Evolution of Evaluation

Scott Rozelle
Freeman Spogli Institute, Stanford University

Rural Education Action Project (REAP)
My assignment:
“In particular, we would encourage you to reflect on how impact evaluation work in development has changed in the last decade.”

Some questions relevant to this topic include:

- Are more studies being conducted? [Q1] ✓
- Is the quality improving? ✓
- How are the methods that are utilized changing over time? [Q2] ✓
- Is capacity growing in foreign aid agencies and developing countries?
- Are study findings making their way into policy decisions in developing countries and at organizations like the multilateral development banks? ✓
- What problems might be emerging in terms of ethics, capacity, quality of evidence, and dissemination or use? Q3
Q1: Are more studies being conducted?
Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments

Patrick J. McEwan*
Wellesley College

May 2013
Dates of treatment and data collection in 76 “learning” experiments
Dates of treatment and data collection in 76 experiments

Twice as many studies on learning in the ten years after 2000 … than in the 25 years before 2000
Treatments with teacher training

Effect size = 0.123
SE = 0.025
Treatments with computers or technology

Effect size = 0.15
SE = 0.028
Treatments that modify class size, small-group instruction, or group composition

Effect size = .117
SE = .015
Treatments with information

Effect size = 0.049
SE = 0.03
Treatments with food, beverages, and/or micronutrients

Effect size = 0.035
SE = 0.016
Treatments with de-worming drugs

Effect size = 0.013
SE = 0.016

Mean = 0.013
s.e. = 0.016
Are more studies being conducted?

Summary

• A lot more studies are being conducted …
• Accelerating rate …
  [By the way: significant share supported by 3ie]
• We are beginning to learn about what works and what does not
  – Average size of effects are zero to small to medium sized
  – Large heterogeneity ➔ by country / by type of intervention
• What works to bring kids into school / does not necessarily work to improve learning
Q2: How are the methods that are utilized changing over time?
Our first experiment in China
Pilot study: 2007

- Conducted a survey of 4158 fourth grade students in nine counties in rural Shaanxi province
- Tested students’ hemoglobin levels
Initial Results

- Found that 39% of students had anemia
In fact, students in poor rural areas are poor students!!
What can be done?
Will iron supplements \( \rightarrow \) less anemia?

Lower anemia \( \rightarrow \) Better school performance?

**Stage 1**

**Baseline Survey (Oct. 2008):**
- Anemia (Hb) Test
- Standardized Math Test

**POLICY EXPERIMENT**

- **treated**
- **control**

**Evaluation survey**
Locations of 60 sample schools in Shaanxi Province
Using Hemocue 201+ technology gives Hb levels in 45 seconds (Oct. 2008)
Baseline TIMMS test (October 2008)
All fourth grade students
Locations of sample schools in Shaanxi Province

(●) Treatment Schools
(○) Control Schools
Stage 2 ➔ The Intervention

Will iron supplements ➔ less anemia?

Lower anemia ➔ Better school performance?

Baseline survey ➔ POLICY EXPERIMENT ➔ Evaluation survey

Stage 2
School Type A (30 schools)

“Centrum / Day”

Give students one multi-vitamin with iron per day (5 mg of iron) … from November 2008 to May 2009

(≈4 US cents/day)
30 control schools

Zero: no vitamins
Evaluation Survey (stage 3)

Will iron supplements \( \rightarrow \) less anemia?

Lower anemia \( \rightarrow \) Better school performance?

Baseline survey \( \rightarrow \) POLICY EXPERIMENT

RCT’s

treated

critical

Evaluation survey

Stage 3
Evaluation survey
(June 2009)

Re-taking the standardized academic tests …

… after 5 months of vitamins + letter to parent.
Re-taking the Hb Test (June 2009)
Still anemic or not?
Results
Impact of vitamin on students:

**Hemoglobin Points**

- Control: 0
- Treatment (Vitamin/day): 3

**Anemia Rates (%)**

- Control: 0
- Treatment (Vitamin/day): -25%

**Math Test Scores (std. dev.)**

- Control: 0
- Treatment (Vitamin/day): 0.4
Great results ...

Publishable (of course: we got statistically significant results)

Policy impact (??)

-- who did it effect?
-- compliance?
-- external validity?
-- cost effectiveness?
But, even “worse”

What happens if something like this happens?
The PROBLEM: High School Gap in China today

China in the 2010

Percent of students that go to High School

Large cities in China

Poor rural areas
Reasons

• $$$  ???

  – Yes ➔ but, decentralized fiscal system ➔ local government will not spend $$$ on making high school free ➔ more educated people are, faster they leave the county!!

  … or …

• Information  ???

  – High school conveys high benefits on families … and will likely be even higher in the future ➔ “Let’s try this!!”
  – And, it is “inexpensive”
Step 1: Design

Baseline Survey of 132 Junior High Schools (JHSs), 19,832 students

Information, 44 JHSs → Teacher Training → Student Lessons
Career Counseling, 44 JHSs
Control Group, 44 JHSs

Follow-up Survey of 132 JHSs

October, 2010
Random Assignment; November 2010
November, 2010
December, 2010
May, 2011

3ie ➔ Open Window 2
Step 2: Baseline Survey
(2010.10)

• Student survey:
  — standardized math test;
  — household assets;
  — family characteristics;
  — schooling characteristics.

• Teacher and school survey

Students filling in student forms. [Right above]

Teachers filling in teacher forms. [Left below]
Step 3: Interventions (2010.12)

Information Intervention:
Professional counselor trained teachers to give a scripted 45 minute lesson on the wages and costs associated with different levels of schooling.

Career Counseling Intervention
Professional counselor trained teachers to give 4 scripted lessons (45 minutes each) on:

a. The world of work & your place in it
b. Careers overview & career planning
c. The information intervention lesson
d. Choices after junior high school
Step 4: Evaluation Survey (2011.05)

- Near identical survey forms as the baseline:
  - Standardized math test
  - Student forms
  - Teacher forms

Special attention given to identifying the dropped out students (distinguishing them from those who transferred out, repeated a grade or were temporarily absent).
Results
## Results: Information and Counseling

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>dropout</td>
<td>math 2011</td>
<td>any HS</td>
<td>acad. HS</td>
<td>voc. HS</td>
</tr>
<tr>
<td>information</td>
<td>.01</td>
<td>-.01</td>
<td>-.01</td>
<td>-.03</td>
<td>-.01</td>
</tr>
<tr>
<td></td>
<td>[.01]</td>
<td>[.04]</td>
<td>[.01]</td>
<td>[.02]</td>
<td>[.02]</td>
</tr>
<tr>
<td>counseling</td>
<td>.01**</td>
<td>-.07</td>
<td>.00</td>
<td>.02</td>
<td>.02</td>
</tr>
<tr>
<td></td>
<td>[.01]</td>
<td>[.04]</td>
<td>[.01]</td>
<td>[.02]</td>
<td>[.02]</td>
</tr>
<tr>
<td>N</td>
<td>12,082</td>
<td>10,848</td>
<td>10,837</td>
<td>10,837</td>
<td>10,837</td>
</tr>
<tr>
<td>$R^2$</td>
<td>.08</td>
<td>.36</td>
<td>.14</td>
<td>.17</td>
<td>.29</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses, *** p<.01, ** p<.05, * p<.1
Covariate-adjusted analyses
Impacts of:

• Information:
  – On drop out ➔ none
  – On plans for high school ➔ none
  – One learning ➔ none

• Counseling:
  – On drop out ➔ more drop out!!
  – On plans for high school ➔ none
  – One learning ➔ none
Which journal?

JNR
[journal of non-results]

Which policy maker?
[who wants to push something that failed?]
But, that was before …

• In response to critique of Impact Evaluation’s focus on: “impact/no impact” only …

→ Theory of Causal Chain Analysis
CCA helps address one a major criticism of the current wave of IE studies ... They only tell us: what works ... and not much else!

- If research teams can address this criticism, help answer a number of challenges:
  - Why would anyone want to be told that their project does not work
    - World Bank employee?
    - Government official?
    - NGO? Teach for China
  - If you only know it does not work, what is the implication? Eliminate the program ... or fix it? But, how? How to make it work better?
International Initiative for Impact Evaluation (3ie) is an international organization trying to put the “how” in rigorous IE with “theory-based evaluation” or “causal chain analysis”

Example: a nutrition project in Bangladesh

• Bangladesh Integrated Nutrition Project (BINP) … a World Bank Project

• Problem: lots of malnutrition … difficult to solve in traditional institutional structures

→

• Growth monitoring, nutritional counselling and supplementary feeding (based on a program in Tamil Nadu, which was successful)

• According to the design of the project, implemented by NGOs at field level, used Community Nutrition Practitioners (CNPs)
Instead of: Simple Search for Impact

Will MCH program → improve nutrition?

Stage 1
Baseline survey

Stage 2
POLICY EXPERIMENT
RCT's

treated
control

Stage 3
Evaluation survey
Program design (theory of change)

Target group participate in program (mothers of young children)

- Target group for nutritional counselling is the relevant one
- Children are correctly identified to be enrolled in the program
- Exposure to nutritional counselling results in knowledge acquisition and behaviour change
- Behaviour change sufficient to change child nutrition
- Supplementary feeding is supplemental, i.e. no leakage or substitution
- Food is delivered to those enrolled
- Food is of sufficient quantity and quality

Improved nutritional outcomes

Policy experiment

- A targeting participants
  - B1 mothers
  - B2 children
  - C1 counselling
  - C2 supplemental feeding
  - D1 change behavior in child nutrition
  - D2 no leakage / substitute
  - D3 sufficient qnty / qlty
  - E improved nutrition outcome
The evaluation story

• Looked like it was working – all bits in place and **outcome monitoring** data showed fall in severe malnutrition

• Bank agreed to scale up (this is an expensive program ... funded at expense of other projects)

• But Save the Children UK critical, though Bank’s M&E team was positive

• Bank’s evaluation department (IEG) did a more rigorous evaluation – found little or no impact

• Theory-based approach explains why
Measuring outcomes and impacts (M&E)

Project M&E: Post treatment control differences

Height for Age Scores single differences (between treatment and controls)

After the project
Mid term

-0.1
-0.05
0
0.05
0.1
0.15
The evaluation story

- Looked like it was working – all bits in place and outcome monitoring data showed fall in severe malnutrition
- Bank agreed to scale up (this is an expensive program … funded at expense of other projects)
- But Save the Children UK critical, though Bank’s M&E team was positive
- Bank’s evaluation department (IEG) did a more rigorous evaluation – found little or no impact

- Theory-based approach explains why
Measuring outcomes and impacts

Two points:

1.) need for control group that is similar to treatment group (this is what PSM does)

2.) demands an explanation of the problems with the program
The evaluation story

• Looked like it was working – all bits in place and **outcome monitoring** data showed fall in severe malnutrition

• Bank agreed to scale up (this is an expensive program … funded at expense of other projects)

• But Save the Children UK critical, though Bank’s M&E team was positive

• Bank’s evaluation department (IEG) did a more rigorous evaluation – found little or no impact

• Theory-based approach explains why
## Implementing theory-based analysis

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Provide nutritional counseling to caregivers</td>
<td>Mothers are not decision makers, especially if they live with their mother-in-law</td>
</tr>
<tr>
<td>2. Women know about sessions and attend</td>
<td>90% participation, lower in more conservative areas</td>
</tr>
<tr>
<td>1. Malnourished and growth faltering children correctly identified</td>
<td>No – community nutrition practitioners (CNPs) cannot interpret growth charts</td>
</tr>
<tr>
<td>4. Women acquire knowledge</td>
<td>Those attending training do so</td>
</tr>
<tr>
<td>5. And knowledge is turned into practice</td>
<td>No there is a substantial knowledge-practice gap</td>
</tr>
<tr>
<td>6. Supplementary feeding is additional food for intended beneficiary</td>
<td>No, considerable evidence of substitution and leakage</td>
</tr>
<tr>
<td>Adopted changes are sufficient to improve intended outcomes</td>
<td>Only sometimes ➔ for mother/caregivers</td>
</tr>
</tbody>
</table>
Target group participate in program (mothers of young children)

Target group for nutritional counselling is the relevant one

Exposure to nutritional counselling results in knowledge acquisition and behaviour change

Behaviour change sufficient to change child nutrition

Improved nutritional outcomes

Children are correctly identified to be enrolled in the program

Food is delivered to those enrolled

Supplementary feeding is supplemental, i.e. no leakage or substitution

Food is of sufficient quantity and quality

Policy experiment targeting participants

mothers counselling change behavior in child nutrition

B1 C1 D1 E

B2 C2 D2 D3

supplemental feeding

no leakage / substitute

sufficient qnty / qlty

improved nutrition outcome
Impacts when mother participated

Project M&E: Post treatment control differences

OED: propensity score matching (PSM)

Height for Age Scores single differences (between treatment and controls)

-0.1
-0.05
0
0.05
0.1
0.15

When examining just mothers that did not live with their mother in laws ... and babies were supposed to be in the project ...
How to implement “Theory of Causal Chain”?

• It needs to be developed BEFORE / AS PART OF ➔ the Project Design …

• It is the plan around which all activities are based:
  – [Preanalysis Plans]
  – Sampling / Identification of Outcome
  – Baseline Survey
  – Design of the Intervention
  – Endline Survey
  – Analysis
Testing 19,500 children in Gansu and Shaanxi Provinces

5000 (≈25%) were myopic (or nearsighted).
Testing 19,500 children in Gansu and Shaanxi Provinces

5000 (≈25%) were myopic (or nearsighted).

Only 650 had eyeglasses (≈ 3%)
Main Question of Project

- Can we improve educational performance by providing eyeglasses to students?

Two approaches

- Free glasses / exam in schools / dispense in school
- Vouchers / exam in schools / make family + student undertake an ordeal (up to 4 hour bus ride) to pick up glasses
Sites of 50 study schools in 4 sample counties ...

Tianshui

All schools are centralized, K-6 elementary schools
Affects survey design ...
… and analysis …

… with solid Causal Chain Analysis, can do the following:

• Does project work or not?

• Who does it work for?

• If it does not work, why does it not work?

• If it does work, what could make it work better?

• Answer a lot more questions that policy makers are interested in …
Q3: Are study findings making their way into policy decisions in developing countries and at organizations like the multilateral development banks?
In part in response to the availability of evidence-based research, China has created a system for connecting research (evaluators) with the top leadership...
Formal notification that there has been a “policy directive” directing MoEdu and MoHealth to move our information into the policy discussion.
Formal notification that there has been a “policy directive” directing MoEdu and MoHealth to move our information into the policy discussion.
REAP has submitted 17 policy briefs to State Council
Policy Action

These policy directives directly respond to our brief and are signed by:

Li Keqiang: Premier

Liu Yandong: Vice Premier

Li Bin: Minister of Health
But, it does not happen without a concerted effort (pre-thought out plan) and a lot of effort

- We start with P.A.C.’s [not political action committees, like here in DC]: Policy Action Committees
  - 2 to 5 local officials that will be our “advisors”:
    - Get input during experimentation design …
    - More get to understand the project
      - E.g., in the case of RCT’s … truly randomly assigned
    - If successful, PAC-members are given “ownership” and have an incentive to try to upscale …
    - They help us push policy … and often sign Policy Briefs …
    - Policy Briefs can “give political cover.”
Policy suggestions (with the support of a lot of other/complimentary work—of course) → associated with:

- 20 billion dollar “nutritious lunch” program
  - 26 million children x 200 lunches per year x 3.5 yuan per lunch x 10 years
- New financial aid fund for poor kids at high school level
- Deworming campaign in 62 poor counties

Ironic thing: can’t run a DiD regression and show causal impact of our impact evaluation … but, the correlations are strong …
Thank You!

http://reap.stanford.edu