USING EVIDENCE IN POLICY MAKING: IMPACT EVALUATION WORKSHOP

The Abdul Latif Jameel Poverty Action Lab (J-PAL) &
Innovations for Poverty Action (IPA) Malawi

16 – 17 July 2014: Malawi Institute of Management (MIM), Lilongwe

ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

ipa
INNOVATIONS FOR
POVERTY ACTION

BOOKLET CONTENTS

Using Evidence in Policymaking:
Impact Evaluation Workshop

<table>
<tr><th></th><th colspan="2">Wednesday, 16 July</th><th>Thursday, 17 July</th></tr>
<tr><td>8:30</td><td colspan="2">Registration</td><td>Registration</td></tr>
<tr><td>9:00</td><td colspan="2">Introduction to J-PAL/IPA &
Impact Evaluation
<em>Thom Munthali</em></td><td>Recap of Day 1
<em>Laura Poswell</em></td></tr>
<tr><td>9:30</td><td colspan="2" rowspan="2">Lecture 1:
<strong>Why Evaluate?</strong>
<em>Laura Poswell</em></td><td rowspan="2">Lecture 3:
<strong>Theory of Change</strong>
<em>Rachna Chowdhuri</em></td></tr>
<tr><td>10:00</td></tr>
<tr><td>10:30</td><td colspan="3">Tea Break</td></tr>
<tr><td rowspan="2">11:00</td><td colspan="2"><strong>Overview of Recent Evidence (parallel sessions):</strong></td><td rowspan="3">Case Study:
<strong>Theory of Change</strong></td></tr>
<tr><td><strong>EDUCATION</strong>
<em>Laura Poswell</em></td><td rowspan="2"><strong>AGRICULTURAL FINANCE</strong>
<em>Emily Cupito</em>
<em>Anna Yalouris</em></td></tr>
<tr><td>11:30</td><td><strong>HEALTH</strong>
<em>Rachna Chowdhuri</em></td></tr>
<tr><td>12:00</td><td colspan="2" rowspan="2">Case Study:
<strong>Why Randomize?</strong></td><td rowspan="2">Lecture 4:
<strong>Cost-Effectiveness Analysis</strong>
<em>Anna Yalouris</em></td></tr>
<tr><td>12:30</td></tr>
<tr><td>13:00</td><td colspan="3" rowspan="2">Lunch Break</td></tr>
<tr><td>13:30</td></tr>
<tr><td>14:00</td><td colspan="2" rowspan="2">Lecture 2:
<strong>Why Randomize?</strong>
<em>Emily Cupito</em></td><td rowspan="2">Group Work:
<strong>Research Design</strong></td></tr>
<tr><td>14:30</td></tr>
<tr><td>15:00</td><td colspan="3">Tea Break</td></tr>
<tr><td>15:30</td><td colspan="2" rowspan="3">Group Work:
<strong>Choosing a Research Question</strong></td><td rowspan="2">Lecture 5:
<strong>Evaluation from Start to Scale</strong>
<strong>Case Study from Malawi</strong>
<em>Thom Munthali</em></td></tr>
<tr><td>16:00</td></tr>
<tr><td>16:30</td><td>Group Presentations &
Workshop Closing Remarks</td></tr>
</table>

COURSE OVERVIEW

USING EVIDENCE IN POLICY MAKING
IMPACT EVALUATION WORKSHOP

Innovations for Poverty Action (IPA) Malawi and the Abdul Latif Jameel Poverty Action Lab (J-PAL) at the University of Cape Town, South Africa, jointly present a custom training workshop intended to build capacity in understanding methods of impact evaluation and critically using evidence in the policy decision-making process.

This two-day workshop will draw on the expertise of J-PAL and IPA's large academic research network to provide participants with practical guidance for understanding impact evaluation, as well as share evidence from the large body of research conducted in Malawi and elsewhere on what works to reduce poverty and improve livelihoods. The workshop will focus on the most effective policies and programs in the areas of Agriculture, Education, Finance and Health.

MOTIVATION

Impact evaluation has emerged in recent years as a powerful instrument for enhancing policy effectiveness. The growing importance of impact evaluations is linked to the increased focus on outcomes, as embodied in the Millennium Development Goals. Impact evaluations are also increasingly being used for diverse purposes: strategic learning, transparency and accountability, program design and policy formulation. More important has been the need by policymakers and practitioners to directly link outcomes to interventions (projects, programs, initiatives). This calls for rigorous impact evaluation methods that are capable of doing so - the randomized evaluation (RE) is one such method. As the demand for rigorous analysis rises, it is important to build the capacity of government policymakers and local researchers in collecting, critiquing, and taking decisions upon the relevant research.

WORKSHOP METHODOLOGY

The workshop will incorporate the following:

- ***Lectures*** from experienced J-PAL and IPA staff about key topics in impact evaluation and research design from experts in the field of monitoring and evaluation.
- ***Case studies*** to allow participants a chance to apply their knowledge to a case from the field.
- ***Small group exercises*** reinforce the material covered in the plenary and parallel tracks. Expert moderators will work with each group to guide the conversation and provide technical support.

OBJECTIVES

The workshop will provide participants engaged in the formulation of policy with the opportunity to do the following:

- Reflect upon the importance of including rigorous evidence in the policy decision-making process.
- Learn about methods of impact evaluation, with a focus on randomized evaluations.
- Gain a better understanding of the existing body of evidence in the relevant sectors.
- Develop more technical skills in designing randomized evaluations during an additional day of skills building.

WORKSHOP PRESENTERS

THOMAS CHATAGHALALA MUNTHALI is the Malawi Country Director for IPA. Thomas has a good grounding of Randomized Control Trials (RCTs) and holds a Masters Degree and PhD in Economics from the University of Leeds in England with a specialty on public and private investment interactions in Southern Africa. He is a seasoned Economist with extensive international experience having worked with Ministry of Economic Planning, the World Bank, and UNFPA. He is the past President of the Economics Association of Malawi (and remains its Executive member). He has also been sitting on the Presidential National Advisory Council for Strategic Planning from November 2009 until September 2011.

EMILY CUPITO works as a Policy Manager for J-PAL Africa at the University of Cape Town. She leads outreach to practitioners and policymakers across the continent. She helps policymakers interpret research results and think strategically about how these results can be translated into effective programs. Prior to her work at J-PAL, Emily spent more than two years working in Uganda with Innovations for Poverty Action, where she supported financial inclusion research by leading dissemination efforts, developing new projects, and working to build the capacity of researchers in Africa and South Asia. Emily received a Master's in Public Policy from Duke University and a BA from the University of North Carolina at Chapel Hill.

RACHNA NAG CHOWDHURI is the Country Director at Innovations for Poverty Action (IPA) Zambia. As the Country Director, Rachna collaborates with policy makers and researchers in Zambia on agriculture, health and education sectors to disseminate research findings as well as to initiate new research. Previously she worked with J-PAL South Asia (2010-2014) working in different parts of the country with both Government and NGO partners on conducting RCTs in health, education, governance and gender. She has also worked on impact evaluations in India (with the World Bank), Vietnam and Laos where she used quasi-experimental methods. She graduated from University of Sussex with a MSc. in Development Economics in 2009.

LAURA POSWELL is the Executive Director for J-PAL Africa at SALDRU at the University of Cape Town. Her role involves working with governments and NGOs in Africa to decipher policy lessons, and collaborating with researchers to conduct randomized evaluations that address policy questions facing African decision-makers. Her last role with FUEL Trust involved working in close partnership with South Africa's Department of Basic Education to implement a service delivery enhancement program with the National School Nutrition Program. She previously worked as a researcher for the Development Policy Research Unit at the University of Cape Town. Laura has an M.BusSc from the University of Cape Town.

ANNA YALOURIS is a Senior Policy Associate at J-PAL Africa at the University of Cape Town. Anna has worked as the manager for J-PAL's Finance & Microfinance Program and the support staff for the Health Program. Anna's responsibilities include conducting outreach to disseminate lessons from J-PAL evaluations to the policy community, with a focus on the African continent. Anna graduated magna cum laude with a BA in Economics from Bates College, where she received the 2008 Stangle Family Award in Economics. Anna brings experience working on an agricultural impact evaluation in Sierra Leone, and an interest in financial product design, preventive healthcare delivery, and nutrition.

RESOURCES FOR FINDING GOOD EVIDENCE

Resources from J-PAL and Partners on:
FINDING EVIDENCE and **CONDUCTING RANDOMIZED EVALUATIONS**

Part I: Resources for Finding Evidence

1. J-PAL Website: Evaluation Summary Database

Available from: www.povertyactionlab.org/evaluations

J-PAL
EVALUATION SUMMARY

ABDUL LATIF JAMEEL
Poverty Action Lab

WWW.POVERTYACTIONLAB.ORG

BALANCING HEALTH BENEFITS AND RISKS OF ACT SUBSIDIES FOR AFRICA

RESEARCHERS: Jessica Cohen
Pascaline Dupas
Simone Schaner
PARTNERS: Clinton Health Access Initiative (CHAI)
Novartis Pharmaceuticals
FIELDWORK IMPLEMENTED BY: Innovations for Poverty Action (IPA)
LOCATION: Western Province, Kenya
SAMPLE: 2,789 households in rural Kenya
TIMELINE: 2009
THEMES: Health
POLICY GOALS: Pricing Health Products

POLICY ISSUE:
Malaria is one of the world's foremost public health concerns, killing n people every year. In many malaria-endemic regions, resistance has developed to all but one class of antimalarial drugs, called artemisinin combination therapies (ACTs). ACTs sold in the retail sector are unaffordable for the poor, and although heavy subsidies can make them accessible, the benefits of treating more people and lowering transmission rates must be balanced against the risk of overtreatment, which can hasten the development of drug resistance. A new malaria testing technology, the rapid diagnostic test (RDT), has made it possible to perform malaria testing in the retail sector, but these tests are not in demand at prevailing prices. How can policymakers tailor the prices of subsidized drugs and diagnostic tests to target those truly sick with malaria and prevent those who do not need ACTs from taking them?

CONTEXT OF THE EVALUATION:
In Kenya, ACTs are now the only effective class of antimalarial drugs. The incidence of malaria in Western Kenya is very high, with nearly 70 percent of households self-reporting an episode of malaria in the month before baseline. ACTs at government health centers in Kenya are nominally free, but health centers feature long waits and limited hours, are often stocked out of medication, and the remotely-located poor often cannot afford to travel the distance to get there. Consequently, many people opt to purchase cheaper, less effective anti-malarials over-the-counter at drug shops located closer to home.

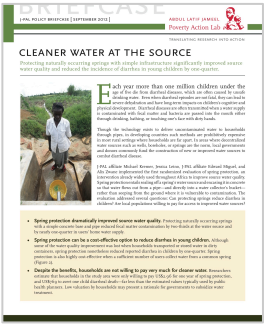
The Affordable Medicines Facility for malaria (AMFm) is an initiative currently being considered by major international aid agencies in which ACTs would be heavily subsidized to first line buyers throughout Africa (the subsidy policy is currently being piloted in seven countries). But such a high subsidy could lead to overtreatment. If people who do not have malaria presumptively buy ACTs without a diagnosis, this can contribute to drug-resistance, waste subsidy money, and delay appropriate treatment for the true cause of illness. Improving

J-PAL's network of 100 affiliated researchers have over 500 completed or ongoing randomized evaluations of programs and policies aimed at improving the well-being of the poor. This research covers diverse topics in the fields of Agriculture, Education, Environment & Energy, Finance & Microfinance, Governance, Health, and Labor Markets. Over 150 of these evaluations were conducted in Africa.

This body of research can be freely accessed through J-PAL's searchable evaluation database. Each online record contains details and resources such as a brief policy-oriented summary of the research, links to academic publications, news coverage, data, and more.

2. J-PAL Website: Policy Publications

Available from: http://www.povertyactionlab.org/policy-lessons/publications



BRIEFCASE
J-PAL POLICY BRIEFCASE [SEPTEMBER 2012]

ABDUL LATIF JAMEEL
Poverty Action Lab

TRANSLATING RESEARCH INTO ACTION

CLEANER WATER AT THE SOURCE

Protecting naturally occurring springs with simple infrastructure significantly improved source water quality and reduced the incidence of diarrhea in young children by one-quarter.

Each year more than one million children under the age of five die from diarrheal diseases, which are often caused by unsafe drinking water. Even when diarrheal episodes are not fatal, they can lead to severe dehydration and have long-term impacts on children's cognitive and physical development. Diarrheal diseases are often transmitted when a water supply is contaminated with fecal matter and bacteria are passed into the mouth either through drinking, bathing, or touching one's face with dirty hands.

Though the technology exists to deliver uncontaminated water to households through pipes, in developing countries such methods are prohibitively expensive in most rural settings where households are far apart. In areas where decentralized water sources such as wells, boreholes, or springs are the norm, local governments and donors commonly fund the construction of new or improved water sources to combat diarrheal disease.

J-PAL affiliate Michael Kremer, Jessica Leino, J-PAL affiliate Edward Miguel, and Alix Zwane implemented the first randomized evaluation of spring protection, an intervention already widely used throughout Africa to improve source water quality. Spring protection entails sealing off a spring's water source and encasing it in concrete so that water flows out from a pipe—and directly into a water collector's bucket—rather than seeping from the ground where it is vulnerable to contamination. The evaluation addressed several questions: Can protecting springs reduce diarrhea in children? Are local populations willing to pay for access to improved water sources?

- **Spring protection dramatically improved source water quality.** Protecting naturally occurring springs with a simple concrete base and pipe reduced fecal matter contamination by two-thirds at the water source and by nearly one-quarter in users' home water supply.
- **Spring protection can be a cost-effective option to reduce diarrhea in young children.** Although some of the water quality improvement was lost when households transported or stored water in dirty containers, spring protection nonetheless reduced reported diarrhea in children by one-quarter. Spring protection is also highly cost-effective when a sufficient number of users collect water from a common spring (Figure 2).
- **Despite the benefits, households are not willing to pay very much for cleaner water.** Researchers estimate that households in the study area were only willing to pay US$2.96 for one year of spring protection, and US$769 to avert one child diarrheal death—far less than the estimated values typically used by public health planners. Low valuation by households may present a rationale for governments to subsidize water treatment.

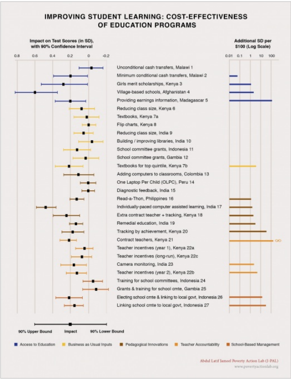
J-PAL's policy group produces policy publications to accompany the most successful or policy-relevant studies. Policy *briefcases* discuss a single study, while *bulletins* synthesize evidence from multiple studies and often accompany cost-electiveness analyses.

Policymakers can use this more in-depth policy discussion of the research to help decide if a program is appropriate in a new context.

3. J-PAL Website: Cost-Effectiveness Analysis (CEA)

Available from: www.povertyactionlab.org/policy-lessons

The cost-effectiveness analyses presented on J-PAL's website show the impact against a specific policy goal that can be achieved for a given expenditure (e.g. additional years of education per $100 spent). All the impact estimates are based on evidence from randomized evaluations.



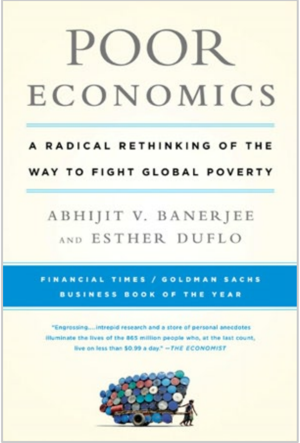
IMPROVING STUDENT LEARNING: COST-EFFECTIVENESS OF EDUCATION PROGRAMS

Abdul Latif Jameel Poverty Action Lab (J-PAL)
www.povertyactionlab.org

Full details of J-PAL's cost-effectiveness methodology, including assumptions on measuring costs and benefits, are included in the 2012 paper *Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries*, available at: www.povertyactionlab.org/publication/cost-effectiveness

Cost-effectiveness analysis, combined with an understanding of the problem being addressed and of other contextual factors such as current input prices and local institutions, can provide important insights into which programs are likely to provide the greatest value for money in a particular situation, and to identify the key factors to which these outcomes are most sensitive.

4. **POOR ECONOMICS: A Radical Rethinking of the Way to Fight Global Poverty**
Additional resources available from: www.pooreconomics.com


POOR
ECONOMICS
A RADICAL RETHINKING OF THE
WAY TO FIGHT GLOBAL POVERTY
ABHIJIT V. BANERJEE
AND ESTHER DUFLO
FINANCIAL TIMES / GOLDMAN SACHS
BUSINESS BOOK OF THE YEAR
"Engrossing....intrepid research and a store of personal anecdotes illuminate the lives of the 865 million people who, at the last count, live on less than $0.99 a day." —THE ECONOMIST

Abhijit Banerjee and Esther Duflo, two of J-PAL's founding directors, present a radical rethinking in the way to fight global poverty.

POOR ECONOMICS argues that so much of anti-poverty policy has failed over the years because of an inadequate understanding of poverty. Through a careful analysis of a rich body of evidence, including hundreds of randomized evaluations, the authors show why the poor, despite having the same desires and abilities as anyone else, end up with entirely different lives. The battle against poverty can be won, but it will take patience, careful thinking and a willingness to learn from evidence.

Website provides supporting material: informative slideshows, material for teaching the book, supporting data, and links to researcher and organization websites.

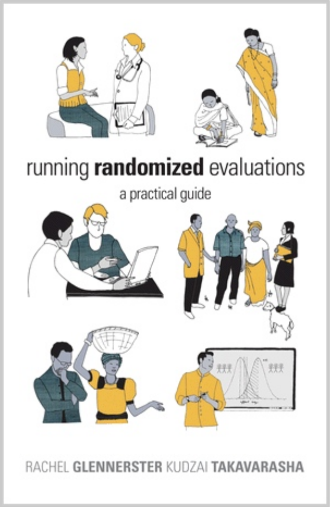
5. **Resources from Partner Organizations**
J-PAL's partner organizations include numerous research centers and program implementers. Key partner organizations are listed below. J-PAL's full partner database can be accessed from:
www.povertyactionlab.org/search/apachesolr_search?filters=type:partner

- Agricultural Technology Adoption Initiative: *www.atai-research.org*
- Center for Effective Global Action (CEGA) – University of California, Berkley*:* *www.cega.berkeley.edu*
- Centre for Micro Finance – IFMR: *www.centre-for-microfinance.org*
- CLEAR Initiative: *www.theclearinitiative.org*
- Evidence for Policy Design (EPoD) - Harvard Kennedy School: *www.hks.harvard.edu/centers/cid/programs/evidence-for-policy-design*
- Evidence Action: *www.evidenceaction.org*
- Deworm the World: *www.dewormtheworld.org*
- Ideas42: *www.ideas42.org*
- Innovations for Poverty Action (IPA): *www.poverty-action.org*
- International Initiative for Impact Evaluation (3ie): *www.3ieimpact.org*
- The Development IMpact Evaluation (DIME) Initiative - World Bank

Part II: Resources for Conducting Randomized Evaluations

1. RUNNING RANDOMIZED EVALUATIONS: A practical guide

Additional resources available from: www.runningres.com


running randomized evaluations
a practical guide
RACHEL GLENNERSTER KUDZAI TAKAVARASHA

Executive Director Rachel Glennerster, along with Kudzai Takavarasha, present a new practical guide for conducting research.

RUNNING RANDOMIZED EVALUATIONS gives evaluators and practitioners the know-how they need to do valid randomized impact evaluations of social programs in developing countries.

The book takes the evaluator step by step through the process of doing a randomized evaluation. They cover the choice of randomization technique, planning for data collection, designing the evaluation to have high statistical power, addressing threats to the validity of the experiment, and analyzing the data. They also explain the role the evaluator plays in program design, and how evaluators can choose the right time and context for conducting an evaluation. Final chapters provide an overview of how to interpret and draw policy conclusions from the results of randomized evaluations or generalize the results from one context to another.

2. J-PAL Executive Education & Custom Evaluation Workshops

J-PAL seeks to build the capacity of others to conduct randomized evaluations through Executive Education training courses. This five-day program on evaluating social programs provides a thorough understanding of randomized evaluations and pragmatic step-by-step training for conducting one's own evaluation. The J-PAL Training Course is held annually in several locations worldwide. General course details, including upcoming courses, are available from: *http://www.povertyactionlab.org/course*

A free online version of the Executive Education course, taught by MIT professors, is available from: *http://ocw.mit.edu/resources/res-14-002-abdul-latif-jameel-poverty-action-lab-executive-training-evaluating-social-programs-2011-spring-2011/*

Resources on J-PAL's custom impact evaluation workshop with the IPA Malawi, *Using Evidence in Policy Making,* are available from: *www.povertyactionlab.org/event/malawi-capacity-building-workshop*

3. J-PAL Website: Methodology Overview

Available at: www.povertyactionlab.org/methodology

The methodology section on J-PAL website provides a detailed overview of randomized evaluations and other impact evaluation methods. These pages cover the what, why, who, when, and how of randomized evaluations. Numerous academic and policy resources are also available, along with detailed descriptions and resources for the following topics: Needs Assessment Program, Theory Assessment, Process Evaluation, Impact Evaluation, Cost-Benefit, Cost-Effectiveness, and Cost-Comparison Analysis, Goals, Outcomes, and Measurement.

CHECKLIST FOR REVIEWING A RANDOMIZED IMPACT EVALUATION

Checklist For Reviewing a Randomized Controlled Trial of a Social Program or Project, To Assess Whether It Produced Valid Evidence

Coalition for Evidence-Based Policy

A NONPROFIT, NONPARTISAN ORGANIZATION

Updated February 2010

This publication was produced by the Coalition for Evidence-Based Policy, with funding support from the William T. Grant Foundation, Edna McConnell Clark Foundation, and Jerry Lee Foundation.

We welcome comments and suggestions on this document (jbaron@coalition4evidence.org).

Checklist For Reviewing a Randomized Controlled Trial of a Social Program or Project, To Assess Whether It Produced Valid Evidence

This is a checklist of key items to look for in reading the results of a randomized controlled trial of a social program, project, or strategy ("intervention"), to assess whether it produced valid evidence on the intervention's effectiveness. This checklist closely tracks guidance from both the U.S. Office of Management and Budget (OMB) and the U.S. Education Department's Institute of Education Sciences (IES)[1]; however, the views expressed herein do not necessarily reflect the views of OMB or IES.

This checklist limits itself to key items, and does not try to address all contingencies that may affect the validity of a study's results. It is meant to aid – not substitute for – good judgment, which may be needed for example to gauge whether a deviation from one or more checklist items is serious enough to undermine the study's findings.

A brief appendix addresses *how many* well-conducted randomized controlled trials are needed to produce strong evidence that an intervention is effective.

Checklist for overall study design

D **Random assignment was conducted at the appropriate level – either groups (e.g., classrooms, housing projects), or individuals (e.g., students, housing tenants), or both.**

Random assignment of individuals is usually the most efficient and least expensive approach. However, it may be necessary to randomly assign groups – instead of, or in addition to, individuals – in order to evaluate (i) interventions that may have sizeable "spillover" effects on nonparticipants, and (ii) interventions that are delivered to whole groups such as classrooms, housing projects, or communities. (See reference 2 for additional detail.[2])

D **The study had an adequate sample size – one large enough to detect meaningful effects of the intervention.**

Whether the sample is sufficiently large depends on specific features of the intervention, the sample population, and the study design, as discussed elsewhere.[3] Here are two items that can help you judge whether the study you're reading had an adequate sample size:

- If the study found that the intervention produced *statistically-significant* effects (as discussed later in this checklist), then you can probably assume that the sample was large enough.

- If the study found that the intervention did *not* produce statistically-significant effects, the study report should include an analysis showing that the sample was large enough to detect meaningful effects of the intervention. (Such an analysis is known as a "power" analysis.[4])

Reference 5 contains illustrative examples of sample sizes from well-conducted randomized controlled trials conducted in various areas of social policy.[5]

Checklist to ensure that the intervention and control groups remained equivalent during the study

D **The study report shows that the intervention and control groups were highly similar in key characteristics prior to the intervention (e.g., demographics, behavior).**

D **If the study asked sample members to consent to study participation, they provided such consent *before* learning whether they were assigned to the intervention versus control group.**

If they provided consent afterward, their knowledge of which group they are in could have affected their decision on whether to consent, thus undermining the equivalence of the two groups.

D **Few or no control group members participated in the intervention, or otherwise benefited from it (i.e., there was minimal "cross-over" or "contamination" of controls).**

D **The study collected outcome data in the same way, and at the same time, from intervention and control group members.**

D **The study obtained outcome data for a high proportion of the sample members originally randomized (i.e., the study had low sample "attrition").**

As a general guideline, the studies should obtain outcome data for at least 80 percent of the sample members originally randomized, including members assigned to the intervention group who did not participate in or complete the intervention. Furthermore, the follow-up rate should be approximately the same for the intervention and the control groups.

The study report should include an analysis showing that sample attrition (if any) did not undermine the equivalence of the intervention and control groups.

D **The study, in estimating the effects of the intervention, kept sample members in the original group to which they were randomly assigned.** This even applies to:

- Intervention group members who failed to participate in or complete the intervention (retaining them in the intervention group is consistent with an "intention-to-treat" approach); and

- Control group members who may have participated in or benefited from the intervention (i.e., "cross-overs," or "contaminated" members of the control group).[6]

Checklist for the study's outcome measures

D **The study used "valid" outcome measures – i.e., outcome measures that are highly correlated with the true outcomes that the intervention seeks to affect.** For example:

- Tests that the study used to measure outcomes (e.g., tests of academic achievement or psychological well-being) are ones whose ability to measure true outcomes is well-established.

- If sample members were asked to self-report outcomes (e.g., criminal behavior), their reports were corroborated with independent and/or objective measures if possible (e.g., police records).

- The outcome measures did not favor the intervention group over the control group, or vice-versa. For instance, a study of a computerized program to teach mathematics to young students should not measure outcomes using a computerized test, since the intervention group will likely have greater facility with the computer than the control group.[7]

D **The study measured outcomes that are of policy or practical importance – not just intermediate outcomes that may or may not predict important outcomes.**

As illustrative examples: (i) the study of a pregnancy prevention program should measure outcomes such as actual pregnancies, and not just participants' attitudes toward sex; and (ii) the study of a remedial reading program should measure outcomes such as reading comprehension, and not just the ability to sound out words.

D **Where appropriate, the members of the study team who collected outcome data were "blinded" – i.e., kept unaware of who was in the intervention and control groups.**

Blinding is important when the study measures outcomes using interviews, tests, or other instruments that are not fully structured, possibly allowing the person doing the measuring some room for subjective judgment. Blinding protects against the possibility that the measurer's bias (e.g., as a proponent of the intervention) might influence his or her outcome measurements. Blinding would be important, for example, in a study that measures the incidence of hitting on the playground through playground observations, or a study that measures the word identification skills of first graders through individually-administered tests.

D **Preferably, the study measured whether the intervention's effects lasted long enough to constitute meaningful improvement in participants' lives (e.g., a year, hopefully longer).**

This is important because initial intervention effects often diminish over time – for example, as changes in intervention group behavior wane, or as the control group "catches up" on their own.

Checklist for the study's reporting of the intervention's effects

D **If the study claims that the intervention has an effect on outcomes, it reports (i) the size of the effect, and whether the size is of policy or practical importance; and (ii) tests showing the effect is statistically significant (i.e., unlikely to be due to chance).**

These tests for statistical significance should take into account key features of the study design, including:

- Whether individuals (e.g., students) or groups (e.g., classrooms) were randomly assigned;

- Whether the sample was sorted into groups prior to randomization (i.e., "stratified," "blocked," or "paired"); and

- Whether the study intends its estimates of the intervention's effect to apply only to the sites (e.g., housing projects) in the study, or to be generalizable to a larger population.

D **The study reports the intervention's effects on all the outcomes that the study measured, not just those for which there is a positive effect.**

This is so you can gauge whether any positive effects are the exception or the pattern. In addition, if the study found only a limited number of statistically-significant effects among many outcomes measured, it should report tests showing that such effects were unlikely to have occurred by chance.

Appendix: How many randomized controlled trials are needed to produce strong evidence of effectiveness?

To have strong confidence that an intervention would be effective if faithfully replicated, one generally would look for evidence including the following:

D **The intervention has been demonstrated effective, through well-conducted randomized controlled trials, in more than one site of implementation.**

Such a demonstration might consist of two or more trials conducted in different implementation sites, or alternatively one large multi-site trial.

D **The trial(s) evaluated the intervention in the real-world community settings and conditions where it would normally be implemented** (e.g., community drug abuse clinics, public schools, job training program sites).

This is as opposed to tightly-controlled conditions, such as specialized sites that researchers set up at a university for purposes of the study, or settings where the researchers themselves administer the intervention.

D **There is no strong countervailing evidence, such as well-conducted randomized controlled trials of the intervention showing an absence of effects.**

References

[1] U.S. Office of Management and Budget (OMB), What Constitutes Strong Evidence of Program Effectiveness, http://www.whitehouse.gov/omb/part/2004_program_eval.pdf, 2004; U.S. Department of Education's Institute of Education Sciences, Identifying and Implementing Educational Practices Supported By Rigorous Evidence, http://www.ed.gov/rschstat/research/pubs/rigorousevid/index.html, December 2003; What Works Clearinghouse of the U.S. Education Department's Institute of Education Sciences, Key Items To Get Right When Conducting A Randomized Controlled Trial in Education, prepared by the Coalition for Evidence-Based Policy, http://ies.ed.gov/ncee/wwc/pdf/guide_RCT.pdf.

[2] Random assignment of groups rather than, or in addition to, individuals may be necessary in situations such as the following:

(a) The intervention may have sizeable "spillover" effects on individuals other than those who receive it.

For example, if there is good reason to believe that a drug-abuse prevention program for youth in a public housing project may produce sizeable reductions in drug use not only among program participants, but also among their peers in the same housing project (through peer-influence), it is probably necessary to randomly assign whole housing projects to intervention and control groups to determine the program's effect. A study that only randomizes individual youth within a housing project to intervention versus control groups will underestimate the program's effect to the extent the program reduces drug use among both intervention and control-group students in the project.

(b) The intervention is delivered to groups such as classrooms or schools (e.g., a classroom curriculum or schoolwide reform program), and the study seeks to distinguish the effect of the intervention from the effect of other group characteristics (e.g., quality of the classroom teacher).

For example, in a study of a new classroom curriculum, classrooms in the sample will usually differ in two ways: (i) whether they use the new curriculum or not, and (ii) who is teaching the class. Therefore, if the study (for example) randomly assigns individual students to two classrooms that use the curriculum versus two classrooms that don't, the study will not be able to distinguish the effect of the curriculum from the effect of other classroom characteristics, such as the quality of the teacher. Such a study therefore probably needs to randomly assign whole classrooms and teachers (a sufficient sample of each) to intervention and control groups, to ensure that the two groups are equivalent not only in student characteristics but also in classroom and teacher characteristics.

For similar reasons, a study of a schoolwide reform program will probably need to randomly assign whole schools to intervention and control groups, to ensure that the two groups are equivalent not only in student characteristics but also school characteristics (e.g., teacher quality, average class size).

[3] What Works Clearinghouse of the U.S. Education Department's Institute of Education Sciences, *Key Items To Get Right When Conducting A Randomized Controlled Trial in Education*, op. cit., no. 1.

[4] Resources that may be helpful in reviewing or conducting power analyses include: the William T. Grant Foundation's free consulting service in the design of group-randomized trials, at http://sitemaker.umich.edu/group-based/consultation_service; Steve Raudenbush et. al., *Optimal Design Software for Group Randomized Trials*, at http://sitemaker.umich.edu/group-based/optimal_design_software; Peter Z. Schochet, *Statistical Power for Random Assignment Evaluations of Education Programs* (http://www.mathematica-mpr.com/publications/PDFs/statisticalpower.pdf), prepared for the U.S. Education Department's Institute of Education Sciences, June 22, 2005; and Howard S. Bloom, "Randomizing Groups to Evaluate Place-Based Programs," in *Learning More from Social Experiments: Evolving Analytical Approaches*, edited by Howard S. Bloom. New York: Russell Sage Foundation Publications, 2005, pp. 115-172.

[5] Here are illustrative examples of sample sizes from well-conducted randomized controlled trials in various areas of social policy: (i) 4,028 welfare applicants and recipients were randomized in a trial of Portland Oregon's Job Opportunities and Basic Skills Training Program (a welfare-to work program), to evaluate the program's effects on employment and earnings – see http://evidencebasedprograms.org/wordpress/?page_id=140; (ii) between 400 and 800 women were randomized in each of three trials of the Nurse-Family Partnership (a nurse home visitation program for low-income, pregnant women), to evaluate the program's effects on a range of maternal and child outcomes, such as child abuse and neglect, criminal arrests, and welfare dependency – see http://evidencebasedprograms.org/wordpress/?page_id=57; 206 9th graders were randomized in a trial of Check and

Connect (a school dropout prevention program for at-risk students), to evaluate the program's effects on dropping out of school – see http://evidencebasedprograms.org/wordpress/?page_id=92; 56 schools containing nearly 6000 students were randomized in a trial of LifeSkills Training (a substance-abuse prevention program), to evaluate the program's effects on students' use of drugs, alcohol, and tobacco – see http://evidencebasedprograms.org/wordpress/?page_id=128.

[6] The study, after obtaining estimates of the intervention's effect with sample members kept in their original groups, can sometimes use a "no-show" adjustment to estimate the effect on intervention group members who actually participated in the intervention (as opposed to no-shows). A variation on this technique can sometimes be used to adjust for "cross-overs." See Larry L. Orr, *Social Experimentation: Evaluating Public Programs With Experimental Methods*, Sage Publications, Inc., 1999, p. 62 and 210; and Howard S. Bloom, "Accounting for No-Shows in Experimental Evaluation Designs," *Evaluation Review*, vol. 8, April 1984, pp. 225-246.

[7] Similarly, a study of a crime prevention program that involves close police supervision of program participants should not use arrest rates as a measure of criminal outcomes, because the supervision itself may lead to more arrests for the intervention group.

GLOSSARY OF IMPACT EVALUATION VOCABULARY

Evaluation Glossary

Sources: 3ie and The World Bank

Attribution
The extent to which the observed change in outcome is the result of the intervention, having allowed for all other factors which may also affect the outcome(s) of interest.

Attrition
Either the drop out of subjects from the sample during the intervention, or failure to collect data from a subject in subsequent rounds of a data collection. Either form of attrition can result in biased impact estimates.

Baseline
Pre-intervention, ex-ante. The situation prior to an intervention, against which progress can be assessed or comparisons made. Baseline data are collected before a program or policy is implemented to assess the “before” state.

Bias
The extent to which the estimate of impact differs from the true value as a result of problems in the evaluation or sample design.

Cluster
A cluster is a group of subjects that are similar in one way or another. For example, in a sampling of school children, children who attend the same school would belong to a cluster, because they share the same school facilities and teachers and live in the same neighborhood.

Cluster sample
Sample obtained by drawing a random sample of clusters, after which either all subjects in selected clusters constitute the sample or a number of subjects within each selected cluster is randomly drawn.

Comparison group
A group of individuals whose characteristics are similar to those of the treatment groups (or participants) but who do not receive the intervention. Comparison groups are used to approximate the counterfactual. In a randomized evaluation, where the evaluator can ensure that no confounding factors affect the comparison group, it is called a control group.

Confidence level
The level of certainty that the true value of impact (or any other statistical estimate) will fall within a specified range.

Confounding factors
Other variables or determinants that affect the outcome of interest.

Contamination
When members of the control group are affected by either the intervention (see “spillover effects”) or another intervention that also affects the outcome of interest. Contamination is a common problem as there are multiple development interventions in most communities.

Cost-effectiveness
An analysis of the cost of achieving a one unit change in the outcome. The advantage compared to cost-benefit analysis, is that the (often controversial) valuation of the outcome is avoided. Can be used to compare the relative efficiency of programs to achieve the outcome of interest.

Counterfactual
The counterfactual is an estimate of what the outcome would have been for a program participant in the absence of the program. By definition, the counterfactual cannot be observed. Therefore it must be estimated using comparison groups.

Dependent variable
A variable believed to be predicted by or caused by one or more other variables (independent variables). The term is commonly used in regression analysis.

Difference-in-differences (also known as double difference or D-in-D)
The difference between the change in the outcome in the treatment group compared to the equivalent change in the control group. This method allows us to take into account any differences between the treatment and comparison groups that are constant over time. The two differences are thus before and after and between the treatment and comparison groups.

Evaluation
Evaluations are periodic, objective assessments of a planned, ongoing or completed project, program, or policy. Evaluations are used to answer specific questions often related to design, implementation and/or results.

***Ex ante* evaluation design**
An impact evaluation design prepared before the intervention takes place. Ex ante designs are stronger than ex post evaluation designs because of the possibility of considering random assignment, and the collection of baseline data from both treatment and control groups. Also called prospective evaluation.

***Ex post* evaluation design**

An impact evaluation design prepared once the intervention has started, and possibly been completed. Unless the program was randomly assigned, a quasi-experimental design has to be used.

External validity
The extent to which the causal impact discovered in the impact evaluation can be generalized to another time, place, or group of people. External validity increases when the evaluation sample is representative of the universe of eligible subjects.

Follow-up survey
Also known as "post-intervention" or "ex-post" survey. A survey that is administered after the program has started, once the beneficiaries have benefited from the program for some time. An evaluation can include several follow-up surveys.

Hawthorne effect
The "Hawthorne effect" occurs when the mere fact that you are observing subjects makes them behave differently.

Hypothesis
A specific statement regarding the relationship between two variables. In an impact evaluation the hypothesis typically relates to the expected impact of the intervention on the outcome.

Impact
The effect of the intervention on the outcome for the beneficiary population.

Impact evaluation
An impact evaluation tries to make a causal link between a program or intervention and a set of outcomes. An impact evaluation tries to answer the question of whether a program is responsible for changes in the outcomes of interest. Contrast with "process evaluation".

Independent variable
A variable believed to cause changes in the dependent variable, usually applied in regression analysis.

Indicator
An indicator is a variable that measures a phenomenon of interest to the evaluator. The phenomenon can be an input, an output, an outcome, or a characteristic.

Inputs
The financial, human, and material resources used for the development intervention.

Intention to treat (ITT) estimate
The average treatment effect calculated across the whole treatment group, regardless of whether they actually participated in the intervention or not. Compare to "treatment on the treated estimate".

Intra-cluster correlation
Intra-cluster correlation is correlation (or similarity) in outcomes or characteristics between subjects that belong to the same cluster. For example, children that attend the same school would typically be similar or correlated in terms of their area of residence or socio-economic background.

Logical model
Describes how a program should work, presenting the causal chain from inputs, through activities and outputs, to outcomes. While logical models present a theory about the expected program outcome, they do not demonstrate whether the program caused the observed outcome. A theory-based approach examines the assumptions underlying the links in the logical model.

John Henry effect
The "John Henry effect" happens when comparison subjects work harder to compensate for not being offered a treatment. When one compares treated units to those "harder-working" comparison units, the estimate of the impact of the program will be biased: we will estimate a smaller impact of the program than the true impact we would find if the comparison units did not make the additional effort.

Minimum desired effect
Minimum change in outcomes that would justify the investment that has been made in an intervention, accounting not only for the cost of the program and the type of benefits that it provides, but also on the opportunity cost of not having invested funds in an alternative intervention. The minimum desired effect is an input for power calculations: evaluation samples need to be large enough to detect at least the minimum desired effects with sufficient power.

Null hypothesis
A null hypothesis is a hypothesis that might be falsified on the basis of observed data. The null hypothesis typically proposes a general or default position. In evaluation, the default position is usually that there is no difference between the treatment and control group, or in other words, that the intervention has no impact on outcomes.

Outcome
A variable that measures the impact of the intervention. Can be intermediate or final, depending on what it measures and when.

Output
The products and services that are produced (supplied) directly by an intervention. Outputs may also include changes that result from the intervention which are relevant to the achievement of outcomes.

Power calculation
A calculation of the sample required for the impact evaluation, which depends on the minimum effect size that we want to be able to detect (see "minimum desired effect") and the required level of confidence.

Pre-post comparison
Also known as a before and after comparison. A pre-post comparison attempts to establish the impact of a program by tracking changes in outcomes for program beneficiaries over time using measures both before and after the program or policy is implemented.

Process evaluation
A process evaluation is an evaluation that tries to establish the level of quality or success of the processes of a program. For example: adequacy of the administrative processes, acceptability of the program benefits, clarity of the information campaign, internal dynamics of implementing organizations, their policy instruments, their service delivery mechanisms, their management practices, and the linkages among these. Contrast with "impact evaluation".

Quasi-experimental design
Impact evaluation designs that create a control group using statistical procedures. The intention is to ensure that the characteristics of the treatment and control groups are identical in all respects, other than the intervention, as would be the case in an experimental design.

Random assignment
An intervention design in which members of the eligible population are assigned at random to either the treatment group (receive the intervention) or the control group (do not receive the intervention). That is, whether someone is in the treatment or control group is solely a matter of chance, and not a function of any of their characteristics (either observed or unobserved).

Random sample
The best way to avoid a biased or unrepresentative sample is to select a random sample. A random sample is a probability sample where each individual in the population being sampled has an equal chance (probability) of being selected.

Randomized evaluation (RE) (also known as randomized controlled trial, or RCT)
An impact evaluation design in which random assignment is used to allocate the intervention among members of the eligible population. Since there should be no correlation between participant characteristics and the outcome, and differences in outcome between the treatment and control can be fully attributed to the intervention, i.e. there is no selection bias. However, REs may be subject to several types of bias and so need follow strict protocols. Also called "experimental design".

Regression analysis
A statistical method which determines the association between the dependent variable and one or more independent variables.

Selection bias
A possible bias introduced into a study by the selection of different types of people into treatment and comparison groups. As a result, the outcome differences may potentially be explained as a result of pre-existing differences between the groups, rather than the treatment itself.

Significance level
The significance level is usually denoted by the Greek symbol, α (alpha). Popular levels of significance are 5% (0.05), 1% (0.01) and 0.1% (0.001). If a test of significance gives a p-value lower than the α-level, the null hypothesis is rejected. Such results are informally referred to as 'statistically significant'. The lower the significance level, the stronger the evidence required. Choosing level of significance is an arbitrary task, but for many applications, a level of 5% is chosen, for no better reason than that it is conventional.

Spillover effects
When the intervention has an impact (either positive or negative) on units not in the treatment group. Ignoring spillover effects results in a biased impact estimate. If there are spillover effects then the group of beneficiaries is larger than the group of participants.

Stratified sample
Obtained by dividing the population of interest (sampling frame) into groups (for example, male and female), then by drawing a random sample within each group. A stratified sample is a probabilistic sample: every unit in each group (or strata) has the same probability of being drawn.

Treatment group
The group of people, firms, facilities or other subjects who receive the intervention. Also called participants.

Treatment on the treated (TOT) estimate
The treatment on the treated estimate is the impact (average treatment effect) only on those who actually received the intervention. Compare to intention to treat.

Unobservables
Characteristics which cannot be observed or measured. The presence of unobservables can cause selection bias in quasi-experimental designs.

ABOUT IPA AND J-PAL

COLLABORATION
Between IPA and J-PAL

IPA and J-PAL are complementary organizations that work together towards the common goal of reducing poverty by ensuring that policy is based on scientific evidence.

Innovations for Poverty Action (IPA) is an international non-profit research organization that has a strong local presence through its 14 country programs.

The Abdul Latif Jameel Poverty Action Lab (J-PAL) is a network of over 80 affiliated professors working through six research centers based at leading universities around the world.

The two organizations work on three core activities, dividing up the work based on their comparative advantages and local presence:

RESEARCH IPA and J-PAL conduct randomized evaluations of social programs in countries where their offices are located, in partnership with implementing organizations. IPA also provides flexible on-the-ground support on randomized evaluation design, measurement, and data collection methods to IPA and J-PAL evaluations worldwide

CAPACITY BUILDING IPA and J-PAL conduct customized trainings on randomized evaluations for key practitioners and policymakers. J-PAL works to build policymakers' capacity to conduct randomized evaluations through its flagship Executive Education course and long-term capacity building partnerships. IPA provides customized training on randomized evaluations through its Country Programs.

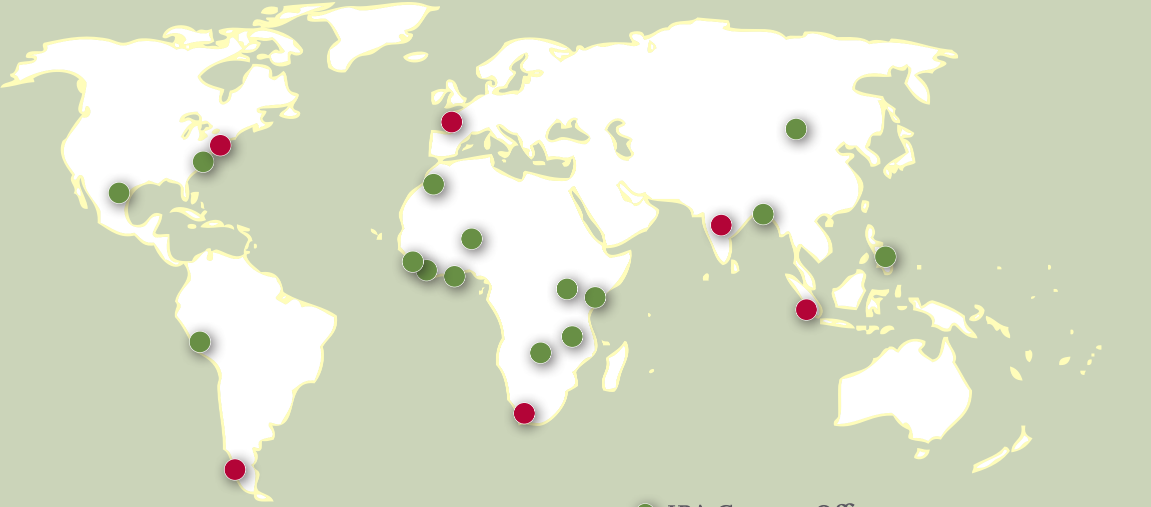
POLICY OUTREACH J-PAL performs policy analysis, disseminates policy lessons through web and printed publications, and engages policymakers at the global and regional level. Capitalizing on J-PAL's policy analyses, IPA uses its country offices to build relationships with local policymakers and practitioners and to inform policy decisions with evidence. Both organizations provide technical assistance to governments and NGOs who want to scale up programs based on rigorous evidence, and IPA also directly implements some promising programs to encourage their wider adoption.

OFFICES AROUND THE WORLD

ipa INNOVATIONS FOR POVERTY ACTION

ABDUL LATIF JAMEEL Poverty Action Lab

TRANSLATING RESEARCH INTO ACTION



● **J-PAL Regional Offices:** **J-PAL Africa** at University of Cape Town (UCT), South Africa; **J-PAL Europe** at Paris School of Economics (PSE), France; **J-PAL LAC** at Pontificia Universidad Católica de Chile (PUC), Chile; **J-PAL South Asia** at Institute for Financial Management and Research (IFMR), India; **J-PAL Southeast Asia** at Universitas Indonesia, Indonesia; **J-PAL Global** (HQ) at Massachusetts Institute of Technology (MIT), USA

● **IPA Country Offices:** Bangladesh, Ghana, Kenya, Liberia, Malawi, Mali, Mexico, Mongolia, Morocco, Peru, Philippines, Sierra Leone, Uganda, Zambia; **IPA HQ in New Haven, USA**

CASE STUDIES & GROUP WORK

ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

Case Study 1: Learn to Read Evaluations

How to Read and Evaluate Evaluations

This case study is based on "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in India," by Abhijit Banerjee (MIT), Rukmini Banerjee (Pratham), Esther Duflo (MIT), Rachel Glennerster (J-PAL), and Stuti Khemani (The World Bank)

J-PAL thanks the authors for allowing us to use their paper

Key Vocabulary

Counterfactual: what would have happened to the participants in a program had they not received the intervention. The counterfactual cannot be observed from the treatment group; can only be inferred from the comparison group.

Comparison Group: in an experimental design, a randomly assigned group from the same population that does not receive the intervention that is the subject of evaluation. Participants in the comparison group are used as a standard for comparison against the treated subjects in order to validate the results of the intervention.

Program Impact: estimated by measuring the difference in outcomes between comparison and treatment groups. The true impact of the program is the difference in outcomes between the treatment group and its counterfactual.

Baseline: data describing the characteristics of participants measured across both treatment and comparison groups prior to implementation of intervention.

Endline: data describing the characteristics of participants measured across both treatment and comparison groups after implementation of intervention.

Selection Bias: statistical bias between comparison and treatment groups in which individuals in one group are systematically different from those in the other. These can occur when the treatment and comparison groups are chosen in a non-random fashion so that they differ from each other by one or more factors that may affect the outcome of the study.

Omitted Variable Bias: statistical bias that occurs when certain variables/characteristics (often unobservable), which affect the measured outcome, are omitted from a regression analysis. Because they are not included as controls in the regression, one incorrectly attributes the measured impact solely to the program.

Introduction

In a large-scale survey conducted in 2004, Pratham discovered that only 39% of children (aged 7-14) in rural Uttar Pradesh could read and understand a simple story, and nearly 15% could not recognize even a letter.

During this period, Pratham was developing the “Learn-to-Read” (L2R) module of its Read India campaign. L2R included a unique pedagogy teaching basic literacy skills, combined with a grassroots organizing effort to recruit volunteers willing to teach.

This program allowed the community to get involved in children’s education more directly through village meetings where Pratham staff shared information on the status of literacy in the village and the rights of children to education. In these meetings, Pratham identified community members who were willing to teach. Volunteers attended a training session on the pedagogy, after which they could hold after-school reading classes for children, using materials designed and provided by Pratham. Pratham staff paid occasional visits to these camps to ensure that the classes were being held and to provide additional training as necessary.

Did this program work? How would you measure the impact?

Did the Learn to Read Project work?

Did Pratham's "Learn to Read" program work? What is required in order for us to measure whether a program worked, or whether it had impact?

In general, to ask if a program works is to ask if the program achieves its goal of changing certain outcomes for its participants, and ensure that those changes are not caused by some other factors or events happening at the same time. To show that the program causes the observed changes, we need to simultaneously show that if the program had not been implemented, the observed changes would not have occurred (or would be different). But how do we know what would have happened? If the program happened, it happened. Measuring what would have happened requires entering an imaginary world in which the program was never given to these participants. The outcomes of the same participants in this imaginary world are referred to as the counterfactual. Since we cannot observe the true counterfactual, the best we can do is to estimate it by mimicking it.

The key challenge of program impact evaluation is constructing or mimicking the counterfactual. We typically do this by selecting a group of people that resemble the participants as much as possible but who did not participate in the program. This group is called the comparison group. Because we want to be able to say that it was the program and not some other factor that caused the changes in outcomes, it is important that the only difference between the comparison group and the participants is that the comparison group did not participate in the program. We then estimate "impact" as the difference observed at the end of the program between the outcomes of the comparison group and the outcomes of the program participants.

The impact estimate is only as accurate as the comparison group is successful at mimicking the counterfactual. If the comparison group poorly represents the counterfactual, the impact is (in most circumstances) poorly estimated. Therefore the method used to select the comparison group is a key decision in the design of any impact evaluation.

That brings us back to our questions: Did the Learn to Read project work? What was its impact on children's reading levels?

case, the intention of the program is to "improve children's reading levels" and the reading level is the outcome measure. So, when we ask if the Learn to Read project worked, we are asking if it improved children's reading levels. The impact is the difference between reading levels after the children have taken the reading classes and what their reading level would have been if the reading classes had never existed.

For reference, Reading Level is an indicator variable that takes value 0 if the child can read nothing, 1 if he knows the alphabet, 2 if he can recognize words, 3 if he can read a paragraph, and 4 if he can read a full story.

What comparison groups can we use? The following experts illustrate different methods of evaluating impact. (Refer to the table on the last page of the case for a list of different evaluation methods).

Estimating the impact of the Learn to Read project

METHOD 1:

News Release: Read India helps children Learn to Read.

Pratham celebrates the success of its "Learn to Read" program—part of the Read India Initiative. It has made significant progress in its goal of improving children's literacy rates through better learning materials, pedagogical methods, and most importantly, committed volunteers. The achievement of the "Learn to Read" (L2R) program demonstrates that a revised curriculum, galvanized by community mobilization, can produce significant gains. Massive government expenditures in mid-day meals and school construction have failed to achieve similar results. In less than a year, the reading levels of children who enrolled in the L2R camps improved considerably.

FIGURE 1

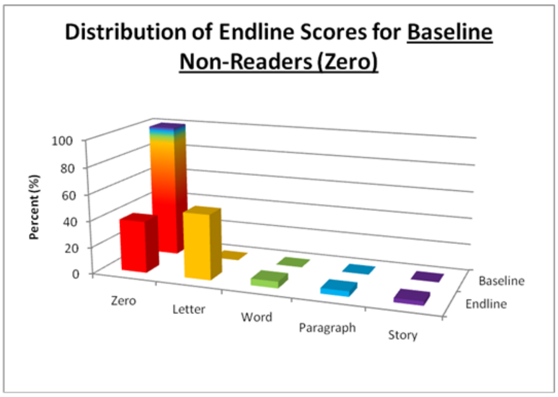
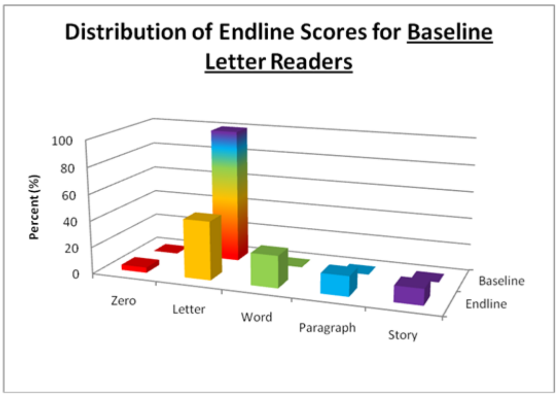


FIGURE 2



Just before the program started, half these children could not recognize Hindi words—many nothing at all. But after spending just a few months in Pratham reading classes, more than half improved by at least one reading level, with a significant number capable of recognizing words and several able to read full paragraphs and stories! *On average, the literacy measure of these students improved by nearly one full reading level during this period.*

DISCUSSION TOPIC 1

Identifying evaluation

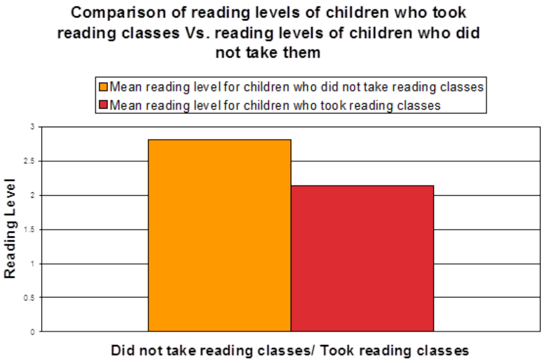
1. What type of evaluation does this news release imply?
2. What represents the counterfactual?
3. What are the problems with this type of evaluation?

METHOD 2:

Opinion: The "Read India" project not up to the mark

Pratham has raised millions of dollars, expanding rapidly to cover all of India with its so-called "Learn-to-Read" program, but do its students actually learn to read? Recent evidence suggests otherwise. A team of evaluators from Education for All found that children who took the reading classes ended up with literacy levels significantly below those of their village counterparts. After one year of Pratham reading classes, Pratham students could only recognize words whereas those who steered clear of Pratham programs were able to read full paragraphs.

FIGURE 3



Notes: Reading Level is an indicator variable that takes value 0 if the child can read nothing, 1 if he knows the alphabet, 2 if he can recognize words, 3 if he can read a paragraph and 4 if he can read a full story.

If you have a dime to spare, and want to contribute to the education of India's illiterate children, you may think twice before throwing it into the fountain of Pratham's promises.

DISCUSSION TOPIC 2

Identifying evaluation

1. What type of evaluation does this news release imply?

2. What represents the counterfactual?
3. What are the problems with this type of evaluation?

METHOD 3:

Letter to the Editor: EFA should consider Evaluating Fairly and Accurately

There have been several unfair reports in the press concerning programs implemented by the NGO Pratham. A recent article by a former Education for All bureaucrat claims that Pratham is actually hurting the children it recruits into its 'Learn-to-Read' camps. However, the EFA analysis uses the wrong metric to measure impact. It compares the reading *levels* of Pratham students with other children in the village—not taking into account the fact that Pratham targets those whose literacy levels are particularly poor at the beginning. If Pratham simply recruited the most literate children into their programs, and compared them to their poorer counterparts, they could claim success without conducting a single class. But Pratham does not do this. And realistically, Pratham does not expect its illiterate children to overtake the stronger students in the village. It simply tries to initiate improvement over the current state. Therefore the metric should be *improvement* in reading levels—not the final level. When we repeated EFA's analysis using the more-appropriate outcome measure, the Pratham kids improved at twice the rate of the non-Pratham kids (0.6 reading level increase compared to 0.3). This difference is statistically very significant.

Had the EFA evaluators thought to look at the more appropriate outcome, they would recognize the incredible success of Read India. Perhaps they should enroll in some Pratham classes themselves.

DISCUSSION TOPIC 3

Identifying evaluation

1. What type of evaluation does this news release imply?
2. What represents the counterfactual?
3. What are the problems with this type of evaluation?

METHOD 4:

The numbers don't lie, unless your statisticians are asleep

Pratham celebrates victory, opponents cry foul. A closer look shows that, as usual, the truth is somewhere in between.

There has been a war in the press between Pratham's supporters and detractors. Pratham and its advocates assert that the Read India campaign has resulted in large increases in child literacy. Several detractors claim that Pratham programs, by pulling attention away from the schools, are in fact causing significant harm to the students. Unfortunately, this battle is being waged using instruments of analysis that are seriously flawed. The ultimate victim is the public who is looking for an answer to the question: is Pratham helping its intended beneficiaries?

This report uses sophisticated statistical methods to measure the true impact of Pratham programs. We were concerned about other variables confounding previous results. We therefore conducted a survey in these villages to collect information on child age, grade-level, and parents' education level, and used those to predict child test scores.

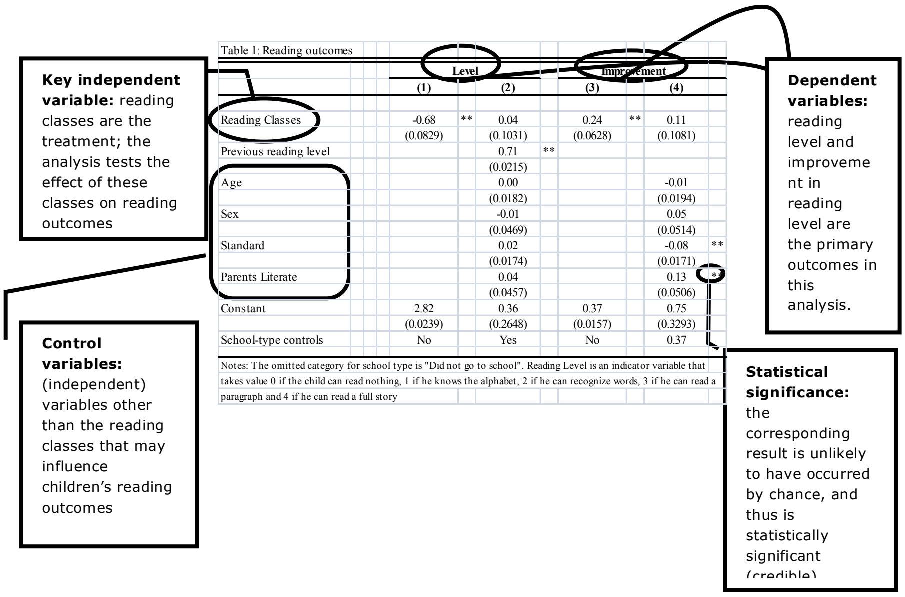


Table 1: Reading outcomes

	Level				Improvement			
	(1)		(2)		(3)		(4)	
Reading Classes	-0.68	**	0.04		0.24	**	0.11	
	(0.0829)		(0.1031)		(0.0628)		(0.1081)	
Previous reading level			0.71	**				
			(0.0215)					
Age			0.00				-0.01	
			(0.0182)				(0.0194)	
Sex			-0.01				0.05	
			(0.0469)				(0.0514)	
Standard			0.02				-0.08	**
			(0.0174)				(0.0171)	
Parents Literate			0.04				0.13	**
			(0.0457)				(0.0506)	
Constant	2.82		0.36		0.37		0.75	
	(0.0239)		(0.2648)		(0.0157)		(0.3293)	
School-type controls	No		Yes		No		0.37	

Notes: The omitted category for school type is "Did not go to school". Reading Level is an indicator variable that takes value 0 if the child can read nothing, 1 if he knows the alphabet, 2 if he can recognize words, 3 if he can read a paragraph and 4 if he can read a full story

Key independent variable: reading classes are the treatment; the analysis tests the effect of these classes on reading outcomes

Dependent variables: reading level and improvement in reading level are the primary outcomes in this analysis.

Control variables: (independent) variables other than the reading classes that may influence children's reading outcomes

Statistical significance: the corresponding result is unlikely to have occurred by chance, and thus is statistically significant (credible)

Looking at Table 1, we find some positive results, some negative results and some "no-results", depending on which variables we control for. The results from column (1) suggest that Pratham's program hurt the children. There is a negative correlation between receiving Pratham classes and final reading outcomes (-0.68). Column (3), which evaluates improvement, suggests impressive results (0.24). But looking at child outcomes (either level or improvement) controlling for initial reading levels, age, gender, standard and parent's education level – all determinants of child reading levels – we found no impact of Pratham programs.

Therefore, controlling for the right variables, we have discovered that on one hand, Pratham has not caused the harm claimed by certain opponents, but on the other hand, it has not helped children learn. Pratham has therefore failed in its effort to convince us that it can spend donor money effectively.

NOTE: Data used in this case are real. "Articles" on the debate were artificially produced for the purpose of the case. Education for All (EFA) never made any of the claims described herein.

DISCUSSION TOPIC 4

Identifying evaluation

1. What type of evaluation does this news release imply?
2. What represents the counterfactual?
3. What are the problems with this type of evaluation?

	Methodology	Description	Who is in the comparison group?	Required Assumptions	Required Data
Quasi-Experimental Methods	**Pre-Post**	Measure how program participants improved (or changed) over time.	Program participants themselves—before participating in the program.	The program was the only factor influencing any changes in the measured outcome over time.	Before and after data for program participants.
	Simple Difference	Measure difference between program participants and non-participants after the program is completed.	Individuals who didn't participate in the program (for any reason), but for whom data were collected after the program.	Non-participants are identical to participants except for program participation, and were equally likely to enter program before it started.	After data for program participants and non-participants.
	Differences in Differences	Measure improvement (change) over time of program participants *relative to* the improvement (change) of non-participants.	Individuals who didn't participate in the program (for any reason), but for whom data were collected both before and after the program.	If the program didn't exist, the two groups would have had identical trajectories over this period.	Before and after data for both participants and non-participants.
	Multivariate Regression	Individuals who received treatment are compared with those who did not, and other factors that might explain differences in the outcomes are "controlled" for.	Individuals who didn't participate in the program (for any reason), but for whom data were collected both before and after the program. In this case data is not comprised of just indicators of outcomes, but other "explanatory" variables as well.	The factors that were *excluded* (because they are unobservable and/or have been not been measured) do not bias results because they are either uncorrelated with the outcome or do not differ between participants and non-participants.	Outcomes as well as "control variables" for both participants and non-participants.
	Statistical Matching	Individuals in control group are compared to similar individuals in experimental group.	Exact matching: For each participant, at least one non-participant who is identical *on selected characteristics*. Propensity score matching: non-participants who have a mix of characteristics which predict that they would be as likely to participate as participants.	The factors that were *excluded* (because they are unobservable and/or have been not been measured) do not bias results because they are either uncorrelated with the outcome or do not differ between participants and non-participants.	Outcomes as well as "variables for matching" for both participants and non-participants.
	Regression Discontinuity Design	Individuals are ranked based on specific, measureable criteria. There is some cutoff that determines whether an individual is eligible to participate. Participants are then compared to non-participants and the eligibility criterion is controlled for.	Individuals who are close to the cutoff, but fall on the "wrong" side of that cutoff, and therefore do not get the program.	After controlling for the criteria (and other measures of choice), the remaining differences between individuals directly below and directly above the cut-off score are not statistically significant and will not bias the results. A necessary but sufficient requirement for this to hold is that the cut-off criteria are strictly adhered to.	Outcomes as well as measures on criteria (and any other controls).
	Instrumental Variables	Participation can be predicted by an incidental (almost random) factor, or "instrumental" variable, that is uncorrelated with the outcome, other than the fact that it predicts participation (and participation affects the outcome).	Individuals who, because of this close to random factor, are predicted not to participate and (possibly as a result) did not participate.	If it weren't for the instrumental variable's ability to predict participation, this "instrument" would otherwise have no effect on or be uncorrelated with the outcome.	Outcomes, the "instrument," and other control variables.
Experimental Method	**Randomized Evaluation**	Experimental method for measuring a causal relationship between two variables.	Participants are randomly assigned to the control groups.	Randomization "worked." That is, the two groups are statistically identical (on observed and unobserved factors).	Outcome data for control and experimental groups. Control variables can help absorb variance and improve "power".

GROUP WORK 1:

CHOOSING A RESEARCH QUESTION

During this session, you will work with your small group to choose a topic for which you would hypothetically like to design a randomized evaluation. You should pick a topic which is both feasible to study and policy relevant to Malawi. Although this is just an exercise, many research ideas from previous J-PAL and IPA trainings have turned into full randomized evaluations. To guide you, we would like to focus on ideas that meet the following criteria:

- **Policy Relevant:** The research idea should fill some gap in answering a policy question. For example, asking the impact of winning the lottery won't give us information about something that's feasible to be scaled up to everyone.

- **Academically Interesting:** Impact evaluations should add to the existing literature on topic. Review the existing research so that you're not answering a question that already had many answers. There are plenty of gaps in knowledge – seek to fill one.

- **Focus on cause and effect:** Descriptive questions (such as, how many people have electricity?) and normative questions (such as, is health care a human right?) are best left to other types of analysis. Impact evaluation can help us see if an intervention leads to a specific outcome.

- **Specific**: For example, a question such as how can we improve a child's diet is not as strong as, will introducing an improved sweet potato decrease anemia in children? Remember to start with an intervention and then identify a particular outcome you would like to evaluate.

In your groups, please discuss:

1. Why are the following research questions NOT suitable for an impact evaluation?

- *What is the impact of expanding the student capacity of the national university?*
- *How can we improve the health of children under five in Malawi?*

2. What makes the following research questions GOOD candidates for an impact evaluation?

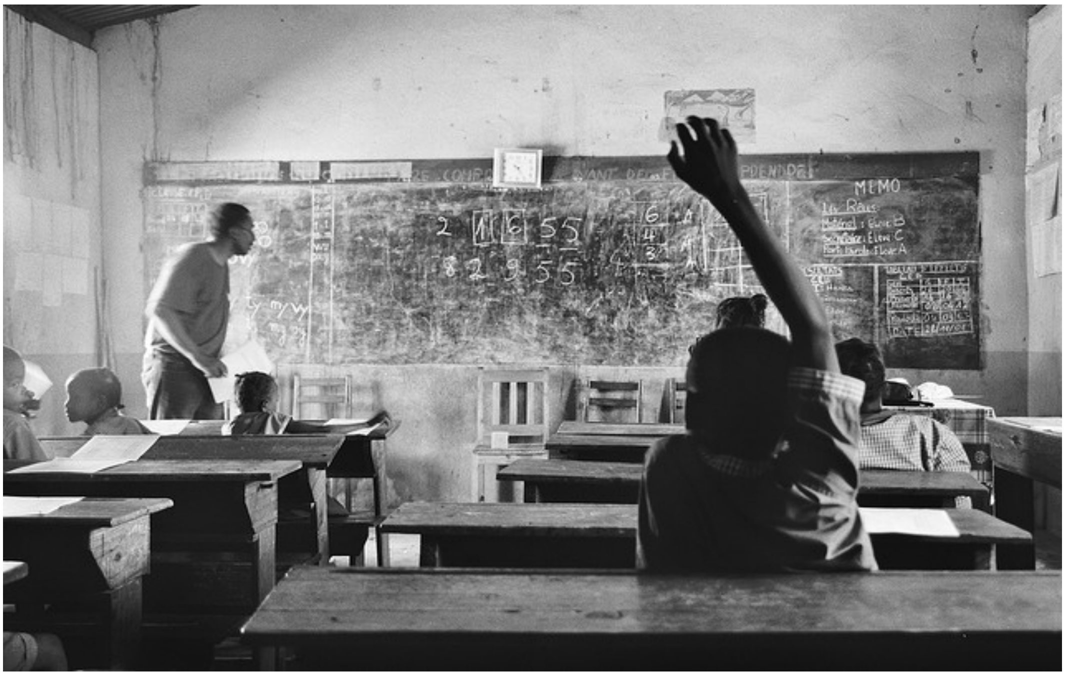
- *Can parental involvement in school committees improve teacher performance in the classroom?*
- *Can teaching parents healthy diet and cooking practices improve child health outcomes?*

Please write down your research question here:

ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

Case 2: Reforming School Monitoring

Program Theory and Measuring Outcomes



This case study is based on the J-PAL Study "Primary Education Management in Madagascar" by Esther Duflo, Gerard Lassibille, and Trang van Nguyen.

J-PAL thanks the authors for allowing us to use their paper.

Key Vocabulary

Hypothesis: a proposed explanation of and for the effects of a given intervention. Hypotheses are intended to be made ex-ante, or prior to the implementation of the intervention.

Indicators: metrics used to quantify and measure specific short-term and long-term effects of a program

Logical Framework (LogFrame): a management tool used to facilitate the design, execution, and evaluation of an intervention. It involves identifying strategic elements (inputs, outputs, outcomes and impact) and their causal relationships, indicators, and the assumptions and risks that may influence success and failure

Theory of Change (ToC): describes a strategy or blueprint for achieving a given long-term goal. It identifies the preconditions, pathways and interventions necessary for an initiative's success

Introduction

Over the last 10 years, low-income countries in Africa have made striking progress in expanding coverage of primary education. However, in many of these countries the education system continues to deliver poor results, putting the goal of universal primary school completion at risk. Incompetent administration, inadequate focus on learning outcomes, and weak governance structures are thought to be some of the reasons for the poor results. This case study will look at a program which aimed to improve the performance and efficiency of education systems by introducing tools and a monitoring system at each level along the service delivery chain.

Madagascar School System Reforms: "Improving Outputs not Outcomes"

Madagascar's public primary school system has been making progress in expanding coverage in primary education thanks in part due to increases in public spending since the late 1990s. As part of its poverty reduction strategy, public expenditure on education rose from 2.2 to 3.3 percent of GDP between 2001 and 2007. In addition to increased funding, the government introduced important reforms such as the elimination of school fees for primary education, free textbooks to primary school students, public subsidies to supplement the wages of non–civil service teachers in public schools (in the past they were hired and paid entirely by parent associations), and new pedagogical approaches.

The most visible sign of progress was the large increase in coverage in primary education in recent years. In 2007, the education system enrolled some 3.8 million students in both public and private schools—more than twice the enrolment in 1996. During the last 10 years, more than 4000 new public primary schools have been created, and the number of primary school teachers in the public sector more than doubled.

While this progress is impressive, enormous challenges remain. Entry rates into grade 1 are high, but less than half of each cohort reaches the end of the five-year primary cycle. Despite government interventions, grade repetition rates are still uniformly high throughout the primary cycle, averaging about 18 percent. Furthermore, test scores reveal poor performance: students scored an average of 30 percent on French and 50 percent on Malagasy and mathematics.

DISCUSSION TOPIC 1

Madagascar school system reforms

1. Would you regard the reforms as successful? Why or why not?
2. What are some of the potential reasons for why the reforms did not translate into better learning outcomes?

Problems remain...

As the starting point of the study, researchers worked with the Ministry of Education to identify the remaining constraints in the schooling system. A survey conducted in 2005 revealed the following key problems:

1. Teacher absenteeism: At 10 percent, teacher absenteeism remains a significant problem. Only 8 percent of school directors monitor teacher attendance (either by taking daily attendance or tracking and posting a monthly summary of attendance), and more than 80 percent fail to report teacher absences to sub-district and district administrators.

2. Communication with parents: Communication between teachers and parents on student learning is often perfunctory, and student absenteeism is rarely communicated to parents.

3. Teacher performance: Essential pedagogical tasks are often neglected: only 15 percent of teachers consistently prepare daily and biweekly lessons plans while 20 percent do not prepare lesson plans at all. Student academic progress is also poorly monitored: results of tests and quizzes are rarely recorded and 25 percent of teachers do not prepare individual student report cards.

Overall, many of problems seem to be result of a lack of organization, control and accountability at every stage of the system, all of which are likely to compromise the performance of the system and lower the chance of the reforms being successful.

Intervention

In order to address these issues, the Madagascar Ministry of Education seeks to tighten the management and accountability at each point along the service delivery chain (see Figure 1) by making explicit to the various administrators and teachers what their responsibilities are, supporting them with teaching tools, and increasing monitoring.

The ministry is considering two approaches to evaluate[1]:

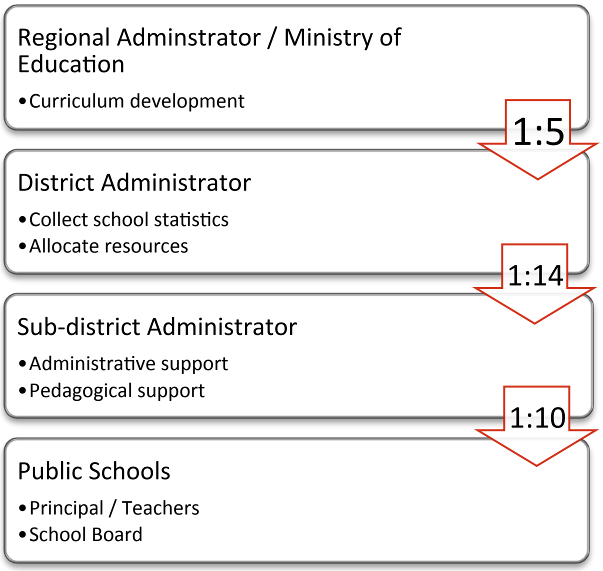
1. Top-Down

Operational tools and guidebooks which outline their responsibilities are given to the relevant administrators. During a meeting, administrators are trained on how to carry out their tasks, and their performance criteria are clarified. This is followed up by regular monitoring of their performance, which is communicated through (sub-) district report cards to higher levels.

2. Bottom-Up

This program promotes the ability of parents to monitor their schools and hold teachers accountable when they perform below expectation. Report cards with easy-to-understand content are given to parents and members of poor rural communities. They contain a small set of performance indicators, information on enrolments and school resources, as well as data that allow a school's performance to be compared that of other schools (see Appendix). In addition, greater community participation in school-based management is encouraged through structured school meetings in which staff of the school, parents, and community members review the report card and discuss their school improvement plan.

FIGURE 1: EDUCATION SYSTEM



DISCUSSION TOPIC 2

Intermediate and final outcomes

1. Before setting up the RCT, researchers carefully analyzed the existing problem. Why do you think this is important as a starting point of an evaluation?
2. What are the intermediate and ultimate goals that this program hopes to achieve?
3. What is the key hypothesis being tested through this impact evaluation?

Theory of Change

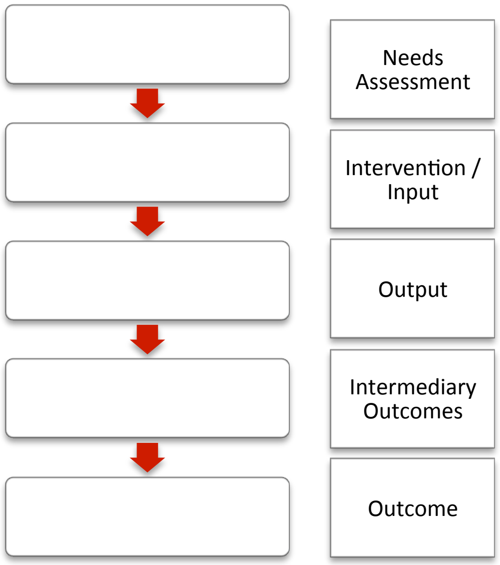
A theory of change (ToC) identifies the causal link between the intervention and the final outcome. Figure 2 shows one way in which a ToC can be structured.

For example, a program or intervention is implemented to address a specific problem identified in the needs assessment (e.g. low literacy levels). The intervention (e.g. text books) may lead to outputs (e.g. students usage of textbooks) through which intermediary outcomes (e.g. reading skills) could be affected. These may lead to longer-term outcomes (e.g. drop-out rates, employment

[1] The actual evaluation included further interventions such as training of teachers. For more details, please refer to the paper. For pedagogical reasons, we focus only on two approaches in this case study.

outcomes). An underlying assumption of this ToC is that students do not already have text books.

FIGURE 2: THEORY OF CHANGE



DISCUSSION TOPIC 3

Theory of change

1. Draw out the causal chain using the format in Figure 2 for each of the Bottom-up and Top-down interventions (use a separate ToC for each).

2. What are the necessary conditions/assumptions underlying these ToCs?

What data to collect? Data collection and measurement

Before deciding which data to collect, you need to be very clear on the outcome you are targeting and in what way the intervention is theorized to impact this outcome. In other words, identifying a key hypothesis and theory of change at the beginning of an evaluation helps you to decide what information to collect.

For each step of the theory of change, we need to identify **indicators** (what to measure) and **instruments** (how to collect data). Continuing with the example of the text book program, an indicator could be reading level of students and the instrument could be standardized reading tests. In addition, we need to collect data on our assumptions to see whether or not they hold true.

DISCUSSION TOPIC 4

Measuring outcomes and indicators

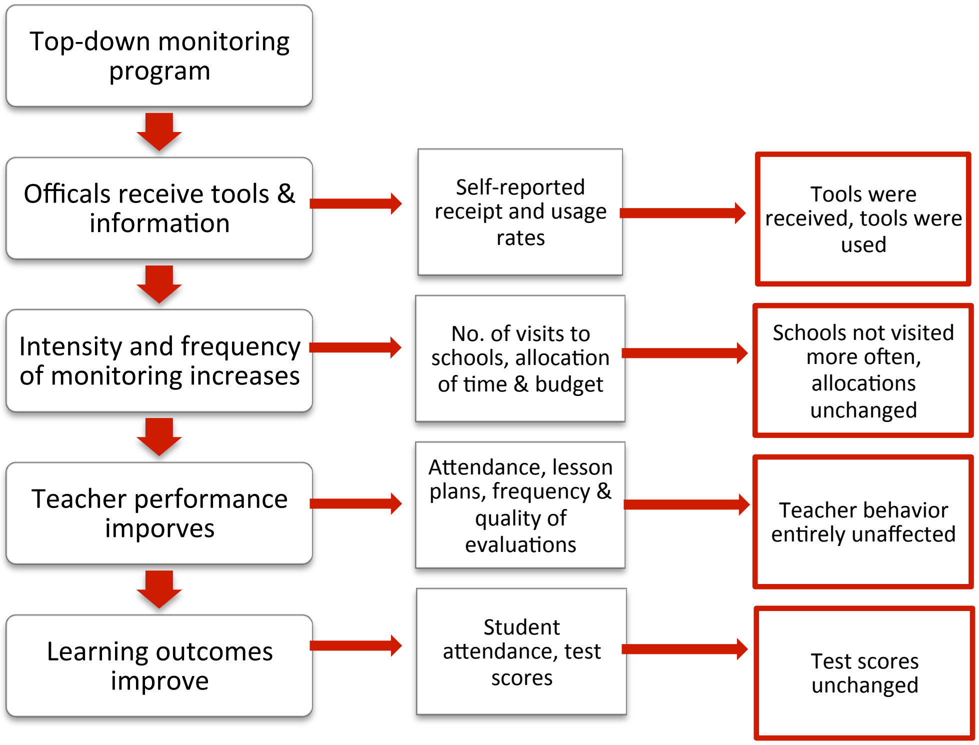
1. Which indicators would you measure at each step in the ToCs you drew up?

2. How would you collect data for these indicators? In other words, what instruments would you use? Do you foresee challenges with these forms of data collection?

How to interpret the results

The evaluation found that the bottom-up approach led to successful results. Attendance at meetings between teachers and community members was high, and although communication between teachers and parents did not change, teachers improved the quality of teaching as shown by an increase in lesson plans and test scores.

However, the findings of the top-down intervention were quite different:



Theory of Change	Indicators	Results
Top-down monitoring program		
Officals receive tools & information	Self-reported receipt and usage rates	Tools were received, tools were used
Intensity and frequency of monitoring increases	No. of visits to schools, allocation of time & budget	Schools not visited more often, allocations unchanged
Teacher performance imporves	Attendance, lesson plans, frequency & quality of evaluations	Teacher behavior entirely unaffected
Learning outcomes improve	Student attendance, test scores	Test scores unchanged

DISCUSSION TOPIC 5

Interpreting the results

1. How do you interpret the results of the Top-down intervention?
2. Why is it important to interpret the results in the context of a program theory of change?
3. What are the policy implications? How might you respond to these findings?

GROUP WORK 2:

RESEARCH DESIGN

In this section, you will work on the research design for the topic your group chose on day 1 of the workshop. Your group should discuss the following topics:

1. What is the population for which your project will focus? Be specific!
   For example, if you were going to do a study on text message reminders to save, your population might be everyone who has their phone number listed at a certain bank.

2. How will you develop a sample from this population? Identify both the sample for the treatment and the control. How will you ensure the control group remains exactly the same as the treatment (besides for the receipt of the treatment)?
   Following the example above, you might advertise to have individuals register their phone numbers to get your population. Then, you would randomly select your treatment and control group from this population and monitor the savings levels of both groups.

3. How will you administer the treatment?
   Continuing the example, you might want to send text messages to all the groups so that you look at the impact of reminders rather than of text messages. The control group can receive a text about an unrelated topic, while the treatment gets reminders to save.

4. What type of data will you collect? From whom? On what timeline?
   Regarding text messages, you might monitor the savings balances in the bank after 6 months and one year using administrative data.

ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

Case Study 2: Theory of Change

Measuring the effects of your intervention
Thinking about measurement and outcomes

Key Vocabulary

Theory of Change: describes a strategy or blueprint for achieving a given long-term goal. It identifies the preconditions, pathways, and interventions necessary for an initiative's success.

Logical Framework: a management tool used to facilitate the design, execution, and evaluation of a range of projects, including large-scale interventions. It involves identifying strategic elements (inputs, outputs, outcomes, and impacts) and their causal relationships, choosing indicators, and acknowledging the assumptions and risks that may influence the success and/or failure of the intervention.

Outputs: what an intervention produces or provides to program participants. They are direct products of program activities/inputs and may include services delivered by the program. Outputs will be tracked through monitoring and process evaluation.

Outcomes: effects or changes that are anticipated to occur as a result of the intervention. These consequences of the intervention can be intended or unintended, positive or negative, as well as short-term or long-term. It is important to think of each type of possible outcome.

Counterfactual: what would have happened to the participants in a program had they not participated in the intervention. The counterfactual cannot be observed, as by definition it is the state of the world that does not occur.

Comparison Group: members of the study's population that are compared to the group that received a particular intervention in order to estimate the impact of the intervention. The accuracy of an impact evaluation is based on how well this group represents the counterfactual.

Introduction

Work with your group to decide upon a policy question which can be answered using an impact evaluation. Remember that a good research question is one which helps policymakers make a choice between two or more options.

Theory of Change

Write your research question here:

DISCUSSION TOPIC 1

As a group, brainstorm the areas you think your program might impact (for example, household income, school attendance or test scores, disease rates, etc.)

DISCUSSION TOPIC 2

Discuss the theory of change of this intervention. In the table on the next page, fill in the inputs, outputs, outcomes, and impact. (Check the Key Vocabulary section to see definitions for theory of change, outputs, and outcomes.)

DISCUSSION TOPIC 3

What assumptions are needed to get from the inputs to outputs, the outputs to outcomes, and from outcomes to impacts? Discuss why these assumptions might not always hold.

DISCUSSION TOPIC 4

In what time frame would you expect each outcome and impact to occur? Discuss outcomes which you think might be different in the short-run than in the long-run.

Theory of Change

	Objective	Indicator	Assumptions	Time Frame
Impact				
Outcomes				
Outputs				
Inputs				