difference</td>
<td>6.82*</td>
</tr>
<tr>
<td>(4) Regression</td>
<td>1.92</td>
</tr>
<tr>
<td>(5) Randomized Experiment</td>
<td>5.87*</td>
</tr>
</tbody>
</table>

*: Statistically significant at the 5% level

**Bottom Line: Which method we use matters!**
Case Study

- Learn to Read evaluations
How to randomise?

• Randomizing at the individual level
• Randomizing at the group level: “Cluster Randomized Trial”
• Which level to randomize? How to choose?
  – Nature of the treatment
  – How is the intervention administered?
  – What is the catchment area of each “unit of intervention”
  – How wide is the potential impact?
• Generally, best to randomize at the level at which the treatment is administered.
Constraint 1: Political/ethical

• For sure – certain things are not ethical.

• Lotteries as a solution
  – Simple lottery is useful when there is no a priori reason to discriminate and a resource constraint
  – Lotteries are simple, common and transparent
  – Randomly chosen from applicant pool
  – Participants know the “winners” and “losers”
  – Perceived as fair
  – Transparent
Constraint 2: Resources

• Most programs have limited resources
• Results in more eligible recipients than resources will allow services for
• Resource constraints are often an evaluator’s best friend
  – May distinguish recipients by arbitrary criteria
Constraint 3: contamination
Spillovers/Crossovers

- Remember the counterfactual! If control group is different from the counterfactual, our results can be biased
- Can occur due to
  - Spillovers
  - Crossovers
Constraint 4: logistics

• Need to recognize logistical constraints in research designs.
  – Ex: de-worming treatment by health workers
  – Suppose administering de-worming drugs was one of many responsibilities of a health worker
  – Suppose the health worker served members from both treatment and control groups
  – It might be difficult to train them to follow different procedures for different groups, and to keep track of what to give whom
Constraint 5: fairness

- Randomizing at the child-level within classes
- Randomizing at the class-level within schools
- Randomizing at the community-level
Constraint 6: sample size

- The program is only large enough to serve a handful of communities
- Primarily an issue of statistical power
Nice approaches

1. Phase-in: takes advantage of expansion
2. Randomization near the threshold
3. Encouragement design - randomize encouragement to receive treatment
<table>
<thead>
<tr>
<th>Design</th>
<th>Most useful when...</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic Lottery</td>
<td>• Program oversubscribed</td>
<td>• Familiar</td>
<td>• Control group may not cooperate</td>
</tr>
<tr>
<td></td>
<td>• Expanding over time</td>
<td>• Easy to understand</td>
<td>• Differential attrition</td>
</tr>
<tr>
<td></td>
<td>• Everyone must receive treatment eventually</td>
<td>• Easy to implement</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Can be implemented in public</td>
<td></td>
</tr>
<tr>
<td>Phase-In</td>
<td>• Program has to be open to all comers</td>
<td>• Easy to understand</td>
<td>• Anticipation of treatment may impact short-run behavior</td>
</tr>
<tr>
<td></td>
<td>• When take-up is low, but can be easily improved with an incentive</td>
<td>• Constraint is easy to explain</td>
<td>• Difficult to measure long-term impact</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Control group complies because they expect to benefit later</td>
<td></td>
</tr>
<tr>
<td>Encouragement</td>
<td>• Program has to be open to all comers</td>
<td>• Can randomize at individual level even when the program is not administered at that level</td>
<td>• Measures impact of those who respond to the incentive</td>
</tr>
<tr>
<td></td>
<td>• When take-up is low, but can be easily improved with an incentive</td>
<td></td>
<td>• Need large enough inducement to improve take-up</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Encouragement itself may have direct effect</td>
</tr>
</tbody>
</table>
More nice approaches

4. Multiple treatment groups: Perfectly possible but a control group still required

5. Cross-cutting treatments: Useful for testing different components of treatment in different combinations

6. Varying levels of treatment: Different dosage levels
Mechanics of randomization

• Need sample frame
• Pull out of a hat/bucket
• Use random number generator in spreadsheet program to order observations randomly
• Stata program code
Points to consider

- Attrition
- Partial Compliance
- Spillovers
- Threat to external validity:
  - Treatment group behavior changes: Hawthorne effect
  - Comparison group behavior changes: John Henry effect
- Generalizability of results
  - Can it be nationally replicated?
  - Is the sample representative?
  - Special implementation?
Estimation

*Population*
We wish to learn about this

*Sample*
But we only see this

- The *sample average* is our estimate of the *population average*
What estimate to be Right On Average

Which sampling strategy will give us a more accurate estimate?
Estimation

• When we do estimation
  – Sample size allows us to say something about the variability of our estimate
  – But it doesn’t ensure that our estimate will be close to the truth on average

➢ Only randomization ensures accuracy.
➢ Then control precision with sample size.
Accuracy versus Precision

Accuracy (Randomization)

Precision (Sample Size)

truth estimates

truth estimates

truth estimates

truth estimates
Measuring Significance: Scientific Method

• Does the scientific method apply to social science?

• The scientific method involves:
  – 1) proposing a hypothesis
  – 2) designing experimental studies to test the hypothesis

• How do we test hypotheses?
Hypothesis testing: Legal analogy

• In criminal law, most institutions follow the rule: “innocent until proven guilty”
• The prosecutor wants to prove their hypothesis that the accused person is guilty
• The burden is on the prosecutor to show guilt
• The jury or judge starts with the “null hypothesis” that the accused person is innocent
Hypothesis testing

• In program evaluation, instead of “presumption of innocence,” the rule is: “presumption of insignificance”

• Policymaker’s hypothesis: the program improves learning

• Evaluators approach experiments using the hypothesis:
  – “There is zero impact” of this program
  – Then we test this “Null Hypothesis” ($H0$)

• The burden of proof is on the program
  – Must show a statistically significant impact
Hypothesis testing

• If our measurements show a difference between the treatment and control group, our first assumption is:
  – In truth, there is no impact (our $H_0$ is still true)
  – There is some margin of error due to sampling
  – “This difference is solely the result of chance (random sampling error)”

• We (still assuming $H_0$ is true) then use statistics to calculate how likely this difference is in fact due to random chance
Hypothesis testing: conclusions

• If it is very unlikely (less than a 5% probability) that the difference is solely due to chance:
  – We “reject our null hypothesis”

• We may now say:
  – “our program has a statistically significant impact”
Hypothesis testing: conclusions

• Are we now 100 percent certain there is an impact?
  – No, we may be only 95% confident
  – And we accept that if we use that 5% threshold, this conclusion may be wrong 5% of the time
  – That is the price we’re willing to pay since we can never be 100% certain

• Because we can never see the counterfactual, We must use random sampling and random assignment, and rely on statistical probabilities
Back to the Balsakhi example: Baseline test score data

This was the distribution of test scores in the baseline. The test was out of 100. Some students did really well, most, not so well. Many actually scored zero.
Now, look at the improvement. Very few scored zero, and many scored much closer to the 40-point range...
Post-test: control & treatment

- Stop! That was the control group. The treatment group is green.
Average difference: 6 points

- What’s the probability that the 6 point difference is due to chance? (Testing statistical significance)
“Significance level” (5%)
Q: How many children would we need to randomly sample to detect that the difference between the two groups is statistically significantly different from zero?
That probability depends on sample size (here: N=2)
Significance: Sample size = 8
Significance: Sample size = 18
Significance: Sample size = 100

![Graph showing treatment mean vs control mean with a significance of 95% and a sample size of 100.](image)
Significance: Sample size = 100

Treatment Mean – Control Mean

Difference under null
Observed difference
$\alpha=95\%$
$N=6,000$

Significance: Sample size = 6,000

Observe $N=6,000$
Hypothesis testing: conclusions

- What if the probability is greater than 5%?
  - We can’t reject our null hypothesis
  - Are we 100 percent certain there is no impact?
  - No, it just didn’t meet the statistical threshold to conclude otherwise
  - Perhaps there is indeed no impact
  - Or perhaps there is impact
    - But not enough sample to detect it most of the time
    - Or we got a very unlucky sample this time
    - How do we reduce this error?
Hypothesis testing: conclusions

• When we use a “95% confidence interval”
• How frequently will we “detect” effective programs?
• That is **Statistical Power**
Hypothesis testing: 95% confidence

<table>
<thead>
<tr>
<th>THE TRUTH</th>
<th>YOU CONCLUDE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Effective</strong></td>
<td><strong>Effective</strong> ☺</td>
</tr>
<tr>
<td><strong>No Effect</strong></td>
<td><strong>Type II Error</strong> ☹</td>
</tr>
<tr>
<td><strong>Effective</strong></td>
<td><strong>Type I Error</strong> ☹</td>
</tr>
<tr>
<td><strong>No Effect</strong></td>
<td><strong>No Effect</strong> ☻</td>
</tr>
</tbody>
</table>

Power: How frequently will we “detect” effective programs?
Power: main ingredients

• **Variance**
  - The more “noisy” it is to start with, the harder it is to measure effects

• **Effect Size to be detected**
  - The more fine (or more precise) the effect size we want to detect, the larger sample we need
  - Smallest effect size with practical / policy significance?

• **Sample Size**
  - The more children we sample, the more likely we are to obtain the true difference
TIME Example Projects

- Migration, Remittances, and Information.
- Powering Education: the impact of solar lamps on educational attainment in rural Kenya.
- Repayment flexibility, contract choice and investment decisions among Indian microfinance borrowers.
- So fresh and so clean: urban community engagement to keep streets trash-free and improve the sustainability of drainage infrastructure in Senegal.